ORIGINAL RESEARCH ARTICLE

Empowering Graph Neural Network-Based Computational Drug Repositioning with Large Language Model-Inferred Knowledge Representation
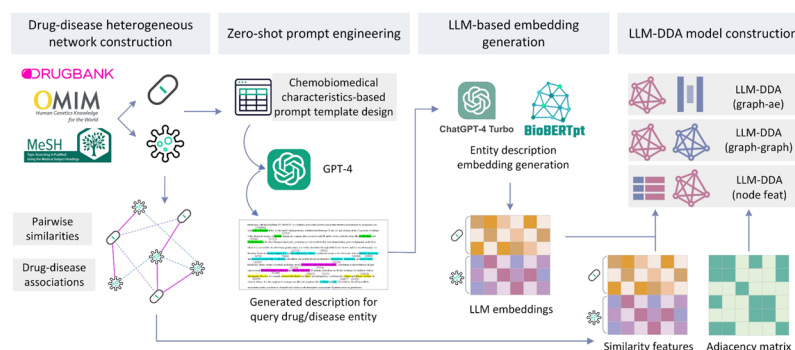
Yaowen Gu[1] · Zidu Xu[2] · Carl Yang[3]

Abstract

Computational drug repositioning, through predicting drug-disease associations (DDA), offers significant potential for discovering new drug indications. Current methods incorporate graph neural networks (GNN) on drug-disease heterogeneous networks to predict DDAs, achieving notable performances compared to traditional machine learning and matrix factorization approaches. However, these methods depend heavily on network topology, hampered by incomplete and noisy network data, and overlook the wealth of biomedical knowledge available. Correspondingly, large language models (LLMs) excel in graph search and relational reasoning, which can possibly enhance the integration of comprehensive biomedical knowledge into drug and disease profiles. In this study, we first investigate the contribution of LLM-inferred knowledge representation in drug repositioning and DDA prediction. A zero-shot prompting template was designed for LLM to extract high-quality knowledge descriptions for drug and disease entities, followed by embedding generation from language models to transform the discrete text to continual numerical representation. Then, we proposed LLM-DDA with three different model architectures (LLM-DDA$_{Node Feat}$, LLM-DDA$_{Dual GNN}$, LLM-DDA$_{GNN-AE}$) to investigate the best fusion mode for LLM-based embeddings. Extensive experiments on four DDA benchmarks show that, LLM-DDA$_{GNN-AE}$ achieved the optimal performance compared to 11 baselines with the overall relative improvement in AUPR of 23.22%, F1-Score of 17.20%, and precision of 25.35%. Meanwhile, selected case studies of involving Prednisone and Allergic Rhinitis highlighted the model's capability to identify reliable DDAs and knowledge descriptions, supported by existing literature. This study showcases the utility of LLMs in drug repositioning with its generality and applicability in other biomedical relation prediction tasks.

Graphical abstract

Keywords Drug repositioning · Drug-disease association prediction · Heterogeneous graph neural network · Large Language Model

---

Yaowen Gu and Zidu Xu contributed equally.

Extended author information available on the last page of the article

1 Introduction

Drug development has always been costly and time consuming, averagely taking 3 billion dollars over a 13-year cycle usually end with low chance of success [1–5]. Computational drug repositioning has been recognized as a promising alternative to overcome this substantial challenge [6, 7], which 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], and Thalidomide for Erythema Nodosum Leprosum and Multiple Myeloma [15].

The core idea of computational drug repositioning is to identify the new associations between known drugs and diseases. Currently, the computational drug repositioning methods include three main types [1, 16]: machine learning methods, matrix factorization/completion methods, and deep learning methods. Conventional machine learning methods predict drug-disease associations (DDAs) using drug and disease information as features, using classical classification models such as Support Vector Machines, Regularized Least Squares, and Random Forests. For instance, Gao et al. developed an approach called DDA-SKF, which enhances prediction by combining a Laplacian regularized least squares algorithm with a similarity kernel fusion method [17]. The main issue with these methods is their low performance, which is largely attributable to their reliance on high-quality features that depend heavily on specific domain knowledge and experience. Matrix factorization/completion methods reconstruct a DDA matrix into lower-dimensional matrices to uncover latent factors [18]. As an example, SCPMF identifies new drug-virus interactions by projecting a heterogeneous drug-virus interaction network into latent feature matrices for drugs and COVID-19 viruses and incorporated weighted similarity constraints [19]. Despite the competitive performances, it suffers from limited effectiveness in representing drugs and diseases, especially in sparse association networks.

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 DDAs. Such framework allows accurate predictions for query DDAs at the same time, without the need for extensive manual feature engineering. Intuitively, graph neural network (GNN) [20–22], is readily embedded in end-to-end architectures to perform specific tasks with graph data inputs, captures structural information of graphs via message passing between the nodes of graphs. Contributing to its applicability for graph and network data, GNN architecture has been widely used in drug discovery-related tasks, such as property prediction [23–25] and virtual screening [26–28]. For drug repositioning, most of deep learning-based DDA prediction methods were designed based on GNN architectures [29–33]. For instance, PSGCN was proposed to transform DDA prediction into a graph classification problem by converting DDAs into partner-specific graphs using the SortPool strategy to handle variable-sized graph data effectively [30]. REDDA integrated GNN, graph attention, and layer attention mechanisms to learn drug and disease representations, and was trained on a multifaceted network to enhance drug-protein-gene-pathway-disease relationships through sequential learning blocks [33]. Although GNNs have significantly advanced DDA predictions with well-structured models and input data, their performance is often limited by the richness of input features, which typically rely on drug and disease pairwise similarities. These similarities heavily rely on topology, neglecting abundant related biomedical knowledge, which is widely stored in multiple data source but hard to collect exhaustively.

One promising solution to this challenge could be the large language models (LLMs) like BERT (Bidirectional Encoder Representations from Transformers) [34] and GPT (Generative Pre-trained Transformer) [35]. They are well-known for learning billions of parameters through the large-scale multi-source data training process. The advanced neural network architecture leveraging self-attention also allows them to excel in deep contextual understanding and text generation [36, 37]. LLMs have shown great promise in biomedical data synthesis, knowledge retrieval, and reasoning, with applications in tasks like biomedical knowledge graph entity and relation extraction [38] and clinical trial matching [39]. The accessibility of these models through user-friendly interfaces like ChatGPT has made them even more popular, potentially revolutionizing biomedical knowledge enrichment and representation augmentation by providing contextually rich descriptions. A recent study finetuned BERT on biomedical literature data for similarity calculation in drug-disease heterogeneous network and DDA prediction [40]. However, it neglects the use of LLM for drug and disease representation augmentation, which is more important and valuable than network similarity calculation.

To address the above challenge, and to investigate the effectiveness of LLM in improving the DDA predictions, we proposed a comprehensive framework for LLM-based DDA prediction: a zero-shot prompting template using GPT-4 is designed to generate precise descriptions of drugs and diseases. These descriptions were then converted into entity description embeddings, utilizing both GPT-4 and BioBERT. Subsequently, we integrated these

embeddings into a GNN-based DDA prediction model, termed LLM-DDA, exploring the optimal mode for such integration. Specifically, three model architectures were developed: LLM-DDA$_{Node\ Feat}$, LLM-DDA$_{Dual\ GNN}$, LLM-DDA$_{GNN-AE}$. Comprehensive experiments conducted on four benchmark datasets demonstrated the superiority of LLM-DDA for DDA prediction compared to 11 competitive baseline methods. Meanwhile, the best integration mode (LLM-DDA$_{GNN-AE}$) and LLM embedding generator of LLM-based embedding and GNN-based model were determined by architecture analysis and performance comparison. Furthermore, case studies also emphasized the applicability of LLM-DDA in practical drug repositioning that discover novel DDAs. Our investigational study provides computational evidence for the potential of LLM-inferred knowledge representation for computational drug repositioning and more general biomedical network association prediction tasks.

2 Materials and Methods

The workflow for this study is illustrated in Fig. 1 and encompasses several key phases: DDA benchmark collection, drug-disease heterogeneous network construction, prompt engineering, LLM-based embedding generation, and LLM-DDA model construction. Detailed descriptions of each phase are provided in the subsequent sections.

2.1 Problem Formulation

The DDA prediction problem is formulated as a link prediction task within a heterogeneous network $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V} = \{\mathcal{V}_r, \mathcal{V}_d\}$ is the node set comprising N drugs ($\mathcal{V}_r$) and M 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 $\mathcal{E}_{r-r}$, drug-disease $\mathcal{E}_{r-d}$, and disease-disease $\mathcal{E}_{d-d}$ associations. The goal is to model a function $f_{DDA}(H_i, H_j, \mathcal{E} \not\ni v_i - v_j)$ that estimates the association probability p for a given drug-disease
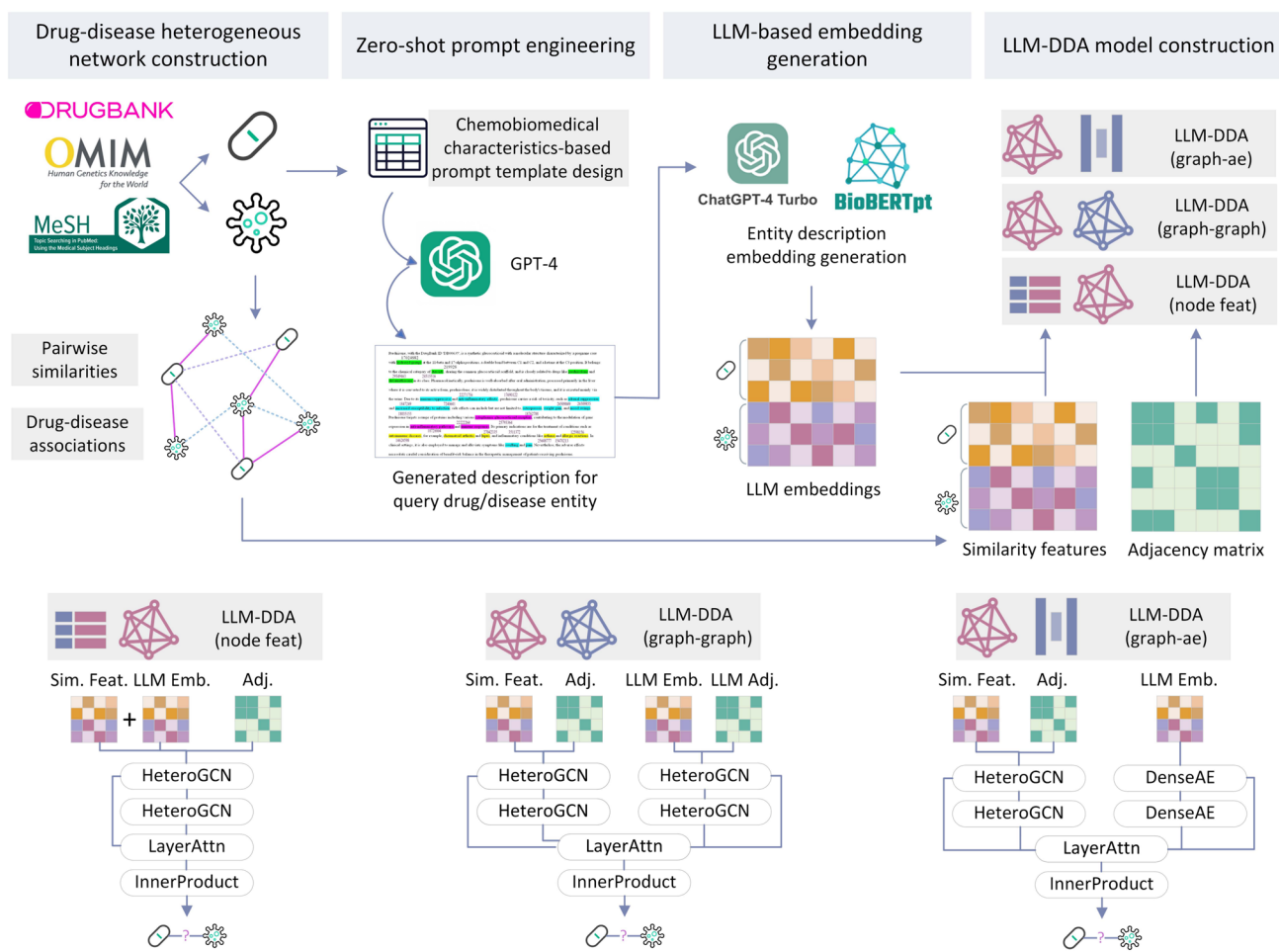

Fig. 1 Overview of study methods

pair "$v_i - v_j$" while $v_i \in \mathcal{V}_r$, $v_j \in \mathcal{V}_d$, and with the known edge set without the query drug-disease pair linked, also with H_i and H_j as their respective feature embeddings for prediction.

