Multimodal Contrastive Representation Learning in Augmented Biomedical Knowledge Graphs

Tien Dang [1] Viet Thanh Duy Nguyen [1] Minh Tuan Le [2] Truong-Son Hy [1, 3]

[1]University of Alabama at Birmingham [2]Washington University in St. Louis

[3]Correspondence to thy@uab.edu

Abstract

Biomedical knowledge graphs (BKGs) illustrate complex relationships among biological entities, revealing key connections like drug-disease interactions. This study proposes:

- **Node Representation:** Integrate Language Models (LMs) with Graph Contrastive Learning (GCL) for better node representations.
- **PrimeKG++:** Introduce PrimeKG++, an augmented knowledge graph enriching existing BKGs with more features and data.
- **Experimental Validation:** Validate our approach with PrimeKG++ and the DPI benchmark, showing significant improvements in link prediction accuracy.

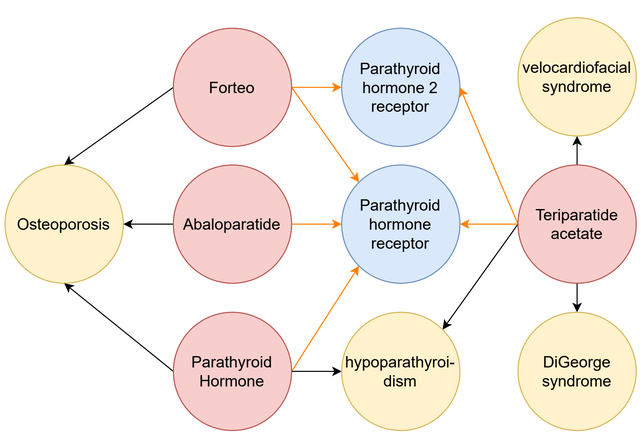

Figure 1. The subgraph illustrates the interactions surrounding the Parathyroid hormone receptor and its connections to related drugs and diseases. Different entity types are color-coded: red nodes represent drugs, blue nodes indicate genes or proteins, and yellow nodes denote diseases. Black arrows depict drug-treatment relationships with diseases, while orange arrows represent drug-receptor interactions.

PrimeKG++: An Augmented Biomedical Knowledge Graph

PrimeKG is a multimodal biomedical KG for precision medicine with **100K+ nodes**, **4M+ edges** (29 types), and rich descriptors for diseases and drugs. Yet, it **lacks contextual gene/protein data**.

PrimeKG++ addresses this by:

- Adding **biological sequences**: genes/proteins (DNA, AA), drugs (SMILES, antibodies)
- Including **text descriptions** for all key node types
- Linking to authoritative sources: **Entrez Gene** (genes/proteins), **DrugBank** (drugs)

These enrichments enable **expressive LM-based embeddings** and richer biomedical relation modeling.

