Large Language Models for Biomedical Network Augmentation: An Example in Drug–Disease Associations Prediction through A Dual-Channel Layer Attention Graph Convolutional Network

E4040.2023Fall.DGCN.report
Zidu Xu znx2000
Columbia University

Abstract

Drug repositioning provides promising prospects for accelerating drug discovery by identifying potential drug-disease associations (DDAs) for existing drugs/diseases. There are plenty of computational methods proposed for DDA prediction on drug-disease heterogeneous networks. However, they often overlook the enhancement of network construction from unstructured data sources, which could provide a more holistic view of DDA indications and improve model performance. Despite the promises, it is time-consuming and expensive to leverage traditional natural language processing (NLP) methods to exhaustively search through numerous biomedical knowledge base, intentionally extract useful information, and synthesis them for specific DDA usage. Notably, the rise of latest generation NLP methods - large language models (LLMs)- has opened the door to improved biomedical data exploitation and augmentation. Their sophisticated NLP capabilities allow efficient generation of contextually appropriate drug and disease descriptions, which are well-positioned to be a more informative input feature to enhance the power of DDA prediction models. As such, the author introduced a novel dual-channel layer attention graph convolutional network named LAGCN-LLM that incorporates LLM-generated descriptions as model inputs. LAGCN-LLM is an extension of baseline LAGCN. LAGCN-LLM contains a heterogeneous GCN-based node encoder to acquire node-level representations for drugs and diseases pairwise similarities and word embeddings of drug and diseases descriptions, two attention layers to integrate all useful structural information from multiple graph convolution layers, and a bidirectional instance encoder followed by a linear layer with pre-defined score function for DDA decisions. Our primary objective is to evaluate how LLM-generated embeddings enhance drug-disease prediction performances. We firstly utilized GPT-4 Turbo for text generation of drug and disease characteristics in a zero-shot manner, and then redesigned the GCN to incorporate dual-channel inputs that can leverage both topological and LLM-derived features. We compared baseline LAGCN with LAGCN-LLM on Comparative Toxicogenomics Database. Overall, we demonstrate our hypothesis, that the integration of LLM-generated feature can enhance the DDA through heterogenous GCN. The results show that LAGCN-LLM significantly outperforms LAGCN in AUPR (0.521 vs. 0.497, P=0.0009), accuracy (0.874 vs. 0.870, P=0.029), and precision (0.459 vs. 0.448, P=0.025), demonstrating the viability of LLMs in enhancing biomedical network augmentation and streamlining DDA identification for expedited drug discovery. The source code is available at the https://github.com/ecbme4040/e4040-2023fall-project-DGCN-znx2000/tree/master.

1. Introduction

New drug development has always been costly and time consuming, which on average takes 3 billion dollars over a 13-year cycle to examine the drug efficacy, toxicity, pharmacokinetic, and pharmacodynamic properties in vitro and human subject clinical trials, but usually end with low chance of success[1-5]. To overcome these challenges, computational drug repositioning has been recognized as a promising alternative [6, 7]. Aiming to identify new uses for existing drugs whose dosing, toxicity, and safety profiles are well-established, computational drug repositioning significantly reduces the drug safety examination cost and shortening the period of drug approval and launch[8-11]. There have been quite a few successful examples demonstrating the effectiveness of computational drug repositioning in accelerating the drug discovery process[9-13], such as repurposing Metformin for various neoplasm[14], which were originally for Type 2 Diabetes, and repurposing Thalidomide for Erythema Nodosum Leprosum and Multiple Myeloma[15], which was originally marketed as a sedative and for morning sickness during pregnancy.

The core idea of computational drug repositioning is to identify the new associations between known drugs and diseases. Statistical analysis and machine learning are two predominant computational approaches. Currently, machine learning-based computational drug repositioning has gained their popularity through the analysis of large scale heterogeneous data and identifying complex, non-linear patterns and relationships, which includes three main types[1, 16]: similarity/interaction network, matrix factorization/completion methods, and deep learning. Similarity/Interaction Networks are formed by connecting nodes (representing entities) with edges that represent either similarity (e.g., genetic similarity, structural similarity in drugs) or interactions (e.g., protein-protein interactions)[17]. The main issue of this method is the representation bias for nodes with high degrees in heterogeneous networks, which limits the performance. Matrix factorization/completion methods reconstruct a drug-disease association matrix into lower-dimensional matrices to uncover latent factors[18]. Despite the competitive performances, it suffers from limited effectiveness in representing drugs and diseases, especially in sparse association networks. Conversely, deep learning methods use neural networks to construct end-to-end frameworks for the representation learning of drugs and diseases in an integrated manner, enabling accurate predictions for drug-disease associations. Such framework allows enables accurate predictions for query drug-disease associations at the same

time, without the need for extensive manual feature engineering. Intuitively, graph convolutional network (GCN) [19-21], an extending convolutional neural networks for processing graph data, is readily embedded in end-to-end architectures to perform specific tasks with graph inputs, captures structural information of graphs via message passing between the nodes of graphs and retains high interpretability. For instance, LAGCN [22] developed a graph convolutional network with a layer attention mechanism to aggregate the output features in each layer for DDA prediction. DRHGCN [23] establishes a heterogeneous graph neural network for improved representation learning of drug and disease nodes.

GCNs have significantly advanced drug-disease association predictions with well-structured models and input data. However, their performance is often limited by the richness of input features, which typically rely on drug and disease pairwise similarities. These similarities are based on categorical characteristics or hierarchical classifications like the MeSH tree, necessitating richer input features. Thus, we propose integrating textual data, particularly drug and disease descriptions, to leverage semantic information for more accurate predictions. Traditional NLP in drug discovery, involving entity and relation extraction, and document classification, is often labor-intensive and depends on manually selected high-quality data sources[24]. Large language models (LLMs) like BERT[25] and GPT[26] excel in deep contextual understanding and text generation. The accessibility of these models through user-friendly interfaces like ChatGPT has made them even more popular. Their sophisticated NLP capabilities and understanding of biomedical terminologies could revolutionize biomedical data augmentation[27, 28], providing contextually rich descriptions. However, their application in computational drug repositioning remains unexplored.

To address this knowledge gap, we introduced LAGCN-LLM, an advanced version of the layer attention graph convolutional network (LAGCN). This method enhances drug-disease association predictions by incorporating a heterogeneous GCN-based node encoder, which processes drug-drug and disease-disease similarities along with drug and disease word embeddings. It also features dual attention layers for integrating structural information from multiple GCN layers, a bidirectional instance encoder for instance-level representation learning, and a linear layer with a predefined score function for predicting drug-disease associations.

LAGCN-LLM uniquely combines traditional topological data with text features from Large Language Models (LLMs) like GPT-4 Turbo, utilizing zero-shot settings for testing.