2.2 Benchmark Dataset Preparation

Four drug-disease association benchmarks were adopted in our study for model performance comparisons, which include: B-dataset [18], C-dataset [41], F-dataset [42], and R-dataset [33]. These datasets have been extensively used in previous drug repositioning studies. Basic statistical descriptions of these datasets are provided in Table 1. The datasets exhibit variations in label imbalance and data sparsity, enabling a comprehensive performance evaluation across both general and data-scarce scenarios.

- B-dataset: Comprises 269 drugs, 598 diseases, and 18,416 DDAs, sourced from the Comparative Toxicogenomics Database (CTD)) [43]. Drug-drug and disease-disease similarities were assessed through multi-source interactions (such as substructures and target enzymes) and MeSH (Medical Subject Headings) semantic similarities.
- C-dataset: Contains 663 drugs, 409 diseases, and 2532 DDAs, generated by integrating the Dndataset [44] with F-dataset [42], as described by Luo et al. [41]. Similarities between drugs were calculated using Anatomical Therapeutic Chemical (ATC) codes, and disease similarities were derived from Disease Ontology (DO) terms.
- F-dataset: Includes 593 drugs, 313 diseases, and 1933 DDAs, originating from OMIM and processed using the MetaMap tool. This dataset's similarities were calculated based on comprehensive similarity measurements [42].
- R-dataset: Features 894 drugs, 454 diseases, and 2704 DDAs, reorganized as a combination of C-dataset, F-dataset, and additional data from the KEGG database. Similarities were measured using molecular fingerprint similarities and MeSH semantic similarities.

Table 1 Summary of four benchmark datasets

Dataset	Drugs	Diseases	Drug-disease associations	Density	Pos-Neg Ratio
B-dataset	269	598	18,416	0.114	11.45%
C-dataset	663	409	2,532	0.009	1.57%
F-dataset	593	313	1,933	0.010	1.05%
R-dataset	894	454	2,704	0.007	0.67%

Details on the overlap of common drugs, diseases, and DDAs between these datasets are provided in the Supplementary Materials.

2.3 Drug-Disease Heterogeneous Network Construction

For GNN-based model, a drug-disease association network is essential for effective DDA prediction. To construct this network, we first derived drug-drug and disease-disease associations from the pre-calculated similarity matrices S_{r-r} and S_{d-d}, respectively. We applied a Top15 filtering method to select the most significant associations, ensuring the relevance and strength of the interactions within our model. Then, taking drug and disease entities as nodes, drug-drug interactions, disease-disease interactions, and drug-disease associations A_{r-d} were adopted as edges to construct drug-disease heterogeneous networks. By representing the network as a node feature matrix $\boldsymbol{H}_{\text{Sim}}^{(0)} \in \mathbb{R}^{(N+M)\times(N+M)}$ and an adjacency matrix $\boldsymbol{A}_{\text{Sim}} \in \mathbb{R}^{(N+M)\times(N+M)}$, the drug-disease network can be formulated as

$$\boldsymbol{H}_{\text{Sim}}^{(0)} = \begin{bmatrix} \boldsymbol{S}_{r-r} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{S}_{d-d} \end{bmatrix} \tag{1}$$

$$\boldsymbol{A}_{\text{Sim}} = \begin{bmatrix} \text{Top15}(\boldsymbol{S}_{r-r}) & \boldsymbol{A}_{r-d} \\ (\boldsymbol{A}_{r-d})^T & \text{Top15}(\boldsymbol{S}_{d-d}) \end{bmatrix} \tag{2}$$

2.4 Prompt Design for Description Generation

We utilized the principle of GPT-4's zero-shot prompting, which lies in its ability to understand and generate appropriate responses to tasks without needing explicit prior examples or fine-tuning for those specific tasks. This technique harnesses the core capabilities of LLMs—comprehension, reasoning, and explanation—thereby ensuring efficient and effective description generation. As illustrated in Tables 2 and 3, GPT-4 (version: gpt-4-0125-preview) was configured to mimic the expertise typical of scientists in relevant fields, optimizing it for the generation of chemo-biomedical descriptions crucial for the DDA prediction task. By calling upon specific databases (i.e., DrugBank, OMIM, and SMILES) to provide domain-specific knowledge, the zero-shot prompts guide the model to generate coherent responses that encompass key information beneficial to subsequent link prediction tasks.

The text generation prompts focused on:

1) Domain-specific knowledge enrichment: Prompts were crafted to guide the LLM to produce responses enriched with domain-specific knowledge, including information about genes, signaling pathways, related diseases,

Table 2 Description generation prompt

Prompt task	Disease description generation	Drug description generation
Prompt beginning	"Generate a single, cohesive, narrative paragraph for the disease '**{disease_name}**' associated with OMIM ID '**{omim_id}**'." The response should include **9** key information as follows	"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
Prompt 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	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)
Prompt 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."	"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."

Table 3 Drug disease association prediction prompt

Prompt component	Content
Introduction	The Online Mendelian Inheritance in Man (OMIM) database serves as a comprehensive and authoritative repository of human genes and genetic phenotypes. Simultaneously, the DrugBank database merges detailed drug information with extensive drug target data. Our research focuses on identifying associations between drugs and diseases. In this network model, both diseases and drugs (including certain chemicals not traditionally used as human drugs) are represented as nodes, with edges depicting the relationships between them. This includes associations like the **link between arsenic and diseases such as prostatic neoplasms and myocardial ischemia**
Query	Considering the information provided, does the drug identified by the name "**{drug_name}**" and DrugBank ID "**{drug_db_id}**" have any known associations with the disease listed as "**{disease_name}**" with OMIM ID "**{omim_id}**"?

and drugs. This was facilitated by a detailed template specifying the essential elements to be included in the responses.

2) Minimization of hallucinated information [45, 46]: To reduce the generation of inaccurate or fabricated information, specific constraints were incorporated into the prompts. Terms such as "precise" "with examples" and "evidence-based" were used to direct the model's responses. Additionally, the model was programmed to respond with "not available" when encountering queries beyond its scope of knowledge. This approach was intended to enhance the confidence and relevance of the LLM's outputs.

For the DDA prediction using GPT-4, our study adopted a few-shot prompting technique with GPT-4 as outlined in Table 2, which includes: (1). Introduction Section: Provides GPT-4 with contextual background through examples of known drug-disease associations. (2). Answer Query: Elicits direct predictions for specific drug-disease pairs. This approach primes GPT-4 with relevant examples before querying, enhancing the accuracy and relevance of its predictions. These predictions serve as the "DirectPred" baseline in our comparative analysis.

2.5 Entity Description-Based Embedding Generation

We transformed drug and disease-related descriptions into LLM-based embeddings to facilitate the mapping from discrete semantic spaces to continuous hidden vector spaces. This process enables the incorporation of high-order semantic information into deep neural network architectures for

drug-disease association (DDA) prediction. Specifically, we adopted two LLMs as the embedding generator: GPT-4 [47] (version: text-embedding-ada-002) is a general-purpose LLM which achieved the most competitive performance among all LLMs; BioBERT [48] is a BERT-based LLM tailored for biomedical applications with smaller parameter size [48]. For embedding generation, each description, denoted as $\mathcal{D}$, is processed into an embedding vector with dimension E. The resulting embeddings $\boldsymbol{H}_{\text{LLM}}^{(0)} \in \mathbb{R}^{(N+M) \times E}$ can be obtained by

$$\boldsymbol{H}_{\text{LLM}}^{(0)} = \text{LLM}(\mathcal{D}), \text{LLM} \in \{\text{GPT4}, \text{BioBERT}\} \tag{3}$$

2.6 LLM-DDA Model Architectures

To explore the best method for integrating LLM-based embeddings into current GNN-based framework for computational drug repositioning, we designed three distinct model architectures, each differing in how LLM embeddings are incorporated: (1). LLM-DDA$_{\text{Node Feat}}$: This architecture incorporates LLM embeddings directly as node features within the graph. These enriched node features are designed to enhance the node's representational learning directly through the GNN's processing layers; (2). LLM-DDA$_{\text{Dual GNN}}$: LLM-based embeddings serve as inputs to a dual-channel GNN. This model leverages a novel drug-disease heterogeneous graph recalculated based on LLM embeddings, effectively creating a more informed network topology for the GNN to process; (3). LLM-DDA$_{\text{GNN-AE}}$: LLM-based embeddings are fed into an Autoencoder (AE) within a dual GNN-AE channel. The AE's role is to refine and reconstruct the LLM embeddings, aiming to capture and utilize complex patterns more effectively. Each model variant is represented in Fig. 1, and detailed descriptions of their methodologies are provided in subsequent sections.

