



# Improving Knowledge Graph Embeddings through Contrastive Learning with Negative Statements

Rita T. Sousa

Data and Web Science Group  
Universität Mannheim  
Mannheim, Germany  
rita.sousa@uni-mannheim.de

Heiko Paulheim

Data and Web Science Group  
Universität Mannheim  
Mannheim, Germany  
heiko.paulheim@uni-mannheim.de

## Abstract

Knowledge graphs represent information as structured triples and serve as the backbone for a wide range of applications, including question answering, link prediction, and recommendation systems. A prominent line of research for exploring knowledge graphs involves graph embedding methods, where entities and relations are represented in low-dimensional vector spaces that capture underlying semantics and structure. However, most existing methods rely on assumptions such as the Closed World Assumption or Local Closed World Assumption, treating missing triples as false. This contrasts with the Open World Assumption underlying many real-world knowledge graphs. Furthermore, while explicitly stated negative statements can help distinguish between false and unknown triples, they are rarely included in knowledge graphs and are often overlooked during embedding training.

In this work, we introduce a novel approach that integrates explicitly declared negative statements into the knowledge embedding learning process. Our approach employs a dual-model architecture, where two embedding models are trained in parallel, one on positive statements and the other on negative statements. During training, each model generates negative samples by corrupting positive samples and selecting the most likely candidates as scored by the other model. The proposed approach is evaluated on both general-purpose and domain-specific knowledge graphs, with a focus on link prediction and triple classification tasks. The extensive experiments demonstrate that our approach improves predictive performance over state-of-the-art embedding models, demonstrating the value of integrating meaningful negative knowledge into embedding learning.

## Keywords

Knowledge Graph, Knowledge Graph Embedding, Negative Statements, Link Prediction, Triple Classification

## ACM Reference Format:

Rita T. Sousa and Heiko Paulheim. 2025. Improving Knowledge Graph Embeddings through Contrastive Learning with Negative Statements. In *Knowledge Capture Conference 2025 (K-CAP '25)*, December 10–12, 2025, Dayton, OH, USA. ACM, New York, NY, USA, 8 pages. <https://doi.org/10.1145/3731443.3771343>



This work is licensed under a Creative Commons Attribution 4.0 International License. *K-CAP '25*, Dayton, OH, USA

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-1867-0/25/12

<https://doi.org/10.1145/3731443.3771343>

## 1 Introduction

A knowledge graph (KG) [12] is a structured representation of knowledge that encodes information as a set of triples  $(h, r, t)$ . For each triple, the head  $h$  is an entity, the tail  $t$  is either an entity or a literal (represented as nodes), and the relation  $r$  (represented as an edge) describes the relationship or property connecting them. Over the years, public and commercial KGs have been developed [36], reflecting the increasing importance of structured knowledge in diverse domains. One of the key strengths of KGs is their ability to be understood by both humans and machines, allowing them to support a wide range of applications, including question answering [16], link prediction [26], and recommendation systems [11].

A variety of approaches have been developed for KG-based tasks. Among these, KG embedding (KGE) methods have gained significant attention, with a wide range of methods proposed in the literature. KGE methods encode entities and relationships as low-dimensional vectors within a latent embedding space, aiming to capture and preserve the underlying structural information and graph properties [37]. The majority of these methods learn embeddings by defining a scoring function that measures the likelihood of a given triple being true. During training, the model adjusts the embeddings to assign higher scores to triples that exist in the KG and lower scores to negative triples. The process of generating negative triples, known as negative sampling, typically involves corrupting true triples by randomly replacing their head or tail entity with another entity from the KG [18].

Although most KGs are built and interpreted under the Open World Assumption (OWA), which means that the non-stated facts may represent unknown facts or true negative statements, many KGE models adopt the Closed World Assumption (CWA), or more commonly, the Local Closed World assumption (LCWA) [8], especially during the generation of negative triples. Under LCWA, if a triple  $(h, r, t)$  exists, then the KG is considered to include all possible tails for any statements of the form  $(h, r, ?)$ . This works reasonably well for relations expected to have a single tail, such as *hasCapital*. However, it breaks down for relations that can link a head to multiple tails, like *worksAt*, which is typical for most real-world relations [9]. Recognizing these assumptions is critical for many KG-related tasks, namely link prediction. To predict new links accurately, it is first necessary to distinguish between links that are false and those that are merely unknown. This highlights the relevance of explicitly defining negative statements in the KG.

Negative statements have long been underrepresented in KGs, despite the growing recognition of their importance. They are either filtered out during construction or treated no differently than positive relations, merely encoded as another edge type. As a result, only some studies have explored how to incorporate negative

knowledge into representation learning [31]. More recently, however, some research initiatives focus on enriching existing KGs with interesting and meaningful negative statements. Given the infinite pool of possible negative statements, the most valuable are typically those that people might reasonably expect or believe to be true, but which are ultimately shown to be false. Notable examples include efforts to enrich Wikidata with negative statements [3], generate informative negative commonsense statements [1], or extend a widely used biomedical ontology with statements indicating that a protein does not perform a given function [39].