The study's main goal is to investigate the effectiveness of LLM generated embeddings in improving the drug-disease association predictions. Accordingly, we adopted the latest interactive LLM, GPT-4 in zero-shot settings to generate drug and disease descriptions, then add the converted text embeddings as a new model input for LAGCN-LLM. A comparative study using the Comparative Toxicogenomics Database assesses this approach, highlighting two key contributions: integrating LLM-generated text for domain-specific input and redesigning the GCN architecture for dual-channel integration of topological and LLM features. More details, source code, and datasets are available at https://github.com/ecbme4040/e4040-2023fall-project-DGCN-znx2000/tree/master.

2. Summary of the Original Papers

2.1 Methodology of the Original Paper

The original study proposes the Layer Attention Graph Convolutional Network (LAGCN) to predict drug-disease associations (**Figure 1**).

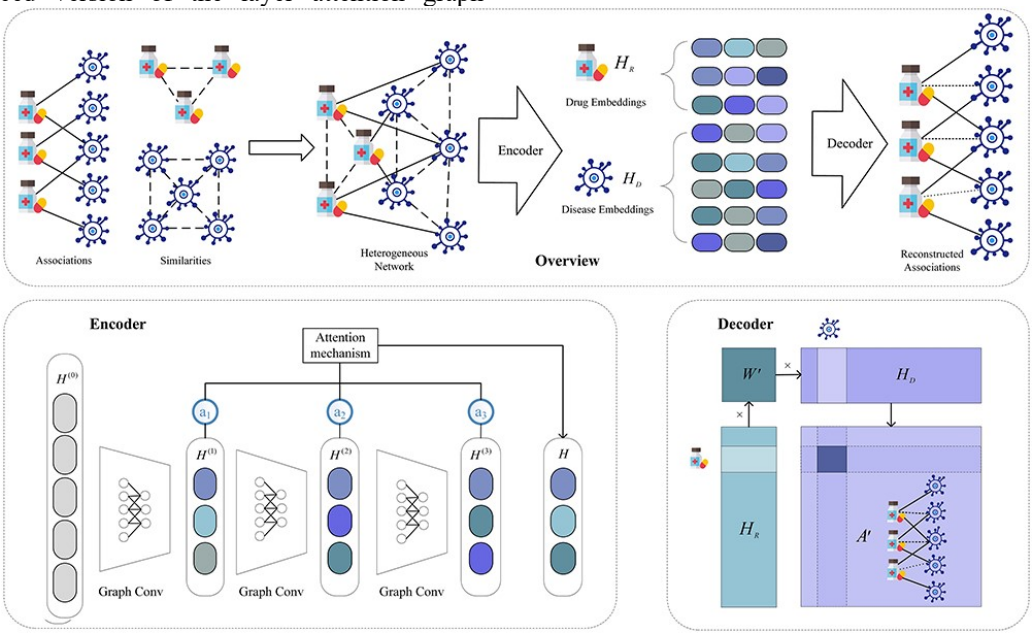

Figure 1. LAGCN model architecture.

2.1.1 Dataset

(1) a main dataset from the Comparative Toxicogenomics Database (CTD) with 18,416 drug-disease associations across 269 drugs and 598 diseases, complemented by detailed drug information from DrugBank[29] and disease terms from medical subject headings vocabulary (MeSH).

(2) Therapeutic dataset includes 6244 therapeutic associations annotated in CTD extracted from the main dataset.

2.1.2 Model Architecture

The LAGCN is an end-to-end deep learning method based on a heterogeneous GCN that captures structural information from a heterogeneous network of drug-disease, drug-drug, and disease-disease relationships. An attention mechanism integrates multi-layer embeddings for enriched node representation. The attention mechanism was used to resort and integrate all useful structural information from multiple graph convolution layers. Finally, the predictive scores for unobserved drug–disease associations are given by a well-defined score function based on the integrated embeddings. The workflow as well as the block diagrams of the original LAGCN are shown in **Figure 1**.

2.2 Key Results of the Original Paper

LAGCN achieved predictive performance metrics on the main dataset (CTD) being an AUPR of 0.3168, AUC of 0.8750, recall (RE) of 0.3600, specificity (SP) of 0.9760, accuracy (ACC) of 0.9605, and F1 score of 0.3150.

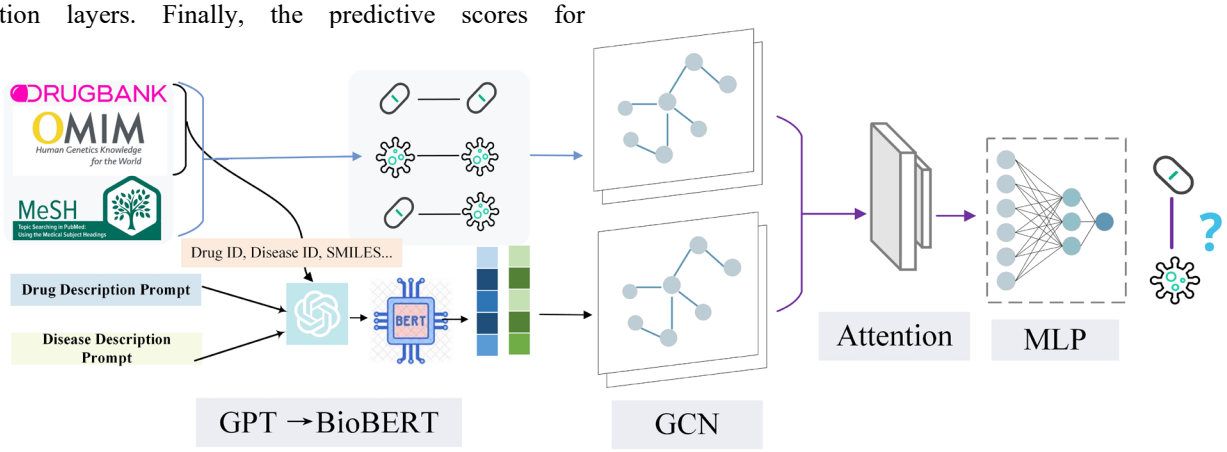

Figure 2. Flowchart of LAGCN-LLM.

3. Methodology

With the overarching goal to advance drug-disease association predictions by integrating LLM generated embeddings, the author proposed a novel dual-channel GCN architecture. Specifically, the added channel receives LLM generated embeddings derived from candidate drug and disease descriptions for DDA predictions. Distinct from the original LAGCN framework, the current methodology incorporates GPT-4 Turbo for textual feature generation and establishing a multi-view learning channel with an attention-based feature aggregation mechanism. The methods overview is shown in **Figure 2**.

3.1 Objectives and Technical Challenges

The primary objective is to evaluate the effectiveness of LLM-generated embeddings in improving the DDA prediction performances. To achieve this, we have two sub-aims: (1) to employ LLMs to create novel input features, and (2) to evolve the existing GCN structure into a more sophisticated dual-channel model.

