STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **x**, Month 202x, pp 0–12.
Published online in International Academic Press (www.IAPress.org)

Relation type-aware knowledge graph embeddings for biomedical data: a semantically adaptive framework

Senhaji Yassine [1], El Moutaouakil Karim [1,*], Oulaika Abdelfattah [1], EL Marnissi Boujemaa [2], Hafidi Youssef [3]

[1]*Laboratory of Mathematics and Data Science, Sidi Mohamed Ben Abdellah University Multidisciplinary Faculty of Taza, Taza, Morocco*
[2]*Higher Institute of Nursing Professions and Health Techniques of Fez-Morocco*
[3]*hospital pharmacy department, Hassan II University Hospital of Fez-Morocco*

Abstract Knowledge graphs (KGs) are increasingly used in biomedicine to integrate and reason over heterogeneous data such as genes, proteins, drugs, and diseases. However, existing knowledge graph embedding (KGE) methods typically rely on a single, fixed scoring function for all relation types, which limits their ability to model the semantic diversity of biomedical interactions. In this work, we propose a relation type-aware KGE framework that dynamically adapts scoring functions to the structural nature of relations hierarchical, symmetric, or asymmetric thereby improving the semantic fidelity of embeddings. We improve training with a constraint-aware negative sampling approach that creates realistic false examples. This forces the model to focus on learning true biomedical relationships instead of dismissing easy, nonsensical ones. We built our knowledge graph by grabbing biomedical papers from PubMed. Then, we ran them through some natural language processing tools. After that, we used the SemRep server to make subject-relation-object triples, which gave us the graph structure for studies on colorectal cancer. Our KG includes entities and relations at the molecular, clinical, and pharmacology levels. We set up the learning objective as a minimization problem with regularization. This combines margin-based ranking loss with relation-specific changes. We checked our work using MR, MRR, Hits@1, and Hits@10, which let us do a good comparison with baselines like TransE, ComplEx, RESCAL, and ConvE. The results showed our system improved MRR by 9-14% and Hits@10 by as much as 12% compared to other good systems. It was especially good at understanding things like subtype_of and interacts_with. In short, this research gives a way that can be scaled, is based on math and biology for filling in the blanks in biomedical knowledge graphs. By putting together relation-aware modeling, constraint-guided learning, and several ways of measuring, our method moves forward knowledge graph-based in cancer research and gives a method that can be copied for use in other bio fields.

Keywords Colorectal Cancer, Knowledge Graph Embeddings, Triple Classification, Negative Sampling

AMS 2010 subject classifications 68T05, 92B20

DOI: 10.19139/soic-2310-5070-3706

1. Introduction

Biomedical knowledge graphs (KGs) are a systematic way of representing biological entities that differ from one another and their relationships generally showing complex associations between each of genes, proteins, diseases, and drugs [1]. KGs structure information into machine-readable triples which are essential to support a number of important applications such as drug repurposing, disease subtyping, treatment recommendation and personalized medicine [2]. KGs allow for the integration of evidence from a variety of biomedical information sources, thus enabling systematic knowledge discovery and predictive inference at scale.

[*] Correspondence to: El Moutaouakil Karim (Email: karim.elmoutaouakil@usmba.ac.ma). Laboratory of Mathematics and Data Science, Sidi Mohamed Ben Abdellah University Multidisciplinary Faculty of Taza, Taza, Morocco.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

Even with these benefits, one major challenge relates to the representation learning of biomedical KGs. Knowledge graph embedding (KGE) techniques project entities and relations into continuous vector spaces to provide for efficient reasoning and link predictions. Classic methodologies such as TransE [3] and RotatE [4] work well in general domains; nevertheless, these approaches treat all relations the same and use the same scoring function regardless of the relation's semantic meaning. This practice ignores important differences like hierarchical (e.g., *subtype_of*) and symmetric (e.g., *interacts_with*) relations, which are both commonly found in biomedicine. Consequently, traditional KGE models do not effectively capture the structural and functional characteristics of biomedical interactions, and reduce biological plausibility and interpretability [5].

To remedy these limitations, we present a relation type–aware KGE approach that allows the scoring functions to adjust dynamically according to the relation types (i.e., hierarchical, symmetric, and general). Aligning the low-dimensional representations with the properties of relations leads to directional continuity for disease taxonomies while maintaining bidirectionality for drug–drug interactions and improving the semantic fidelity of representations in the embedding process as a whole [6, 16]. This design enhances the capacity of the embeddings to capture biologically meaningful patterns and supports downstream reasoning tasks and helps with improved accuracy and interpretability with downstream tasks.

The second contribution of our work is a constraint-aware negative sampling design. Conventional negative sampling applies a random sampling strategy to corrupt triples, often generating biologically implausible examples as examples (e.g., ”Gene X treats Democracy”), which weakens the training signal and the trustworthiness of the model [7, 15]. Our approach constrains the negative sampling to type-consistent entities and therefore generates hard but biologically valid negatives (e.g., ”Gene X treats Disease Y” only corrupted with other valid disease entities). This loss function compels the model to discriminate between true and plausible-but-false statements, thereby strengthening its generalized ability and improving potential performance on downstream link prediction and inference reasoning tasks.

Third, we utilize a PubMed-based relation typing workflow to build the biomedical knowledge graph. Initially, we collect biomedical articles from PubMed based on PubMed semantic articles and processes the information through natural language processing (NLP) tools (e.g., SciSpacy[8]) to determine and normalize entities. Then, we upload the extracted text to the SemRep server[17], which extracts subject-relation-object triples based on biomedical semantics. The two-step process accomplishes a few important tasks. First, the knowledge graph is completely based on empirical biomedical evidence. Second, it minimizes manual annotation. Finally, it scales to large corpora while maintaining the correct biomedical semantics of extracted relations.