In this work, we address the gap left by most existing KGE methods by proposing an approach that incorporates explicitly declared negative statements and adapts the negative sampling process using a contrastive learning strategy. Our approach trains separate KGE models on positive and negative statements, and generates negative samples by corrupting triples and selecting the most plausible corrupted triples based on the scores from the opposing model. We assess the effectiveness of our approach on different KGs, including a general-purpose KG like Wikidata and a domain-specific biomedical KG such as the Gene Ontology (GO). Our evaluation covers two tasks: link prediction and triple classification for protein-protein interaction (PPI) prediction. Our contributions are as follows:

- (1) The design of a dual-model training framework that learns separate embeddings for positive and negative statements.
- (2) A novel negative sampling strategy based on contrastive learning that iteratively refines negative samples, leading to more challenging negatives.
- (3) Extensive experiments to demonstrate the effectiveness, robustness, and broad applicability of the proposed approach on diverse KGs and tasks.

## 2 Related Work

There is a substantial body of work on KGEs, as highlighted by several surveys [38]. Broadly, research in this area has evolved along two main directions [24]. The first focuses on creating embeddings of KG entities for downstream tasks, with an emphasis on capturing semantic similarity. Methods in this category, such as RDF2Vec [25], are typically based on natural language processing and random walks, and are evaluated on external ground truth data outside the KG. The second direction centers on link prediction, where the objective is to distinguish between correct and incorrect triples using a scoring function. This line of work includes translational distance models [4] or semantic matching models [40].

To the best of our knowledge, TrueWalks [31] is the only KGE-based approach that incorporates explicitly defined negative statements during embedding learning for ontology-rich KGs. TrueWalks generates two distinct embeddings for each entity: one capturing the positive semantics and another capturing the negative semantics. For the positive embedding, it generates walks based on positive and *subClassOf* relationships, then inputs these walks into a skip-gram model to learn the embedding. Similarly, for the negative embedding, it generates walks using negative and *superClassOf* relationships, feeding them into a skip-gram model to obtain the negative embedding. This dual embedding strategy allows TrueWalks to account for the semantic implications of negation in ontology-rich KGs. Despite its effectiveness, TrueWalks has

some limitations. First, its applicability is restricted to ontology-rich KGs, which are prevalent in the biomedical domain but less common in other areas. Additionally, by learning positive and negative embeddings independently, TrueWalks does not explicitly model how contradictory statements affect entity dissimilarity.

Beyond the KGE literature, some works have explored explicit negative information using graph neural networks (GNNs), particularly to model whether a user is or is not interested in an item in recommendation systems. SiGAT [14] uses social theories (balance and status theory) to categorize neighbor nodes into motifs. Then, it applies a GAT-based aggregation to combine information from those categorized neighbors. SiGRec [15] creates two separate embeddings per node (positive and negative) and combines them by concatenation. It also introduces the sign cosine loss, a loss function designed to handle various types of negative feedback. Lastly, PANE-GNN [21] partitions the graph into two distinct bipartite graphs based on positive and negative feedback and then generates an interest embedding and a disinterest embedding with positive and negative edges. For the negative graph, a distortion is introduced to denoise the negative feedback.

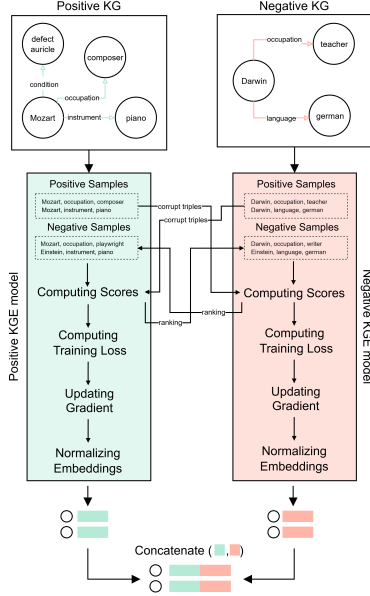
In summary, although prior work has explored negative information in KG representation, several gaps remain. GNN-based approaches are designed for homogeneous graphs and struggle with the heterogeneous, richly typed relations in KGs. As a result, directly applying such architectures to KGs often overlooks complex semantics. KGE models, on the other hand, are capable of capturing such semantic information, but the only existing approach that explicitly handles negative statements is path-based and cannot be extended to link prediction approaches. The proposed approach overcomes these limitations by training separate link prediction KGE models on positive and explicitly declared negative statements, using each to guide the other's negative sampling.

## 3 Methodology

An overview of the proposed approach for generating contrastive negative samples is illustrated in Figure 1. First, two separate KGs are built: one from positive statements and another from explicitly defined negative statements. For each KG, a KGE model is then initialized and trained. In standard KGE training, triples from the KG are used directly as positive samples while negative samples are generated through random corruption (e.g., replacing the head or tail entity of a triple with a random entity). In contrast, our approach introduces a contrastive learning strategy: for each triple in one KG, we replace either the head or the tail entity with all possible entities from the KG, then compute the plausibility scores using the model trained on the other KG. The corrupted triple with the highest score, which represents the most plausible according to the other model, is selected as the negative example. After training, each node in the KG is represented by the concatenation of its embeddings from both positive and negative KGEs, yielding a dual perspective that captures both positive and negative aspects.