Method

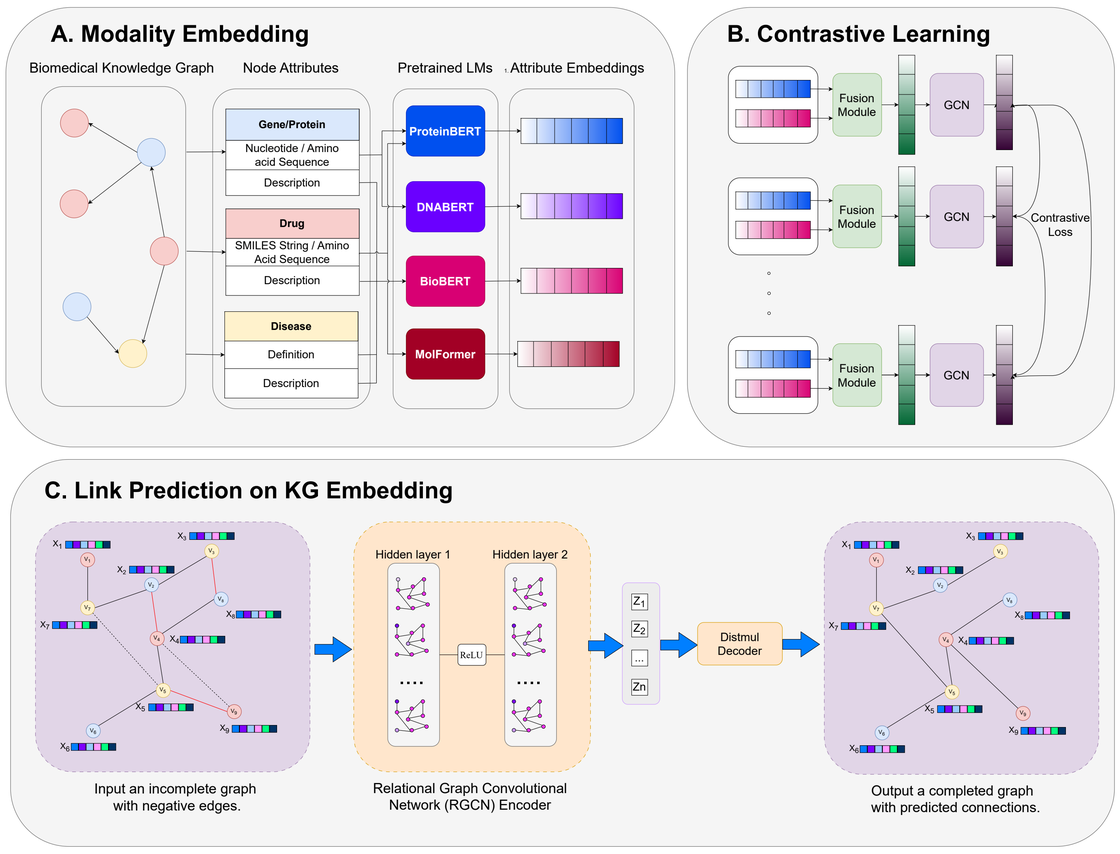

Figure 2. Overview of our proposed framework. **A. Modality Embedding**: Creating node attribute embeddings through domain-specific LMs. **B. Contrastive Learning:** Enhancement of LM-derived embeddings for specific node attributes of the same type through Fusion Module and Contrastive Learning. **C. Link Prediction on KG Embedding:** Utilizing the enhanced embeddings to perform link prediction tasks through a Knowledge Graph Embedding (KGE) model that learns relationships and enhances semantic information across distinct node types.

Experiments

- **Comparative Evaluation:** Tested various GCL models (GGD, GRACE, DGI) with different fusion strategies (None, Attention, ReDAF) on PrimeKG++. Compared against Random Init. and LM-derived embeddings.
- **Cross-Dataset Transfer Evaluation:** Pretrained GCL models on PrimeKG++ were fine-tuned on DrugBank DTI dataset to assess transferability and robustness across KGs with distinct structures and attributes.
- **Negative Sampling:** Used 1:1, 1:3, and 1:5 ratios to simulate sparse interaction scenarios and evaluate robustness under harder conditions.

Dataset	Total	Train	Val	Test
PrimeKG++	3,527,861	2,116,717	705,572	705,572
DrugBank (DTI)	18,678	13,448	1,494	3,736

Table 1. Statistics of dataset splits for training, validation, and testing.

Results and Discussion

Table 2. Link prediction performance on the PrimeKG++ dataset with varying negative sampling ratios.