Finally, we focus on biomedical applications in the context of cancer research, specifically colorectal cancer research. Cancer progression and treatment planning imply hierarchical relationships (e.g., tumor subtype classification) and symmetric relationships (e.g., drug-drug and/or protein-protein relationships). The application of our relation-aware embedding framework to colorectal cancer examples shows improvements in predicting novel drug-drug interactions, informing subtyping types of the disease, and identifying potential therapeutic targets.

In conclusion, the contributions of this study can be summarized into three items: 1) a relation type–aware embedding framework that customizes scoring functions to the semantics of biomedical relations, 2) a constraint-aware negative sampling strategy for more robust and plausible KG training, and 3) a PubMed–NLP pipeline to scale relation extraction and KG construction. Taken together, these contributions represent progress in the area of biomedical KG modeling, offering a biologically motivated, semantically rich, and computationally scalable solution for knowledge generation in oncology and beyond.

2. Related Work

Utilization of knowledge graphs (KGs) within the biomedical field is receiving increasingly high interest from academics and research professionals. This is primarily due to KGs' ability to combine facets of heterogenous datatypes and types of analysis for discovery and insight generation. This section reviews applied research for colorectal cancer studies done with KGs and advances that have been made into knowledge graph embedding. KGs are increasingly part of biomedicine to illustrate the complex relationships between genes, diseases, drugs, and clinical phenotypes. Some represent disease classification or drug discovery approaches for individualized treatment recommendations, allowing multimodal data sources. For example, QAnalysis[9], exemplifies the translational power of a biomedical KG by converting natural language questions into graph-based structured queries, in order to help speed the derivation of insights from literature and clinical data bases. The application to analysis of colorectal cancer has allowed findings reviews to occur and realize milestones to breakthroughs: created clinical guideline-molecular data recommendations framework to allow for personalized therapies[11]. predicted cancer subtypes relating mutational profiles to disease progression pathways[10]. The success of these kinds of applications rely on success with knowledge graph embedding (KGE) techniques that help leverage and represent the different entities and relations in a low-dimensional vector spaces.

Translational models, such as TransE, indicated baseline performance using distance-based scoring ($\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|$); however, they do not work with biomedical complexity, especially one-to-many relationships, such as gene pleiotropy, and many-to-many drug interactions. Subsequent improvements to translational models have addressed this issue. RESCAL[22] uses tensor factorization to capture pairwise interactions across multiple entities RESCAL models have been shown to accurately predict polypharmacy side effects. ComplEx[21] modifies embedding to include complex-valued representation, which allows for modeling of asymmetric relations, prevalent in networks of information describing disease progression. ConvE[19] combines convolutional neural networks with knowledge graph embeddings to capture local patterns of interaction from drug-target binding interaction graphs.

The applied KGE approaches described above used methods which advance biomedical applications—such as how ComplEx captures asymmetric relations that improves the accuracy of colorectal cancer subtype predictions in ontologies[10], and how ConvE's feature learning ability underpins predictive analytics capabilities (and resolutions of ambiguous entity linkages) from literature-derived biomedical knowledge graphs, like QAnalysis. Nevertheless, current models struggle with mapping biomedical-specific domain constraints, such as the directional logic built into the relation *treats*, as well as the transitivity of the *biomarker_for* relation hierarchy. These gaps in biomedical knowledge graph representation learning are addressed specifically by our relation type-aware adaptation framework.

Although knowledge graphs (KGs) provide a great deal of opportunity for application in the biomedical domain, the current state of the art encounters multiple important limitations. Most work studies only a small number of specific disease classes or homogeneous data, limiting how well their proposed methodology can generalize. As an example, cardiovascular disease-specific models are seldom appropriate for application in any oncology use case [18]. Furthermore, embeddings produced through commonly used methods for KGs have been pushed to scale sufficiently to large noisy biomedical graphs, which is remarkably affected when the data is incomplete or inconsistent between literature (e.g., when different literature sources report conflicting gene-disease relationships) [15]. These limitations reveal a tension between producing biological-rich pipelines using domain knowledge, while also focusing on scalability, which introduces important challenges [5]. To address these challenges, this paper builds off of existing KGE state of the art to introduce a relation type-aware KGE framework for application in colorectal cancer, which captures hierarchical, symmetric, and compositional relations that are inherent to the use case of oncology KGs.

In contrast to generic methods, our adaptive scoring functions maintain directional reasoning in taxonomies (*subtype_of* hierarchies in tumor classifications), and symmetry in drug interactions (*combines_with* relationships) [16]. We address data noise through biomedical constraint-aware negative sampling, which implies rules and restrictions during the training time; specifically, that corrupted triples must satisfy ontological constraints (e.g., replacing only anatomically reasonable entities in *located_in* relations) [7]. To improve generalizability, we introduce PubMed-driven relation typing that automates relation extraction and typing through *SemRep* [17] and *SciSpacy* [8] to ground KG semantics against evidence from over 250,000 PubMed abstracts.

3. Materials and Methods

3.1. Study Design and Setting

The current study is a retrospective computational design approach on building and analyzing a Biomedical Knowledge Graph (KG) related to colorectal cancer. The primary goal of this study is to create a relation type-aware Knowledge Graph Embedding (KGE) framework that overcomes semantic deficiencies present in previous models by incorporating heterogeneous biomedical data from the molecular, pharmacological and clinical datasets. A directed multi-graph called a Knowledge Graph aggregates and organizes factual information.

