Knowledge Graph-Driven Approach to Understanding Cancer Risks and Treatment Recommendations

Buz Galbraith wbg231@nyu.edu
Naveen Krishna Kodali nk3564@nyu.edu
Jon Xufan Ma xm595@nyu.edu
Hailie Nguyen hln2020@nyu.edu

Abstract

Recent increases in the volume of interconnected biomedical data have led to a surge in methods aimed at jointly utilizing these data for research purposes. Due to its prevalence and complexity, cancer is one particularly active vector of research in this area. We investigate methods for the construction and embedding of knowledge graphs, a heterogeneous data structure capable of incorporating various sources of data, and compare two methods for knowledge graph completion. RotatE is a framework for generating triplet embeddings in the complex plane, and TXGNN is a deep learning model designed to generate network embeddings capturing biomedical information surrounding a given node. Investigations of such methods and their performance confirm that knowledge graph embedding models are an effective method for multi-task inference, utilizing heterogeneous data in the medical domain.

1 Introduction

According to the Centers for Disease Control and Prevention, cancer is the second most common cause of death in the US.[1] The study of its causes and pathology as well as treatment is essential. Recent advancements in medical technologies have led to an explosion in the volume of biomedical data available for cancer and treatment development research. The interconnection of these heterogeneous data types within the medical domain resulted in a surge of research into the development of computational methods jointly utilizing these data sources. Knowledge graphs aim to integrate such data sources by organizing data in *(head, relation, tail)* triplets, where relations can describe the connection between various node types. In this paper, we aim to investigate knowledge graph embedding methods and their capabilities in facilitating downstream prediction tasks related to cancer, such as potential gene mutations or suitable treatments.

2 Related Work

Knowledge graph embedding methods aim to learn informative vector representations of a knowledge graph's nodes and edges in the same space. This allows embedding-based models to make inference across multiple triplet types. Among this class of methods, two particularly methods of interest are RotatE and TXGNN.

2.1 RotatE

The RotatE framework, proposed by Sun et al. (2019), aims to learn informative representations of a knowledge graph through embeddings in a high-dimensional complex space. Each entity and relation embedding vector is represented by $h, r, t \in \mathbb{C}^k$. Relations in this space are conceptualized

[1] NCHS Leading Causes of Death in the US

Preprint. Under review.

as normalized rotations on the complex plane. This naturally allows the model to define the score of any given (h, r, t) triplet (representing the likelihood) as the ℓ_1 distance between the tail and the head rotated by the relation, that is:

$$d_r(h, t) = ||h \circ r - t||_1^2 \tag{1}$$

This framework allows RotatE to exhibit a number of key computational properties including symmetry, anti-symmetry, inversion and composition. These properties give RotatE flexibility to model and make inference across a wide range of triplet patterns. Negative triplets are weighed by the estimated probability of their current triplet embeddings:

$$p(h_j^{'}, r, t_j') = \frac{e^{\alpha d_r(h_j^{'}, t_j^{'})}}{\sum_{i=1}^{n} e^{\alpha d_r(h_i^{'}, t_i')}} \tag{2}$$

where α is the temperature of the soft-max. The loss function is defined as:

$$\ell(h, r, t) = -\log(\sigma(\gamma - d_r(h, t))) - \sum_{i=1}^{n} \frac{1}{k} p(h_i^{'}, r, t_i') \log(\frac{1}{k} \sigma(d_r(h_i^{'}, t_i^{'}) - \gamma)) \tag{3}$$

where γ is a fixed margin, $h^{'}, t^{'}$ are arrays of length n containing negative samples of the head and tail for a given relation, and σ is the sigmoid function.

2.2 TXGNN

The TXGNN framework, developed by Huang et al. (2023), is a geometric deep learning approach designed for "zero-shot" drug predictions, meaning the model can extend drug predictions to diseases without any existing treatment. This model has two key steps:

1. **Pre-training:** TXGNN is pretrained on a medical knowledge graph, using a graph neural network, to produce biological embeddings in a latent representation space for any entity.
2. **Finetuning:** After pretraining, TXGNN is finetuned to predict relationships between drug candidates and diseases.