During the implementation phase, this study faced several technical challenges. The author has listed them below, attached with corresponding solutions:

Challenge 1: Limited Training Data. Solution: Focused Scope with 5-Fold Cross-Validation. This study only used the main dataset as the original paper did, meanwhile conducting 5-fold cross-validation to maximize the utility of the available data.

Challenge 2: Large Language Model Implementation. Solution: ChatGPT API with GPT-4 Turbo. This study utilized the ChatGPT API, which offers an interactive platform that is manageable for teams with constrained computational capacity. GPT-4 Turbo was selected to ensure the ideal balance between performance and resource availability.

Challenge 3: GCN Implementation in older TensorFlow. Solution: Update to TensorFlow 2.X. The original LAGCN framework was built using TensorFlow1.X. This project converted GCN implementation from TensorFlow 1.X to the more contemporary TensorFlow 2.X framework, specifically version 2.10, under Python 3.9.

3.2 Problem Formulation and Design Description

3.2.1 Problem Formulation

The DDA prediction problem is formulated as a link prediction task within a heterogeneous graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{\mathcal{V}_r, \mathcal{V}_d\}$ is the node set comprising nodes representing drugs ($\mathcal{V}_r$) and diseases $\mathcal{V}_d$, and $\mathcal{E} = \{\mathcal{E}_{r-r}, \mathcal{E}_{r-d}, \mathcal{E}_{d-d}\}$ is the edge set, including edges denoting drug-drug, drug-disease, and disease-disease associations. The goal is to model a function $f_{DDA}(H_i, H_j)$ that estimates the association probability p for a given drug-disease pair, with H_i and H_j as their respective feature embeddings and $v_i \in \mathcal{V}_r, v_j \in \mathcal{V}_d$ are given drug-disease pair for prediction.

3.2.2 Algorithmic Implementation with Pseudo Code

As detailed in **Figure 2** and Algorithm Tables. The methodology includes a two-step data augmentation with LLM-generated and BioBERT-derived embeddings (**Algorithms 1**). This process involves constructing prompts for the LLM, generating textual descriptions, creating word embeddings, and incorporating these enhancements into the dataset. For model architecture, the GCN employs a dual-channel, layer attention, and a bilinear decoder to synthesize

and predict DDA associations, marking an improvement over the original LAGCN by introducing a multi-view learning channel and attention-based feature aggregation for more comprehensive representation learning (**Algorithms 2**).

Algorithm 1: EnhanceDataset - Data Augmentation and Embedding Process

Input: Dataset D with columns (Entity_Name, Entity_ID, Additional_Info), pre-trained Large Language Model (LLM) M, BioBERT model B.

Output: Enhanced dataset $D_{enhanced}$.

Initialize augmented dataset D_{aug} as an empty list ;
for *Channel 1 GCN layer* $l \leftarrow 1, ..., L$ **do**
  $H^{(l)} \leftarrow GCN(H^{(l-1)})$;
endfor
$H^{(L+1)} \leftarrow LayerAttention(\{H^{(1)}, H^{(2)}, ..., H^{(L)}\})$;
for *each row* r *in* D **do**
  Construct prompt for LLM M using r ;
  Generate description $Desc$ for r using M ;
  Generate word embeddings Emb for $Desc$ using BioBERT B ;
  Create augmented row r_{aug} by adding $Desc$ and Emb to r ;
  Append r_{aug} to D_{aug} ;
endfor
$D_{enhanced} \leftarrow D_{aug}$;
Output $D_{enhanced}$.

Algorithm 1. Data augmentation process.

Algorithm 2: The LAGCN_LLM algorithm

Input: Adjacency matrix A, initialized node similarity feature $H^{(0)}$ and node description embedding $H^{(LLM)}$.

Output: Reconstructed drug-disease association probability matrix $\hat{A}$.

Initialize the trainable parameters in LAGCN_LLM ;
for *Channel 1 GCN layer* $l \leftarrow 1, ..., L$ **do**
  $H^{(l)} \leftarrow GCN(H^{(l-1)})$;
endfor
$H^{(L+1)} \leftarrow LayerAttention(\{H^{(1)}, H^{(2)}, ..., H^{(L)}\})$;
for *Channel 2 GCN layer* $l \leftarrow 1, ..., L$ **do**
  $H^{(l)_{LLM}} \leftarrow GCN(H^{(l-1)_{LLM}})$;
endfor
$H^{(L+1)_{LLM}} \leftarrow LayerAttention(\{H^{(1)_{LLM}}, H^{(2)_{LLM}}, ..., H^{(L)_{LLM}}\})$;
$H^{(Node)} \leftarrow ChannelAttention(\{H^{(L+1)}, H^{(L+1)_{LLM}}\})$;
Reconstruct the drug-disease association probability matrix $\hat{A} \leftarrow Decoder(H_r^{(Node)}, H_d^{(Node)})$;
Update parameters ;
Ouput $\hat{A}$.

Algorithm 2. Computing flow of LAGCN-LLM.

4. Implementation

This section firstly elaborates the process to create the descriptions through zero- prompt applied to GPT-4 Turbo, then details the implementation steps LAGCB-LLM architecture (**Fig. 3)** comprising three core modules: dual-channel HGNN-based node encoder, layer attention, and bidirectional instance encoder, finally describes the validation settings and evaluation process.

4.1 Data

4.1.1 Biomedical Network Data from Literature

The biomedical network data include drug-drug associations determined by drug-drug a similarity matrix, disease-disease associations determined by a disease-disease similarity matrix, and drug-disease associations collected from wet-lab experiment articles. These data were manually obtained and extracted informed by Yu et al's work [30]. The biomedical network data encompasses 269 drugs and 598 diseases. The number of drug-drug and disease-disease associations can vary based on the binarization threshold value used.

4.1.2 Drug and Disease Descriptions

Two datasets were created containing names and IDs of 269 drugs and 598 diseases from Online Mendelian Inheritance in Man (OMIM)[31] and DrugBank [29]. In addition, each drug and disease entity is accompanied by a unique description, outlining its characteristics like associated drugs, genes and signaling pathways. These descriptions were transformed into word embeddings using BioBERT to enhance domain-specific representation learning in the model.