A KG can be defined formally as a collection of triples $\mathcal{T} \subseteq \mathcal{E} \times \mathcal{R} \times \mathcal{E}$, where: $\mathcal{E}$ is a finite collection of entities (node) and $\mathcal{R}$ includes all of the possible relation types (the directed edges). A factual statement is represented by each triple $(h, r, t) \in \mathcal{T}$, where the head entity is $h \in \mathcal{E}$, the tail entity is $t \in \mathcal{E}$, and the relation between them is $r \in \mathcal{R}$. The triple $(\text{BRCA1}, \text{associated\_with}, \text{Breast Cancer})$, for example, captures a known gene-disease association in a biomedical KG. One of the main goals of KG analysis is link prediction, which is to use the current graph structure to infer missing triples. The task is to predict the most likely head or tail entity from the set $\mathcal{E}$ given an incomplete query like $(h, r, ?)$ or $(?, r, t)$.

3.2. Study Participants and Sampling

This study consist of biological entities that are represented as nodes in the constructed knowledge graph. The sampling of training examples from the KG is performed through various sampling methods. The info captured in the knowledge graph (KG) comes from the representation of the knowledge or data of each biological entity $\mathcal{E}$; where the entity may exist in more than one biological class (e.g., a gene, drug, protein or disease). The types of relationships $\mathcal{R}$ connecting each of those entities are categorized into three distinct categories depending on the types of characteristics (i.e., structural features) of the relationship: hierarchical (e.g., *subtype_of*), symmetric (e.g., *interacts_with*), and directional (e.g., *treats*) relationships.

To train the model effectively, we employ a specific constraint-aware negative sampling strategy. For every positive triple $(h, r, t) \in \mathcal{T}$, the model generates a set of negative (corrupted) triples $\mathcal{N}(h, r, t)$. Naive random sampling often produces unrealistic examples, such as replacing a gene with a city, which provides weak training signals. We counteract this by using type-constrained negative sampling. Let $\tau : \mathcal{E} \to \mathcal{T}_{\text{sem}}$ be a function that associates each entity with its semantic type. For a given relation r, we pre-define the set of permitted head types $\mathcal{A}_r^{(h)}$ and tail types $\mathcal{A}_r^{(t)}$. Negative samples (h', r, t') are produced by substituting entities that strictly meet these constraints. Specifically, a corrupted head h' is sampled from $\{e \in \mathcal{E} \mid \tau(e) \in \mathcal{A}_r^{(h)}\}$, and a corrupted tail t' is sampled from $\{e \in \mathcal{E} \mid \tau(e) \in \mathcal{A}_r^{(t)}\}$. This process ensures that the model learns to discriminate between true associations and biologically plausible but false hypotheses.

3.3. Data Collection Tools and Technique

A biomedical knowledge graph will be developed by collecting and pre-processing existing data on Colon Cancer in an organized way. The Biomedical publications will first be imported from the PubMed and DrugBank databases that are publicly accessible and are focused on Colon Cancer keywords. We will utilize Natural Language Processing techniques to convert the unstructured text into structured data, utilizing *SciSpacy*. This process will establish a set of preprocessing techniques that will create a normalized representation of biological terms for the unstructured text via Tokenization, Sentence splitting, and Named Entity Recognition (NER). The preprocessed texts will then be sent to the *SemRep* server for extracting semantic relationships and types of relationships. Using the extracted relationships we will create a structured dataset containing Subject-Relation-Object triples which include associations between diseases, genes and drugs (e.g., a Gene associated with a Disease and the Gene's association to a Drug). This dataset will become the foundation of our knowledge graph and the inputs to the knowledge graph embedding process.

3.4. Data Analysis

Because the basic topologies of biomedical relations are different, the scoring functions applied to them vary. For example, the subtype_of hierarchical relation includes taxonomical structures, and the graph's depth is kept whole with transitive and anti-symmetric relations. The interacts_with symmetric relations show mutual actions (i.e., a switch of head/tail changes nothing) while the directionality of the edge is irrelevant. The directional (i.e., treats) relations depict the functional/causal flow of events from source to target. When these categories are assigned their respective scoring geometries, the model's internal logic matches the actual biological event(s).

Data analysis centers on the formulation of the scoring functions, the optimization of the loss function, and the statistical evaluation of link prediction performance.

Scoring Functions and Model Architecture. Our approach differs from standard models that use just one scoring process across the entire graph. By mixing methods, we stop 'semantic collapse.' If one process is used for varied relationships, the model might not tell apart a hierarchical link from a symmetric interaction. We use TransE for direction, hyperplane projections for hierarchy, and ComplEx for symmetry, making an adaptive mix. This makes sure that the features of one relationship do not mess with another, improving MRR by 9-14%.

The core analysis relies on a hybrid scoring function $s : \mathcal{E} \times \mathcal{R} \times \mathcal{E} \to \mathbb{R}$ that adapts to the relation type. For general directional relations ($r \in \mathcal{R}_{\text{dir}}$), we adopt a standard translational model where relationships are interpreted as translation operations. The score is defined as $f_{\text{r\_dir}}(h,t) = -\|\mathbf{h} + \mathbf{r} - \mathbf{t}\|_p$, where the norm $p \in \{1,2\}$. For hierarchical relations ($r \in \mathcal{R}_{\text{hier}}$), we apply a projection-based approach. A normal vector $\mathbf{w}_r \in \mathbb{R}^d$ defines a hyperplane, and a translation vector $\mathbf{d}_r \in \mathbb{R}^d$ lies on that hyperplane. The head and tail embeddings are first projected onto this plane via $\mathbf{h}_\perp = \mathbf{h} - \mathbf{w}_r^\top \mathbf{h} \mathbf{w}_r$ and $\mathbf{t}_\perp = \mathbf{t} - \mathbf{w}_r^\top \mathbf{t} \mathbf{w}_r$. The scoring function then becomes:

$$f_{\text{r\_hier}}(h,t) = -\|\mathbf{h}_\perp + \mathbf{d}_r - \mathbf{t}_\perp\|_p.$$