The proposed approach is KGE-agnostic and can be integrated into any KGE model that defines a scoring function and employs negative sampling during training. The implementation is available on GitHub<sup>1</sup>.

<sup>1</sup>[https://github.com/ritatsousa/KGE\\_with\\_CL](https://github.com/ritatsousa/KGE_with_CL)



**Figure 1: Overview of the proposed approach that trains two models, one on the positive KG and another on the negative KG, using each to guide the other’s negative sampling.**

### 3.1 Building the Positive and Negative Knowledge Graphs

The positive and negative KGs are constructed as RDF graphs. An RDF graph is a directed, labeled graph in which edges represent the named relations between two nodes, and are represented as triples in the form  $(h, r, t)$ . The nodes may correspond to individuals, classes, literals, or other RDF resources. When the KG is backed by an ontology, the transformation follows the OWL to RDF Graph Mapping guidelines defined by W3C<sup>2</sup>. Simple OWL axioms are directly translated into RDF triples. More complex axioms, especially those involving nested class expressions, are decomposed into multiple RDF triples and may include blank nodes to preserve structure and semantics.

### 3.2 Initializing the Embedding Models

Once the KGs are built, the KGEs models can be initialized to learn low-dimensional vector representations of entities and relations. Numerous KGE approaches have been proposed in recent years [37], particularly for link prediction tasks, to which our approach can be integrated. Despite their differences, these KGE models typically share common components, including the definition of a scoring function, embedding initialization, optimization strategy, and loss function. The scoring function quantifies the plausibility of a given triple  $(h, r, t)$ . Training involves learning entity and relation embeddings by minimizing a loss function over several epochs. The goal is to assign higher scores to positive samples (triples that exist in the KG) and lower scores to negative samples (triples that are absent from the KG and are presumed false). Standard loss functions include margin-based ranking loss or binary cross-entropy,

**Table 1: Scoring functions for the employed KGE models.**

KGE	Scoring Function
TransE	$f(h, r, t) = -\ h + r - t\ _p$
DistMult	$f(h, r, t) = \sum_i h_i \cdot r_i \cdot t_i$
Complex	$f(h, r, t) = \text{Re}(\sum_i h_i \cdot r_i \cdot t_i)$

and optimization is usually carried out using stochastic gradient descent or its variants.

In this work, we focus on integrating the proposed approach into three widely adopted KGE models. TransE [4] is a translational model that represents relations as translation vectors in the embedding space. It assumes that for a triple  $(h, r, t)$ , the vector of the tail  $t$  should be close to the vector of the head  $h$  plus the vector of the relation  $r$ . DistMult [40] models the score of a triple using a bilinear dot product of the entity and relation embeddings. ComplEx [35] extends DistMult into the complex vector space by incorporating both real and imaginary components in the embeddings. Table 1 summarizes their scoring functions.

### 3.3 Contrastive Learning for Generating Negative Samples

Generating negative samples is critical for KGE training [18]. The most widely adopted strategy is random negative sampling, where negative samples are created by corrupting positive triples, typically by replacing either the head or the tail entity with a randomly chosen entity from the KG. However, since the majority of the KGs reside under the OWA, such random corruption can often result in false negatives, as many triples that are not present in the KG may still represent valid facts. Additionally, replacing entities indiscriminately can yield trivial negatives that are semantically invalid. More advanced sampling strategies have since been proposed, such as (i) replacing heads or tails of positive triples with entities seen in the same relation [30], (ii) selecting replacement entities of the same type as the original [20], (iii) assuming one relation per entity pair and picking entities connected through different relations than the current one [19], (iv) using a pre-trained embedding model to select nearest neighbors of entities in vector space as negatives [19], or (v) using the highest scoring triples of a model’s previous iteration which are not contained in the KG as negatives [28]. These advanced sampling strategies incorporate entity type constraints to generate harder negatives, yet they still fall short in addressing the challenges posed by the OWA. Beyond the KGE domain, negative sample generation has also been explored in active learning, where various strategies, such as contrastive learning-based methods [22], are employed to selectively acquire informative examples that enhance model training efficiency.

In the proposed approach, we employ a two-stage approach to negative sampling. In the initial phase of training, negative samples are generated using the standard random corruption strategy. However, after a predefined number of epochs, denoted as *cl\_phase*, we switch to a contrastive learning-based strategy (see Algorithm 1). In the contrastive phase, for each positive triple from one KGE model, we generate candidate negative triples by replacing either the head or the tail entity with every other entity in the KG. Each candidate triple is then scored by the KGE model trained on the other KG. For instance, given the triple  $(Mozart, occupation, composer)$  in