4.1.3 Data Augmentation with ChatGPT

ChatGPT, an autoregressive language model built on transformer decoder blocks[32], is pre-trained in an unsupervised manner on a dataset $X = \{x_1, x_2, \dots, x_n\}$, with each sample x_i consists of m tokens $x_i = \{s_1, s_2, \dots, s_m\}$. The pre-training objective is to maximize the following likelihood:

$$L(x_i) = \sum_{i=1}^{m} \log P(s_i | s_1, \dots, s_{i-1}; \theta) \quad (1)$$

Here, θ represents ChatGPT trainable parameters. The model uses token embeddings and position embeddings expressed as:

$$h_0 = x_i W_e + W_p \quad (2)$$

where W_e is the token embedding matrix; W_p is the position embedding matrix. The embeddings are processed through N transformer blocks to extract features from the sample:

$$h_n = transformer\_blocks(h_{n-1}) \quad (3)$$

where $n \in [1, N]$. Finally, ChatGPT generates predictions on the final hidden state:

$$s_i = softmax(h_N W_e^T) \quad (4)$$

Differing from open-source LLMs like GPT-3, ChatGPT's training process involves AI trainers acting as both users and assistants to refine responses via prompts[28]. For this study, a zero-shot prompting technique was chosen for efficiency, designed to exploit LLMs' core capabilities of comprehension, reasoning, and explanation. As shown in **Table 1- 2**, ChatGPT is instructed to mimic the expertise of scientists in relevant domains. The bullet points of key information are tailored to each task, ensuring relevancy to the drug-disease prediction task.

The prompts aim to: (1) direct the LLM to construct responses infused with domain-specific knowledge, and (2) prevent the generation of hallucinated information[33, 34]. To achieve aim 1, a detailed template is implemented to list the expected elements contained in responses, emphasizing gene, signaling pathway, and associated diseases and drugs, along with a 'role' parameter directing to focused context. To achieve aim 2, the author uses constraint words like "precise", "with examples", and "evidence-based," and instructed the model to default to "not available" for queries beyond its knowledge scope, thus enhancing the response confidence and relevancy.

Table 1. Disease Description Prompt.

Prompt Part	Content
Beginning	"Generate a single, cohesive, narrative paragraph for the disease '**{disease_name}**' associated with OMIM ID '**{omim_id}**'." The response should include **10** key information as follows:
Key Information	1) Associated **genes, proteins, or mutations** (3 examples). 2) Associated **signal pathway** (key molecular/cellular components). 3) Associated **drugs** for treatment (3 examples with mechanisms of action) 4) Linked **comorbidities** and **complications**. 5) **Nature** of the disease. 6) Typical clinical **symptoms** and signs. 7) **Types** of the disease. 8) **Inheritance** patterns and **genetic** components (examples). 9) **Diagnostic criteria** and testing methods.
End	"If no specific answer, just return not available. The information does not need to be current or from a live database. Ensure the final summary is precise, evidence-based, suitable for a professional medical audience, and condenses all the points above into a coherent narrative."

Note: comorbidity and complication are two key points.

4.1.4 Embedding Generation with BioBERT

Upon obtaining drug and disease descriptions, the author utilizes BioBERT[35], a BERT variant trained on extensive biomedical literature, to generate embeddings. BioBERT, pre-trained on a vast corpus of biomedical literature, is adept at capturing the nuanced, context-rich representations that are essential for biomedical domain.

Table 2. Drug Description Prompt.

Prompt Part	Content
Beginning	"Generate a single, comprehensive paragraph for the drug '**{drug_name}**' associated with its DrugBank ID '**{drug_id}**', and its SMILES (Simplified Molecular Input Line Entry System) notation '**{SMILES_note}**'." The response should include **10** key information as follows:
Key Information	1) Detailed description of its **chemical structure**. 2) **Chemical category**. 3) **Chemical scaffold**. 4) Known similar **drugs** (examples). 5) **Pharmacokinetics** (absorption, distribution, metabolism, excretion). 6) **Toxicity** details (examples). 7) List of target **proteins**. 8) **Indications** (diseases/symptoms examples). 9) **Side effects** (examples). 10) Clinical **usage** (examples).
End	"If no specific answer, just return not available. The information does not need to be current or from a live database. Ensure the final summary is precise, evidence-based, suitable for a professional medical audience, and condenses all the points above into a coherent narrative."

4.2 Construction of Drug-Disease Network

Denoting drug as r and disease as d, the drug-disease network requires the drug-disease networks to be trained on, which include drug-drug associations S_{r-r}, drug-disease associations A_{r-d}, and disease-disease associations S_{d-d}.

4.2.1 Drug-drug Similarities

The drug-drug similarities were obtained from the original work. Specifically, the similarity was first calculated by measuring the pairwise similarities between drug features (Target, Enzyme, Drug–drug interactions, Pathway, Substructure) using the Jaccard similarity index.

For each pair of drugs r_i and r_j, the drug-drug similarity is calculated as:

$$S_{ij}^R = \frac{|F_{r_i} \cap F_{r_j}|}{|F_{r_i} \cup F_{r_j}|} \quad (5)$$

F_{r_i} and F_{r_j} are feature sets for drugs r_i and r_i, respectively.

4.2.2 Disease-Disease Similarities

The disease-disease associations were also from original work. They are calculated based on MeSH semantic

similarities represented as a hierarchical directed acyclic graph (DAG). Given a disease d, the DAG can be represented as $\mathrm{DAG}(d) = (\mathcal{N}(d), \mathcal{E}(d))$, where $\mathcal{N}(d)$ denotes the set of nodes including d and its ancestral nodes, and $\mathcal{E}(d)$ denotes the parent-child relation links among $\mathcal{N}(d)$. The semantic contribution of a node $n \in \mathcal{N}(d)$ for d can be formulated as:

$$C_d(n) = \begin{cases} 1, & if\ n = d \\ max\{\Delta \cdot C_d(n') | n' \in children\ of\ n\}, & otherwise \end{cases} \quad (6)$$

where Δ is a contribution factor. The overall semantic contribution $DV(d) = \sum_{n \in \mathcal{N}(d)} C_d(n)$. The disease pairwise similarity of d_i and d_j can be measured by the number of common ancestral nodes and the semantic contribution proportion of these ancestral nodes in $DV(\mathrm{d}_i)$ and $DV(\mathrm{d}_j)$:

$$S_{ij}^{D} = \frac{\sum_{n \in \mathcal{N}(\mathrm{d}_i) \cap \mathcal{N}(\mathrm{d}_j)} \left(C_{\mathrm{d}_i}(n) + C_{\mathrm{d}_j}(n) \right)}{DV(\mathrm{d}_i) + DV(\mathrm{d}_j)} \quad (7)$$

Here, $S^D \in \{0,1\}$.