2.6.1 LLM-DDA$_{\text{Node Feat}}$

LLM-based embeddings, generated from descriptions of each drug and disease entity, are utilized as node features within a GNN-based framework for DDA prediction. We incorporate these embeddings into the existing node feature matrix to enhance the representational capacity of the nodes. As for GNN-based model design, we employed a heterogeneous graph convolutional network (HeteroGCN) equipped with a layer attention module. This configuration helps mitigate the "over-smoothing" issue often encountered in multilayer GNNs, a challenge noted in several prior studies [29, 32, 33]. The integration process involves merging the similarity-based node feature matrix $\boldsymbol{H}_{\text{Sim}}^{(0)}$ with the

LLM-based node embedding matrix $\boldsymbol{H}^{(0)} \in \mathbb{R}^{(N+M) \times E}$ as the input node features:

$$\boldsymbol{H}^{(0)} = \text{Concat}\left(\boldsymbol{H}_{\text{Sim}}^{(0)}, \boldsymbol{H}_{\text{LLM}}^{(0)}\right) \tag{4}$$

Then, L-layered HeteroGCNs were constructed to calculate updated node features by firstly considering homogeneous neighbor node sets in a general Graph Convolutional Network (GCN) manner, and then employ a summation process for heterogeneous aggregation. The updating process at each l-th HeteroGCN layer is be formulated as

$$\boldsymbol{H}^{(l)} = \text{HeteroGCN}\left(\boldsymbol{H}^{(l-1)}, \boldsymbol{A}_{\text{Sim}}\right) \tag{5}$$

$$\boldsymbol{H}^{(l)} = \hat{\boldsymbol{A}}_{\text{Sim}} \begin{bmatrix} \tilde{\boldsymbol{H}}_r^{(l)} \\ \tilde{\boldsymbol{H}}_d^{(l)} \end{bmatrix} = \hat{\boldsymbol{A}}_{\text{Sim}} \begin{bmatrix} \sigma\left(\hat{\boldsymbol{D}}_r^{-\frac{1}{2}} \hat{\boldsymbol{A}}_{\text{Sim}(r-r)} \hat{\boldsymbol{D}}_r^{-\frac{1}{2}} \boldsymbol{H}_r^{(l-1)} \boldsymbol{W}_r^{(l)}\right) \\ \sigma\left(\hat{\boldsymbol{D}}_d^{-\frac{1}{2}} \hat{\boldsymbol{A}}_{\text{Sim}(d-d)} \hat{\boldsymbol{D}}_d^{-\frac{1}{2}} \boldsymbol{H}_d^{(l-1)} \boldsymbol{W}_d^{(l)}\right) \end{bmatrix} \tag{6}$$

where $\widehat{\boldsymbol{A}}_{\text{Sim}} \in \mathbb{R}^{(N+M) \times (N+M)}$ represents the adjacency matrix augmented with the identity matrix, which can be decomposed into drug-drug homogeneous adjacency matrix $\widehat{\boldsymbol{A}}_{\text{Sim}(r-r)} \in \mathbb{R}^{N \times N}$ and disease-diesease homogeneous adjacency matrix $\widehat{\boldsymbol{A}}_{\text{Sim}(d-d)} \in \mathbb{R}^{M \times M}$. This setup allows for the application of graph convolutional operations separately on drug and disease entities thus capturing intra-type interactions. Given the dimension for the hidden vector as K, the intermediate node features $\begin{bmatrix} \tilde{\boldsymbol{H}}_r^{(l)} \\ \tilde{\boldsymbol{H}}_d^{(l)} \end{bmatrix} \in \mathbb{R}^{(N+M) \times K}$ represents the aggregated node features from homogeneous graphs. For homogeneous GCN, σ is the ReLU activation function, $\hat{\boldsymbol{D}}_r \in \mathbb{R}^{N \times N}$ and $\hat{\boldsymbol{D}}_d \in \mathbb{R}^{M \times M}$ are the degree matrix for drug and disease homogeneous graphs. $\boldsymbol{W}_r^{(l)}$ and $\boldsymbol{W}_d^{(l)}$ are the trainable parameter matrix for l-th HeteroGCN layer.

Subsequently, a layer attention mechanism, proposed by Yu et al. [29], was introduced to dynamically aggregate output node embeddings from different layers of the HeteroGCN, which helps alleviate the issue of over-smoothing observed in deep GNNs. For each l-th HeteroGCN layer, the node embedding $\boldsymbol{H}^{(l)}$ undergo a process where normalized attention coefficients are calculated to determine the significance of each layer's output for both drug and disease nodes, formulated as

$$\alpha_r^{(l)} = \frac{\exp\left(\boldsymbol{H}_r^{(l)} \boldsymbol{W}_r \boldsymbol{q}_r^{\text{T}}\right)}{\sum_{l \in L} \exp\left(\boldsymbol{H}_r^{(l)} \boldsymbol{W}_r \boldsymbol{q}_r^{\text{T}}\right)}, \alpha_d^{(l)} = \frac{\exp\left(\boldsymbol{H}_d^{(l)} \boldsymbol{W}_d \boldsymbol{q}_d^{\text{T}}\right)}{\sum_{l \in L} \exp\left(\boldsymbol{H}_d^{(l)} \boldsymbol{W}_d \boldsymbol{q}_d^{\text{T}}\right)} \tag{7}$$

where $\boldsymbol{q}_r, \boldsymbol{q}_d \in \mathbb{R}^{1 \times K}$ and $\boldsymbol{W}_r, \boldsymbol{W}_d \in \mathbb{R}^{K \times K}$ are trainable parameter matrixes. The output node embedding $\boldsymbol{H} \in \mathbb{R}^{(N+M) \times K}$ can be represented as

$$H = \begin{bmatrix} H_r \\ H_d \end{bmatrix} = \begin{bmatrix} \sum_{l \in L} \alpha_r^{(l)} H_r^{(l)} \\ \sum_{l \in L} \alpha_d^{(l)} H_d^{(l)} \end{bmatrix} \tag{8}$$

Finally, a bilinear inner product decoder was introduced to reconstruct and predict the drug-disease association matrix $\hat{A}$ based on node embeddings:

$$\hat{A} = \sigma\left(H_r W H_d^{\mathrm{T}}\right) \tag{9}$$

where σ is the Sigmoid activation function and W is a trainable parameter matrix.

2.6.2 LLM-DDA$_{\text{Dual GNN}}$

LLM-DDA$_{\text{Dual GNN}}$ leverages LLM-based embeddings to encode biomedical knowledge into a high-order vector space for drugs and diseases. We developed a drug-disease heterogeneous graph that is knowledge-intensive and can enhance the GNN-based DDA prediction method by integrating network topology with embedded biomedical knowledge. LLM-DDA$_{\text{Dual GNN}}$ is designed as a dual HeteroGCN model. Specifically, the first channel of LLM-DDA$_{\text{Dual GNN}}$ utilizes initial similarity-based embeddings as inputs to generate topology-based representations $H_{\text{Sim}}^{(l)}$ based on L-layered HeteroGCNs, which can be represented as **Eqs.** (4)-(6). Concurrently, the second channel utilizes LLM-based dense embeddings to compute drug-drug and disease-disease similarities by cosine similarity. Top15 filtering is then applied to refine these similarities, selecting the most significant associations to construct the adjacency matrix A_{LLM} for the LLM-based drug-disease heterogeneous graph as

2.6.3 LLM-DDA$_{\text{GNN-AE}}$

Inspired by previous DDA prediction studies utilizing the Auto-Encoder for feature deduction and proved its contribution to for multi-source drug and disease representation learning [49, 50], LLM-DDA$_{\text{GNN-AE}}$ was proposed as a dual channel with an AE used for generating deduced LLM-based embeddings and a GNN-based channel used for generating network topology-based embeddings. Specifically, similar to LLM-DDA$_{\text{Node Feat}}$ and LLM-DDA$_{\text{Dual GNN}}$, L-layered HeteroGCNs were constructed in the GNN-based channel to obtain H_{Sim} based on Eqs. (4), (5), (6). Then, a two-layered dense neural network was employed as the AE that takes the initial LLM-based embeddings as the input and produces higher-order embeddings $H_{\text{LLM}}^{(1)} \in \mathbb{R}^{(N+M) \times E}$ and $H_{\text{LLM}}^{(2)} \in \mathbb{R}^{(N+M) \times E}$:

$$H_{\text{LLM}}^{(1)} = \begin{bmatrix} H_{\text{LLM}(r)}^{(0)} W_r^{(1)} + b_r^{(1)} \\ H_{\text{LLM}(d)}^{(0)} W_d^{(1)} + b_d^{(1)} \end{bmatrix}$$
$$= \begin{bmatrix} \left(H_{\text{LLM}(r)}^{(0)} W_r^{(1)} + b_r^{(1)}\right) W_r^{(2)} + b_r^{(2)} \\ \left(H_{\text{LLM}(d)}^{(0)} W_d^{(1)} + b_d^{(1)}\right) W_d^{(2)} + b_d^{(2)} \end{bmatrix} \tag{12}$$

$$H_{\text{LLM}}^{(2)} = \begin{bmatrix} H_{\text{LLM}(r)}^{(2)} W_r^{(2)} + b_r^{(2)} \\ H_{\text{LLM}(d)}^{(2)} W_d^{(2)} + b_d^{(2)} \end{bmatrix}$$
$$= \begin{bmatrix} \left(H_{\text{LLM}(r)}^{(0)} W_r^{(1)} + b_r^{(1)}\right) W_r^{(2)} + b_r^{(2)} \\ \left(H_{\text{LLM}(d)}^{(0)} W_d^{(1)} + b_d^{(1)}\right) W_d^{(2)} + b_d^{(2)} \end{bmatrix} \tag{13}$$

$$A_{\text{LLM}} = \begin{bmatrix} \text{Top15}\left(\tilde{H}_{\text{LLM}(r)}^{(0)}\left(\tilde{H}_{\text{LLM}(r)}^{(0)}\right)^{\mathrm{T}}\right) & A_{r-d} \\ (A_{r-d})^{\mathrm{T}} & \text{Top15}\left(\tilde{H}_{\text{LLM}(d)}^{(0)}\left(\tilde{H}_{\text{LLM}(d)}^{(0)}\right)^{\mathrm{T}}\right) \end{bmatrix} \tag{10}$$

where $\tilde{H}_{\text{LLM}(r)}^{(0)} \in \mathbb{R}^{N \times E}$ and $\tilde{H}_{\text{LLM}(d)}^{(0)} \in \mathbb{R}^{M \times E}$ represent the normalized LLM-based embedding matrixes for drug and disease, respectively. Subsequently, another L-layered HeteroGCNs are used to generate updated LLM-based embedding $H_{\text{LLM}}^{(l)}$ based on $H_{\text{LLM}}^{(0)}$ and A_{LLM}, which was as represented in Eqs. (4), (5), (6). Then, the layer attention block combined these embedding into a final integrated one H for further DDA prediction:

$$H = \text{LayerAttn}\left(H_{\text{Sim}}^{(0)}, .., H_{\text{Sim}}^{(L)}, H_{\text{LLM}}^{(0)}, \dots, H_{\text{LLM}}^{(L)}\right) \tag{11}$$

Finally, the predicted drug-disease association matrix was reconstructed by a bilinear inner product decoder (Eq. (9)).

where W and b are trainable parameter matrixes in each layer. Similarly, a layer attention was adopted to aggregate output embeddings from each HeteroGCN and AE layers based on Eqs. (11). Finally, a bilinear inner product decoder predicted the final drug-disease association probability matrix based on Eqs. (8) and (9).

2.7 Optimization

The above three variants of LLM-DDA were optimized by a weighted cross-entropy loss function to balance different categories and focused on known drug-disease associations. The loss function is formulated as

$$\mathcal{L} = -\frac{1}{N}\left(\gamma \sum_{(i,j)\in\mathcal{S}^+} \log\hat{A}_{ij} + \sum_{(i,j)\in\mathcal{S}^-} \left(1 - \log\hat{A}_{ij}\right) \right) \quad (14)$$

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 is for model optimization and the trainable parameters in each layer are initialized by Xavier [51]. Moreover, the dropout layer and batch normalization layer are also adopted to inhibit overfitting.

2.8 Experimental Settings

We employed fivefold cross-validations to evaluate the performance of the LLM-DDA and to facilitate comparison with baseline methods. In this setup, known drug-disease associations (DDAs) were treated as positive samples, while all unknown DDAs were considered negative. Each validation fold was composed of 20% positive and 20% negative samples, with four folds used for training and one reserved for validation. This strategy ensured comprehensive validation of all samples within the datasets. To prevent data leakage, any DDAs present in the test set were excluded from the training graphs. Given the inherent label imbalance in DDA prediction benchmarks, we utilized several metrics for performance assessment: area under the receiver operating curve (AUC), area under Precision-Recall curve (AUPR), F1-score, and Precision.