<sup>2</sup><https://www.w3.org/TR/owl2-mapping-to-rdf/>

**Algorithm 1** Training strategy for the positive KG ( $kg\_pos$ ) and the negative KG ( $kg\_neg$ ), incorporating a contrastive learning phase. Initially, both models (one trained on the  $kg\_pos$ , the other on the  $kg\_neg$ ) generate random negative samples. After a predefined epoch ( $cl\_phase$ ), each model generates contrastive negative samples by scoring entity perturbations using the other model.

```

1: function TRAIN( $kg\_pos, kg\_neg$ )
2:    $epochs \leftarrow$  number of training epochs
3:    $cl\_phase \leftarrow$  epoch to start contrastive learning
4:    $pos\_model.INIT(kg\_pos)$ 
5:    $neg\_model.INIT(kg\_neg)$ 
6:   for  $epoch$  in  $epochs$  do
7:     if  $epoch > cl\_phase$  then
8:        $pos\_model.GET\_CONTRASTIVE\_SAMPLES(kg\_pos, neg\_model)$ 
9:        $neg\_model.GET\_CONTRASTIVE\_SAMPLES(kg\_neg, pos\_model)$ 
10:    else
11:       $pos\_model.GET\_RANDOM\_NEGATIVE\_SAMPLES()$ 
12:       $neg\_model.GET\_RANDOM\_NEGATIVE\_SAMPLES()$ 
13:       $loss\_pos \leftarrow pos\_model.TRAIN\_STEP()$ 
14:       $loss\_neg \leftarrow neg\_model.TRAIN\_STEP()$ 
15: function GET\_CONTRASTIVE\_SAMPLES( $self, kg, contr\_model$ )
16:    $ents \leftarrow$  GETS UNIQUE ENTITIES( $kg$ )
17:    $contr\_negative\_samples \leftarrow []$ 
18:   for all  $(h, r, t)$  in  $kg$  do
19:      $perturb\_heads\_flag \leftarrow$  random flags for head/tail perturbation
20:     if  $perturb\_heads\_flag$  is True then
21:       for all candidate  $h$  in unique  $ents$  do
22:          $score \leftarrow$  COMPUTE SCORE( $candidate\_h, r, t, contr\_model$ )
23:          $best\_h \leftarrow$  candidate with highest score
24:          $negative\_samples.append(best\_h, r, t)$ 
25:     else
26:       for all candidate  $t$  in unique  $ents$  do
27:          $score \leftarrow$  COMPUTE SCORE( $h, r, candidate\_t, contr\_model$ )
28:          $best\_t \leftarrow$  candidate with highest score
29:          $contr\_negative\_samples.append(h, r, best\_t)$ 
30:    $self.negative\_samples \leftarrow contr\_negative\_samples$ 

```

the positive KG, we replace the tail entity with all other entities to produce candidates such as (*Mozart, occupation, playwright*) or (*Mozart, occupation, scientist*). The negative model scores all candidate triples, and the one with the highest score (i.e., the most plausible according to the model trained on the explicit negative statements) is selected as the negative sample for the positive model.

The proposed contrastive learning strategy enables the dynamic generation of negative samples that are more meaningful and challenging as training progresses, thus improving the KGE model’s ability to learn robust representations under the OWA.

### 3.4 Generating the Final Representation

After training the positive and negative KGE models simultaneously, each node and relation in the KG is associated with two distinct representations: one learned from positive statements and another learned from negative statements. As already shown in [31] and [15], these dual perspectives capture complementary semantic information. To construct the final entity representation, we combine these two embeddings using vector concatenation.

## 4 Evaluation

We evaluate the proposed approach on two KGs: Wikidata KG, a general-purpose KG covering a broad range of domains, and the GO KG, a biomedical domain-specific KG focused on representing protein functions. Both KGs are enriched with negative statements and are described in the following sections.

The evaluation tasks differ for each KG. For Wikidata, we perform link prediction, aiming to predict missing links. In contrast, for

the GO KG, we perform triple classification, specifically framed as predicting whether a protein interacts with another protein. In both tasks, our approach generates lower-dimensional representations for KG entities. For link prediction on Wikidata, these embeddings are used to compute scores for candidate triples, which are then ranked to identify new links. For the PPI prediction task on the GO KG, it is cast as a binary classification problem, where we combine the embeddings of the entities and relation to create a composite representation of the triple, which is then fed into a supervised classifier to determine whether the interaction is likely to exist.

As a baseline, we compare our approach against KGE models trained separately on the positive and negative statements. To ensure a fair comparison, the baseline models are trained using embeddings with twice the dimensionality of those in the proposed approach. This setup guarantees that the final embedding sizes of both methods are equivalent.

### 4.1 Wikidata Knowledge Graph

Wikidata [36] is a free collaborative knowledge base supported by a large community of active users as well as automated bots. It contains a vast collection of statements describing millions of entities, with each entity typically associated with multiple statements. The majority of these statements are positive assertions under the OWA. Some exceptions, such as deleted or deprecated statements, may imply negative information. Examples of positive statements include assertions like (*Barack\_Obama, occupation, Politician*). Regarding its taxonomy, although it is built through collective efforts, the community faces challenges in deciding between *instanceOf* or *subclassOf* to model inheritance in Wikidata [29].