4.2.3 Drug-Disease Associations

For heterogeneous graph neural network, the drug-disease network is a heterogeneous graph data. The drug-disease graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{\mathcal{V}_r, \mathcal{V}_d\}$ is the node set and $\mathcal{E} = \{\mathcal{E}_{r-r}, \mathcal{E}_{r-d}, \mathcal{E}_{d-d}\}$ is the edge set. $\mathcal{V}$ and $\mathcal{E}$ can be represented as a node feature matrix $H^{(0)} \in \mathbb{R}^{(N+M)\times(N+M)}$ and an adjacency matrix $A \in \mathbb{R}^{(N+M)\times(N+M)}$:

$$A = \begin{bmatrix} S_{r-r} & A_{r-d} \\ (A_{r-d})^{\mathrm{T}} & S_{d-d} \end{bmatrix} \quad (8)$$

$$H^{(0)} = \begin{bmatrix} H_r^{(0)} \\ H_d^{(0)} \end{bmatrix} = \begin{bmatrix} S_{r-r} & 0 \\ 0 & S_{d-d} \end{bmatrix} \quad (9)$$

Here, $H^{(0)}$ and A are taken as the input of GCN for model training and DDA prediction. An entry in A is "1" if there is an association, and "0" otherwise.

The added value from this study, in this step, is to introduce a new input channel for $H^{(0)} \in \mathbb{R}^{(N+M)\times(768)}$, with LLM and BioBERT generated embeddings.

4.2.4 Dataset Splitting

During the training phrase, positive associations in A (entries marked as '1') are masked as '0' to prevent data leakage. During the testing phase, the model predicts the masked associations using the indices of the test data. This process forces the model to infer unseen drug-disease relationships that were not available during the training phase.

4.3 LAGCN-LLM Mode Architecture

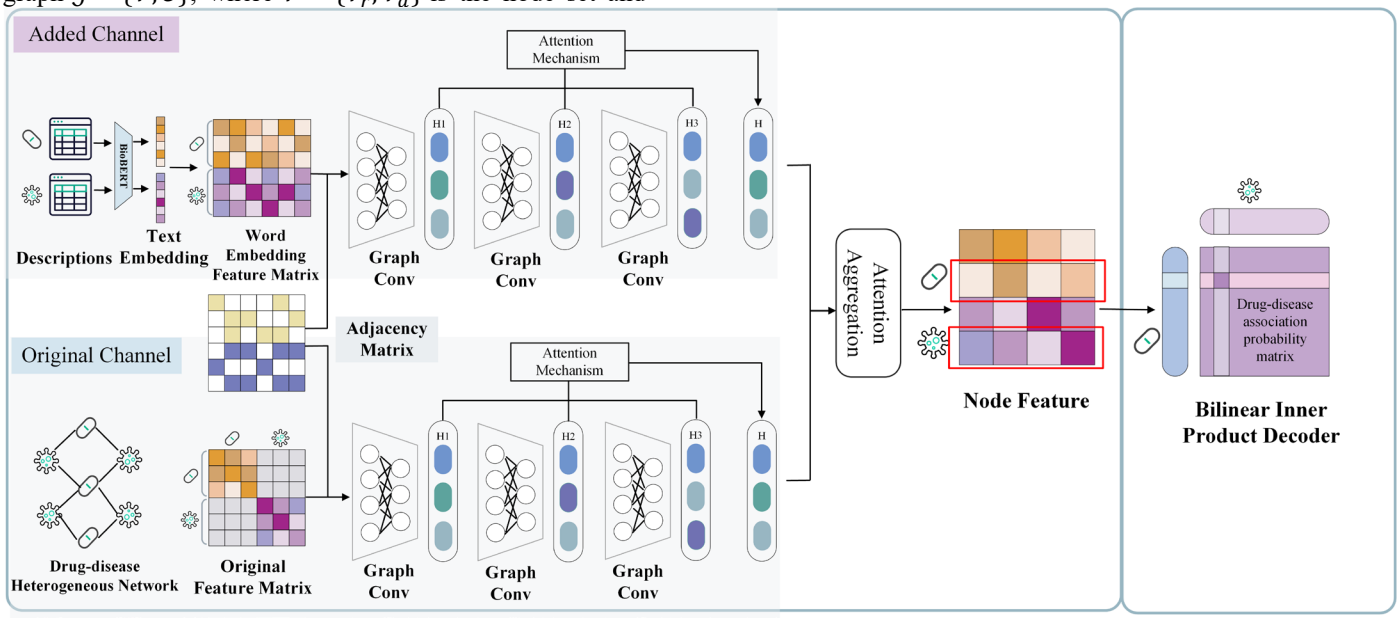

Figure 3. LAGCN-LLM model architecture.

4.3.1 Heterogeneous Graph Convolutional Network

GCN is a neural network architecture designed for learning node representations in graph-structured data. In heterogeneous GCNs, nodes (either drugs or diseases) in homogeneous graphs aggregate information from their neighboring nodes of various types. This aggregation is achieved through layer-wise propagation using the network's adjacency matrix, enabling the learning and updating of node embeddings in heterogeneous graphs.

Given a homogeneous graph $\mathcal{G}^{\sigma-\sigma}$ whose node type is σ, node representations $\widehat{H}^{\sigma}$ are learned by a classic GCN in $\mathcal{G}^{\sigma-\sigma}$. The GCN layer computation is expressed as:

$$\widehat{H}^{\sigma} = GCN(A^{\sigma-\sigma}, H^{\sigma}, W^{\sigma}) \quad (10)$$

where $A^{\sigma-\sigma}$ is the adjacency matrix of the σ-σ interaction network, H^{σ} is the input node embedding, and W^{σ} is the trainable parameter matrix.

Graph convolution in each GCN layer is performed on every homogeneous graph, including drug-drug and disease-

disease graphs, facilitating message passing and node embedding aggregation. The computation for the l-th HGNN layer and a homogeneous graph $\mathcal{G}_{v-v}$, where $v \in \{r, d\}$ is given by:

$$\widetilde{H}_v^{(l)} = \sigma(\widehat{D}_v^{-\frac{1}{2}} \hat{A}_{v-v} \widehat{D}_v^{-\frac{1}{2}} H_v^{(l-1)} W) \quad (11)$$

$W \in \mathbb{R}^{D \times D}$ is a trainable parameter matrix. In the dual-channel GCN, the graph convolution operations are distinctly applied through two channel inputs. For the original channel from LAGCN, $\hat{A}_{v-v} \in \mathbb{R}^{(N+M)\times(N+M)}$ denotes the adjacency matrix plus the identify matrix, $\widehat{D}_v \in \mathbb{R}^{(N+M)\times(N+M)}$ is the degree matrix. For the new channel with LLM-generated text embeddings, graph convolution incorporates text embeddings with network topology to learn semantic features. This is because the unsupervised embeddings created by generative AI need to add supervised signal to enhance representation, graph convolution operation was applied: $\hat{A}_{v-v} \in \mathbb{R}^{(N+M)\times(768)}$ and $\widehat{D}_v \in \mathbb{R}^{(N+M)\times(768)}$ is the degree matrix.