Mathematically, the model takes a drug-disease pair as input and provides the likelihood of the drug acting on the disease as output. For disease i, drug j, and relation r, the predicted likelihood p is calculated as:

$$p_{i,j,r} = \frac{1}{1 + \exp\left(-\sum h_i \cdot w_r \cdot h_j\right)} \tag{4}$$

where h_i represents the embedding for node i, and w_r is a trainable weight vector for relation type r.

The training loss is calculated via binary cross entropy loss:

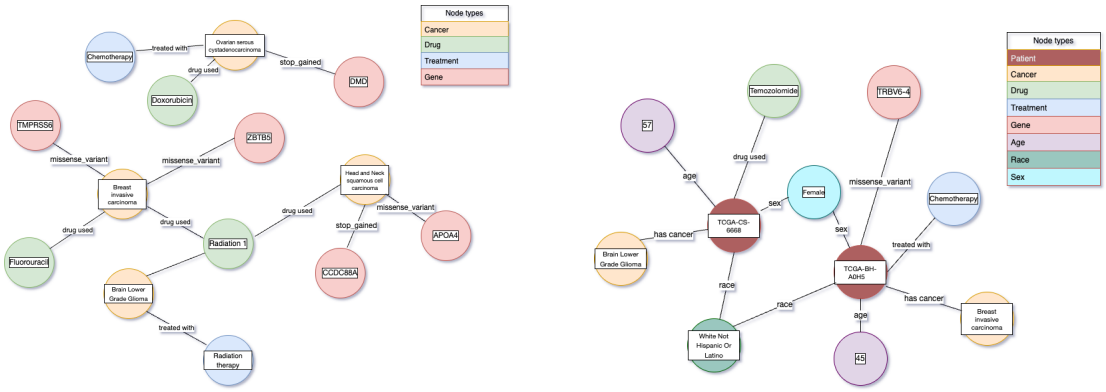
$$\mathcal{L} = \sum_{i,r,j} y_{i,r,j} \cdot \log(p_{i,r,j}) + (1 - y_{i,r,j}) \cdot \log(1 - p_{i,r,j}) \tag{5}$$

where $y_{i,r,j} = 1$ represents all pairs (i, j) with relation types r in the positive and negative samples. TXGNN capitalizes on the insight that diseases are intrinsically related by leveraging disease molecular mechanisms to learn disease similarities through their signature vectors and transfer knowledge between these diseases.

2.3 Method Comparison

It is worth highlighting two key dissections between TXGNN and RotatE.

1. **Type of learning:** RotatE is not a "model" in the traditional sense. It is a framework to construct embeddings within the complex plane, and calculate the score of each (h, r, t) triplet given these embeddings. On the other hand, TXGNN is a deep learning model, which learns a numerical vector for each node by transforming initial node embeddings through several layers of local graph-based non-linear function transformations.



(a) Cancer-focused (b) Patient-focused

Figure 1: Segments of constructed knowledge graphs

2. **Generalizability:** These two methods differ in terms of their inference generalizability. As RotatE learns the embeddings of all nodes and edges from a graph, its inference process functions like a lookup table. That is, if it has yet to observe a node or edge type, it cannot make predictions on it. On the other hand, TXGNN is designed to excel at making zero-shot predictions for diseases not observed during training.

These dissections serve to highlight the specific use cases for each method. RotatE is relatively simple, makes few assumptions about its inference task, and is constrained by its learned representations. TXGNN is comparatively complex and is less constrained by its input data.

3 Problem Definition and Algorithm

3.1 Task

The primary motivation for our project is the investigation of machine learning methods to utilize heterogeneous biomedical data to make cancer-focused predictions. To this end, we construct biomedical knowledge graphs by leveraging multimodal patient information and investigate knowledge graph embedding models' capabilities in facilitating downstream predictive tasks.