Finally, for symmetric relations ($r \in \mathcal{R}_{\text{sym}}$), we employ complex-valued embeddings where $\mathbf{h}, \mathbf{r}, \mathbf{t} \in \mathbb{C}^d$. The scoring function is defined as the real part of the Hermitian product:

$$f_{\text{r\_sym}}(h,r,t) = \text{Re}(\langle \mathbf{h}, \mathbf{r}, \overline{\mathbf{t}} \rangle) = \text{Re}\left(\sum_{k=1}^{d} h_k r_k \overline{t_k} \right),$$

The unified scoring function for any triple (h,r,t) is then selected according to the type of relation:

$$f_r(h,t) = \begin{cases} -\|h + r - t\|_p & \text{if } r \in \mathcal{R}_{\text{dir}} \\ -\|h_\perp + d_r - t_\perp\|_p & \text{if } r \in \mathcal{R}_{\text{hier}} \\ \text{Re}\left(\langle h, r, \bar{t} \rangle\right) & \text{if } r \in \mathcal{R}_{\text{sym}} \end{cases}$$

Using a logical process, and the three categories as shown in table 1, the 93 relations were divided into these three categories for the reasons outlined below. The Hierarchical category included relations such as (ISA) and (PART_OF) that contained part/whole or taxonomic-form structural properties. This was due to the necessity of hyperplane projections in order to maintain transitive directionality. The Symmetric category included relations such as (INTERACTS_WITH) which were inherently bidirectional as they had a similarity between the two directions in terms of a generalization (i.e. $Score(h,r,t) \approx Score(t,r,h)$). Directional categories included all other relations such as (TREATS) (CAUSES) and (COMPARES) that were determined by their functional, casual or comparative properties, and their relative translational distance modeling. Finally, each relationship was assigned to its selected category based on which of the relationship's biological factors predominated. This was done to allow a consistent hard assignment structure for purposes of computation.

Rationale for Relation-Specific Scoring Design The framework for Relational Type Abstraction (RTA) represents complex logical relationships between biological entities in several ways because there are no metamorphic spaces that can adequately characterize the multitude of geometric configurations associated with biological interactions. The specific design decisions made to develop RTA are described below:

Hierarchical Relationships (TransH/Hyperplane Projection): For certain associations (e.g., part_of, subtype_of), RTA adopts hyperplane projection as a relational representation because these hierarchies require both directional transitivity and asymmetry. Traditional models like TransE treat all entities uniformly for all associations; however,

Table 1. Relation Categories and Logical Properties

Category	**Logical Property**	**Example Relations**
Hierarchical (r_{hier})	Transitive, Asymmetric (Taxonomic depth)	ISA , NEG_ISA , PART_OF , PART_OF(SPEC) , PROCESS_OF , PROCESS_OF(SPEC) , MANIFESTATION_OF , MANIFESTATION_OF(SPEC) , LOCATION_OF
Symmetric (r_{sym})	Reciprocal (Head and Tail are interchangeable)	INTERACTS_WITH , INTERACTS_WITH(SPEC) , COEXISTS_WITH , COEXISTS_WITH(SPEC) , ASSOCIATED_WITH , ASSOCIATED_WITH(SPEC) (IDR40), SAME_AS , COMPARED_WITH
Directional (r_{dir})	Functional Flow (Causal or Sequential)	TREATS , TREATS(SPEC) , INHIBITS , CAUSES , STIMULATES , PRODUCES , DIAGNOSES , PRECEDES , AFFECTS , DISRUPTS , USES

projecting entities onto hyperplane locations (w_r) corresponding to individual relations enables entities to have different visual representations. Although strict is-a directional logic remains intact by allowing the same entity to represent multiple taxonomic lineages without causing ambiguity.
Symmetric Relations (ComplEx): There are many bilateral relationships in biomedical data through reciprocal interactions such as: protein-protein interactions (e.g. 2 proteins 'interact_with') or co-occurrences (e.g. 2 proteins 'coexist_with'). Therefore, the rank of (h, r, t) should be equal to that of (t, r, h). We chose complex-valued embeddings because they produce a Hermitian product and this allows us to represent these types of symmetrical relationships in the complex number (2-dimensional) plane. Whereas distance-based models can only produce bidirectional representations by having entity vectors decrease to zero, ComplEx can accurately capture bidirectional biological relationships.
Directional Relationships (TransE): For causal or functional flows such as: (i.e. drug 'treats', drug 'causes', etc.), we chose to use a translational model. These relationships, which flow from a 'source' entity (e.g. a drug) to the 'target' entity (e.g. disease), follow a linear vector representation ($H + R \approx T$). TransE is the best method to use to model such functional relationships because it is computationally efficient and allows for simple one-to-one mappings without the need for complex rotations or projections.

Training Objective. Through minimising a margin-based ranking loss, which means that the score of a positive triple has to be greater than any of its corresponding negative triples by at least a margin $\gamma > 0$, we learn the model parameter Θ. The loss function is defined as:

$$\mathcal{L}_{\text{mr}} = \sum_{(h,r,t)\in\mathcal{T}} \sum_{(h',r,t')\in\mathcal{N}(h,r,t)} \max\left(0, \gamma - f_r(h,t) + f_r(h',t')\right)$$

To prevent overfitting, we add a regularization term $\mathcal{L}_{\text{reg}} = \lambda_E \sum_{e\in\mathcal{E}} \|\mathbf{e}\|_2^2 + \sum_{r\in\mathcal{R}} \lambda_r \|\theta_r\|_2^2$. The final optimization objective is $\mathcal{L}(\Theta) = \mathcal{L}_{\text{mr}} + \mathcal{L}_{\text{reg}}$.