Enriching Wikidata with negative statements has been addressed through a statistical inference method called peer-based inference [2]. The core idea of this method is to identify, for an entity  $e$ , a group of closely related entities referred to as peers. These peers provide a context from which negative statements about the entity  $e$  can be formulated. The approach assumes that if a certain attribute is consistently present for most peers but missing for the entity  $e$ , the absence is interpreted as a negative fact rather than incomplete data. For example, if several U.S. presidents in the knowledge base are recorded as having a legal background or having worked as lawyers, but Barack Obama is not listed with this occupation, it is assumed that it is false for Obama and not a missing statement. Thus, the method infers the candidate negative statement (*Barack\_Obama, occupation, Lawyer*). A learning-to-rank model then scores candidate negative statements using features like predicate and object prominence, frequency, and textual context. These negative statements can be found on the Wikinegata platform<sup>3</sup>.

To construct the link prediction dataset, we filter the positive statements to retain only those involving entities that are also present in the set of negative statements. Formally, we include a positive triple  $(h, r, t)$  in our dataset only if entity  $h$  participates in relation  $r$  in at least one negative statement, thereby ensuring that each entity is represented with both positive and negative statements. Subsequently, we split the positive statements into training and test sets, so that no entity or relation type is exclusive to the test set. The resulting dataset comprises 743 798 positive

<sup>3</sup><https://d5demos.mpi-inf.mpg.de/negation/index.html>

statements and 453 325 negative statements in the training set, as well as a test set with 1 251 positive statements used to evaluate link prediction performance through ranking metrics.

## 4.2 Gene Ontology Knowledge Graph

GO KG integrates both the GO itself and the GO annotation data [10]. GO is structured as a directed acyclic graph, where nodes represent GO classes that describe gene product functions, and edges capture semantic relationships such as *isA* or *partOf*. For example, the triple (*GO\_0055089*, *isA*, *GO\_0055088*) expresses that fatty acid homeostasis (*GO\_0055089*) is a subclass of lipid homeostasis (*GO\_0055088*). In addition to the ontology, GO annotations provide associations between gene products and GO classes, linking biological entities to their functional descriptions. In triple form, this can be represented as (*Insulin*, *hasFunction*, *GO\_0055089*), indicating that insulin is involved in fatty acid homeostasis (*GO\_0055089*).

Recent efforts have enriched the GO KG by incorporating biologically meaningful negative GO annotations, which explicitly indicate that a protein does not perform a particular function. In this work, we employ the negative GO associations introduced by [39] derived from phylogenetic analysis of protein families. By tracking the evolutionary history of proteins across species, the phylogenetic trees reveal where functions are gained or lost over time. This enables the identification of functions that are absent in specific lineages and thus suitable to be recorded as negative annotations. For instance, if proteins from the insulin family once exhibited D-glucose transmembrane transporter activity but lost this function, the corresponding negative annotation can be represented as the triple (*Insulin*, *hasFunction*, *GO\_0055056*), where *GO\_0055056* denotes D-glucose transmembrane transporter activity. As a result, the GO KG comprises 91 701 positive statements derived from GO and GO annotations and 32 959 negative statements reflecting the negative GO annotations.

Predicting PPIs is essential for understanding cellular processes and disease mechanisms, yet experimentally determining these interactions is expensive and time-consuming. Consequently, computational prediction methods aim to rank protein pairs based on their likelihood of interaction, thereby guiding experimental efforts effectively. Namely, functional information represented in the GO KG has been widely explored to predict PPIs in recent studies [32]. We evaluate our approach using a PPI dataset introduced by [33] and used to assess TrueWalks [31], constructed from the STRING database [34]. This dataset comprises 440 proteins, 1 024 interacting protein pairs (corresponding to valid examples), and another 1 024 non-interacting protein pairs (corresponding to invalid examples).

## 5 Results and Discussion

From subsections 5.1 to 5.3, we first present and analyze the effectiveness of the proposed approach when *cl\_phase* is set to 350. In terms of embedding dimensions, baseline models generate 100-dimensional embeddings, whereas our approach simultaneously trains positive and negative models to create embeddings of 50 dimensions each. To thoroughly evaluate our approach, we first report performance results for the two tasks (link prediction on Wikidata and triple classification on the GO KG). Next, we investigate the quality of the embeddings on both KGs, Wikidata and GO,

**Table 2: Rank-based metrics (MRR, Hits@10, and Hits@1) on the Wikidata KG. The reported metrics are averaged over head and tail predictions. Higher values reflect better results. Baselines are shaded in grey, while bold text highlights the best approach for each metric within each KGE model.**

KGE Model		MRR	Hits@10	Hits@1
TransE	pos	10.06%	18.20%	5.52%
	neg	9.34%	17.28%	5.08%
	ours (pos)	10.15%	18.56%	5.52%
	ours (pos-concat-neg)	<b>10.40%</b>	<b>19.32%</b>	5.36%
DistMult	pos	6.86%	12.64%	3.76%
	neg	4.052%	7.480%	2.240%
	ours (pos)	7.90%	17.52%	3.40%
	ours (pos-concat-neg)	<b>9.63%</b>	<b>20.56%</b>	<b>4.64%</b>
ComplEx	pos	4.66%	10.320%	1.840%
	neg	4.77%	8.60%	2.96%
	ours (pos)	8.86%	19.12%	4.16%
	ours (pos-concat-neg)	<b>10.07%</b>	<b>20.80%</b>	<b>4.84%</b>

by applying clustering metrics to determine how well KG entities of the same semantic type are grouped in the embedding space. Finally, subsection 5.4 presents ablation studies to analyze the impact of varying two hyperparameters (*cl\_phase* and embedding size).

