# MCL-DMD: Multi-modal Contrastive Learning for Drug-Microbe-Disease Association Prediction

Niecia Say
Georgia Gwinnett College
Lawrenceville, Georgia, USA
deleesay@gmail.com

Moctar Keita
University of Health Sciences and Pharmacy in St. Louis
St Louis, Missouri, USA
Moctar.Keita@uhsp.edu

Farhan Tanvir
Georgia State University
Atlanta, Georgia, USA
ftanvir@gsu.edu

Lilia Chebbah
Georgia State University
Atlanta, Georgia, USA
lchebbah1@student.gsu.edu

Muhammad Ifte Khairul Islam
Georgia State University
Atlanta, Georgia, USA
mislam29@student.gsu.edu

Esra Akbas
Georgia State University
Atlanta, Georgia, USA
eakbas1@gsu.edu

## Abstract

Modeling interactions between drugs, microbes, and diseases is essential for advancing drug discovery and precision medicine. Although most existing computational approaches focus on pairwise association prediction, such as drug–microbe or microbe–disease associations, they often overlook the interdependencies among all three entities. In real-world biomedical systems, drug–microbiome interactions can modulate treatment efficacy, influence toxicity, and shape disease trajectories. Moving beyond binary relationships to explore triplet-level associations is essential for uncovering drugs' mechanisms of action (MoAs). Recent advances in Graph Neural Networks (GNNs) have significantly improved the modeling of complex molecular structures, enabling more accurate property prediction. However, molecular and biomedical data extend beyond graph structures, encompassing diverse modalities such as molecular sequences, taxonomic hierarchies, and ontological descriptors—features that GNNs cannot fully capture. To address these limitations, we propose `MCL-DMD` (Multi-modal Contrastive Learning for Drug-Microbe-Disease Association Prediction), a novel framework that combines heterogeneous graph modeling with domain-specific biomedical knowledge. `MCL-DMD` employs HeTAN (Heterogeneous Triple Attention Network) to model the interconnectedness of all entities in a heterogeneous graph and augments it with a biomedical knowledge encoder that leverages SMILES representations, microbial taxonomies, and MeSH disease descriptors. Through multi-modal contrastive learning, `MCL-DMD` aligns cross-modal representations to improve semantic consistency and predictive robustness. Experimental results demonstrate that `MCL-DMD` significantly outperforms state-of-the-art baselines in both classification and ranking metrics, offering a powerful framework to uncover novel drug–microbe–disease associations.

## CCS Concepts

• **Information systems** → **Data mining**; **Computing platforms**.

## Keywords

Multi-modal Learning, Contrastive Learning, Drug-Microbe-Disease

**ACM Reference Format:**
Niecia Say, Moctar Keita, Farhan Tanvir, Lilia Chebbah, Muhammad Ifte Khairul Islam, and Esra Akbas. 2018. MCL-DMD: Multi-modal Contrastive Learning for Drug-Microbe-Disease Association Prediction. In *Proceedings of 16th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB '25).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

## 1 Introduction

The intricate interplay among drugs, microbes, and diseases plays a crucial role in disease progression and therapeutic outcomes. Microbes significantly influence drug metabolism, efficacy, and toxicity, while drugs can alter microbial diversity and function, thus affecting disease states [1]. Understanding these interactions is critical for drug repurposing, personalized medicine, and minimizing adverse drug reactions.

Recent advances in machine learning have facilitated the study of drug mechanisms of action (MoAs) through various tasks such as drug behavior analysis and disease modeling [2]. Computational methods have been developed to predict pairwise associations, including drug-microbe [1], microbe-disease [3], and drug-disease interactions [4]. However, such pairwise approaches do not capture the full complexity of biological systems. Many diseases are influenced by a combination of microbial activity and drug interactions, highlighting the need to model these relationships as triplets rather than as isolated pairs. Neglecting higher-order dependencies can obscure key insights into disease mechanisms and potential treatment synergies [5]. Although recent studies have highlighted the importance of predicting drug-microbe-disease (DMD) associations, this remains an area needing additional research [6]. Therefore, a more comprehensive approach that models triplet-level interactions among drugs, microbes, and diseases is essential to fully capture their biological interplay.

Traditional methods for biomedical triplet prediction primarily rely on tensor factorization, representing interactions as multi-dimensional tensors and applying decomposition techniques to

*BCB '25, Philadelphia, PA*

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

infer missing associations [7]. Although models like Canonical Polyadic (CP) and Tucker decomposition can uncover latent patterns, they suffer from linearity assumptions and data sparsity, limiting their applicability to complex biomedical data. Nonlinear tensor approaches offer improvements but often depend on Gaussian processes, which are challenging to estimate [8]. Moreover, these tensor models do not leverage graph-based machine learning techniques, which has become central in drug discovery.

In contrast, Graph Neural Networks (GNNs) effectively capture structured, nonlinear relationships via iterative message passing [9], and have been widely applied in predicting chemical properties [10, 11]. Hypergraph Neural Networks (HGNNs) further extend this capability to model high-order interactions in triplet-level tasks. For instance, Liu et al. [12] proposed a multi-view hypergraph contrastive learning model for robust DMD prediction. However, despite their strengths, GNNs and HGNNs often neglect critical non-graph biomedical attributes, such as molecular structures (e.g., SMILES), taxonomic hierarchies, and disease ontologies (e.g., MeSH). Since biomedical entities are inherently multi-faceted, ignoring such domain knowledge leads to incomplete representations and suboptimal predictions. This highlights the need for a unified framework that integrates graph-based structures with domain-specific knowledge — an area still underexplored in triplet association prediction.

To address this challenge, we propose MCL-DMD (**M**ulti-modal **C**ontrastive **L**earning for **D**rug-**M**icrobe-**D**isease Association Prediction), a unified framework that integrates heterogeneous graph representations with biomedical domain knowledge-based representations. Our model captures complex triplet interactions by leveraging Heterogeneous Triple Attention Network (HeTAN) [13], which employs triplet message passing and triplet attention mechanisms to learn from heterogeneous graph structures. HeTAN employs a novel triplet attention mechanism to capture higher-order interactions within the drug-microbe-disease triplets. To encode domain-specific knowledge, we incorporate three distinct biomedical modalities: (i) drug features derived from molecular graphs using Graph Isomorphism Networks (GIN) [14], (ii) microbe features based on pairwise functional and taxonomic similarity, and (iii) disease features computed from semantic similarities and MeSH-based classifications. These features are processed through a dedicated encoder to construct domain knowledge-based representations. To align the representations learned from heterogeneous graphs and biomedical descriptors, we incorporate a multi-modal contrastive learning (CL) module. CL aligns semantically similar entities across different modalities while pushing apart dissimilar ones in the latent space, thereby enhancing representation quality and model generalization. This integrated approach enables MCL-DMD to capture complex, multi-faceted interactions among drugs, microbes, and diseases, ultimately improving the accuracy and robustness of triplet prediction in biomedical applications. Our main contributions are summarized as follows:

- **Multi-modal integration for triplet prediction:** We propose MCL-DMD, the first multi-modal CL framework for drug-microbe-disease (DMD) association prediction. By integrating heterogeneous graph structures with biomedical domain knowledge - covering molecular features (SMILES), taxonomic similarities, and ontological descriptors (MeSH), our model effectively captures complex, high-order dependencies that underpin DMD interactions.
- **Contrastive Learning (CL) across modalities:** We develop a novel CL module that aligns representations of entities across heterogeneous graphs and biomedical knowledge sources. This approach improves generalization by enforcing semantic consistency between graph-based and domain-specific embeddings.
- **Extensive experimental validation:** We conduct comprehensive experiments on benchmark datasets and demonstrate that MCL-DMD consistently outperforms state-of-the-art methods in both classification and ranking tasks. Our model achieves significant gains in F1-score, ROC-AUC, and Hit@N metrics, highlighting the effectiveness of multi-modal representation learning in biomedical triplet prediction.