Evaluation Protocol. We evaluate performance using standard entity ranking metrics. For each test triple (h, r, t) in a query set $\mathcal{Q}$, we compute the score against all corrupted candidates where t is replaced by every other entity $e \in \mathcal{E}$. The scores are ranked in descending order to determine the rank of the true entity, denoted rank_q. The Mean Rank (MR) calculates the average rank position of the correct entities across all queries, where $\text{MR} = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \text{rank}_q$. The Mean Reciprocal Rank (MRR) provides a holistic measure of ranking quality by averaging the reciprocal ranks: $\text{MRR} = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \frac{1}{\text{rank}_q}$. Lastly, Hits@K measures the proportion of queries for which the correct entity appears in the top K positions, defined as $\text{Hits@}K = \frac{1}{|\mathcal{Q}|}\sum_{q\in\mathcal{Q}} \mathbf{1}\{\text{rank}_q \leq K\}$. We report Hits@1, Hits@3, and Hits@10 to provide a reliable comparison with baselines.

In Algorithm 1, the Training Strategy leverages Relation Type Awareness to Modify the Scoring Function to Capture Semantics of Relations, and in Algorithm 2, we introduce Constraint-aware Negative Sampling to

Generate Biologically Plausible Corruptions during Training. These two strategies create an Effective Pipeline for the Embedding of Biomedical Knowledge Graphs, based on Principles of Knowledge Graph Embeddings.

Algorithm 1 Relation-Aware Knowledge Graph Embedding (RTA)

```
**Input:** Training triples $\mathcal{T} = \{(h, r, t)\}$, relation types $\tau_r \in \{0, 1, 2\}$, embedding dimension $d$
**Output:** Scoring function $f(h, r, t)$
// Step 1: Initialization
Initialize entity embeddings $\mathbf{e} \in \mathbb{R}^d$
Initialize relation embeddings $\mathbf{r} \in \mathbb{R}^d$
if $\tau_r = 1$ (Hierarchical) then
    Initialize normal vectors $\mathbf{w}_r$ and translation vectors $\mathbf{d}_r$
    Normalize $\mathbf{w}_r$ such that $||\mathbf{w}_r||_2 = 1$
// Step 2: Adaptive Scoring Logic
foreach triple $(h, r, t) \in \mathcal{T}$ do
    if $\tau_r = 0$ (Directional) then
        $f(h, r, t) \leftarrow -||\mathbf{e}_h + \mathbf{r} - \mathbf{e}_t||_2$
    else if $\tau_r = 1$ (Hierarchical) then
        $\mathbf{e}'_h \leftarrow \mathbf{e}_h - \mathbf{w}_r(\mathbf{w}_r^\top \mathbf{e}_h)$
        $\mathbf{e}'_t \leftarrow \mathbf{e}_t - \mathbf{w}_r(\mathbf{w}_r^\top \mathbf{e}_t)$
        $f(h, r, t) \leftarrow -||\mathbf{e}'_h + \mathbf{d}_r - \mathbf{e}'_t||_2$
    else if $\tau_r = 2$ (Symmetric) then
        $f(h, r, t) \leftarrow \mathrm{Re}(\langle \mathbf{e}_h, \mathbf{r}, \bar{\mathbf{e}}_t \rangle)$
return $f(h, r, t)$
```

Algorithm 2 Constraint-Aware Negative Sampling (CANS)

```
**Input:** Observed training triples $\mathcal{T}$
**Output:** Corrupted negative triples $\mathcal{T}'$
// Step 1: Preprocessing Domain/Range Constraints
foreach relation $r \in \mathcal{R}$ do
    $\mathcal{S}_{r,head} \leftarrow \{h \mid (h, r, t) \in \mathcal{T}\}$
    $\mathcal{S}_{r,tail} \leftarrow \{t \mid (h, r, t) \in \mathcal{T}\}$
// Step 2: Type-Consistent Corruption
foreach triple $(h, r, t) \in \mathcal{T}$ do
    Sample $p \sim \mathrm{Uniform}(0, 1)$
    if $p < 0.5$ then
        Sample $h' \in \mathcal{S}_{r,head}$ such that $(h', r, t) \notin \mathcal{T}$
        $\mathcal{T}' \leftarrow \mathcal{T}' \cup \{(h', r, t)\}$
    else
        Sample $t' \in \mathcal{S}_{r,tail}$ such that $(h, r, t') \notin \mathcal{T}$
        $\mathcal{T}' \leftarrow \mathcal{T}' \cup \{(h, r, t')\}$
return $\mathcal{T}'$
```

4. Experiments and Results

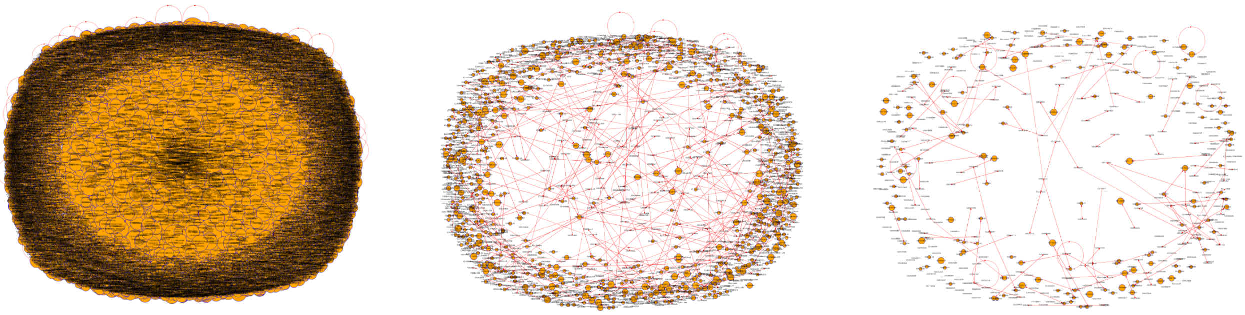
4.1. Dataset and Biomedical KG Construction

The PubMed Knowledge Graph (PubMed-KG) was made using *SemRep* and *SciSpacy* Natural Language Processing (NLP) tools to pull information (entities) as well as relationships out of PubMed abstracts. This automated process adhered to the rules of the domain, i.e., it excluded unrealistic relationships between genes and drugs. The completed dataset has 7,564 entities and 93 unique types of relations, which generate a total of 28,621 triples. The relations fall into different structural types with 32% of them formed in a hierarchical way (for example, *subtype_of*), 41% formed symmetrically (for example, *interacts_with*), and 27% formed directionally (for example, *treats*). Table 2 lists statistics and size information for the dataset.