3.2 Data & Knowledge Graph Construction

The Cancer Genome Atlas (TCGA) is a joint effort between NCI and the National Human Genome Research Institute, which molecularly characterizes over 20,000 primary cancer and matches normal samples spanning 33 cancer types.[2] Using data spanning demographics, mutational profiles, treatment history of 10,104 cancer patients, we construct two knowledge graphs as described below.

3.2.1 Cancer-Type Focused Knowledge Graph

A cancer-focused knowledge graph was constructed with 22,848 entities, 33 relation types, 4 (h, r, t) triplet types. The triplet types considered in this graph are: *(cancer, drug used, drug)*, *(cancer, treated with, treatment)*, *(cancer, mutation, gene)*, *(gene, upregulate/downregulate, cancer)*. This knowledge graph is illustrated in Figure 1a. Models trained on this type of knowledge graph are suitable for cancer-focused inference tasks, such as predicting therapeutic candidates given a cancer type, or genes and mutations that are likely to cause certain types of cancer.

3.2.2 Patient-Focused Knowledge Graph

The second type of knowledge graph we constructed is patient-centered, illustrated in Figure 1b. We primarily employ this knowledge graph to generate patient embeddings and observe patient clustering

[2]The Cancer Genome Atlas Program

patterns by cancer type. The (h, r, t) triplet types considered in this graph are: *(patient, age, years)*, *(patient, race, racial background)*, *(patient, biological sex, sex assigned at birth)*, *(patient, drug used, drug)*, *(patient, treated with, treatment)*, *(patient, mutation, gene)*. Excluding cancer from the patient embedding process enhances the model's ability to capture nuanced and clinically relevant latent features. The assessment of model performance and exploration of alternative algorithms suitable for executing inference tasks on this type of knowledge graphs is deferred to future work.

3.3 Algorithms

We principally investigated the use of RotatE and TXGNN, described previously in **Related work**.

3.3.1 RotatE

RotatE can be directly applied to produce embeddings of all entities in our knowledge graph. It is worth noting, however, that RotatE initializes the probabilities of drawing negative samples uniformly across all node types. This initialization approach presents two challenges. First, even though $p(h', r, j')$ is expected to approach zero for triplets not observed in the training data over the course of training, it still introduces noise during early training epochs. Second, in cases where there is a large number of triplet types or rare triplet types, the probabilities may become very small, affecting numerical stability. To address this, we explore a stratified negative sampling procedure that draws negative samples from nodes belonging to the same node type. This aligns with the negative sampling approach employed by TXGNN, facilitating meaningful result comparisons between the two models. Negative samples are drawn following the same distribution outlined in Equation 3.

3.3.2 TXGNN

TXGNN was specifically designed to leverage biological knowledge, including disease protein profiles and disease-disease relationships, to make predictions on drug indications and contraindications. The team met with the model's author to discuss the model's methodology and came to a consensus that TXGNN is best suited to perform disease-centric inference tasks, such as predicting therapeutic candidates given a cancer type.

3.4 Experiments conducted

We conducted four main types of experiments.

1. **RotatE Negative Sampling Method Comparisons:** During this experiment, we train and validate RotatE on our cancer-focused knowledge graph using both uniform negative sampling initialization across all nodes and stratified negative sampling strategies, and compare the results.
2. **RotatE Patient Embedding Clustering by Cancer Type:** We plot the patient embeddings produced by implementing RotatE on our patient-focused knowledge graph and color them by cancer type to visualize potential patient clustering patterns using UMAP.
3. **TXGNN Parameter Explorations:** We experiment with two disease embedding aggregation mechanisms, *heuristic* and *rarity* (Huang et al., 2023), which serve to integrates similar diseases into an auxiliary embedding that subsidizes the original disease embeddings.
4. **RotatE and TXGNN Comparisons:** We train both models on our cancer-focused knowledge graph, and compare similar metrics to benchmark their performance. Preprocessing is performed to format the knowledge graph into a suitable structure for each model.