The remainder of this paper is organized as follows. Section 2 reviews related work in multi-modal CL and triplet prediction. Section 3 introduces our proposed MCL-DMD framework, including its heterogeneous graph encoder, biomedical knowledge encoder, and CL strategy. Section 4 presents experimental results and comparative analysis with baseline models. Finally, Section 5 concludes the paper and outlines future research directions.

## 2 Related Work

In this section, we present an overview of related research on multi-modal CL and triplet prediction across various domains.

### 2.1 Multi-Modal Contrastive Learning

CL is a widely adopted self-supervised paradigm that generates augmented views of the same data instance and applies contrastive loss to maximize agreement between representations in the latent space [15]. Traditional single-stream models have limited generalization due to their inability to independently encode different modalities. Multi-modal CL, exemplified by CLIP [16], addresses this by learning aligned representations across modalities using separate encoders. Recent models like SLIP [17] and DeCLIP [18] integrate self-supervised objectives to improve efficiency, while FILIP [19] introduces fine-grained contrastive objectives for localized feature correspondences. Additionally, Yang et al. [20] proposed Dual Space Graph CL, exploring graph representations in feature and structure spaces, highlighting CL's adaptability across domains.

Despite successes in computer vision and NLP, multi-modal CL for biomedical applications, particularly triplet-based association prediction, remains underexplored. Biomedical data inherently include structural representations such as SMILES strings, taxonomic hierarchies, and ontology-based disease descriptors. Existing methods often overlook cross-modal alignment crucial for capturing synergistic effects between entities. To address this, our framework employs multi-modal CL to align heterogeneous biomedical representations, enabling holistic and accurate predictions of drug-microbe-disease associations.

## 2.2 Triplet Prediction

Triplet prediction has been extensively explored in natural language processing (NLP), computer vision (CV), and drug discovery. In NLP, it enables relation extraction by identifying associations among entities, often using attention mechanisms to enhance the interpretability of sentence representations. For example, Ji et al. [21] demonstrated its effectiveness in distant supervision to extract structured knowledge, while Zhou et al. [22] proposed a transformer-based framework that models query-key-value interactions to improve relational understanding in textual data. In the CV domain, triplet learning plays a central role in tasks such as face recognition and person re-identification. Notably, Schroff et al. [23] introduced FaceNet, which optimizes embeddings using a triplet loss function to ensure that representations of the same identity are close, while those of different identities are pushed apart.

Although triplet-based learning has proven to be effective in multiple domains, many computational models in biomedicine still focus predominantly on pairwise relationships rather than higher-order associations. For instance, in drug discovery, existing works often predict drug-disease or drug-microbe associations. Liang et al. [24] proposed NMGMDA, a framework that integrates nuclear norm minimization with graph attention networks to infer microbe-drug interactions. Similarly, Zhu et al. [25] developed the Nearest Neighbor Attention Network (NNAN), which leverages similarity-based network aggregation to predict drug-microbe associations. Although effective in modeling direct interactions, these approaches fail to capture the intricate dependencies among drugs, microbes, and diseases, which are critical for understanding complex biomedical processes.

To address these limitations, triplet-based prediction models have recently gained traction in biomedical research, offering a more comprehensive representation of biological relationships. For example, Zitnik and Zupan [26] applied collective matrix factorization to integrate heterogeneous biomedical data and model drug-target-disease interactions. Chen et al. [27] introduced a neural tensor network for drug-target-disease prediction, outperforming traditional tensor factorization techniques. More recently, HeTAN [13] introduced a triplet attention mechanism within a heterogeneous graph framework, significantly improving predictive accuracy.

Despite these advancements, many biomedical triplet-based approaches still rely heavily on graph-based representations and often neglect complementary modalities, such as molecular sequences, taxonomies, and ontologies. Although graph neural networks (GNNs) are effective at modeling interactions, they struggle to incorporate non-graph information, limiting their capacity to learn holistic representations. To overcome these challenges, we propose MCL-DMD, a framework that integrates heterogeneous graphs with molecular features and biomedical ontologies. Using CL, MCL-DMD aligns multimodal representations, improving both predictive accuracy and generalization in drug-microbe-disease association tasks.

## 3 Methodology

In this work, we integrate two complementary sources to predict drug-microbe-disease associations. First, we construct a heterogeneous graph representing drugs, microbes, and diseases via pairwise edges from confirmed triplets. We then apply a heterogeneous triple attention network to capture higher-order interactions among triplet-based message passing and attention. Simultaneously, we extract biomedical features from multiple modalities. Drugs are represented using SMILES strings transformed into molecular graphs via GIN, while microbes and diseases are encoded from taxonomic and ontological descriptors using fully connected networks. We then align the heterogeneous graph embedding and the domain specific features using multi-modal CL, ensuring that similar entities are close in the latent space. Finally, the integrated embeddings are concatenated and fed into an MLP to predict the association likelihood.

### 3.1 Problem definition and Dataset

Given a triplet consisting of a drug, a microbe, and a disease, our goal is to learn a predictive function that estimates the likelihood of an interaction. Let $D$, $M$, and $N$ represent the sets of drugs, microbes, and diseases, respectively. The complete set of potential drug-microbe-disease (DMD) triplets can be defined as the Cartesian product:

$$\mathcal{S} = D \times M \times N$$

Each triplet $(d, m, n) \in \mathcal{S}$ is assigned a binary label $p \in \{0, 1\}$, where $p = 1$ denotes a known association and $p = 0$ indicates an unverified or unknown interaction. Importantly, a label of 0 does not necessarily indicate the absence of an interaction, but reflects the current lack of experimental or clinical evidence.

The objective is to learn a function

$$f : \mathcal{S} \to [0, 1],$$

that predicts the probability of interaction for any given triplet $(d, m, n)$.

**Dataset Construction.** To build our dataset, we integrate drug-microbe-disease (DMD) associations from multiple publicly available sources. Drug-microbe interactions are obtained from MDAD [28], aBiofilm [29], and DrugVirus [30], while microbe-disease associations are sourced from HMDAD [31], Disbiome [32], gutMDisorder [33], and Peryton [34]. These associations are merged to construct triplets of the form $\langle d, m, n \rangle$, resulting in a dataset of 2,763 confirmed triplets involving 270 drugs, 58 microbes, and 167 diseases.

Drug molecular information is represented using SMILES (Simplified Molecular Input Line Entry System) [35] strings obtained from PubChem [36]. SMILES provide concise, human-readable representations of molecular structures, which we transform into molecular graphs for downstream processing. Similarly, microbial taxonomic data and disease ontological descriptors are extracted from the NCBI Taxonomy [37] and MeSH [38] databases, respectively. These resources capture hierarchical relationships essential for modeling biological similarities between microbes and diseases.

Compared to the total number of possible DMD combinations $(270 \times 58 \times 167)$, the number of confirmed associations is extremely sparse, representing only 0.11% of all possible triplets. To address this imbalance, we generate negative samples by randomly shuffling drug-microbe-disease combinations that lack confirmed associations. This results in a balanced dataset with a 1:1 ratio of positive to negative samples, which prevents the model from being biased toward confirmed interactions. Mitigating this class imbalance enables the model to better differentiate true associations from unknown ones, thereby improving generalization to unseen data.

## 3.2 Multi-modal CL

In this paper, we propose MCL-DMD, a novel **M**ulti-modal **C**ontrastive **L**earning framework for **D**rug-**M**icrobe-**D**isease Association Prediction. Our model integrates representations of entities (drug, microbe, disease) derived from a heterogeneous graph with representations based on biomedical domain knowledge within an end-to-end architecture. MCL-DMD employs CL to align entity representations in different modalities, thus enhancing the consistency of the features and improving the predictive ability. By combining the relational structure among entities with modality-specific domain knowledge, MCL-DMD offers a robust solution for uncovering novel drug-microbe-disease (DMD) associations.