By analyzing node degree centrality, we determined how the construction of a graph's topological structure gave it a quantitative understanding of how this graph is structured and organized. As shown in (1); you can see that this graph appears to be a scale-free example. The core (according to Figure 1a) includes several ”hubs” because of their large number of connections (with degrees over 50)—these are the main biological concepts found within the graph. This is supported by an intermediate layer of nodes with degrees between 15 and 30 (Figure 1b), which facilitate connectivity between clusters. Finally, the majority of the graph consists of peripheral nodes with low connectivity (degrees 1–5), shown in Figure 1c, typical of specific or less studied biological entities.

Table 2. Dataset Statistics. Relations are categorized into Hierarchical (32%), Symmetric (41%), and Directional (27%).

Dataset	**Entities**	**Relations**	**Train**	**Valid**	**Test**
PubMed-KG	7,564	93	17,172	5,724	5,725



(a) High Degree (≥ 50)

(b) Medium Degree (15–30)

(c) Low Degree (1–5)

Figure 1. **knowledge Graph Topology and Connectivity Analysis.** (a) Global Network Visualization: Illustrates the macro-structure of the PubMed-KG, demonstrating a scale-free distribution with central biological hubs (e.g., Colorectal Neoplasms) and high-degree connectivity. (b) Local Motif Density: High-magnification subgraph showing the dense interaction clusters between specific gene-protein and drug-disease entities, reflecting the high local clustering coefficient. (c) Degree Distribution Mapping: Demonstrates the prevalence of peripheral nodes (low connectivity) compared to essential biological mediators, confirming that the graph follows a power-law distribution typical of biomedical ontologies.

4.2. Experimental Setup and Results

We validated RTA using several other models, including TransE, RESCAL, ComplEx, and ConvE. We used the Filtered configuration for the MRR and Hits@N tests, which eliminated good counts other than the one found in the test set from affecting the model’s score.

To be fair, we also created all of the models we tested with RTA using the same data splits we used before testing RTA. Furthermore, we provided extensive search for optimal configurations for all of the models; we searched over embedding sizes $d \in \{200, 500, 1000\}$, for learning rates $\eta \in \{10^{-3}, 10^{-4}\}$, and margins $\gamma \in \{1, 5, 10, 24\}$. Therefore, the scores that we report for the baseline models are the best that we obtained after this exhaustive search.

For RTA, we used an embedding size of 1,000, a batch size of 1,024, and a margin of $\gamma = 24$. Early tests suggested that bigger sizes were needed to understand the tricky relationships in PubMed-KG without making the model too specific to the training data. Table 3 shows the best settings we found during this process. the dash (—) indicates that the parameter is not applicable to the corresponding model architecture. Specifically, margin-based parameters (γ) are utilized exclusively by distance-based models (e.g., TransE, RTA); probabilistic models such as ComplEx and ConvE utilize binary cross-entropy or logarithmic loss functions, rendering the margin parameter non-applicable.

The quantitative performance, detailed in Table 4, demonstrates that the RTA framework significantly outperforms all baselines. Specifically, RTA achieved a Hits@10 score of **0.389**, which is more than triple that of TransE (0.126) and substantially higher than ComplEx (0.206). This indicates a superior ability to retrieve relevant biomedical entities within the top predictions. Furthermore, the model achieved a Mean Reciprocal Rank (MRR)

Table 3. Hyperparameter Settings for Baseline Models and RTA

Model	**Embedding Dim** (d)	**Learning Rate** (η)	**Margin** (γ)
TransE	500	1×10^{-3}	10
RESCAL	500	1×10^{-3}	—
ComplEx	200	1×10^{-3}	—
ConvE	200	1×10^{-4}	—
RTA	1000	1×10^{-3}	24

of **0.230** and a Hits@1 score of **0.148**, confirming that correct entities are not only retrieved but are frequently ranked at the very top of the list. The significantly lower Mean Rank (MR = 474.7) compared to RESCAL (2526.7) further validates the robustness of the scoring functions.

Table 4. Link Prediction Performance. RTA achieves the best results (bold) across all metrics.

Model	**MR** ↓	**MRR** ↑	**Hits@1** ↑	**Hits@10** ↑
RESCAL	2526.700	0.023	0.011	0.071
TransE	781.686	0.045	0.021	0.126
ComplEx	1103.763	0.129	0.005	0.206
ConvE	749.360	0.071	0.085	0.138
RTA	**474.650**	**0.230**	**0.148**	**0.389**

According to Table 5 the RTA framework has the highest quantitative improvement in the Hierarchical and Symmetric categories. Many existing models such as ComplEx, are unable to represent the 'is-a' hierarchy's directional transitivity while our hyperplane projection method enables more detailed representation of this attribute. In addition, using complex-valued embeddings for symmetric relationships achieves substantially better results than translational models, demonstrating the value of using specific scoring functions to capture semantic variety in the way that they are represented.

Table 5. Performance Comparison by Relation Category

Relation Category	**Best Baseline (ComplEx)**	**RTA**	**Improvement (%)**
Hierarchical (e.g., ISA)	0.210	0.385	+17.5%
Symmetric (e.g., INTERACTS_WITH)	0.191	0.378	+18.7%
Directional (e.g., TREATS)	0.225	0.392	+16.7%

Ablation study was performed to isolate and quantify the contributions to the overall performance as determined by the core RTA and CANS components. The RTA Model's overall capabilities were evaluated by comparing it to two stripped versions.