4 Experimental Evaluation

4.1 Methodology

- **RotatE** ranks test triplets against all other candidate triplets not appearing in the training, validation, or test set, where candidates are generated by corrupting subjects or objects.
- **TxGNN** aims to to predict whether or not a relation holds given two entities in the knowledge graph, which can be formulated as a binary classification task for each relation.

Table 1: RotatE results

	Negative sampling from all node types			Stratified negative sampling		
	MRR	**Hits@1**	**AUROC**	**MRR**	**Hits@1**	**AUROC**
(Cancer, Gene)	0.853	0.836	0.740	0.786	0.753	0.703
(Cancer, Treatment)	0.937	0.880	0.701	1.000	1.000	0.731
(Cancer, Drug)	0.757	0.722	0.844	0.733	0.704	0.881

Table 2: TXGNN results

	Heuristic				Rarity			
	Micro AUROC	**Macro AUROC**	**Micro AUPRC**	**Macro AUPRC**	**Micro AUROC**	**Macro AUROC**	**Micro AUPRC**	**Macro AUPRC**
Pretraining	0.798	0.852	0.790	0.863	0.816	0.809	0.800	0.848
Finetuning								
Drug	0.884	0.884	0.863	0.856	0.883	0.884	0.861	0.856
Treatment	0.870	0.862	0.838	0.831	0.869	0.863	0.836	0.829

- **Metrics:**
  - MRR $= \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\text{rank}_i}$
  - Hits@N $= \frac{\text{Number of relevant items in top } N}{\text{Total number of relevant items}}$
  - AUROC = $\int_0^1$ True Positive Rate (Sensitivity) d(False Positive Rate (1-Specificity))
  - AUPRC = $\int_0^1$ Precision d(Recall)

4.2 Results

1. **RotatE Negative Sampling Comparison:** See Table 1.
2. **RotatE Patient Embedding Clustering by Cancer Type:** See Figure 2.
3. **TXGNN Aggregation Method Comparison:** See Table 2.
4. **Method Compassion:** See Table 3.

4.3 Discussion

1. **RotatE Negative Sampling Comparisons:** As can be seen Table 1, the embedding produced by RotatE with stratified negative sampling perform comparably or slightly better than those produced by RotatE with uniform negative sampling initialization across all nodes.
2. **Cancer Type Clustering:** As can be seen from Figure 2, our clustering indicates that it is likely that patients can be embedded into distinct clusters in some high-dimensional space.
3. **TXGNN Aggregation Methods:** The two experimented aggregation methods appear to deliver comparable results.
4. **Method Comparisons:** Compared to RotatE, TXGNN appears to deliver consistently high quality results on treatment and drug predictions.

Table 3: Model comparison (stratified negative sampling)

	RotatE AUROC	**TXGNN AUROC**	
		Heuristic	**Rarity**
(Cancer, Treatment)	0.731	0.899	0.899
(Cancer, Drug)	0.881	0.874	0.873

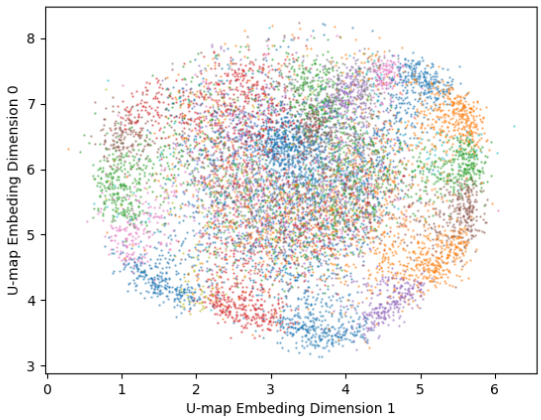


Figure 2: UMAP visualization of patient embeddings by cancer type

5 Conclusions

We can draw the following conclusions about our data, limitations and directions for future work.