The overall architecture of the MCL-DMD model is illustrated in Figure 1. In the following sections, we provide a detailed overview of each component.

### A. Heterogeneous Graph Triplet Attention Network

We construct a heterogeneous graph $G = (V, E)$ to represent the complex relationships among drugs, microbes, and diseases. The set of nodes is defined as

$$V = D \cup M \cup N,$$

where $D$, $M$, and $N$ denote the sets of drugs, microbes, and diseases, respectively. In this graph, we model each confirmed drug-microbe-disease interaction—represented as a triplet $(d, m, n)$ with label $p = 1$) is modeled by adding pairwise edges between the corresponding nodes. Specifically, for each confirmed triplet, we add the following edges to the edge set $E$:

$$(d, m), \quad (d, n), \quad \text{and} \quad (m, n).$$

To initialize the node features, we incorporate domain-specific representations rather than random initialization. For drugs, we use their chemical substructures derived from SMILES strings [39], processed using the Explainable Substructure Partition Fingerprint (ESPF) algorithm [40]. ESPF decomposes SMILES strings and amino acid sequences into frequent substructures and selects the most informative ones based on a frequency threshold, producing rich molecular fingerprints for drug and target nodes. Microbes and diseases are represented using one-hot encoded vectors, providing distinct identifiers for each entity type.

We adopt HeTAN [13] as our relational module to learn expressive node representations within heterogeneous graphs. HeTAN leverages a triplet-attention encoder to capture the higher-order dependencies among drugs, microbes, and diseases. Unlike traditional message-passing mechanisms, which rely primarily on pairwise interactions, HeTAN employs triplet-level attention. This approach allows the model to dynamically assign varying importance to different node pairs within a given triplet, thus enhancing its capability to accurately model the underlying complex relationships.

Since nodes and edges belong to distinct types in heterogeneous graphs, each with its own feature space, directly comparing or aggregating their information is challenging. To address this, as an initial step, we introduce a type-specific transformation matrix $T$, which projects the node features into a shared latent space, ensuring compatibility between different types of entities. The transformation is defined as $h_i^{'} = T \circ h_i$

Traditional Graph Convolutional Networks (GCNs) rely on pairwise message passing, where each node aggregates information independently from its neighbors. While effective in many settings, this approach is insufficient for capturing higher-order dependencies—particularly in biomedical networks, where the interaction between two entities (e.g., a drug and a microbe) is often conditioned on the presence of a third (e.g., a disease).

To address this limitation, we introduce the **Triplet Message Passing (TMP)** mechanism, which explicitly incorporates triplet-level context into representation learning. Instead of aggregating messages from individual neighbors, TMP considers neighbor pairs $(j, k)$ and learns their joint influence on a central node $i$. For a node of type $i$, its triplet neighborhood is defined as:

$$N_i = \{(j_1, k_1), \dots, (j_n, k_n)\},$$

where each $(j_t, k_t)$ is a semantically meaningful pair (e.g., microbe–disease pairs for a drug node). By passing messages from these neighbor pairs to the central node, TMP captures relational patterns that cannot be observed through pairwise aggregation alone.

This design allows the model to represent complex biomedical interactions more faithfully, enriching the learned node embeddings and improving predictive performance in drug–microbe–disease association tasks. Formally, the TMP update rule is:

$$z_i^l = \text{TMP}\left(z_i^{l-1}, N_i\right),$$

where $z_i^{l-1}$ is the embedding of node $i$ at layer $l - 1$, and TMP$(\cdot)$ denotes the aggregation over all neighbor pairs in $N_i$.

Not all neighboring pairs contribute equally to the central node's representation. To account for this, we introduce a triplet attention mechanism, which assigns importance scores based on the features of all three nodes in a triplet. The attention coefficient $e_{ijk}$ is computed as:

$$e_{ijk} = \text{LeakyReLU}(\text{NN}(h_i^{'}||h_j^{'}||h_k^{'})), \tag{1}$$

where NN is a neural network capturing intricate dependencies among triplet components. To ensure comparability across different nodes, attention coefficients are normalized via a softmax function:

$$\alpha_{ijk} = \frac{\exp(e_{ijk})}{\sum_{l,m \in N(i)} \exp(e_{ilm})}. \tag{2}$$

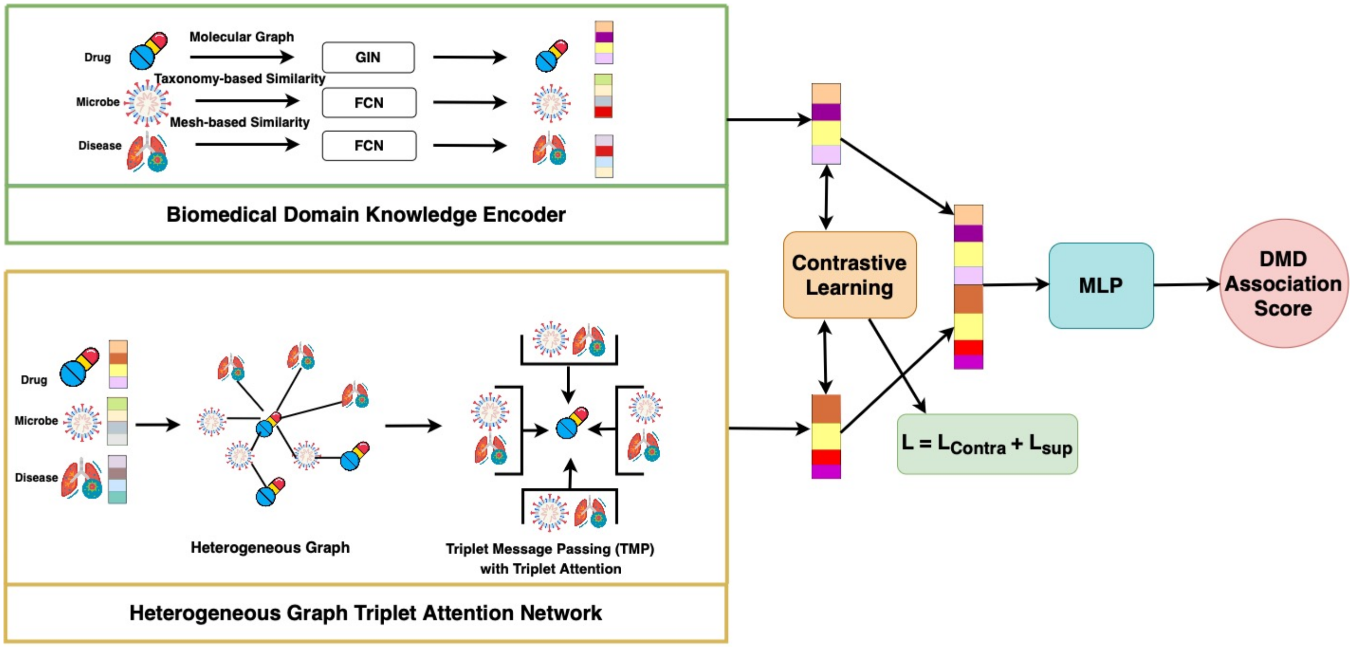

Figure 1: The MCL-DMD workflow comprises four key steps: heterogeneous graph learning, biomedical domain knowledge encoder representation learning, Contrastive Learning, and model training. A heterogeneous graph is built with drug, microbe, and disease nodes, refined via triplet message passing with triplet attention. Simultaneously, structural features from molecular graphs, taxonomies, and biomedical ontologies are extracted using GIN and FCNs. These representations are enhanced through CL, ensuring consistency across modalities. Finally, an MLP classifier predicts drug-microbe-disease associations from the integrated embeddings.