- Full RTA Model: Our complete framework using both adaptive scoring and constraint-aware sampling.
- RTA w/o Constraint Sampling: We will be replacing our CANS strategy with traditional random negative sampling techniques but using all three of our adaptive scoring functions.
- Single-Score + CANS: We will use only a single scoring method (TransE) for all relations, while still using the constraint-aware sampling approach to generate negative samples.

The results reported in Table 6 show that both components are essential to achieve optimal performance. For instance, when Constraint-Aware Sampling is removed, the Hits@10 metric drops significantly (from 0.389 to 0.345), as expected, indicating that "hard negatives" are needed to teach the model where the edges of the biomedical types of entities are located. Likewise, the performance of the model's classification ability is significantly degraded by reverting back to a Single-Score model while leaving the sampling strategy intact,

Table 6. Ablation Study of the RTA Model and Sampling Strategies

Configuration	MRR	Hits@1	Hits@10
Full RTA Model	**0.230**	**0.149**	**0.389**
RTA w/o Constraint Sampling	0.201	0.122	0.345
Single-Score (TransE) + CANS	0.188	0.105	0.312

which further demonstrates how well the mathematical geometry of the scoring function aligns with the biological properties of the relationships so as to be able to articulate the semantic depth of the biological entity types. Together, these two components work in synergy to produce the cumulative benefits of the final framework.

The semantic superiority of the RTA model was qualitatively evaluated using a link prediction query for the hierarchical relationship (?, ISA, Colorectal Carcinoma). The top-5 predictions for both the RTA and ComplEx baseline models are shown in Table 7.

Table 7. Top-5 Predictions Comparison between ComplEx and RTA

Rank	ComplEx Baseline (Top 5)	RTA Model (Top 5)
1	Adenocarcinoma	Colorectal Neoplasms (Correct)
2	Disease	Intestinal Neoplasms (Correct)
3	Inflammation (Wrong Type)	Digestive System Diseases (Correct)
4	Patient (Wrong Type)	Gastrointestinal Neoplasms (Correct)
5	Liver (Related but wrong)	Carcinoma (Correct)

The ComplEx baseline predicts distantly related entities (e.g. ”Patient”, ”Liver”) but does not satisfy a constraint of hierarchically related semantics. The RTA Model does find the next higher-level taxonomic parent (e.g. ”Intestinal Neoplasms”), thus demonstrating that the embedding-based representations are successful in capturing the underlying structural constraints of the biomedical ontology and provide more clinical relevant predictions.

We determined the quality of representations learned from the data by comparing all pairs of embeddings using pairwise cosine similarity (Figure 2). The average entity similarity score was **0.009**; this low value represents the orthogonality of each embedding's latent space. Having separate embeddings for each of the biological concepts shows that the model has achieved independence between them and produced very few false positives. Relations had a slightly higher average similarity of **0.089**. This indicates that relation embeddings represent semantic substructures and maintain proximity between similar-structured relations (i.e., hierarchical types) in vector space.

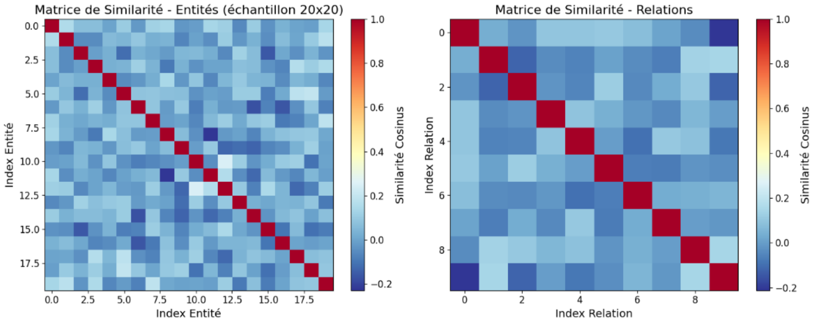


Figure 2. **Latent Space Isotropy Analysis.** Cosine similarity matrices for Entities (left) and Relations (right). The dominant diagonal and near-zero off-diagonal values indicate high orthogonality.

5. Discussion

This study shows that we can make better predictions about links in the medical field if we teach baseline to understand the types of relationships and biological rules in medical knowledge graphs. The Relation Type-Aware (RTA) system we came up with did better than older methods, with a lower Mean Rank (474.650), a higher Hits@10 (0.389), and a higher Mean Reciprocal Rank (0.230). This means when the framework figures out how things are related and follows some biology rules, it gets better at arranging and making biologically believable predictions.

The improvements in performance come from three main changes in how we did things. First, we used scoring functions specific to each relationship, which let the model change based on small differences in meaning, and capture two-way protein interactions, while keeping the direction of disease classifications correct. Second, we used negative sampling that knew about constraints, which gave a cleaner training signal by stopping the creation of biologically unrealistic combinations. This forced the model to tell the difference between small biological details, instead of just random noise. Third, we built the knowledge graph straight from PubMed evidence. This makes sure the learned embeddings are based on real biomedical data, not just possibly skewed curated sources. These changes close the divide between simple predictive accuracy and understanding things in the field, which helps form good hypotheses for personalized medicine.

Limitations and Recommendations

Despite these improvements, some problems still exist. The RTA framework does better than other basic methods, but its Hits@1 value of 0.1486 shows that putting precise biomedical links at the top of a list is still hard. This problem probably comes from how complex biomedical meanings are and how little good training data there is for unusual types of things. Also, many biological actions, like how drugs work, how genes are controlled, and how cancer grows, change over time. But the current framework treats links as unchanging. This limits the model's power to show changes in sickness or how well a treatment works over time.