After finishing the graph convolution on homogeneous graphs, node embeddings are aggregated heterogeneously using a summation process. Given node v_i and its neighbor node set $\mathcal{N}_{v_i}$, the node embedding $H_{v_i}^{(l)} \in \mathbb{R}^D$ in l-th heterogeneous GCN layer is:

$$\widehat{H}_{v_i}^{(l)} = \sum_{v_j \in \mathcal{N}_{v_i}} \widetilde{H}_{v_j}^{(l)} \quad (12)$$

4.3.2 Layer Attention Mechanism

Similar to LAGCN, an attention mechanism is incorporated to prioritize information from different layers. This is crucial to address the varying importance of embeddings across layers. To prevent over-smoothing, like residual connections in CNNs, a layer attention mechanism dynamically aggregates node embeddings from each layer. The importance weight $w_{v_i}^{(l)}$ for node embedding $H_{v_i}^{(l)}$ from l-th HGNN layer can be formulated as:

$$w_{v_i}^{(l)} = \frac{1}{L} \sum_{l \in L} q^{\mathrm{T}} W \widehat{H}_{v_i}^{(l)} \quad (13)$$

where $q \in \mathbb{R}^D$ and $W \in \mathbb{R}^{D \times D}$ are trainable parameter matrixes. The normalized attention coefficient $\alpha_{v_i}^{(l)} \in (0,1)$ is obtained using a SoftMax function:

$$\alpha_{v_i}^{(l)} = \frac{\mathrm{Exp}\left(w_{v_i}^{(l)}\right)}{\sum_{l=1}^{L} \mathrm{Exp}\left(w_{v_i}^{(l)}\right)} \quad (14)$$

The final node embedding $H_{v_i}^{(Node)} \in \mathbb{R}^D$ for node v_i is a weighted sum of layer embeddings:

$$H_{v_i}^{(Node)} = \sum_{l=1}^{L} \alpha_{v_i}^{(l)} \widehat{H}_{v_i}^{(l)} \quad (15)$$

4.3.3 Bilinear Inner Product Decoder

The decoder layer employs a dot product between the drug feature matrix D $\in \mathbb{R}^{(Ndrug)\times(Ndim)}$, and the transpose of the disease feature matrix E$\in \mathbb{R}^{(Ndim)\times(Ndisease)}$, followed by a linear layer to reconstruct the drug-disease association matrix A $\in \mathbb{R}^{(Ndrug)\times(Ndisease)}$. This reconstruction is described by the equation:

$$\hat{A} = f(H^R, H^D) = sigmoid(H^R W (H^D)^T) \quad (16)$$

where $W \in \mathbb{R}^{K \times K}$ is the trainable parameter matrix and $\hat{A}$ is the reconstructed drug-disease association matrix.

4.4 Optimization

The model uses a weighted cross-entropy loss function to balance different categories and focus on known drug-disease associations. This function is formulated as:

$$\mathcal{L} = -\frac{1}{N}\left(\gamma \sum_{(i,j)\in S^+} log\hat{A}_{ij} + \sum_{(i,j)\in S^-} (1 - log\hat{A}_{ij})\right) \quad (17)$$

where $\gamma = \frac{|S^-|}{|S^+|}$ is the balance weight, $|S^+|$ and $|S^-|$ are the number of known/unknown drug-disease associations in the training set, and $\hat{A}_{ij}$ is the predicted probability of drug i and disease j.

The Adam optimizer for model optimization and initialize the trainable parameters in each layer by Xavier [36]. Moreover, the dropout layer and batch normalization layer are also adopted to inhibit overfitting. Dropout is included to prevent overfitting, and a cyclic learning rate scheduler enhances training stability.

4.5 Experimental settings

To estimate the performance of our updated model and the original LAGCN, we execute 5-fold cross-validations then repeated 10 times to minimize random error from data splitting. Due to label imbalance in drug-disease association predictions, various metrics including AUC, AUPR, F1-Score, Accuracy, Recall, Specificity, and Precision were used for assessment. Model hyperparameter settings were sourced from original literature to evaluate model performance.

5. Results

5.1 Project Results

The author used a model comparison experiment to demonstrate the superiority of LAGCN-LLM. The AUC, AUPR, F1-Score, Accuracy, Recall, Specificity, and Precision of 5-fold cross validations from 10-time experiment were listed in **Table 3** in the format of mean (standard error). Two-sample t-tests were conducted to compare the difference of metrics across two groups. The corresponding ROC and Precision-Recall curves along with the 95% CI were shown in **Fig. 4**. The performance metrics of 10 individual experiments are provided in the supplement Table III and Table IV.

Overall, LAGCN-LLM outperformed the original LAGCN on all average metrics except AUC. Among the superior performed metrics, there were statistically significant differences in AUPR between (0.521 vs. 0.497, P =0.0009), Accuracy (0.874 vs. 0.870, P = 0.029), and

Precision LAGCN-LLM and LAGCN (0.459 vs. 0.448, P = 0.025). There was no statistically significant difference in AUC across two groups (0.848 vs. 0.857). The results of average metrics and statistical tests demonstrated that LAGCN-LLM continuously performed with solidarity and superiority.

5.2 Comparison of the Results Between the Original Paper and Students' Project

Compared to the original LAGCN, this study introduced new input features using GPT-4 Turbo and redefining the model architecture with a dual-channel approach. The comparison experiment demonstrated the potential benefits of this modification. The use of narrative format features from the large language model and BioBERT-processed text embeddings, can add significant value to traditional tabular features, enriching input data with substantial biomedical domain knowledge and semantic context. Secondly, the refined dual-channel architecture in the new model was demonstrated to effectively integrate both topological and LLM features for a more comprehensive analysis.

5.3 Discussion / Insights Gained

This study has several strengths, such as using LLM-generated embeddings for feature augmentation, which likely improved drug-disease association predictions. Additionally, the attention mechanism for aggregating embeddings from dual channels could efficiently capture key predictors for these associations. However, we acknowledge there are several limitations. The main concern is the generalizability of results due to reliance on a single dataset for cross-validation. Without testing on broader datasets, claiming the superiority of LAGCN_LLM remains tentative. Additionally, as a preliminary study, only zero-shot prompts for LLM effectiveness was tested, and the current model architecture may have not fully optimize the integration of LLM embeddings. Further experimentation with diverse datasets and prompt formats is needed to confirm the updated model's superiority and optimal solutions.

Table 3. The AUC, AUPR, and F1-Score Results of LAGCN-LLM and LAGCN on Comparative Toxicogenomics Database in 5-fold Cross Validations.

Metric	LAGCN	LAGCN_LLM	Statistical Difference
AUPR	0.497 (0.002)	**0.521 (0.019)**	3.972 (P =0.0009)
AUC	**0.857 (0.000)**	0.848 (0.023)	-1.154 (P =0.264)
F1-Score	0.498 (0.001)	**0.506 (0.012)**	1.934 (P =0.069)
Accuracy	0.870 (0.003)	**0.874 (0.004)**	2.378 (P =0.029)
Recall	0.562 (0.012)	**0.564 (0.019)**	0.284 (P =0.779)
Specificity	0.910 (0.005)	**0.914 (0.004)**	1.801 (P =0.088)
Precision	0.448 (0.008)	**0.459 (0.012)**	2.448 (P =0.025)

Footnote: The best result in each row is in **bold faces**. Statistically significant differences are underlined.