For the LLM-DDA model, we configured the number of HeteroGCN layers to two, set the hidden vector dimensions at 128, applied a dropout rate of 0.4, and ran the models for 5000 epochs with a learning rate of 0.01. The hyperparameter settings for baseline methods were adopted directly from their respective original literature to ensure fairness in comparisons.

3 Results

This study conducted computational experiments to evaluate several aspects of integrating LLM embeddings into GNN-based models for drug-disease association (DDA) prediction. We specifically focused on and addressed the following questions: (Q1). Model Architecture: Which model architecture is most effective when incorporating LLM embeddings, such as those generated by GPT-4 or BioBERT, into traditional GNN-based models? (Q2). Performance and Stability: How do the performances and stabilities of the enhanced models (LLM-DDA) compare across four different datasets? (Q3). Comparative

Analysis: Is the LLM-DDA approach competitive with existing DDA prediction baselines? (Q4). Embedding Impact: To what extent do LLM-based embeddings contribute to the accuracy of DDA predictions within the LLM-DDA framework? (Q5). Application Potential: Can LLM embeddings be effectively applied to the discovery of new indications and the extraction of knowledge for query drugs and diseases?

3.1 For Q1: Model Performance Comparison in LLM-DDA

Our fivefold cross-validation experiments evaluated the efficacy of three distinct model architectures integrated with two types of LLM-based embeddings for drug-disease association (DDA) prediction.

Architecture comparison: The model performances of LLM-DDA$_{\text{Node Feat}}$, and LLM-DDA$_{\text{Dual GNN}}$, LLM-DDA$_{\text{GNN-AE}}$ on four benchmark datasets are presented in Fig. 2. The results showed that LLM-DDA$_{\text{GNN-AE}}$ achieved the best performance compared to LLM-DDA$_{\text{Node Feat}}$ and LLM-DDA$_{\text{Dual GNN}}$ on four datasets. Among them, LLM-DDA$_{\text{Node Feat}}$ performed worst, which indicates simply combining LLM-based embeddings with network similarity features could not increase the model capacity. Therefore, it requires elaborate architecture design for the integration of LLM-based embeddings to the general GNN-based DDA prediction methods; LLM-DDA$_{\text{Dual GNN}}$ performed moderately with mild performance gaps compared to LLM-DDA$_{\text{GNN-AE}}$. One possible explanation is, LLM-based embeddings already imply high-order association information for drugs and diseases based on the knowledge description overlaps between similar drugs/diseases. Therefore, when updating such high-order features based on a lower-order topology network, the representation for drugs and disease could degrade. This could cause model failing to fully utilize the complex relationships and patterns inherent in LLM features. Regarding the best-performed LLM-DDA$_{\text{GNN-AE}}$, the results indicate the Autoencoder is more effective for LLM-based embedding updating and integrating, which can maintain the high-order association information within these high-order embeddings.

We assessed LLM-DDA$_{\text{Node Feat}}$, LLM-DDA$_{\text{Dual GNN}}$, and LLM-DDA$_{\text{GNN-AE}}$ across four benchmark datasets, as shown in Fig. 2. The LLM-DDA$_{\text{GNN-AE}}$ model demonstrated superior performance over the other architectures, effectively leveraging the high-order association information provided by LLM-based embeddings. In contrast, LLM-DDA$_{\text{Node Feat}}$ exhibited the weakest performance, indicating that simply merging LLM embeddings with network similarity features does not sufficiently enhance model capacity. This underscores the need for more sophisticated integration techniques in GNN-based DDA prediction methods. LLM-DDA$_{\text{Dual GNN}}$
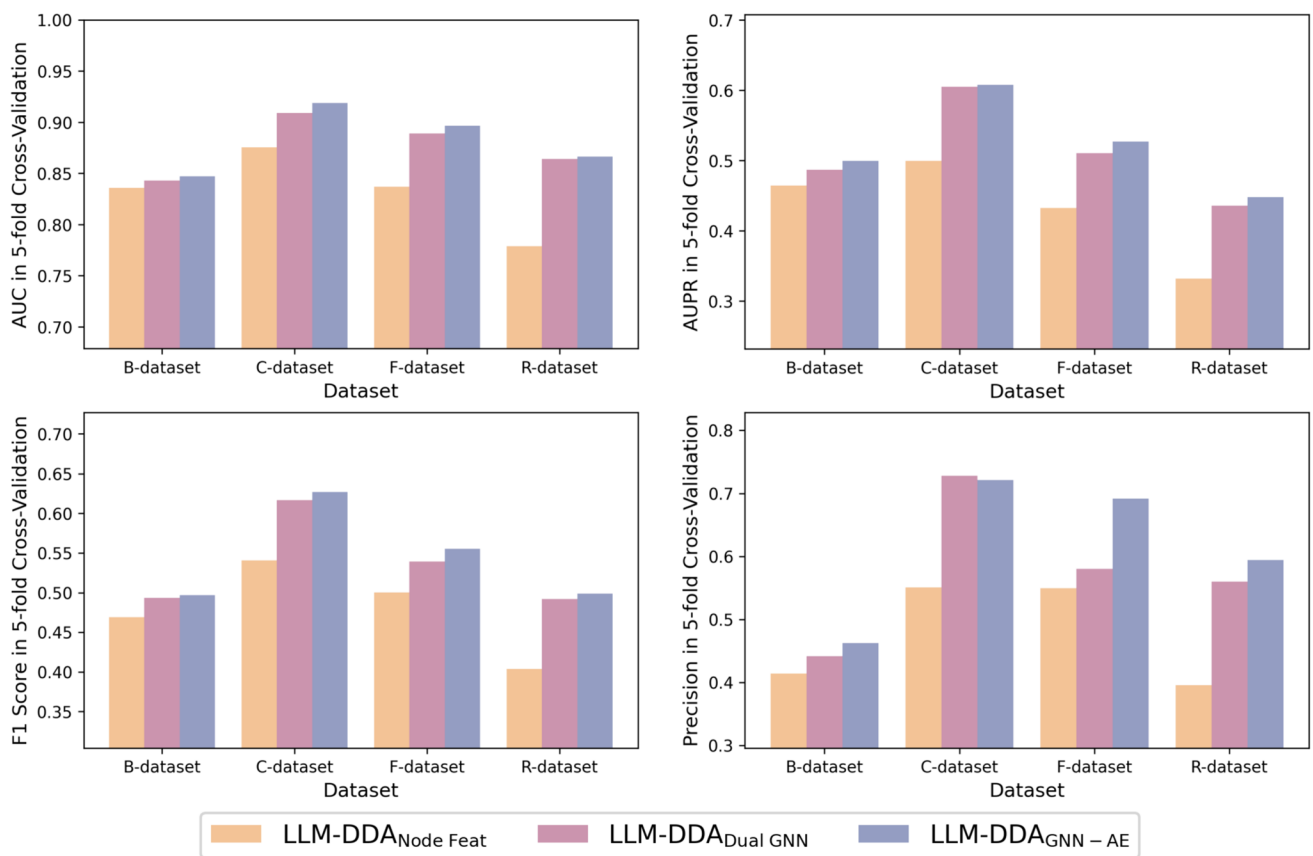
Fig. 2 Performance comparison for three LLM-DDA architectures on four datasets in fivefold cross-validation

displayed moderate performance, suggesting that the full potential of LLM embeddings might not be realized when updated through simpler network topologies. This degradation in feature representation could hinder the model's ability to capitalize on the complex relationships encoded in the LLM features. Our findings highlight that an autoencoder framework, such as that used in LLM-DDA$_{GNN-AE}$, is more adept at maintaining and updating high-order embeddings effectively.

Embedding Generator Comparison: Using LLM-DDA$_{GNN-AE}$ as the reference model, we compared the performance impacts of different LLM embedding generators—GPT-4 and BioBERT (Fig. 3). The performance differences between models using GPT-4 and BioBERT were minimal, indicating that both a general large-scale LLM like GPT-4 and a domain-specific LLM like BioBERT are effective at encoding biomedical knowledge into usable vectors for DDA prediction. This suggests that the choice between these embedding generators can be based on other factors such as computational resources or specific model requirements, rather than efficacy alone.

3.2 For Q2: Cross-Validation on Four Datasets

We utilized fivefold cross-validation results to evaluate the predictive performance and stability of LLM-DDA$_{GNN-AE}$ across four benchmark datasets. Following the experimental settings outlined earlier, each dataset was divided into training and validation sets in an 8:2 ratio for each fold, ensuring no overlap among validation sets. We plotted the AUC and AUPR curves for each fold and for the aggregate of all validation sets (Overall), as shown in Fig. 4. The results demonstrate that LLM-DDA$_{GNN-AE}$ consistently delivered strong AUC performance across all folds, with minimal deviation, indicating high stability of the model. In terms of AUPR, we observed larger fluctuations across different folds, which is attributable to the impact of dataset splitting on performance metrics in imbalanced datasets. Notably, the AUPR was more consistent across folds of the better-balanced B-dataset, showing smaller performance variations compared to other, more imbalanced datasets.