### 5.1 Link Prediction on Wikidata

Link prediction involves inferring a missing entity in a triple, such as predicting the head entity *h* given (*r*, *t*), or the tail entity *t* given (*h*, *r*). For each test triple (*h*, *r*, *t*), we generate corrupted triples by replacing either the head or the tail entity with every entity in the KG. A scoring function  $f(h, r, t)$  is then applied to both the original and corrupted triples. To ensure fair evaluation, we follow previous works and filter out any corrupted triples that appear in the training set, as their inclusion could underestimate the model’s performance. Finally, the remaining triples are ranked by their scores to evaluate the model’s ability to distinguish valid from invalid triples.

Table 2 reports the mean reciprocal rank (MRR) and the proportion of valid triples ranked in the top 10 (Hits@10) and top 1 (Hits@1) for both head and tail predictions on Wikidata KG, using three KGE models. For each KGE, we compare the baselines trained separately on the positive and negative statements with two variations of our approach: one that uses only the embeddings learned from the positive KG, and another that represents entities by concatenating their embeddings from both the positive and negative KGs. Across all models, the proposed contrastive learning approach consistently outperforms the baselines for MRR and Hits@10, with the best performance consistently achieved when combining positive and negative embeddings. The improvements are particularly pronounced for DistMult and ComplEx, with MRR values increasing from 6.86% to 9.63% or from 4.66% to 10.07%. The proposed approach narrows the performance gap between simpler models, such as TransE, and more complex ones, like ComplEx and DistMult, so that they converge to similar MRR, Hits@10, and Hits@1 values despite starting with considerable baseline differences.

Interestingly, and somewhat counterintuitively, training a baseline model on a KG with negative statements and then using it to score positive statements yields competitive results, especially for ComplEx, where training on the negative KG outperforms training on the positive KG alone. This highlights the value of meaningful negative statements and suggests that, because negative statements

**Table 3: Sem@1 on Wikidata KG, reported separately for predicting the head entity, the tail entity, and their average. Baselines are shaded in grey, while bold text highlights the best approach for each metric within each KGE model.**

KGE Model		Head	Tail	Average
TransE	pos	53.12%	<b>62.48%</b>	57.80%
	neg	20.96%	58.72%	39.84%
	ours (pos)	<b>63.36%</b>	57.84%	<b>60.60%</b>
	ours (pos-concat-neg)	53.04%	62.40%	57.72%
DistMult	pos	62.16%	40.00%	51.08%
	neg	25.28%	46.64%	35.96%
	ours (pos)	71.84%	46.48%	59.16%
	ours (pos-concat-neg)	<b>74.40%</b>	<b>58.96%</b>	<b>66.68%</b>
ComplEx	pos	76.72%	34.08%	55.40%
	neg	37.44%	<b>63.20%</b>	50.32%
	ours (pos)	<b>80.24%</b>	39.20%	59.72%
	ours (pos-concat-neg)	74.88%	57.60%	<b>66.24%</b>

are still plausible, training on them helps the model to better distinguish between plausible and implausible triples.

In addition to rank-based metrics, semantic awareness is assessed in Table 3 using the Sem@K metric [17], which measures the proportion of triples that are semantically valid in the first  $K$  top-scored triples. Formally, Sem@K is defined as:

$$\text{Sem@K} = \frac{1}{|B|} \sum_{q \in B} \frac{1}{K} \sum_{q' \in S_q^K} \text{compatibility}(q, q') \quad (1)$$

where for each ground-truth triple  $q = (h, r, t)$ , the set  $S_q^K$  contains the top  $K$  candidate triples scored by the KGE model, either by predicting the tail given  $(h, r)$  or the head given  $(r, t)$ . The compatibility function,  $\text{compatibility}(q, q')$ , evaluates whether a candidate triple  $q'$  is semantically compatible with its ground-truth  $q$ . In this work, semantic compatibility means that the predicted entity (head or tail) belongs to the same type as the corresponding entity in the ground-truth triple. For example, given the ground-truth triple  $(\text{Mozart}, \text{occupation}, \text{Composer})$ , the prediction  $(\text{Mozart}, \text{occupation}, \text{Teacher})$  is semantically valid since both tails are of type profession, whereas  $(\text{Mozart}, \text{occupation}, \text{Germany})$  is invalid because its tail belongs to the type country. Table 3 shows that the proposed dual model approach with contrastive learning yields more semantically valid predictions. The best results for DistMult and ComplEx are obtained when embeddings from both KGs are combined. These results indicate that the proposed approach not only improves ranking metrics but also leads to more semantically plausible predictions.