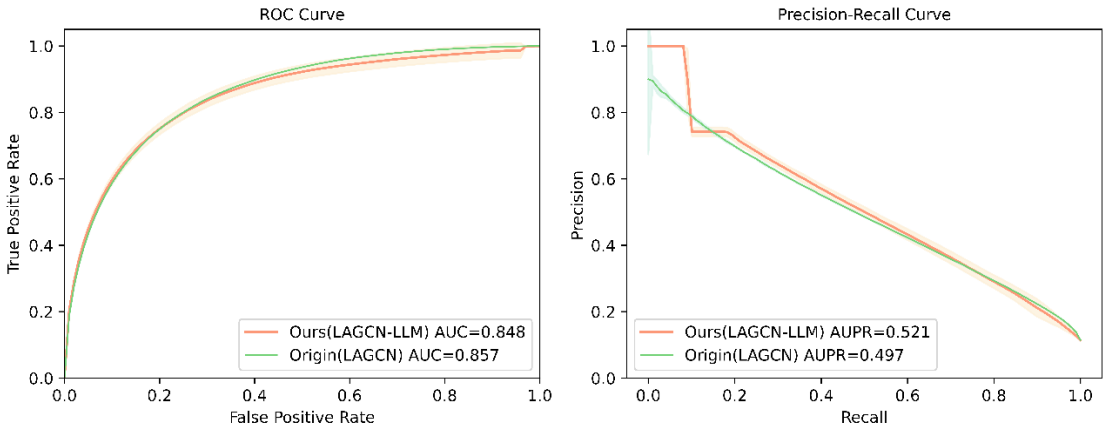

Figure. 4.The AUC and AUPR results of LAGCN-LLM and original LAGCN.

6. Future Work

In alignment with the abovementioned limitations, future work will focus on expanding model training and testing on a broader range of drug and disease datasets, enhancing the robustness of our Heterogeneous GCN. Additionally, exploration of more LLM variations and prompt engineering approaches, particularly focusing on biomedical-specific LLMs, is planned. Also, advanced GNN models will be considered for improved node representation, moving beyond the classic GCN currently used.

7. Conclusion

This study introduces LAGCN-LLM, a dual-channel layer attention graph convolutional network for drug-disease association prediction. It innovatively combines graph node feature data with text embeddings from GPT-4 Turbo and BioBERT-processed descriptions of drugs and diseases. The updated model architecture integrates topological and textual-sourced features, demonstrating superior performance on the CTD dataset. Future efforts will focus on optimizing LLM prompts and expanding dataset trials to bolster the heterogonous GCN's generalizability and efficiency in drug repositioning, ultimately accelerating drug discovery.

8. Acknowledgement

I am thankful to the instructor of E4040 Deep Learning & Neural Networks, Mehmet Kerem Turkcan, and all the teaching assistants for their insightful guidance during the 2023 fall semester. Additionally, I appreciate the LAGCN framework provided at https://github.com/storyandwine/LAGCN, which was crucial for my project's implementation phase.

9. References

[1] N. Zong, A. Wen, S. Moon, S. Fu, L. Wang, Y. Zhao, Y. Yu, M. Huang, Y. Wang, G. Zheng, M. M. Mielke, J. R. Cerhan, and H. Liu, "Computational drug repurposing based on electronic health records: a scoping review," *npj Digital Medicine,* vol. 5, no. 1, pp. 77, 2022/06/14, 2022.

[2] H. C. S. Chan, H. Shan, T. Dahoun, H. Vogel, and S. Yuan, "Advancing Drug Discovery via Artificial Intelligence," *Trends Pharmacol Sci,* vol. 40, no. 8, pp. 592-604, Aug, 2019.

[3] V. Prasad, and S. Mailankody, "Research and Development Spending to Bring a Single Cancer Drug to Market and Revenues After Approval," *JAMA Intern Med,* vol. 177, no. 11, pp. 1569-1575, Nov 1, 2017.

[4] J. A. DiMasi, H. G. Grabowski, and R. W. Hansen, "Innovation in the pharmaceutical industry: New estimates of R&D costs," *J Health Econ,* vol. 47, pp. 20-33, May, 2016.

[5] C. H. Wong, K. W. Siah, and A. W. Lo, "Estimation of clinical trial success rates and related parameters," *Biostatistics,* vol. 20, no. 2, pp. 273-286, Apr 1, 2019.

[6] M. R. Hurle, L. Yang, Q. Xie, D. K. Rajpal, P. Sanseau, and P. Agarwal, "Computational drug repositioning: from data to therapeutics," *Clin Pharmacol Ther,* vol. 93, no. 4, pp. 335-41, Apr, 2013.

[7] J. Li, S. Zheng, B. Chen, A. J. Butte, S. J. Swamidass, and Z. Lu, "A survey of current trends in computational drug repositioning," *Brief Bioinform,* vol. 17, no. 1, pp. 2-12, Jan, 2016.

[8] J. S. Shim, and J. O. Liu, "Recent advances in drug repositioning for the discovery of new anticancer drugs," *Int J Biol Sci,* vol. 10, no. 7, pp. 654-63, 2014.

[9] K. Mohamed, N. Yazdanpanah, A. Saghazadeh, and N. Rezaei, "Computational drug discovery and repurposing for the treatment of COVID-19: A systematic review," *Bioorg Chem,* vol. 106, pp. 104490, Jan, 2021.

[10] J. I. Traylor, H. E. Sheppard, V. Ravikumar, J. Breshears, S. M. Raza, C. Y. Lin, S. R. Patel, and F. DeMonte, "Computational Drug Repositioning Identifies Potentially Active Therapies for Chordoma," *Neurosurgery,* vol. 88, no. 2, pp. 428-436, Jan 13, 2021.

[11] L. Bai, M. K. D. Scott, E. Steinberg, L. Kalesinskas, A. Habtezion, N. H. Shah, and P. Khatri, "Computational drug repositioning of atorvastatin for ulcerative colitis," *J Am Med Inform Assoc,* vol. 28, no. 11, pp. 2325-2335, Oct 12, 2021.

[12] G. Fahimian, J. Zahiri, S. S. Arab, and R. H. Sajedi, "RepCOOL: computational drug repositioning via integrating heterogeneous biological networks," *J Transl Med,* vol. 18, no. 1, pp. 375, Oct 2, 2020.

[13] C. Budak, V. Mençik, and V. Gider, "Determining similarities of COVID-19 - lung cancer drugs and affinity binding mode analysis by graph neural network-based GEFA method," *J Biomol Struct Dyn*, pp. 1-13, Dec 8, 2021.