These findings highlight not only the robustness and effectiveness of the LLM-DDA$_{GNN-AE}$ model but also underscore its consistent performance across various dataset
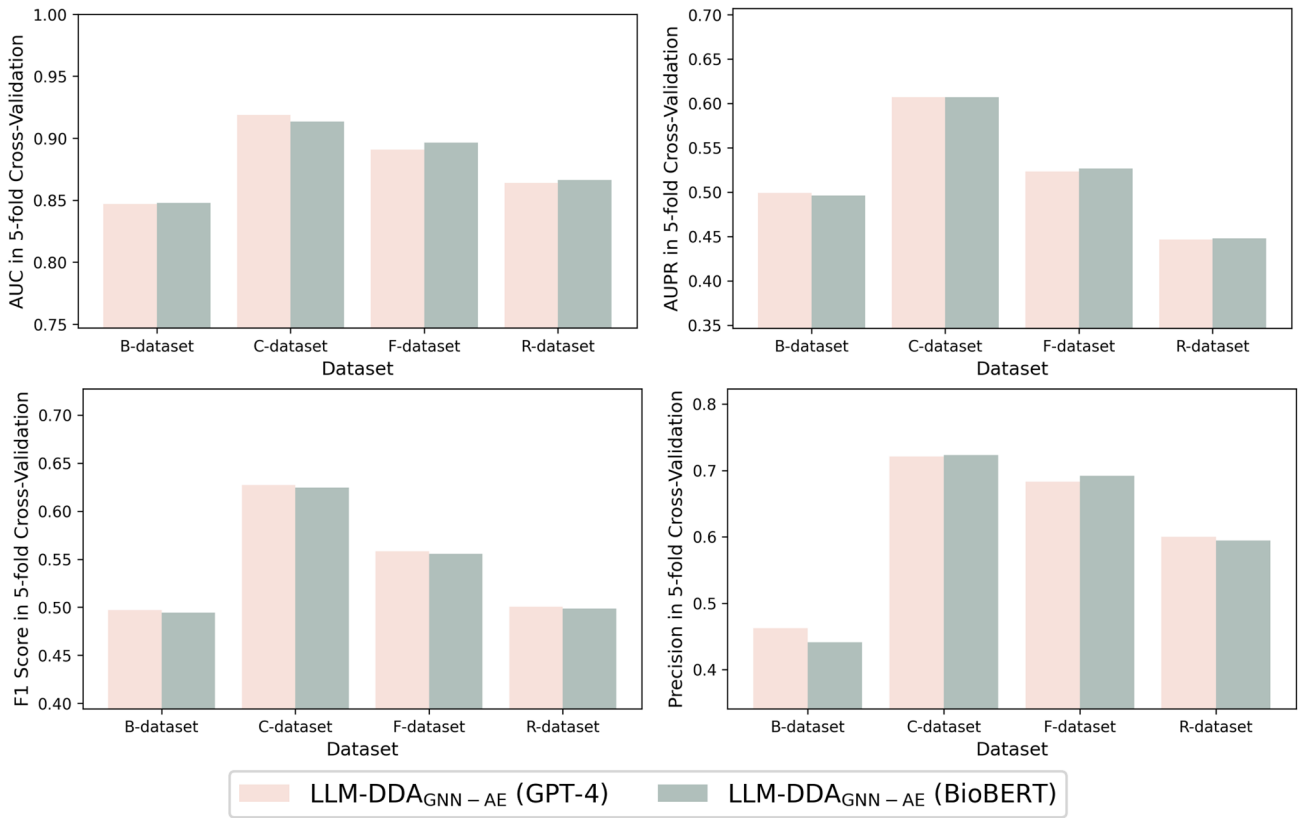
Fig. 3 Performance comparison for two LLM embedding generators on four datasets in fivefold cross-validation
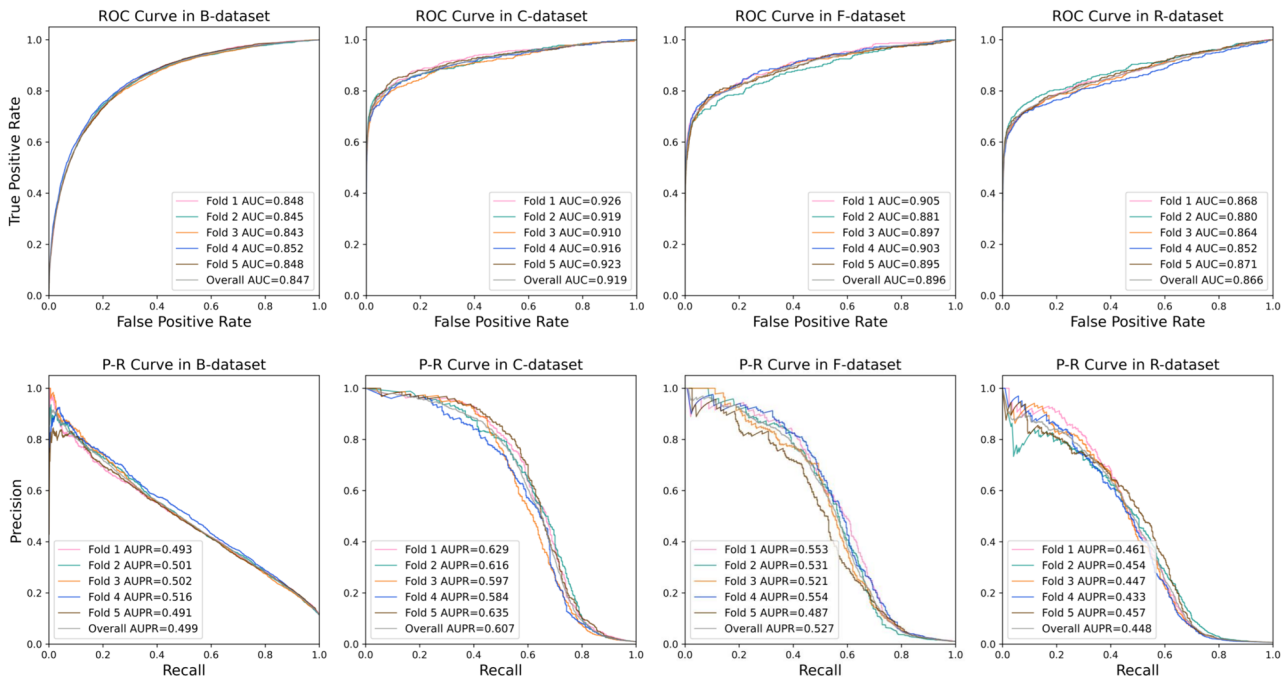

Fig. 4 AUC and AUPR curves of LLM-DDA$_{GNN-AE}$ on four datasets in fivefold cross-validation

Table 4 The AUC, AUPR, F1-Score, and Precision results of LLM-DDA and baseline methods on B-dataset in fivefold cross validation

Model	AUC	AUPR	F1-Score	Precision
DDA-SKF	0.701	0.252	0.328	0.259
SCPMF	**0.859**	**0.511**	**0.509**	**0.468**
NIMCGCN	0.667	0.233	0.290	0.223
HAN	0.695	0.256	0.323	0.258
MHGNN	0.574	0.160	0.222	0.136
DRWBNCF	0.838	0.455	0.474	0.428
REDDA	<u>0.847</u>	0.490	0.494	0.444
PSGCN	0.814	0.392	0.432	0.365
LAGCN	0.811	0.493	0.438	0.370
HDGAT	0.828	0.461	0.461	0.415
DirectPred	0.510	0.171	0.205	0.114
LLM-DDA$_{GNN-AE}$	<u>0.847</u>	<u>0.499</u>	<u>0.497</u>	<u>0.462</u>

The best result in each row is in bold faces. Second best are underlined

Table 5 The AUC, AUPR, F1-Score, and Precision results of LLM-DDA and baseline methods on C-dataset in fivefold cross validation

Model	AUC	AUPR	F1-Score	Precision
DDA-SKF	0.796	0.096	0.154	0.138
SCPMF	0.906	0.423	0.465	0.531
NIMCGCN	0.653	0.049	0.098	0.089
HAN	0.837	0.083	0.137	0.100
MHGNN	0.681	0.036	0.079	0.055
DRWBNCF	0.884	<u>0.541</u>	0.561	0.708
REDDA	0.895	0.490	0.531	0.591
PSGCN	**0.919**	0.334	0.395	0.384
LAGCN	0.914	0.492	0.521	0.564
HDGAT	0.884	0.221	0.352	0.237
DirectPred	0.506	0.124	0.019	0.010
LLM-DDA$_{GNN-AE}$	**0.919**	**0.607**	**0.627**	**0.721**

The best result in each row is in bold faces. Second best are underlined

Table 6 The AUC, AUPR, F1-Score, and Precision results of LLM-DDA and baseline methods on F-dataset in fivefold cross validation

Model	AUC	AUPR	F1-Score	Precision
DDA-SKF	0.775	0.096	0.152	0.148
SCPMF	0.886	0.344	0.392	0.440
NIMCGCN	0.619	0.036	0.077	0.078
HAN	0.806	0.061	0.108	0.071
MHGNN	0.651	0.036	0.072	0.057
DRWBNCF	0.865	0.398	0.433	0.554
REDDA	0.869	0.401	0.455	0.533
PSGCN	**0.904**	0.242	0.324	0.294
LAGCN	<u>0.898</u>	0.381	0.462	0.472
HDGAT	0.877	0.231	0.229	0.196
DirectPred	0.514	0.132	0.022	0.012
LLM-DDA$_{GNN-AE}$	0.896	**0.527**	**0.556**	**0.692**

The best result in each row is in bold faces. Second best are underlined

Table 7 The AUC, AUPR, F1-Score, and Precision results of LLM-DDA and baseline methods on R-dataset in fivefold cross validation

Model	AUC	AUPR	F1-Score	Precision
DDA-SKF	0.802	0.118	0.161	0.125
SCPMF	<u>0.875</u>	0.374	0.433	0.500
NIMCGCN	0.683	0.067	0.136	0.125
HAN	0.828	0.065	0.141	0.107
MHGNN	0.665	0.039	0.089	0.068
DRWBNCF	0.831	0.161	0.227	0.237
REDDA	0.859	0.306	0.381	0.402
PSGCN	<u>0.901</u>	0.165	0.220	0.185
LAGCN	**0.908**	0.312	0.385	0.407
HDGAT	0.791	0.218	0.129	0.079
DirectPred	0.576	0.231	0.020	0.010
LLM-DDA$_{GNN-AE}$	0.866	**0.448**	**0.499**	**0.594**

The best result in each row is in bold faces. Second best are underlined

conditions, reinforcing the reliability of the LLM-DDA approach in handling diverse and imbalanced data.

3.3 For Q3: Model Performance Comparison Against Baseline Methods

To benchmark the LLM-DDA$_{GNN-AE}$ model, we compared its performance with eleven baseline methods derived from previous studies, including two baselines developed by omitting the LLM-based embeddings from LLM-DDA: the reproduced LAGCN model (layer attention graph convolutional network) [29] and DirectPred from GPT-4 turbo. Our selection of baseline methods spanned different computational drug repositioning categories: machine learning-based (DDA-SKF [17]), matrix completion/factorization-based (SCPMF [19]), and deep learning-based (NIMCGCN [52], HAN [53], MHGNN [54], DRWBNCF [31], REDDA [33], PSGCN [30], LAGCN [29], HDGAT [55], and DirectPred). All the hyperparameter settings for the baselines were collected from their original studies or attached codebases. The brief introductions of these methods were in Supplementary Materials.

Utilizing fivefold cross-validation, we assessed various performance metrics (AUC, AUPR, F1-Score, and Precision) across four datasets, presented in Tables 4, 5, 6, 7, 8. LLM-DDA$_{GNN-AE}$ consistently demonstrated superior
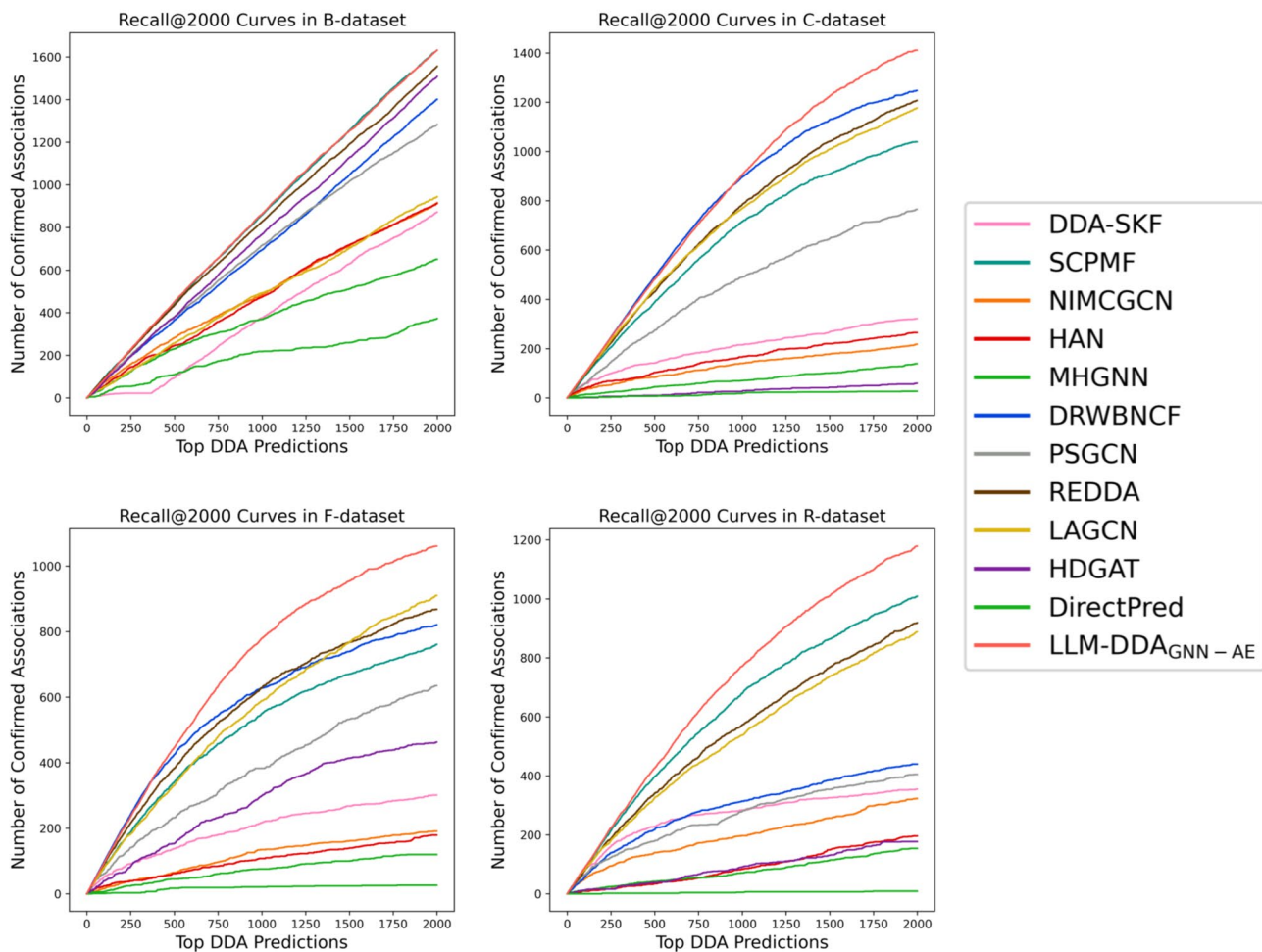
Table 8 The average AUC, AUPR, F1-Score, and Precision results of LLM-DDA and baseline methods on four datasets in fivefold cross validation