To incorporate these attention scores into message passing, we aggregate neighbor pair messages, transforming them via a feed-forward layer and combining them with the central node representation:

$$z_i = \delta\left(h_i' + W \circ \sum_{j,k \in N(i)} (\alpha_{ijk} \circ \mathrm{NN}(h_j' || h_k'))\right), \tag{3}$$

where $W$ is a learnable parameter ensuring effective feature fusion. To enhance representation learning, we employ **multi-head attention**, where multiple attention mechanisms operate in parallel, each capturing different aspects of the data:

$$z_i = ||_{k=1}^{K} \delta\left(h_i' + W \circ \sum_{j,k \in N(i)} (\alpha_{ijk} \circ \mathrm{NN}(h_j' || h_k'))\right). \tag{4}$$

By integrating triplet attention with multi-head aggregation, our model learns enriched representations that effectively capture complex interactions among drugs, targets, and diseases. This enables us to predict novel drug-microbe-disease interactions with improved accuracy.

### B. Biomedical Domain Knowledge Encoder

In biomedical domains, entities such as drugs, microbes, and diseases possess distinct attributes that extend beyond graph topology and arise from domain-specific knowledge. These features are essential for capturing the biological and chemical characteristics necessary for understanding interactions among biomedical entities. For drugs, SMILES strings encode critical information about molecular structure. For microbes, taxonomic hierarchies offer insights into phylogenetic relationships. For diseases, MeSH descriptors provide ontological classifications. Integrating these heterogeneous biomedical features into predictive models is crucial for achieving comprehensive and informative representation learning.

To effectively leverage this domain-specific knowledge, we represent node attributes as drug features ($H_D$), microbe features ($H_M$), and disease features ($H_N$). Specifically, for drugs, SMILES string representations are transformed into molecular graphs $G = (X, A)$ using the DeepChem package [41], where atoms serve as nodes, $X$ denotes the node (atom) feature matrix, and $A$ is the adjacency matrix encoding molecular bonds. We employ a Graph Isomorphism Network (GIN) [14]. The feature extraction process for the $k$-th layer of the GIN encoder is defined as:

$$X^{(k)} = \mathrm{MLP}^{(k)}\left((A + (1+\epsilon)I)Z^{(k-1)}\right) \tag{5}$$

where $\mathrm{MLP}^{(k)}$ is a multi-layer perceptron, $I$ is the identity matrix, $\epsilon$ is a learnable scalar, and the initial node features are given by $X^{(0)} = X$. After applying Global Max Pooling (GMP) [42] to the learned molecular graph representations, we obtain a set of drug embeddings compiled as $Z_D \in \mathbb{R}^{|D| \times F}$.

For microbes and diseases, we construct similarity-based embeddings using established methodologies [43, 44]. Similarity computation is a fundamental concept in bioinformatics and network medicine [45], where it is hypothesized that biologically similar entities often participate in shared events, such as influencing the same protein or contributing to similar disease mechanisms. To compute microbe similarity, we use taxonomy, a hierarchical classification that organizes microbes based on evolutionary relationships. We further improve this by using sequence alignment, which compares the DNA or protein sequences of microbes to understand how similar they are at the genetic level. To compute microbe similarity, we adopt the method from NinimHMDA [43], which integrates taxonomic hierarchy and sequence similarity for microbes. For diseases, we rely on the Medical Subject Headings (MeSH), a hierarchical system that classifies diseases based on shared symptoms and etiologies. Diseases within the same MeSH categories often exhibit similar clinical traits. To enhance this representation, we apply Gene Ontology (GO) term similarity, which compares the molecular functions of genes associated with each disease. We follow the approach of Wang et al. [44], which measures Gene Ontology (GO) semantic similarity. This combination allows the model to capture both clinical and biological aspects of disease similarity.

After applying these similarity methods to the microbe ($H_M$) and disease ($H_N$) features, we construct a microbe similarity matrix $S_M \in 0, 1^{|M|\times|M|}$ and a disease similarity matrix $S_N \in \mathbb{R}^{|N|\times|N|}$, which encode pairwise similarities derived from taxonomic and hierarchical classifications. These matrices are subsequently transformed into low-dimensional embeddings, $Z_M \in \mathbb{R}^{|M|\times F}$ and $Z_N \in \mathbb{R}^{|N|\times F}$, using fully connected networks (FCNs).

### C.Multi-modal Contrastive Learning

Representations of drugs, microbes, and diseases learned from the heterogeneous graph capture topological triplet relationships among them. Furthermore, domain knowledge-based embeddings offer complementary insights into chemical properties, hierarchical classifications, and biomedical semantics. Combining these complementary representations ensures a more holistic embedding space where both interaction-driven and knowledge-driven similarities are effectively preserved.

However, since heterogeneous graph embeddings and structural encodings originate from fundamentally different modalities, their latent spaces may become misaligned. To mitigate this discrepancy, we introduce CL to refine and align embeddings derived from both sources. This strategy enhances the model's capacity to capture meaningful relationships by ensuring that representations of semantically similar entities are drawn closer in the embedding space, while dissimilar ones are pushed apart.

Let $\mathbf{Z}_{\mathrm{D}} \in \mathbb{R}^{|D|\times F}$, $\mathbf{Z}_{\mathrm{M}} \in \mathbb{R}^{|M|\times F}$, and $\mathbf{Z}_{\mathrm{N}} \in \mathbb{R}^{|N|\times F}$ denote the biomedical encodings produced by the BioEncoder for all drugs, microbes, and diseases, respectively. For a given triplet $tr = (d_i, m_i, n_i)$, we first gather the embeddings of the corresponding entity and concatenate them:

$$\mathbf{Z}_{\mathrm{Bio}}^{(tr)} = \big[\, \mathbf{Z}_{\mathrm{D}}[d_i] \parallel \mathbf{Z}_{\mathrm{M}}[m_i] \parallel \mathbf{Z}_{\mathrm{N}}[n_i] \,\big], \quad \mathbf{Z}_{\mathrm{Bio}}^{(tr)} \in \mathbb{R}^{|S|\times 3F}. \tag{6}$$

Concretely, let $\mathbf{Z}_{\mathrm{Bio}} \in \mathbb{R}^{|S|\times 3F}$ and $\mathbf{Z}_{\mathrm{HeTAN}} \in \mathbb{R}^{|S|\times F}$ denote the embeddings produced by the BioEncoder and HeTAN encoder, respectively, for all $|S|$ triplets.

We first define similarity between triplet as a direct cosine-similarity

$$sim(\mathbf{Z}_{\mathrm{Bio}}, \mathbf{Z}_{\mathrm{HeTAN}}^{\top}) \;=\; -\,\frac{\mathbf{Z}_{\mathrm{Bio}}\, \mathbf{Z}_{\mathrm{HeTAN}}^{\top}}{\|\mathbf{Z}_{\mathrm{Bio}}\|_F \;\|\mathbf{Z}_{\mathrm{HeTAN}}\|_F}\,, \tag{7}$$

The goal of our contrastive learning framework is to align embeddings from the two modalities, HeTAN and BioEncoder, so that semantically equivalent triplets are close in the shared embedding space, while unrelated triplets are far apart.

For each triplet $tr = (d_i, m_i, n_i)$ consisting of a drug, microbe, and disease, we treat the embeddings generated by the two modalities, $Z_{\mathrm{Bio}}^{(tr)}$ from BioEncoder and $Z_{\mathrm{HeTAN}}^{(tr)}$ from the HeTAN encoder, as a positive pair. The objective is to maximize the cosine similarity between these two embeddings, encouraging both modalities to produce consistent semantic representations for the same triplet.