- **Conclusions from data:** Our results indicate that knowledge graph embedding methods are an effective way to make healthcare-related inference using heterogeneous biomedical data.
- **Limitations:** Certain cancer types and mutations are more likely than others. As a result, our cancer-focused knowledge graph has some balance issues, which may cause the model to overfit, reducing generalizability.
- **Future Directions:** The main directions for future work are presented as follows:
  1. **Patient-Focused Knowledge Graph:** In this project, we mainly utilize our patient-focused knowledge graph to generate embeddings for patient clustering purposes. In an effort to improve clinical applicability, future models may further investigate patient-focused knowledge graphs. Potential algorithms one may consider on patient-level data are DeepWalk and GraphSAGE.
  2. **Developing Clearer Base Models:** In order to have a more robust model performance understanding, further consideration may be put into developing suitable baseline models, based on non-graphical data.
  3. **Negative Sampling Methods:** It would be worthwhile to further validate the effect of various negative sampling methods in the contrastive loss function for RotatE.
  4. **Supplemental Data for TXGNN:** During the scope of this project, we solely pretrained TXGNN on TCGA data. Future work can consider combining TCGA data and the knowledge graph constructed by Huang et al. (2023) for pretraining purposes.

6 Lessons Learned

The lessons we learned over the course of this project primarily centered around increasing our skills and sharpening our theoretical interests. The project introduced us to the intricacies of knowledge graphs and the handling of both knowledge graphs and biological data—a new experience for some of us. Engaging with Dr. Dey, an expert in deep graph learning, and other experts in the biomedical field further deepened our group's interest in utilizing knowledge graphs in machine learning and addressing the methodological challenges posed by heterogeneous medical data.

Throughout the project, we honed our skills in setting up Singularity environments and executing jobs on High-Performance Computing (HPC) platforms. The process of adapting baseline frameworks and codes to align with our specific data and objectives was particularly insightful. This hands-on involvement in real-world research significantly contributed to our growth as data scientists.

Acknowledgments

We would like to express our most sincere gratitude to our mentor, Dr. Kushal Dey from the Memorial Sloan Kettering Cancer Center, for their invaluable guidance and support throughout the duration of this project. Additionally, we would like to thank Dr. Jacopo Cirrone, Dr. Brian McFee, Dr. Saadia Gabriel, and Dr. Elisha Cohen at New York University, our advisors, for their feedback and support. Finally, we would like to thank Kexin Huang and other members of Dr. Dey's lab for their time and guidance.

Contributions

- **Buz Galbraith (wbg231@nyu.edu):** Primarily responsible for constructing the cancer-focused knowledge graph and patient-focused knowledge graph used in RotatE, and the RotatE method development and implementation. Also was proactive in organizing group logistics such as weekly presentations and discussions to allocate work.
- **Naveen Krishna Kodali (nk3564@nyu.edu):** I'm Responsible for regenerating the models on TCGA Dataset.
- **Jon Xufan Ma (xm595@nyu.edu):** Collaborate with team on visualization, poster report generation, model implementation, result discussions.
- **Hailie Nguyen (hln2020@nyu.edu):** Primarily responsible for implementing TXGNN and constructing cancer-focused knowledge graph used in the TXGNN implementation. Contributed to poster and report writeups and visualizations.

References

Huang, K., Chandak, P., Wang, Q., Havaldar, S., Vaid, A., Leskovec, J., Nadkarni, G., Glicksberg, B., Gehlenborg, N., and Zitnik, M. (2023). Zero-shot prediction of therapeutic use with geometric deep learning and clinician centered design. *medRxiv*.

Sun, Z., Deng, Z.-H., Nie, J.-Y., and Tang, J. (2019). Rotate: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*.