Model	AUC	AUPR	F1-Score	Precision
DDA-SKF	0.769	0.141	0.199	0.168
SCPMF	0.882	0.413	0.450	0.485
NIMCGCN	0.656	0.096	0.150	0.129
HAN	0.792	0.116	0.177	0.134
MHGNN	0.643	0.068	0.116	0.079
DRWBNCF	0.855	0.389	0.424	0.482
REDDA	0.868	<u>0.422</u>	<u>0.465</u>	<u>0.493</u>
PSGCN	**0.885**	0.283	0.343	0.307
LAGCN	<u>0.883</u>	0.420	0.452	0.453
HDGAT	0.845	0.283	0.293	0.232
DirectPred	0.527	0.165	0.067	0.037
LLM-DDA$_{GNN-AE}$	0.882	**0.520**	**0.545**	**0.618**

The best result in each row is in bold faces. Second best are underlined

performance on most metrics. On the B-dataset, it was second only to SCPMF, which marginally outperformed in all evaluated metrics. Across other datasets, LLM-DDA$_{GNN-AE}$ excelled in AUPR, F1-Score, and Precision. When results were averaged (Table 8), LLM-DDA$_{GNN-AE}$ exhibited significant improvements: **23.22%** in AUPR, **17.20%** in F1-Score, and **25.35%** in Precision over the suboptimal model, while maintaining comparable AUC.

When comparing against LAGCN, LLM-DDA$_{GNN-AE}$ outperformed this model in AUPR, F1-Score, and Precision across all datasets, affirming the advantage of integrating LLM-based embeddings into DDA predictions. Regarding DirecPred, this comparison highlighted the limitations of using unmodified LLM outputs for DDA prediction, which resulted in ineffective results. Conversely, our model's sophisticated prompting design and integration strategies effectively harnessed the LLM's knowledge extraction capabilities for enhanced DDA prediction.

Fig. 5 Recall@2000 curves of LLM-DDA$_{GNN-AE}$ and eight baselines on four datasets in fivefold cross-validation

To assess accuracy in high-stakes predictions crucial for screening potential drug repositioning candidates, we calculated the "Recall@K" metric following previous studies [29, 33]. Specifically, the predictions were ranked and the recalled known DDAs among the top predictions were counted. Recall@K reflects the screening and ranking power of DDA methods. The Recall@2000 curves for LLM-DDA$_{GNN-AE}$ and the eight baselines across four datasets were plotted in Fig. 5. LLM-DDA$_{GNN-AE}$ achieved the highest screening and ranking power in three datasets (C-dataset, F-dataset, and R-dataset) and remained among the top performers in the B-dataset. These findings underscore the model's high potential for practical drug repositioning applications.

3.4 For Q4: Attention Visualization

To further elucidate the role of LLM embeddings in enhancing drug and disease representation within our LLM-DDA$_{GNN-AE}$

model, we conducted a visualization analysis of the attention weights from the model's layer attention aggregation layer. We extracted and analyzed the distribution of attention weights allocated to LLM embeddings (H_{LLM}) and similarity embeddings (H_{Sim}), as illustrated in Fig. 6. This analysis revealed that LLM-based embeddings were allocated approximately 20% of the attention for drug representations and 10% for disease representations. These proportions correlate well with the performance improvements observed in LLM-DDA$_{GNN-AE}$ (Tables 4, 5, 6, 7, 8), validating the effectiveness of incorporating LLM embeddings in our model. The relatively low attention proportions assigned to H_{LLM} can be attributed to the presence of noise in the descriptions generated by the LLM. Without corrections based on ground-truth knowledge, these descriptions may introduce inaccuracies. However, our model employs layer attention mechanisms to denoise these embeddings effectively. This approach not only preserves the beneficial information contained within the LLM embeddings but also prevents potential performance deterioration due to noisy data.
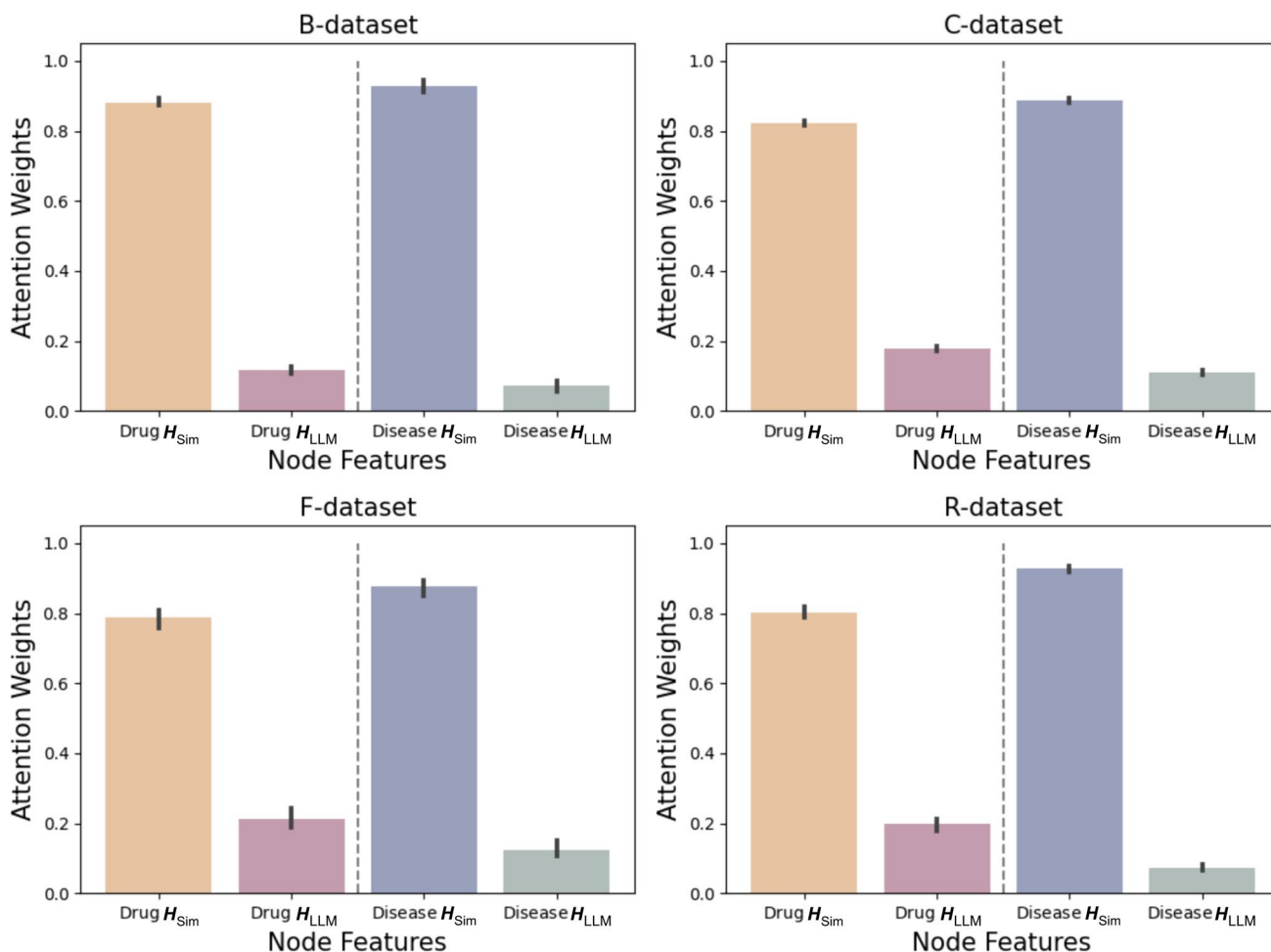

Fig. 6 Attention weight distribution for H_{Sim} and H_{LLM} on four datasets. "Drug Sim. Feat." represents $H_{Sim(r)}$; "Drug LLM Feat." $H_{LLM(r)}$, "Disease Sim. Feat." represents $H_{Sim(d)}$, "Disease LLM Feat." $H_{LLM(d)}$

Table 9 The AUC and AUPR results of LLM-DDA$_{\text{GNN-AE}}$ with different embedding aggregation methods

Aggregation method		B-dataset	C-dataset	F-dataset	R-dataset
AUC	Concat	0.794	0.893	0.854	0.827
	Mean	0.846	0.909	0.887	0.864
	Sum	0.821	0.918	0.891	0.861
	Layer Attention	**0.847**	**0.919**	**0.896**	**0.866**
AUPR	Concat	0.369	0.442	0.396	0.353
	Mean	0.479	0.585	0.514	0.435
	Sum	0.434	0.560	0.499	0.391
	Layer Attention	**0.499**	**0.607**	**0.527**	**0.448**

The best result in each row is in bold faces

We also conducted an ablation analysis to study the impact of different embedding aggregation methods. In addition to our adopted layer attention aggregation, we tested three aggregation methods as alternatives: Concat (concatenating embeddings from different layers), Mean (averaging embeddings from different layers), and Sum (summing embeddings from different layers). The AUC and AUPR performance results of LLM-DDA$_{\text{GNN-AE}}$ on four benchmark datasets are listed in Table 9, indicating layer attention as the embedding aggregation method achieves the best performances among all datasets. This finding demonstrates attention mechanism can be a more efficient method for the aggregation of drug and disease representations than other trivial approaches that inappropriately handle the representation importances for multi-layered and multi-sourced models.