To generate negative examples, we adopt a corruption-based strategy. For each positive triplet $(d_i, m_i, n_i)$, we produce $q$ corrupted triplets by randomly replacing one or more entities (drug, microbe, or disease) with a different entity from the corresponding set, ensuring that the resulting triplet is absent from the known dataset. This approach produces biologically plausible yet incorrect triplets that force the model to learn discriminative embeddings.

We employ a cosine similarity-based InfoNCE contrastive loss function. For each positive pair $(Z_{\mathrm{HeTAN}}^{(tr)}, Z_{\mathrm{Bio}}^{(tr)})$ and its $q$ hard negatives $(Z_{\mathrm{Bio}}^{(j)})$, the contrastive loss is defined as:

$$\mathcal{L}_{\mathrm{contra}} = -\frac{\exp\!\big(sim(\mathbf{Z}_{\mathrm{HeTAN_{tr}}}^{\top}, \mathbf{Z}_{\mathrm{Bio_{tr}}})\big)}{\sum_{j=1}^{q} \exp\!\big(sim(\mathbf{Z}_{\mathrm{HeTAN_{tr}}}^{\top}, \mathbf{Z}_{\mathrm{Bio_{j}}})\big)}\,, \tag{8}$$

### D. Model training

Once the embeddings for the drug, microbe, and disease are obtained, they are concatenated and passed through a Multilayer Perceptron (MLP) to estimate the probability of association. The MLP serves as a scoring function that evaluates the likelihood of interaction among the triplet entities. Formally, the predicted probability $\hat{p}$ is given by:

$$\hat{p} = \mathrm{MLP}(z_d \parallel z_m \parallel z_n) \tag{9}$$

where || denotes the concatenation operator. A higher value of $\hat{p}$ (closer to 1) indicates a strong potential association among the drug, microbe, and disease, whereas a lower value (closer to 0) suggests a weak or absent interaction. This scoring mechanism constitutes the decoder component of our model, evaluating and predicting potential drug-microbe-disease associations based on the learned embeddings.

We train our entire encoder-decoder architecture as a binary classification problem by minimizing a binary cross-entropy loss function specified as follows:

$$\mathcal{L}_{sup} = -\sum_{i=1}^{N} Y_i \log Y_i^{'} + (1 - Y_i)\log(1 - Y_i^{'}) \tag{10}$$

where $N$ is the total number of triplets, $Y_i$ is the actual label indicating the presence or absence of an interaction for the triplet, and $Y_i^{'}$ is the predicted score for the triplet.

Finally, we jointly train our model with a weighted sum of the supervised loss and contrastive loss as follows:

$$\mathcal{L} \ = \ \alpha \, \mathcal{L}_{\text{sup}} \ + \ (1 - \alpha) \, \mathcal{L}_{\text{contra}} \, , \tag{11}$$

where $\alpha$ balances the supervised prediction against the alignment of the BioEncoder and HeTAN embeddings.

# 4 Experiment

To evaluate the performance of our MCL-DMD model, we conduct experiments involving negative sampling and randomly splitting the dataset into training and testing subsets. Model effectiveness is assessed using several evaluation metrics, including Recall, Precision, F1-score, AUROC, and the widely adopted top-n metrics, Hit@n and NDCG@n. This section details our experimental setup, baseline comparisons, and result analysis.

## 4.1 Experimental Setup

**Evaluation protocols.** We evaluate MCL-DMD using the dataset described in Section 3.1, employing a stratified 80% training and 20% testing split. To ensure robustness, this process is repeated ten times with different random seeds, and the average results are reported. The model is implemented using PyTorch and PyTorch Geometric, and optimized with the Adam optimizer. The objective function uses contrastive loss. Hyperparameters are selected via grid search over learning rates 1e-2, 5e-3, 1e-3, 5e-4, dropout rates 0.2, 0.4, 0.6, and hidden dimensions 16, 32, 64, with weight decay fixed at 5e-4 and batch size set to 32. The model is trained for up to 100 epochs, with early stopping applied if validation loss does not improve for 200 consecutive epochs. These parameter values were chosen as they demonstrated a good balance between model performance and computational efficiency. The model does not take a long time to run and shows promising accuracy results, which indicates that the chosen hyperparameters contribute to an efficient and effective DMD association prediction process.

**Implementation details.** We assess the performance of MCL-DMD through a diverse set of metrics encompassing accuracy, precision, F1-score, and AUROC. In addition, to assess the model's effectiveness in identifying meaningful associations, we adopt two widely used top-n evaluation metrics: Hit@N and normalized discounted cumulative gain (NDCG@N) [46, 47]. For each test triplet, we generate an unobserved triplet as negative samples and rank the test triplet among them based on the predicted scores. Hit@N evaluates whether the true triplet appears within the top-n ranked list, while NDCG@n is a position-aware metric that assigns larger weights to higher positions.

## 4.2 Baselines

To evaluate the effectiveness of our proposed model, we conduct a comprehensive, comparative analysis against a diverse set of baseline methods. These baselines are categorized as follows:

- **Tensor Decomposition Methods:** Canonical Polyadic (CP) and Tucker decomposition are widely adopted tensor factorization techniques that have been applied in biomedical data analysis [7]. These methods rely on multilinear assumptions to infer missing associations from high-dimensional tensors.
- **Attention-based Methods:** Transformer-based models, which employ self-attention mechanisms, are used to capture complex interactions among drugs, microbes, and diseases. Transformers have demonstrated strong performance in sequence modeling and representation learning across multiple domains.
- **Graph Neural Networks (GNNs):** We include the Graph Isomorphism Network (GIN) [14], a GNN architecture that effectively aggregates and updates node representations while maintaining permutation invariance. GIN is applied to the heterogeneous graph to learn enriched node embeddings that capture relational dependencies.
- **NeurTN:** The Neural Tensor Network (NeurTN) [27] integrates tensor algebra with deep neural networks to model the nonlinear relationships among drugs, microbes, and diseases more effectively.
- **CoSTCo:** CoSTCo [48] is a CNN-based tensor completion model that captures complex interactions within higher-order tensors to predictsociations.
- **MCHNN:** MCHNN [12] improves the prediction of the association of DMD by employing a hypergraph neural network combined with multi-view CL to extract robust and expressive node features.

This comparative evaluation enables a rigorous assessment of our multi-modal CL framework relative to a wide range of existing state-of-the-art methods.

## 4.3 Comparison with Baselines

We conduct a comprehensive performance evaluation of MCL-DMD against a diverse set of state-of-the-art baseline models. To assess the classification performance, we report five core metrics: F1-score, Precision, Recall, ROC-AUC, and AUPR, as summarized in Table 1. We further evaluate each method's ability to prioritize relevant triplets using ranking metrics, including Hit@n and NDCG@n, as shown in Figure 2.

Our proposed model, MCL-DMD consistently outperforms all baselines across every evaluation metric. It achieves an F1-score of 93.65%, outperforming the strongest baseline MCHNN (84.10%). On precision and recall, MCL-DMD achieves 93.99% and 93.31%, respectively, indicating strong balance between sensitivity and specificity. The model also records the highest ROC-AUC (97.98%) and AUPR (97.56%), demonstrating its superior discrimination capacity in distinguishing positive from negative associations even under class imbalance.

The strength of MCL-DMD is further reflected in ranking-based evaluations. It achieves a Hit@3 score of 99.00%, surpassing MCHNN (95.58%) and CoSTCo (94.65%). Similarly, in terms of NDCG, MCL-DMD attains 100.00% at NDCG@3, outperforming MCHNN (NDCG@3: 94.59%) and other baselines. These results underscore MCL-DMD's ability to effectively rank high-confidence triplets, making it particularly suitable for real-world prioritization tasks in biomedical applications.