Initial Embedding	Attribute Fusion	GCL Models	1:1		1:3		1:5	
			AP	F1	AP	F1	AP	F1
Random Initialization	-	-	0.980	0.960	0.945	0.893	0.909	0.829
Direct LM-derived	None	-	0.993	0.975	0.982	0.934	0.972	0.902
Our Approaches	None	GGD	0.993	0.978	0.979	0.933	0.966	0.895
	Attention	GGD	0.994	0.979	0.982	0.937	0.970	0.901
	ReDAF	GGD	0.993	0.978	0.981	0.934	0.968	0.896
	None	GRACE	**0.996**	**0.983**	0.987	**0.947**	0.979	0.916
	Attention	GRACE	**0.996**	**0.983**	0.982	0.937	**0.980**	**0.917**
	ReDAF	GRACE	**0.996**	**0.983**	**0.988**	**0.947**	0.980	0.916
	None	DGI	0.993	0.979	0.980	0.936	0.968	0.899
	Attention	DGI	0.994	0.979	0.982	0.936	0.970	0.898
	ReDAF	DGI	0.993	0.977	0.979	0.931	0.965	0.891

Table 3. Link prediction performance on the DrugBank DTI dataset with varying negative sampling ratios.

Initial Embedding	Attribute Fusion	GCL Models	1:1		1:3		1:5	
			AP	F1	AP	F1	AP	F1
Random Initialization	-	-	0.834	0.749	0.661	0.513	0.579	0.591
Direct LM-derived	None	-	0.994	0.957	**0.988**	0.884	**0.982**	0.822
Our Approaches	None	GGD	0.985	0.948	0.963	0.862	0.936	0.793
	Attention	GGD	0.9862	0.951	0.964	0.870	0.940	0.803
	ReDAF	GGD	0.9865	0.954	0.965	0.877	0.941	0.813
	None	GRACE	**0.994**	**0.972**	0.985	**0.928**	0.976	**0.887**
	Attention	GRACE	**0.994**	**0.972**	0.986	0.927	0.976	**0.887**
	ReDAF	GRACE	**0.994**	0.969	0.986	0.918	0.977	0.871
	None	DGI	0.986	0.948	0.964	0.863	0.940	0.793
	Attention	DGI	0.986	0.95	0.966	0.870	0.943	0.803
	ReDAF	DGI	0.983	0.946	0.957	0.858	0.928	0.785

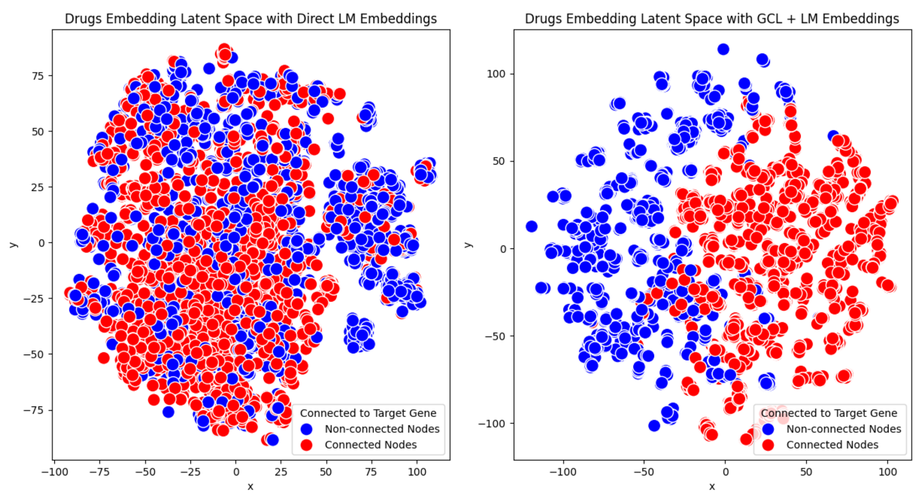

Figure 3. t-SNE visualization of drug embeddings for a single protein with the highest number of interactions in the PrimeKG++ dataset. This comparison illustrates the structural differences in the latent space resulting from the two embedding methods.

Access the Code and Preprint

GitHub Repository

Preprint on arXiv

Scan the QR codes to view the full codebase and manuscript.