3.5 For Q5: Case Study—Evaluating Predictions for Prednisone and Allergic Rhinitis

To demonstrate the practical efficacy of the LLM-DDA model, we conducted a case study using Prednisone (DrugBank ID: DB00635) and Allergic Rhinitis (OMIM ID: D607154) in C-dataset as representative examples for drug and disease,

respectively. We aimed to validate the model's predictions of new drug-disease associations (DDAs) through external literature verification. Focusing on the hit rates of predicted DDAs which also existed in C-dataset, LLM-DDA accurately predicted known DDAs for Prednisone at a rate of 73.68% (28/38) and for Allergic Rhinitis at 87.09% (27/31), with predictions probabilities exceeding 0.5. Focusing on identifying new DDAs which not existed in known DDA set, Table 10 presents the top 10 predictions for both Prednisone and Allergic Rhinitis, achieving a 100% validation rate, as all predicted associations were corroborated by PubMed references. Notably, Prednisone was endorsed by the American Gastroenterological Association (AGA) as the first-line therapy for refractory celiac disease in the absence of budesonide, highlighting its efficacy in improving symptoms and histological outcomes [56]. Another significant finding involved ketotifen and epinastine, which were shown to alleviate symptoms of seasonal allergic conjunctivitis in a controlled clinical study [57]. Meanwhile, Table 11 features LLM-generated description profiles for Prednisone and Allergic Rhinitis as a knowledge extraction case. The drug description outlines the drug family, relevant clinical symptoms, and gene expressions associated with Prednisone. The disease description details the associated diseases, clinical symptoms, and treatment mechanisms for Allergic Rhinitis. Each entity identified in these profiles is substantiated by references available in PubMed, denoted by their respective PMID numbers.

The overall case study results suggest that LLM is capable of generating descriptions with biomedical knowledge and association embedded, and LLM-DDA can discover new DDAs thus contributing to computational drug repositioning.

4 Discussion

This study marks a pioneering effort to introduce LLM into the realm of computational drug repositioning, particularly for DDA prediction task. We explored the potential of LLMs to improve drug and disease representation learning by

Table 10 The top 10 LLM-DDA-predicted associated diseases for Prednisone and Allergic Rhinitis

Prednisone		Evidence (PMID)	Allergic Rhinitis		Evidence (PMID)
OMIM ID	Disease name		DrugBank ID	Drug name	
D266600	inflammatory bowel disease (crohn disease)	34,078,656	DB00920	Ketotifen	12,487,225
D215140	reenberg dysplasia	11,007,214	DB00443	Betamethasone	27,670,203
D607202	celiac disease	36,137,844	DB01234	Dexamethasone	4,180,817
D145590	hyperthermia, cutaneous, with headaches and nausea	16,225,918	DB00557	Hydroxyzine	7,007,473
D109543	leukemia, chronic lymphocytic	34,788,401	DB01003	Cromoglicic acid	3,100,877
D151400	leukemia, chronic lymphocytic; cll	18,760,709	DB00751	Epinastine	18,631,332
D601626	leukemia, acute myeloid	14,743,000	DB01084	Emedastine	18,631,332
D254500	myeloma	29,231,133	DB00768	Olopatadine	19,243,710
D254200	myasthenia gravis	27,509,100	DB01069	Promethazine	24,791,618

Table 11 Generated descriptions with examples of Allergic Rhinitis and Prednisone

Description	Identified entity
Drug: Prednisone, with the DrugBank ID 'DB00635', is a synthetic glucocorticoid with a molecular structure characterized by a pregnane core with hydroxyl groups at the 11-beta and 17-alpha positions, a double bond between C1 and C2, and a ketone at the C3 position. It belongs to the chemical category of steroids, sharing the common glucocorticoid scaffold, and is closely related to drugs like prednisolone and dexamethasone in its class. Pharmacokinetically, prednisone is well-absorbed after oral administration, processed primarily in the liver where it is converted to its active form, prednisolone; it is widely distributed throughout the body's tissues, and it is excreted mainly via the urine. Due to its immunosuppressive and anti-inflammatory effects, prednisone carries a risk of toxicity, such as adrenal suppression and increased susceptibility to infection; side effects can include but are not limited to, osteoporosis, weight gain, and mood swings. Prednisone targets a range of proteins including various cytoplasmic glucocorticoid receptors, contributing to the modulation of gene expression in anti-inflammatory pathways and immune responses. Its primary indications are for the treatment of conditions such as autoimmune diseases, for example, rheumatoid arthritis and lupus, and inflammatory conditions like asthma and allergic reactions. In clinical settings, it is also employed to manage and alleviate symptoms like swelling and pain. Nevertheless, the adverse effects necessitate careful consideration of benefit-risk balance in the therapeutic management of patients receiving prednisone.	Associated drugs or drug family: hydroxyl groups (17024882) steroids (2199128) prednisolone (29549463) dexamethasone (241515516) Associated clinical symptoms: immunosuppressive (22271756) anti-inflammatory effects (17690122) adrenal suppression (1847249) increased susceptibility to infection (7240601) osteoporosis (26509049) weight gain (26509855) mood swings (18035153) swelling (2560877) pain (15470213) Associated gene expressions: cytoplasmic glucocorticoid receptors (18762788) anti-inflammatory pathways (22222264) immune responses (25793364) Associated diseases: autoimmune diseases (10728804) rheumatoid arthritis (27042335) lupus (3511372) asthma (12598156) allergic reactions (14620598)
Disease: Allergic Rhinitis, indexed under OMIM ID 'D607154', is a complex condition characterized by an exaggerated immune response to environmental allergens. Key genes implicated in its pathogenesis include IL4, IL13, and HLA-DRB1, all of which influence the IgE-mediated hypersensitivity pathway central to the disease's etiology. This pathway involves the activation of mast cells and basophils, leading to the release of histamine and other inflammatory mediators that are quintessential for symptom manifestation. The efficacy of treatment options like intranasal corticosteroids (e.g., fluticasone), antihistamines (e.g., cetirizine), and leukotriene receptor antagonists (e.g., montelukast) lies in their capacity to attenuate this inflammatory cascade, showcasing diverse mechanisms of action such as inhibition of inflammatory cell migration, blockade of histamine receptors, and prevention of leukotriene effects, respectively. Commonly associated comorbidities include asthma and atopic dermatitis, underscoring a shared atopic background among these conditions. Allergic Rhinitis is primarily characterized by nasal congestion, sneezing, rhinorrhea, and itching, typifying the body's aberrant reaction to harmless substances.	Associated drugs or drug family: intranasal corticosteroids (28602936) fluticasone (33007780) antihistamines (28602936) cetirizine (3847381) leukotriene receptor antagonists (24229824) montelukast (35608857) Associated clinical symptoms: nasal congestion (25629743) sneezing (25629743) rhinorrhea (33313967) itching (25629743) aberrant reaction to harmless substances (17349011) Associated gene expressions: IL4 (25543037) IL13 (34093555) HLA-DRB1 (27013183) IgE-mediated hypersensitivity pathway (19818484) mast cells (27434218) basophils (27678500) histamine (29074456) Associated diseases: asthma (35695326) atopic dermatitis (32738956) Treatment mechanisms: attenuate this inflammatory cascade (28988769) inhibition of inflammatory cell migration (27196703) blockade of histamine receptors (21346365) prevention of leukotriene effects (9329408)

incorporating network topology and biomedical knowledge into a HeteroGCN-based DDA prediction model. The results clearly show that LLM-based embeddings significantly bolster the performance of DDA prediction models across various benchmarks, markedly enhancing top prediction accuracy and aiding more effective drug repositioning strategies. Comparing to other baseline methods, the reason why LLM-DDA achieves notable improvements can be summarized as follows: (1) The abundant chemobiomedical knowledges extracted from LLM provide relevant information for drugs and diseases. (2) Higher-order dense embeddings for LLM-inferred knowledge are generated by LMs for better feature integration. (3) The GNN-AE architecture guarantees more solid representation learning. Considering the generalizability, our approach, which combines LLM embeddings with traditional GNN-based models, demonstrates versatility and could be adapted to other biomedical networks by simply substituting the drug/disease entities with other biological entities, such as proteins, to facilitate predictions of drug-protein interactions.

While GPT-4 demonstrated performances akin to random guessing in some cases, this underscores the challenge of applying general LLMs in specialized domains without tailored knowledge retrieval mechanisms. Furthermore, the superiority of LLM-DDA$_{\text{GNN-AE}}$ compared to other LLM-DDA models may be explained by the avoidance of graph convolution process. LLM embedding may have already captured high-order associations inherently present in descriptions of similar drugs/diseases. However, updating such high-order features in HeteroGCN might degrade their quality, as HeteroGCN focuses on direct connections, potentially underutilizing complex LLM patterns. Autoencoders, by reconstructing LLM features, can maintain or even enhance the high-order association information within the features. This reconstruction process is independent from simpler topological structure of graph data, thereby improving predictive performance.

A primary limitation of our study is its reliance on zero-shot prompting to evaluate the effectiveness of LLMs. Future research should experiment with diverse prompting techniques to identify more efficacious strategies for integrating LLMs into biomedical prediction tasks. Meanwhile, the LLM-generated descriptions were not extensively analyzed in this study. A more rigorous, quantitative analysis of the text generation process is essential to evaluate the accuracy and relevance of the generated content. Moving forward, we plan to extend our investigations to encompass not only the generalized GNN module used in this study but also other more sophisticated architectures. This will include larger datasets and the exploration of different LLM variations and prompt engineering strategies, with a particular emphasis on domain specific LLMs tailored for the biomedical sector. By extending the evaluation to include more datasets and

scenarios, such as 'unseen' drugs or diseases, the practical value of LLM in drug repositioning and our proposed LLM-DDA framework could be more comprehensively demonstrated.

5 Conclusions

This study explored the incorporation of biomedical knowledge into high-order embeddings using LLMs, aiming at enhancing computational drug repositioning with GNN-based DDA prediction model. We achieved this by innovatively merging network topology embeddings with textual embeddings derived from GPT-4 and BioBERT, which processed LLM-generated descriptions of drug and disease entities. The proposed LLM-DDA approach demonstrated superior performance compared to other baseline methods. Moving forward, our focus will be on refining LLM prompt engineering and extending our dataset trials to validate and improve the efficiency of LLM-DDA in drug repositioning, with the ultimate goal of accelerating drug discovery and development.