Traditional tensor decomposition methods, such as CP and Tucker, perform poorly across all metrics, with F1-scores of 37.97% and 56.09%, respectively. These methods are limited by their assumptions of linearity and struggle with the sparsity and complexity of

**Table 1: Comparison of Different Methods for Predicting Drug-Microbe-Disease Interactions**

| Model | F1-Score | Precision | ROC-AUC | Recall | AUPR |
|---|---|---|---|---|---|
| CP Decomposition | 37.97 | 46.39 | 47.02 | 32.14 | 85.95 |
| Tucker Decomposition | 56.09 | 48.42 | 47.75 | 66.66 | 87.19 |
| NeurTN | 70.2 | 65.32 | 67.28 | 75.87 | 59.71 |
| Transformer | 79.59 | 78.00 | 78.17 | 81.25 | 78.05 |
| CoSTCo | 82.77 | 84.36 | 90.61 | 79.47 | 88.94 |
| GNN | 84.01 | 81.54 | 82.66 | 86.63 | 83.19 |
| MCHNN | 84.1 | 82.71 | 92.46 | 85.54 | 71.24 |
| MCL-DMD | **93.65** | **93.99** | **97.98** | **93.31** | **97.56** |

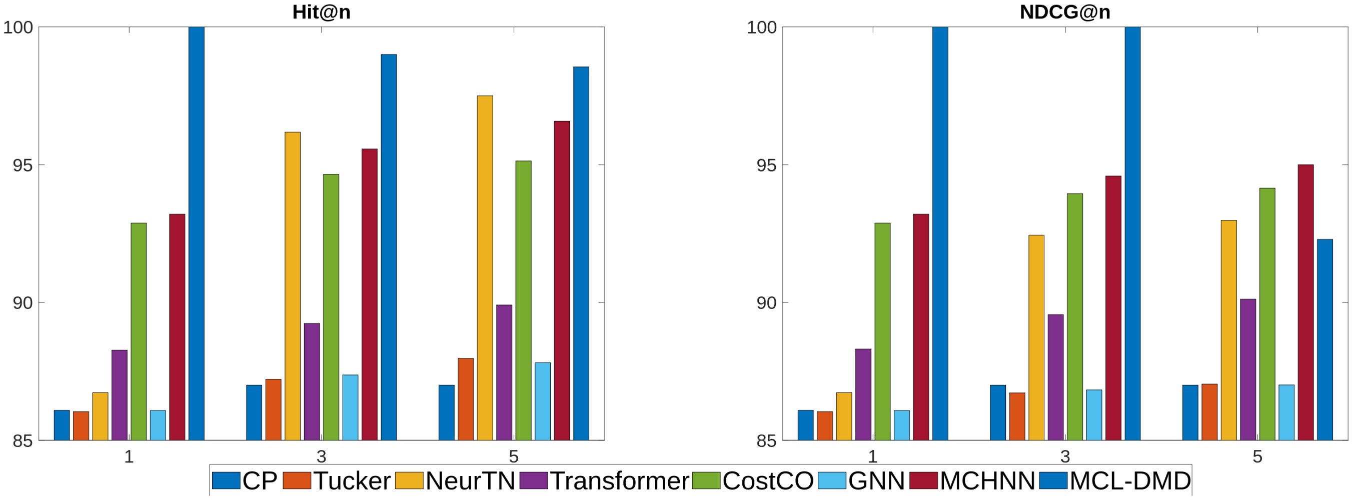

Figure 2: Evaluation of top-n performance for MCL-DMD and other baseline models in terms of Hit@n and NDCG@n

biomedical association data. Even though they can model multi-way interactions, their inability to capture nonlinear dependencies and learn rich representations constrains their performance.

NeurTN, which integrates tensor operations with neural attention mechanisms, improves over traditional decompositions but still trails significantly behind GNN-based models. Its relatively low AUPR (59.71%) and F1 (70.2%) suggest that while attention helps to some extent, it is insufficient to compensate for the lack of rich structural encoding.

Transformer-based architectures leverage self-attention to model complex dependencies, achieving moderate results (F1-score: 79.59%). However, without explicit graph-based relational modeling, Transformers underperform compared to GNN-based models, particularly in structured domains like biomedical graphs.

Among graph neural network methods, GCN and MCHNN stand out. GCN achieves an F1-score of 84.01% and AUPR of 83.19%, while MCHNN slightly surpasses it with an F1 of 84.10%, ROC-AUC of 92.46%, and AUPR of 71.24%. These gains come from their ability to encode node relationships through message passing. MCHNN further integrates pairwise similarity measures—e.g., for microbes and diseases—which enhances biological realism and representation quality. Meanwhile, CoSTCo, a multi-view collaborative learning framework, also performs competitively (F1-score: 82.77%, AUPR: 90.94%) by learning joint embeddings across different relational spaces. However, it lacks a contrastive mechanism to enforce semantic alignment between modalities, which limits its overall generalization.

In contrast, MCL-DMD integrates the heterogeneous graph with the knowledge of the biomedical domain, including molecular graphs, microbial taxonomies and disease ontologies, into a unified architecture. This comprehensive modeling is further enhanced by multi-modal contrastive learning, which aligns graph-based and domain-specific embeddings to reinforce semantic consistency and improve discriminability. The consistent superiority of MCL-DMD across all metrics demonstrates the value of combining heterogeneous graphs with domain knowledge and contrastive learning. This makes MCL-DMD a promising solution for drug-microbe-disease association prediction, offering both accuracy and interpretability in real-world biomedical discovery.

## 4.4 Ablation Study

We perform some ablation studies to assess the individual contributions of each key module in the MCL-DMD framework.

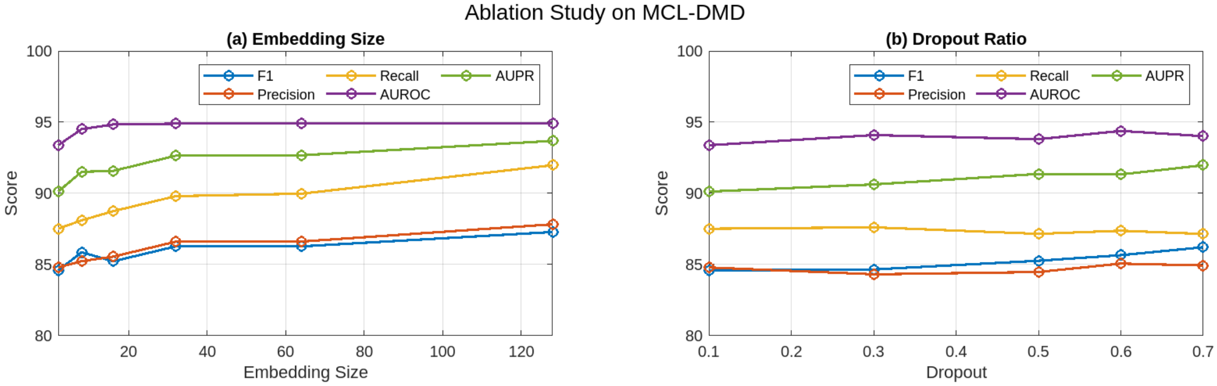

**Figure 3: Ablation study on MCL-DMD showing the impact of embedding size (left) and dropout ratio (right) on key evaluation metrics.**

**Table 2: Ablation study results for selected `MCL-DMD` variants**

| Model | F1-Score | Precision | ROC-AUC | Recall | AUPR |
|---|---|---|---|---|---|
| `MCL-DMD w/o BE` | 44.70 | 39.37 | 59.40 | 57.65 | 58.41 |
| `MCL-DMD w/o HeTAN` | 81.67 | 85.79 | 87.81 | 84.34 | 85.79 |
| `MCL-DMD w/o CL` | 71.97 | 33.97 | 78.01 | 80.20 | 76.58 |
| `MCL-DMD` | **93.65** | **93.99** | **97.98** | **93.31** | **97.56** |