[14] Z. Zhang, L. Zhou, N. Xie, E. C. Nice, T. Zhang, Y. Cui, and C. Huang, "Overcoming cancer therapeutic bottleneck by drug repurposing," *Signal transduction and targeted therapy,* vol. 5, no. 1, pp. 113, 2020.

[15] S. Pushpakom, F. Iorio, P. A. Eyers, K. J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer, C. McNamee, A. Norris, P. Sanseau, D. Cavalla, and M. Pirmohamed, "Drug repurposing: progress, challenges and recommendations," *Nat Rev Drug Discov,* vol. 18, no. 1, pp. 41-58, Jan, 2019.

[16] H. Luo, M. Li, M. Yang, F. X. Wu, Y. Li, and J. Wang, "Biomedical data and computational models for drug repositioning: a comprehensive review," *Brief Bioinform,* vol. 22, no. 2, pp. 1604-1619, Mar 22, 2021.

[17] M. Zhou, C. Zheng, and R. Xu, "Combining phenome-driven drug-target interaction prediction with patients' electronic health records-based clinical corroboration toward drug discovery," *Bioinformatics,* vol. 36, no. Suppl_1, pp. i436-i444, Jul 1, 2020.

[18] W. Zhang, X. Yue, W. Lin, W. Wu, R. Liu, F. Huang, and F. Liu, "Predicting drug-disease associations by using similarity constrained matrix factorization," *BMC Bioinformatics,* vol. 19, no. 1, pp. 233, Jun 19, 2018.

[19] J. Jumper, R. Evans, A. Pritzel, T. Green, M. appemsusuppurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," *Nature,* vol. 596, no. 7873, pp. 583-589, Aug, 2021.

[20] T. Ma, Q. Liu, H. Li, M. Zhou, R. Jiang, and X. Zhang, "DualGCN: a dual graph convolutional network model to predict cancer drug response," *BMC Bioinformatics,* vol. 23, no. Suppl 4, pp. 129, Apr 15, 2022.

[21] Q. Liu, Z. Hu, R. Jiang, and M. Zhou, "DeepCDR: a hybrid graph convolutional network for predicting cancer drug response," *Bioinformatics,* vol. 36, no. Suppl_2, pp. i911-i918, Dec 30, 2020.

[22] Z. Yu, F. Huang, X. Zhao, W. Xiao, and W. Zhang, "Predicting drug-disease associations through layer attention graph convolutional network," *Brief Bioinform,* vol. 22, no. 4, Jul 20, 2021.

[23] L. Cai, C. Lu, J. Xu, Y. Meng, P. Wang, X. Fu, X. Zeng, and Y. Su, "Drug repositioning based on the heterogeneous information fusion graph convolutional network," *Brief Bioinform,* vol. 22, no. 6, Nov 5, 2021.

[24] R. Bhatnagar, S. Sardar, M. Beheshti, and J. T. Podichetty, "How can natural language processing help model informed drug development?: a review," *JAMIA Open,* vol. 5, no. 2, pp. ooac043, 2022.

[25] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[26] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training,"

2018.
[27] C. Coulombe, “Text data augmentation made simple by leveraging nlp cloud apis,” *arXiv preprint arXiv:1812.04718*, 2018.
[28] H. Dai, Z. Liu, W. Liao, X. Huang, Y. Cao, Z. Wu, L. Zhao, S. Xu, W. Liu, and N. Liu, “AugGPT: Leveraging ChatGPT for Text Data Augmentation. arXiv 2023,” *arXiv preprint arXiv:2302.13007,* vol. 10.
[29] D. Online, "DrugBank," 2023.
[30] Z. Yu, F. Huang, X. Zhao, W. Xiao, and W. Zhang, “Predicting drug–disease associations through layer attention graph convolutional network,” *Briefings in Bioinformatics,* vol. 22, no. 4, pp. bbaa243, 2021.
[31] J. H. University, "Online Mendelian Inheritance in Man," 2023.
[32] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, California, USA, 2017, pp. 6000–6010.
[33] T. Guo, K. Guo, B. Nan, Z. Liang, Z. Guo, N. V. Chawla, O. Wiest, and X. Zhang, "What can Large Language Models do in chemistry? A comprehensive benchmark on eight tasks."
[34] J. Wang, E. Shi, S. Yu, Z. Wu, C. Ma, H. Dai, Q. Yang, Y. Kang, J. Wu, and H. Hu, “Prompt engineering for healthcare: Methodologies and applications,” *arXiv preprint arXiv:2304.14670*, 2023.
[35] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “BioBERT: a pre-trained biomedical language representation model for biomedical text mining,” *Bioinformatics,* vol. 36, no. 4, pp. 1234-1240, 2020.
[36] X. Glorot, and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks." pp. 249-256.

10. Appendix

Table 1 Individual Experiment Results of LAGCN-LLM

AUPR	AUC	F1 Score	Accuracy	Recall	Specificity	Precision
0.511536	0.842654	0.4985	0.870081	0.56402	0.90965	0.446618
0.524173	0.856465	0.508347	0.876242	0.558862	0.917274	0.466208
0.475422	0.78718	0.482543	0.871101	0.524978	0.915849	0.446456
0.519282	0.8517	0.501769	0.874793	0.550717	0.916691	0.460811
0.51982	0.854009	0.500129	0.867955	0.576998	0.905571	0.441334
0.520642	0.854132	0.503396	0.870442	0.573577	0.908822	0.448516
0.539174	0.866005	0.522755	0.876335	0.591605	0.913146	0.46826
0.516482	0.843689	0.500919	0.875048	0.54773	0.917365	0.461479
0.536496	0.86464	0.518895	0.877013	0.579333	0.915498	0.469876
0.542908	0.863032	0.521945	0.879338	0.575369	0.918636	0.477598

Table 2 Individual Experiment Results of LAGCN

AUPR	AUC	F1 Score	Accuracy	Recall	Specificity	Precision
0.49716	0.856169	0.498799	0.872872	0.552563	0.914283	0.45457
0.496312	0.856718	0.497548	0.870075	0.561903	0.909917	0.446419
0.499284	0.856657	0.499626	0.867054	0.579768	0.904195	0.438949
0.497435	0.856375	0.496046	0.871642	0.551803	0.912992	0.450523
0.495725	0.856856	0.499041	0.873407	0.550771	0.915119	0.456193
0.496213	0.856654	0.49929	0.872947	0.553323	0.914269	0.45487
0.499697	0.856941	0.498358	0.872723	0.552237	0.914157	0.454058
0.493187	0.856232	0.497065	0.867912	0.570156	0.906407	0.440584
0.493894	0.857552	0.498223	0.865717	0.58232	0.902356	0.435351
0.496902	0.857009	0.498677	0.869298	0.567821	0.908274	0.444544