Declarations

References

1. Zong N, Wen A, Moon S et al (2022) Computational drug repurposing based on electronic health records: a scoping review. NPJ Digit Med 5(1):77. https://doi.org/10.1038/s41746-022-00617-6
2. Chan HCS, Shan H, Dahoun T et al (2019) Advancing drug discovery via artificial intelligence. Trends Pharmacol Sci 40(8):592–604. https://doi.org/10.1016/j.tips.2019.06.004
3. Prasad V, Mailankody S (2017) Research and development spending to bring a single cancer drug to market and revenues after approval. JAMA Intern Med 177(11):1569–1575. https://doi.org/10.1001/jamainternmed.2017.3601
4. DiMasi JA, Grabowski HG, Hansen RW (2016) Innovation in the pharmaceutical industry: new estimates of R&D costs. J Health Econ 47:20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012
5. Wong CH, Siah KW, Lo AW (2019) Estimation of clinical trial success rates and related parameters. Biostatistics 20(2):273–286. https://doi.org/10.1093/biostatistics/kxx069

6. Hurle MR, Yang L, Xie Q et al (2013) Computational drug repositioning: from data to therapeutics. Clin Pharmacol Ther 93(4):335–341. https://doi.org/10.1038/clpt.2013.1

7. Li J, Zheng S, Chen B et al (2016) A survey of current trends in computational drug repositioning. Brief Bioinform 17(1):2–12. https://doi.org/10.1093/bib/bbv020

8. Shim JS, Liu JO (2014) Recent advances in drug repositioning for the discovery of new anticancer drugs. Int J Biol Sci 10(7):654–663. https://doi.org/10.7150/ijbs.9224

9. Mohamed K, Yazdanpanah N, Saghazadeh A et al (2021) Computational drug discovery and repurposing for the treatment of COVID-19: a systematic review. Bioorg Chem 106:104490. https://doi.org/10.1016/j.bioorg.2020.104490

10. Traylor JI, Sheppard HE, Ravikumar V et al (2021) Computational drug repositioning identifies potentially active therapies for chordoma. Neurosurgery 88(2):428–436. https://doi.org/10.1093/neuros/nyaa398

11. Bai L, Scott MKD, Steinberg E et al (2021) Computational drug repositioning of atorvastatin for ulcerative colitis. J Am Med Inform Assoc 28(11):2325–2335. https://doi.org/10.1093/jamia/ocab165

12. Fahimian G, Zahiri J, Arab SS (2020) RepCOOL: computational drug repositioning via integrating heterogeneous biological networks. J Transl Med 18(1):375. https://doi.org/10.1186/s12967-020-02541-3

13. Budak C, Mençik V, Gider V (2023) Determining similarities of COVID-19 - lung cancer drugs and affinity binding mode analysis by graph neural network-based GEFA method. J Biomol Struct Dyn 41(2):659–671. https://doi.org/10.1080/07391102.2021.2010601

14. Zhang Z, Zhou L, Xie N et al (2020) Overcoming cancer therapeutic bottleneck by drug repurposing. Signal Transduct Target Ther 5(1):113. https://doi.org/10.1038/s41392-020-00213-8

15. Pushpakom S, Iorio F, Eyers PA et al (2019) Drug repurposing: progress, challenges and recommendations. Nat Rev Drug Discov 18(1):41–58. https://doi.org/10.1038/nrd.2018.168

16. Luo H, Li M, Yang M et al (2021) Biomedical data and computational models for drug repositioning: a comprehensive review. Brief Bioinform 22(2):1604–1619. https://doi.org/10.1093/bib/bbz176

17. Gao CQ, Zhou YK, Xin XH et al (2022) DDA-SKF: predicting drug-disease associations using similarity kernel fusion. Front Pharmacol 12:784171. https://doi.org/10.3389/fphar.2021.784171

18. Zhang W, Yue X, Lin W et al (2018) Predicting drug-disease associations by using similarity constrained matrix factorization. BMC Bioinformatics 19(1):233. https://doi.org/10.1186/s12859-018-2220-4

19. Meng Y, Jin M, Tang X et al (2021) Drug repositioning based on similarity constrained probabilistic matrix factorization: COVID-19 as a case study. Appl Soft Comput 103:107135. https://doi.org/10.1016/j.asoc.2021.107135

20. Jumper J, Evans R, Pritzel A (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596(7873):583–589. https://doi.org/10.1038/s41586-021-03819-2

21. Ma T, Liu Q, Li H et al (2022) DualGCN: a dual graph convolutional network model to predict cancer drug response. BMC Bioinformatics 23(Suppl 4):129. https://doi.org/10.1186/s12859-022-04664-4

22. Liu Q, Hu Z, Jiang R et al (2020) DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. Bioinformatics 36(Suppl_2):i911–i918. https://doi.org/10.1093/bioinformatics/btaa822

23. Gu Y, Zheng S, Li J (2021) CurrMG: a curriculum learning approach for graph based molecular property prediction. In: 2021 IEEE International Conference on Bioinformatics and

24. Kong W, Zhu J, Shan P et al (2024) iSKIN: integrated application of machine learning and Mondrian conformal prediction to detect skin sensitizers in cosmetic raw materials. SmartMat. https://doi.org/10.1002/smm2.1278

25. Gu Y, Zheng S, Xu Z et al (2022) An efficient curriculum learning-based strategy for molecular graph learning. Briefings Bioinformatics 23(3):bbac099. https://doi.org/10.1093/bib/bbac099

26. Zhu J, Wang J, Wang X et al (2021) Prediction of drug efficacy from transcriptional profiles with deep learning. Nat Biotechnol 39(11):1444–1452. https://doi.org/10.1038/s41587-021-00946-z

27. Kong W, Lian J, Peng C et al (2024) Identification of novel GABAA receptor positive allosteric modulators with novel scaffolds via multistep virtual screening. Acta Physico Chimica Sinica. https://doi.org/10.3866/PKU.WHXB202302044

28. Gu Y, Li J, Kang H et al (2023) Employing molecular conformations for ligand-based virtual screening with equivariant graph neural network and deep multiple instance learning. Molecules 28(16):5982. https://doi.org/10.3390/molecules28165982

29. Yu Z, Huang F, Zhao X et al (2021) Predicting drug–disease associations through layer attention graph convolutional network. Briefings Bioinformatics 22(4):bbaa243. https://doi.org/10.1093/bib/bbaa243

30. Sun X, Wang B, Zhang J et al (2022) Partner-specific drug repositioning approach based on graph convolutional network. IEEE J Biomed Health Inform 26(11):5757–5765. https://doi.org/10.1109/JBHI.2022.3194891

31. Meng Y, Lu C, Jin M et al (2022) A weighted bilinear neural collaborative filtering approach for drug repositioning. Briefings Bioinformatics 23(2):bbab581. https://doi.org/10.1093/bib/bbab581

32. Gu Y, Zheng S, Zhang B et al (2022) Milgnet: a multi-instance learning-based heterogeneous graph network for drug repositioning. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp 430–437 https://doi.org/10.1109/BIBM55620.2022.9995152

33. Gu Y, Zheng S, Yin Q et al (2022) REDDA: integrating multiple biological relations to heterogeneous graph neural network for drug-disease association prediction. Comput Biol Med 150:106127. https://doi.org/10.1016/j.compbiomed.2022.106127

34. Devlin J, Chang MW, Lee K et al (2018) BERT: pre-training of deep bidirectional transformers for language understanding. arXiv. https://doi.org/10.48550/arXiv.1810.04805

35. Radford A, Narasimhan K, Salimans T et al (2018) Improving language understanding by generative pre-training. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

36. Coulombe C (2018) Text data augmentation made simple by leveraging nlp cloud apis. arXiv. https://doi.org/10.48550/arXiv.1812.04718

37. Dai H, Liu Z, Liao W et al (2023) AugGPT: leveraging chatGPT for text data augmentation. arXiv. https://doi.org/10.48550/arXiv.2302.13007

38. Trajanoska M, Stojanov R, Trajanov D (2023) Enhancing knowledge graph construction using large language models. arXiv. https://doi.org/10.48550/arXiv.2305.04676

39. Yuan J, Tang R, Jiang X et al (2023) LLM for patient-trial matching: privacy-aware data augmentation towards better performance and generalizability. In: American Medical Informatics Association (AMIA) Annual Symposium. https://par.nsf.gov/biblio/10448809

40. Kang H, Hou L, Gu Y et al (2023) Drug-disease association prediction with literature based multi-feature fusion. Front Pharmacol 14:1205144. https://doi.org/10.3389/fphar.2023.1205144

41. Luo H, Wang J, Li M et al (2016) Drug repositioning based on comprehensive similarity measures and Bi-Random walk

41. algorithm. Bioinformatics 32(17):2664–2671. https://doi.org/10.1093/bioinformatics/btw228

42. Gottlieb A, Stein GY, Ruppin E et al (2011) PREDICT: a method for inferring novel drug indications with application to personalized medicine. Mol Syst Biol 7:496. https://doi.org/10.1038/msb.2011.26

43. Davis AP, Wiegers TC, Johnson RJ et al (2023) Comparative Toxicogenomics Database (CTD): update 2023. Nucleic Acids Res 51(D1):D1257–D1262. https://doi.org/10.1093/nar/gkac833

44. Martínez V, Navarro C, Cano C et al (2015) DrugNet: network-based drug-disease prioritization by integrating heterogeneous data. Artif Intell Med 63(1):41–49. https://doi.org/10.1016/j.artmed.2014.11.003

45. Guo T, Nan B, Liang Z et al (2023) What can large language models do in chemistry? A comprehensive benchmark on eight tasks. In: Advances in Neural Information Processing Systems, pp 59662–59688. https://proceedings.neurips.cc/paper_files/paper/2023/file/bbb330189ce02be00cf7346167028ab1-Paper-Datasets_and_Benchmarks.pdf

46. Wang J, Shi E, Yu S et al (2023) Prompt engineering for healthcare: methodologies and applications. arXiv. https://doi.org/10.48550/arXiv.2304.14670

47. Achiam J, Adler S, Agarwal S et al (2023) GPT-4 technical report. arXiv. https://doi.org/10.48550/arXiv.2303.08774

48. Lee J, Yoon W, Kim S et al (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 36(4):1234–1240. https://doi.org/10.1093/bioinformatics/btz682

49. Zhao BW, Hu L, You ZH et al (2022) HINGRL: predicting drug-disease associations with graph representation learning on heterogeneous information networks. Briefings Bioinformatics 23(1):bbab515. https://doi.org/10.1093/bib/bbab515

50. Jarada TN, Rokne JG, Alhajj R (2021) SNF–CVAE: computational method to predict drug–disease interactions using similarity network fusion and collective variational autoencoder. Knowl-Based Syst 212:106585. https://doi.org/10.1016/j.knosys.2020.106585

51. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pp 249–256. https://proceedings.mlr.press/v9/glorot10a.html

52. Li J, Zhang S, Liu T et al (2020) Neural inductive matrix completion with graph convolutional networks for miRNA-disease association prediction. Bioinformatics 36(8):2538–2546. https://doi.org/10.1093/bioinformatics/btz965

53. Wang X, Ji H, Shi C et al (2019). Heterogeneous graph attention network. In: The World Wide Web Conference, pp 2022–2032 https://doi.org/10.1145/3308558.3313562

54. Li M, Cai X, Xu S et al (2023) Metapath-aggregated heterogeneous graph neural network for drug-target interaction prediction. Briefings Bioinformatics 24(1):bbac578. https://doi.org/10.1093/bib/bbac578

55. Huang S, Wang M, Zheng X et al (2024) Hierarchical and dynamic graph attention network for drug-disease association prediction. IEEE J Biomed Health Inform. https://doi.org/10.1109/JBHI.2024.3363080

56. Green PHR, Paski S, Ko CW et al (2022) AGA clinical practice update on management of refractory celiac disease: expert review. Gastroenterology 163(5):1461–1469. https://doi.org/10.1053/j.gastro.2022.07.086

57. Borazan M, Karalezli A, Akova YA et al (2009) Efficacy of olopatadine HCI 0.1%, ketotifen fumarate 0.025%, epinastine HCI 0.05%, emedastine 0.05% and fluorometholone acetate 0.1% ophthalmic solutions for seasonal allergic conjunctivitis: a placebo-controlled environmental trial. Acta Ophthalmol 87(5):549–554. https://doi.org/10.1111/j.1755-3768.2008.01265.x

Authors and Affiliations

Yaowen Gu[1] · **Zidu Xu**[2] · **Carl Yang**[3]

✉ Zidu Xu
znx2000@cumc.columbia.edu

1 Department of Chemistry, New York University, New York, NY 10003, USA

2 School of Nursing, Columbia University, 560 W 168th Street, New York, NY 10032, USA

3 Department of Computer Science, Emory College of Arts and Sciences, Emory University, Atlanta, GA 30322, USA