### *4.4.1 Impact of Different Components.*

To evaluate the impact of different components, we examined three model variants by systematically removing specific modules and comparing their performance:

- `MCL-DMD without BioEncoder (w/o BE)`: This variant removes the biomedical domain knowledge encoder, which is responsible for capturing structured features from external knowledge sources such as drug molecular graphs (SMILES), microbe taxonomies, and disease ontologies (MeSH). The model thus relies entirely on the heterogeneous graph encoder (HeTAN) to learn representations based on the structural topology of the drug-microbe-disease network.
- `MCL-DMD without HeTAN (w/o HeTAN)`: In this configuration, we eliminate the heterogeneous graph encoder HeTAN, thereby removing any contribution from topological relationships in the triplet interaction graph. The model is trained solely on the features derived from the biomedical domain knowledge encoder. This variant allows us to isolate the predictive power of non-graph biological features — such as molecular structure, taxonomy, and disease semantics.
- `MCL-DMD without CL (w/o CL)`: In this setting, we retain both the heterogeneous graph encoder and the biomedical domain knowledge encoder, but remove the contrastive learning (CL) module that is responsible for aligning embeddings across modalities. Without CL, the model cannot explicitly enforce consistency between the graph-based and domain knowledge-based representations. This helps quantify the effect of cross-modal alignment on the quality and generalization of learned embeddings, as well as the overall predictive performance.

The results of the ablation study, presented in Table 2, highlight the significance of each module in the `MCL-DMD` framework. The variant w/o BE produces the weakest performance (F1-score: 44.70%), highlighting the indispensable role of the biomedical domain knowledge encoder in capturing complex drug-microbe-disease interactions. Removing CL also results in a substantial drop in performance (F1-score: 71.97%), validating its effectiveness in aligning modality-specific embeddings. Interestingly, even w/o HeTAN, the model still achieves a high performance (F1-score: 81.67%), with the highest precision (85.79%) among all variants, indicating that biomedical features alone offer substantial predictive capability. Nevertheless, the full `MCL-DMD` model achieves the highest scores across most metrics (F1-score: 84.59%, ROC-AUC: 93.38%, Recall: 87.52%, AUPR: 90.13%), demonstrating that the integration of heterogeneous graph structures, biomedical descriptors, and CL provides complementary strengths essential for accurate and robust triplet prediction.

### *4.4.2 Impact of Embedding Size.*

We evaluated the effect of varying the embedding dimensionality to understand how it influences the model's representational capacity. As shown in Figure 3(a), we varied the embedding size from 2 to 128 and observed its impact on performance metrics. The results reveal that performance generally improves with increasing embedding size up to a certain threshold, beyond which it stabilizes. For instance, F1-score increases from 84.59% (dim=8) to 86.28% (dim=64) and then flattens. Similar trends are observed in AUROC and AUPR, which peak around embedding sizes of 64 or 128. This suggests that moderate-dimensional embeddings provide sufficient capacity for capturing complex biological relationships, while further increases yield diminishing returns. Based on this, we adopt an embedding size of 64 as the default configuration, balancing performance and computational efficiency.

*4.4.3 Impact of Dropout Ratio.* We also explore how regularization through dropout affects generalization. Figure 3(b) shows model performance when dropout is varied from 0.1 to 0.7. Interestingly, performance fluctuates more noticeably across dropout values compared to other settings. For instance, AUROC improves with increasing dropout up to 0.6, peaking at 94.39%, and then slightly declines at 0.7. This trend is mirrored in Recall and Precision, suggesting that moderate dropout helps mitigate overfitting, especially when training with limited labeled data. However, too much dropout (e.g., 0.7) may result in underfitting, reducing the model's ability to learn complex associations. We therefore set dropout to 0.5 or 0.6 as the default to achieve optimal generalization.

## 4.5 Generalization to Unseen Entities

To assess the robustness of our framework in realistic discovery scenarios, we conduct experiments under cold-start settings, where certain entities were completely excluded from training and only introduced at test time. This setup evaluates the model's ability to generalize to previously unseen drugs, diseases, and microbes, as well as to entirely new triplets.

We designed four evaluation scenarios:

- **Cold Drug:** All associations involving a held-out set of drugs are removed during training. At test time, the model must predict their interactions with microbes and diseases using only their biomedical representations.
- **Cold Disease:** A subset of diseases is excluded during training, and their associations with drugs and microbes are evaluated at test time.
- **Cold Microbe:** The model is tasked with predicting interactions for microbes absent during training.
- **Cold Triple:** Entire drug–microbe–disease triplets are held out, requiring the model to infer associations without ever seeing the full combination during training.

Table 3 summarizes the performance across these settings.

**Table 3: Performance on unseen entities across different cold-start settings.**

| Scenario | AUC | AUPR | F1 Score | Recall | Precision |
|---|---|---|---|---|---|
| Cold Drug | 79.51 | 78.94 | 70.06 | 65.28 | 75.55 |
| Cold Disease | 78.94 | 82.30 | 67.17 | 57.99 | 79.80 |
| Cold Microbe | 64.10 | 68.30 | 52.06 | 45.73 | 70.67 |
| Cold Triple | 59.65 | 62.47 | 50.80 | 40.41 | 68.40 |

The results reveal several important trends. The model demonstrates strong generalization in the Cold Drug and Disease settings, achieving AUC values of 79.51 and 78.94, respectively. This indicates that the biomedical encoders provide sufficiently rich representations to enable accurate predictions for previously unseen drugs and diseases. Performance in the Cold Microbe and especially the Cold Triple settings is comparatively lower, suggesting that microbe embeddings derived from one-hot identifiers limit the model's ability to capture transferable biological semantics. Despite this, the results remain substantially above random, highlighting the model's capacity to generalize beyond training distributions. These findings underscore both the promise and the limitations of cold-start generalization in biomedical association prediction.

# 5 Conclusion

We introduce MCL-DMD, a multimodal CL framework for the prediction of drug-microbe-disease association that integrates heterogeneous graph structures with biomedical domain descriptors. By aligning representations across modalities, MCL-DMD effectively captures complex biological interactions better than existing methods. Extensive experiments demonstrate that MCL-DMD consistently outperforms state-of-the-art baselines in both classification and ranking tasks. Ablation studies further validate the complementary contributions of its core components—heterogeneous graph encoding, domain-specific knowledge integration, and CL. This work underscores the value of multi-modal fusion in advancing robust biomedical triplet prediction and paves the way for future research in this area. As next steps, we plan to extend MCL-DMD to incorporate temporal dynamics and evaluate its applicability to other triplet prediction tasks, such as drug–gene–disease and circRNA–miRNA–disease associations. Additionally, we aim to develop an open-access implementation of the framework to facilitate broader adoption in drug discovery and personalized medicine.

# References

[1] Yahui Long, Min Wu, Yong Liu, Chee Keong Kwoh, Jiawei Luo, and Xiaoli Li. Ensembling graph attention networks for human microbe-drug association prediction. *Bioinformatics*, 36 Supplement_2:i779–i786, 2020.

[2] Marinka Zitnik, Monica Agrawal, and Jure Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34:i457 – i466, 2018.

[3] Yueyue Wang, Xiu juan Lei, Cheng-Tsung Lu, and Yi Pan. Predicting microbe-disease association based on multiple similarities and line algorithm. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19:2399–2408, 2021.

[4] Ruoqi Liu, Lai Wei, and Ping Zhang. A deep learning framework for drug repurposing via emulating clinical trials on real-world patient data. *Nature Machine Intelligence*, 3:68 – 75, 2021.

[5] Shengbo Wu, Shujuan Yang, Manman Wang, Nan Song, Jie Feng, Hao Wu, Ai ying Yang, Chunjiang Liu, Yanni Li, Fei Guo, and Jianjun Qiao. Quorum sensing-based interactions among drugs, microbes, and diseases. *Science China Life Sciences*, 66:137–151, 2022.