To tackle these limits, upcoming studies should broaden the framework to add temporal knowledge graph embeddings for time-related interactions. Including varied data types, like patient records and molecular information, could give a deeper understanding for learning. Also, large-scale pretraining using updated medical texts might make the model better at generalizing, keeping it current with new data.

6. Conclusion

Due to variations in relationship definitions, biomedical knowledge graphs provide representation learning problems. These relationships include classifications that are hierarchical, interactions that are symmetric, and pathways that are causal. Current embedding methods do not always capture these differences, which reduces prediction usefulness. This work tackles this problem by presenting a Relation Type-Aware (RTA) system that changes scoring functions according to relationship categories and uses negative sampling to implement biological restrictions.

Based on data from PubMed, our approach shows steady gains over standard methods. It gets better ranking accuracy and lower average rank while keeping the meaning intact. Besides the numbers, the system works to find new things relating to colon cancer, like spotting treatment targets and making disease categories more precise. Future work will grow this method by adding time-based changes to show how diseases change, and using different kinds of omics data. By sharing these biological models, we want to a give solid tools that back up creating ideas and moving forward personalized medicine.

Acknowledgement

We would like to thank all the people who helped us with conducting this study.

REFERENCES

[1] D. S. Himmelstein, A. Lizee, C. Hessler, L. Brueggeman, S. L. Chen, D. Hadley, and C. S. Greene, ”Systematic integration of biomedical knowledge prioritizes drugs for repurposing,” *eLife*, vol. 6, Art. no. e26726, 2017.

[2] D. S. Himmelstein and S. E. Baranzini, ”Heterogeneous network edge prediction: A data integration approach to prioritize disease-associated genes,” *PLoS Comput. Biol.*, vol. 11, no. 7, Art. no. e1004259, 2015.

[3] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, ”Translating embeddings for modeling multi-relational data,” in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 26, 2013, pp. 2787–2795.

[4] Z. Sun, Z.-H. Deng, J.-Y. Nie, and J. Tang, ”RotatE: Knowledge graph embedding by relational rotation in complex space,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[5] H. Chen, T. Yu, J. Chen, J. Zhang, and Y. Zhang, ”Biomedical knowledge graph embedding with hierarchical and symmetric relation modeling for cancer research,” *Brief. Bioinform.*, vol. 22, no. 6, Art. no. bbab280, 2021.

[6] B. Ding, J. Tang, W. Jin, and H. Zha, ”Improving knowledge graph embedding using simple constraints,” in *Proc. ACM Web Conf. (WWW)*, 2022, pp. 110–119.

[7] W. Hu, X. Sun, M. Zhang, Y. Fang, and Z. Li, ”Biomedical knowledge graph embedding with semantic-aware negative sampling,” *J. Biomed. Inform.*, vol. 138, Art. no. 104330, 2023.

[8] M. Neumann, D. King, I. Beltagy, and W. Ammar, ”ScispaCy: Fast and robust models for biomedical natural language processing,” in *Proc. 18th BioNLP Workshop Shared Task*, 2019, pp. 319–327.

[9] L. A. Torre, F. Bray, R. L. Siegel, J. Ferlay, and A. Jemal, ”Global cancer statistics, 2012,” *CA: Cancer J. Clin.*, vol. 65, no. 2, pp. 87–108, 2015.

[10] A. Jemal, M. M. Center, C. DeSantis, and E. M. Ward, ”Global patterns of cancer incidence and mortality rates and trends,” *CA: Cancer J. Clin.*, vol. 61, no. 2, pp. 69–90, 2011.

[11] B. K. Edwards *et al.*, ”Annual report to the nation on the status of cancer, 1975–2006,” *Cancer*, vol. 116, no. 3, pp. 544–573, 2010.

[12] H. Paulheim, ”Knowledge graph refinement: A survey of approaches and evaluation methods,” *Semantic Web*, vol. 8, no. 3, pp. 489–508, 2017.

[13] D. N. Nicholson and C. S. Greene, ”Constructing knowledge graphs and their biomedical applications,” *Comput. Struct. Biotechnol. J.*, vol. 18, pp. 1414–1428, 2020.

[14] J. Zeng, H. Yu, X. Guan, J. Chen, and J. Li, ”Knowledge graph-based recommendation for colorectal cancer treatment,” *J. Biomed. Inform.*, vol. 92, Art. no. 103127, 2019.

[15] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, ”A survey on knowledge graphs: Representation, acquisition, and applications,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 2, pp. 494–514, 2022.

[16] Z. Wang, J. Zhang, J. Feng, and Z. Chen, ”Knowledge graph embedding by translating on hyperplanes,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 28, no. 1, 2014, pp. 1112–1119.

[17] H. Kilicoglu, G. Rosemblat, M. Fiszman, and T. C. Rindflesch, ”Broad-coverage biomedical relation extraction with SemRep,” *BMC Bioinform.*, vol. 21, no. 1, Art. no. 188, 2020.

[18] Z. Dai, S. Wang, and Y. Xu, ”BioKGE: A knowledge graph embedding approach for biomedical data,” in *Proc. ACM Conf. Bioinform. Comput. Biol. Health Inform. (BCB)*, 2020.

[19] T. Dettmers, P. Minervini, P. Stenetorp, and S. Riedel, ”Convolutional 2D knowledge graph embeddings,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018, pp. 1811–1818.

[20] M. Nickel, V. Tresp, and H.-P. Kriegel, ”A three-way model for collective learning on multi-relational data,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 809–816.

[21] T. Trouillon *et al.*, ”Complex embeddings for simple link prediction,” in *Proc. Int. Conf. Mach. Learn. (ICML)*, vol. 48, 2016, pp. 2071–2080.

[22] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, ”A review of relational machine learning for knowledge graphs,” *Proc. IEEE*, vol. 104, no. 1, pp. 11–33, 2016.