## 5.2 Triple Classification on Gene Ontology

The triple classification task involves determining whether a given triple  $(h, r, t)$  holds or not and can be framed as a binary classification problem. In the context of the GO KG, this task corresponds to predicting whether one protein interacts with another, i.e., whether  $(p_1, \text{interacts}, p_2)$  is valid. Following the approach in [31], to predict the relation between a pair of proteins  $p_1$  and  $p_2$ , we first obtain their vector embeddings  $v_{p_1}$  and  $v_{p_2}$ . These embeddings are then combined using the Hadamard product [13] to form the pair representation:  $r(p_1, p_2) = v_{p_1} \times v_{p_2}$ . The resulting pair representations  $r(p_1, p_2)$  serve as input to train a Random Forest classifier [5], which has been shown to perform very well in PPI prediction [31]. For

**Table 4: Median of performance metrics (Pr, Re, F1, and AUC) for PPI prediction using the GO KG, computed over 5-fold cross-validation. Bold text highlights the best performing approach for each metric within each KGE model.**

KGE Model		Pr	Re	F1	AUC
TransE	pos	58.85%	58.68%	58.51%	62.80%
	neg	64.07%	64.06%	64.05%	68.49%
	ours (pos)	60.98%	60.98%	60.97%	64.74%
	ours (pos-concat-neg)	<b>67.34%</b>	<b>67.32%</b>	<b>67.31%</b>	<b>74.13%</b>
DistMult	pos	81.67%	81.66%	81.66%	88.91%
	neg	<b>83.70%</b>	<b>82.89%</b>	<b>82.78%</b>	90.58%
	ours (pos)	77.27%	77.07%	77.03%	86.03%
	ours (pos-concat-neg)	83.03%	82.68%	82.64%	<b>90.93%</b>
ComplEx	pos	79.34%	79.02%	78.94%	86.70%
	neg	81.45%	80.20%	79.99%	87.93%
	ours (pos)	78.97%	78.97%	78.97%	87.54%
	ours (pos-concat-neg)	<b>81.91%</b>	<b>81.91%</b>	<b>81.91%</b>	<b>89.55%</b>

the Random Forest classifier, we employ a 5-fold cross-validation strategy. In each fold, the classifier is trained to output predicted labels (0 or 1) for each protein pair. These predicted labels are then used to compute the predictive performance metrics, including precision (Pr), recall (Re), weighted average F-measure (F1), and area under the receiver operating characteristic curve (AUC).

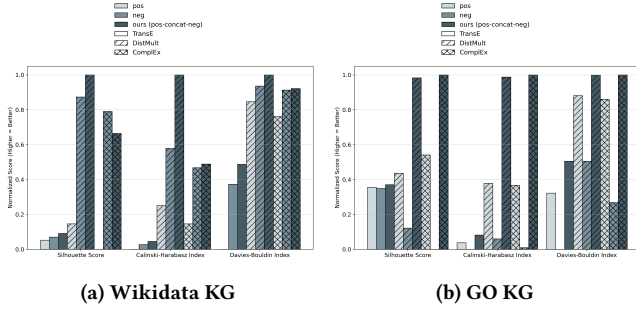
Table 4 reports the median performance metrics for the baselines trained separately on the positive and negative statements, as well as two variations of the proposed approach (one uses only the positive embeddings, and the other concatenates both). For TransE and ComplEx, the proposed approach leads to improvements across all performance metrics compared to training KGE models solely on positive and negative KGs. For example, TransE shows an improvement in F1 from 58.5% to 67.3%. Importantly, the differences between the proposed approach and the baselines for both TransE and ComplEx are statistically significant according to the Kruskal–Wallis test [23] with a significance level of 0.05. The only exception is DistMult, where training solely on the negative KG yields the best performance across most metrics, even outperforming the proposed contrastive learning approach. This behavior mirrors the link prediction results, likely because the model learns from plausible negative statements.

Overall, it appears that weaker KGE models, such as TransE, benefit the most from the proposed contrastive learning approach. For instance, TransE achieves the lowest performance scores but exhibits the largest gains when combined with the proposed contrastive learning approach. In contrast, DistMult is the best performing method, and likely due to its already high baseline performance, the proposed approach offers only small improvements.

## 5.3 Embeddings Evaluation

To gain deeper insights into the distribution of different embeddings, we investigate how entity types are represented using clustering metrics. Rather than performing actual clustering on the data, we use these metrics to assess the degree of natural separation between entities of different types in the embedding space. This provides a proxy for how well the representations obtained with different approaches reflect underlying semantic distinctions. For Wikidata KG, we consider 21 frequent types, whereas for the GO KG only two classes are used, namely the GO classes and proteins.





**Figure 2:** Bar plots display normalized clustering metric values (Calinski-Harabasz, inverted Davies-Bouldin, and Silhouette) comparing the baselines and the proposed approach (distinguished by colors) across three KGE models (distinguished by bar hatchings). Metric values are normalized to [0, 1] within each metric, with higher values consistently indicating better cluster separation. After normalization, the lowest values for each metric are displayed as 0 in the plots.