[6] Lei Wang, Yaqin Tan, Xiaoyu Yang, Linai Kuang, and Pengyao Ping. Review on predicting pairwise relationships between human microbes, drugs and diseases: from biological data to computational models. *Briefings in bioinformatics*, 2022.

[7] Huiyuan Chen and Jing Li. Modeling relational drug-target-disease interactions via tensor factorization with multiple web sources. *The World Wide Web Conference*, 2019.

[8] Shandian Zhe, Kai Zhang, Pengyuan Wang, Kuang chih Lee, Zenglin Xu, Yuan Qi, and Zoubin Ghahramani. Distributed flexible nonlinear tensor factorization. *ArXiv*, abs/1604.07928, 2016.

[9] Thomas Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *ArXiv*, abs/1609.02907, 2016.

[10] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, 2017.

[11] Thomas Gaudelet, Ben Day, Arian Jamasb, Jyothish Soman, Cristian Regep, Gertrude Liu, Jeremy B. R. Hayter, Richard J Vickers, Charlie Roberts, Jian Tang, David Roblin, Tom L. Blundell, Michael M. Bronstein, and Jake P. Taylor-King. Utilizing graph machine learning within drug discovery and development. *Briefings in Bioinformatics*, 22, 2021.

[12] Luotao Liu, Feng Huang, Xuan Liu, Zhankun Xiong, Menglu Li, Congzhi Song, and Wen Zhang. Multi-view contrastive learning hypergraph neural network for drug-microbe-disease association prediction. In *International Joint Conference on Artificial Intelligence*, 2023.

[13] Farhan Tanvir, Khaled Mohammed Saifuddin, Tanvir Hossain, Arunkumar Bagavathi, and Esra Akbas. Hetan: Heterogeneous graph triplet attention network for drug repurposing. *2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, 2024.

[14] Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *ArXiv*, abs/1810.00826, 2018.

[15] Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *ArXiv*, abs/2104.11178, 2021.

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

[17] Norman Mu, Alexander Kirillov, David A. Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. *ArXiv*, abs/2112.12750, 2021.

[18] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *ArXiv*, abs/2110.05208, 2021.

[19] Lewei Yao, Runhu Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *ArXiv*, abs/2111.07783, 2021.

[20] Haoran Yang, Hongxu Chen, Shirui Pan, Lin Li, Philip S. Yu, and Guandong Xu. Dual space graph contrastive learning. *Proceedings of the ACM Web Conference 2022*, 2022.

[21] Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[22] Haoyi Zhou, Jianxin Li, Jieqi Peng, Shuai Zhang, and Shanghang Zhang. Triplet attention: Rethinking the similarity in transformers. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2378–2388, 2021.

[23] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[24] Mingmin Liang, Xianzhi Liu, Qijia Chen, Bin Zeng, and Lei Wang. Nmgmda: a computational model for predicting potential microbe–drug associations based on minimize matrix nuclear norm and graph attention network. *Scientific Reports*, 14, 2024.

[25] Bei Zhu, Yi Xu, Pengcheng Zhao, Siu ming Yiu, Hui Yu, and Jian-Yu Shi. Nnan: Nearest neighbor attention network to predict drug–microbe associations. *Frontiers in Microbiology*, 13, 2022.

[26] Marinka Žitnik and Blaž Zupan. Data fusion by matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*, 37(1):41–53, 2014.

[27] Huiyuan Chen and Jing Li. Learning data-driven drug-target-disease interaction via neural tensor network. In *International Joint Conference on Artificial Intelligence*, 2020.

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

[29] Akanksha Rajput, Anamika Thakur, Shivangi Sharma, and M. Kumar. abiofilm: a resource of anti-biofilm agents and their potential implications in targeting antibiotic drug resistance. *Nucleic Acids Research*, 46:D894 – D900, 2017.

[30] Petter I. Andersen, Aleksandr Ianevski, Hilde Lysvand, Astra Vitkauskienė, Valentyn Oksenych, Magnar Bjørås, Kaidi Telling, Irja Lutsar, Uga Dumpis, Yasuhiko Irie, Tanel Tenson, Anu Kantele, and Denis E. Kainov. Discovery and development of safe-in-man broad-spectrum antiviral agents. *International Journal of Infectious Diseases*, 93:268 – 276, 2019.

[31] Weiyang Bai, Wen Yang, Wenjing Wang, Yang Wang, Can Liu, Qinghua Jiang, Jinlian Hua, and Mingzhi Liao. An analysis of human microbe-disease associations. *Briefings in Bioinformatics*, 18:85–97, 2017.

[32] Yorick Janssens, Joachim Nielandt, Antoon Bronselaer, Nathan Debunne, Frederick Verbeke, Evelien Wynendaele, Filip Van Immerseel, Yves-Paul Vandewynckel, Guy de Tré, and Bart de Spiegeleer. Disbiome database: linking the microbiome to disease. *BMC Microbiology*, 18, 2018.

[33] Liang Cheng, Changlu Qi, Zhuang He, Tongze Fu, and Xue Zhang. gutmdisorder: a comprehensive database for dysbiosis of the gut microbiota in disorders and interventions. *Nucleic Acids Research*, 48:D554 – D560, 2019.

[34] Giorgos Skoufos, Filippos S Kardaras, Athanasios Alexiou, Ioannis Kavakiotis, Anastasia Lambropoulou, Vasiliki Kotsira, Spyros Tastsoglou, and Artemis G Hatzigeorgiou. Peryton: a manual collection of experimentally supported microbe-disease associations. *Nucleic acids research*, 49(D1):D1328–D1333, 2021.

[35] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.

[36] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. Pubchem substance and compound databases. *Nucleic acids research*, 44(D1):D1202–D1213, 2016.

[37] Scott Federhen. The ncbi taxonomy database. *Nucleic acids research*, 40(D1):D136–D143, 2012.

[38] Carolyn E Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265, 2000.

[39] System. System, D. C. I. 2015. Smiles tutorial.

[40] Kexin Huang, Cao Xiao, Lucas Glass, and Jimeng Sun. Explainable substructure partition fingerprint for protein, drug, and more. *NeurIPS Learning Meaningful Representation of Life Workshop*, 2019.

[41] Bharath Ramsundar, Peter Eastman, Patrick Walters, Vijay Pande, Karl Leswing, and Zhenqin Wu. *Deep Learning for the Life Sciences*. O'Reilly Media, 2019. https://www.amazon.com/Deep-Learning-Life-Sciences-Microscopy/dp/1492039837.

[42] Hongyang Gao and Shuiwang Ji. Graph u-nets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:4948–4960, 2019.

[43] Yuanjing Ma and Hongmei Jiang. Ninimhmda: neural integration of neighborhood information on a multiplex heterogeneous network for multiple types of human microbe-disease association. *Bioinformatics*, 2020.

[44] James Zijun Wang, Zhidian Du, Rapeeporn Payattakool, Philip S. Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23 10:1274–81, 2007.

[45] Albert Ĺaszló Barabási, Natali Gulbahce, and Joseph Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12:56–68, 2010.

[46] Jin Huang, Wayne Xin Zhao, Hongjian Dou, Ji rong Wen, and Edward Y. Chang. Improving sequential recommendation with knowledge-enhanced memory networks. *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, 2018.

[47] Jin Huang, Zhaochun Ren, Wayne Xin Zhao, Gaole He, Ji-Rong Wen, and Daxiang Dong. Taxonomy-aware multi-hop reasoning networks for sequential recommendation. *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 2019.

[48] Hanpeng Liu, Yaguang Li, Michael Tsang, and Yan Liu. Costco: A neural tensor completion model for sparse tensors. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019.