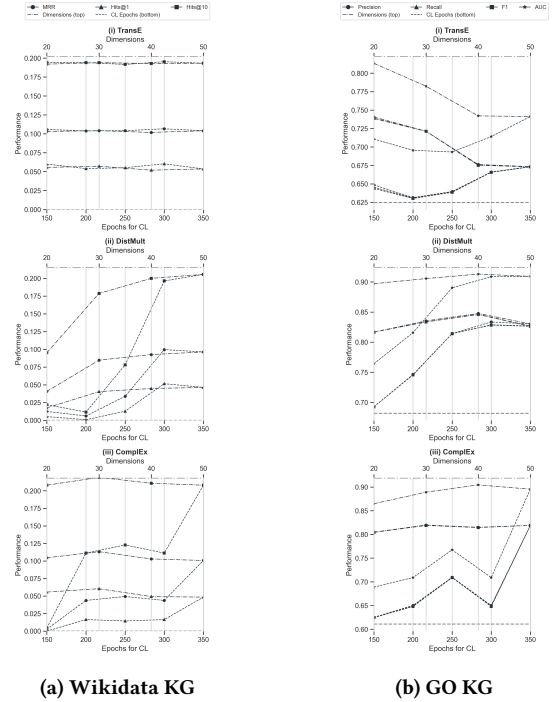
Three clustering metrics are employed. The Calinski-Harabasz score [6] measures the ratio of between-cluster dispersion to within-cluster dispersion. Higher values indicate better-defined clusters with greater separation. The Davies-Bouldin score [7] evaluates the average similarity between each cluster and its most similar peer cluster. Lower values indicate better clustering. The silhouette score [27] assesses how similar an entity is to its own cluster compared to other clusters. Higher values indicate points well-matched to their own cluster and poorly matched to neighboring clusters.

Figure 2 presents bar plots displaying the values of three clustering metrics for baselines trained separately on the positive and negative statements, as well as the proposed approach across three KGE models. Within each metric, all values are normalized to the [0, 1] range, with the Davies-Bouldin scores inverted to maintain a consistent higher-is-better interpretation. Furthermore, this normalization enables direct visual comparison of the relative effectiveness of each approach. Across all metrics, our approach demonstrates superior clustering performance, highlighting the effectiveness of the contrastive learning strategy in increasing inter-type separability.

Interestingly, for Wikidata KG, the baseline trained on the negative KG outperforms the baseline trained on the positive KG in terms of cluster separability. One possible explanation is that the distribution of relations associated with each entity type in Wikidata differs in the negative KG, which may make the clustering task easier. This distributional bias is not observed in the GO KG, where entities typically participate in the same relations (e.g., *hasFunction* isA) across both KGs. An analysis of the two most frequent entity types (*Person* and *Business*) in Wikidata revealed that this bias affects business entities, with negative statements centered around a single relation (*structuredAs*), unlike in positive statements.

## 5.4 Ablation Studies

We conduct ablation studies to investigate the impact of two hyperparameters of the proposed approach: the dimensionality of the embeddings, and the epoch at which the negative sample generation transitions from negative sampling to a contrastive learning-based



**Figure 3:** Line plots illustrating the effect of *cl\_phase* (bottom x-axis) and the embeddings dimensionality (top x-axis) on performance (y-axis) for link prediction on Wikidata KG and triple classification on GO KG.

strategy (*cl\_phase*). Figure 3 shows performance on both the Wikidata and GO KGs for multiple values of *cl\_phase* (100, 150, 200, 250, 300 and 350) and embedding dimensions (20, 30, 40, and 50). The results indicate that, particularly for larger KGs like Wikidata, earlier transitions to contrastive learning (lower *cl\_phase* values) tend to limit the model’s performance, with performance generally increasing as *cl\_phase* approaches 300. Delaying the transition appears to allow embeddings to stabilize before introducing more challenging contrastive samples. It is essential to note that during contrastive learning, for each triple in a KG, either the head or the tail entity is replaced with all possible entities from the KG. This incurs additional computational time per epoch, as the negative samples are dynamically updated rather than being fixed. Therefore, selecting the appropriate *cl\_phase* should balance between training time and performance gains. In terms of embedding dimensionality, the proposed approach demonstrates robustness across different embedding sizes, with relatively stable performance. However, an exception is observed with DistMult on Wikidata KG, where larger dimensions lead to substantial performance gain.

## 6 Conclusion

KGE methods are increasingly applied in a wide array of real-world applications across multiple domains. However, existing KGE models are not designed to handle meaningful negative statements that are explicitly defined, despite their recognized importance in knowledge representation. To overcome this limitation, we propose a novel approach that combines dual-model training with an adapted

negative sampling mechanism grounded in contrastive learning. Our empirical evaluation demonstrates that the proposed approach outperforms state-of-the-art KGE models on both general-purpose (Wikidata) and domain-specific (GO) KGs, yielding substantial gains in link prediction and triple classification tasks. Given its flexible design, the proposed approach can be easily incorporated into any KGE model that employs a scoring function. Moreover, it offers generalizability across different KGs and predictive tasks.

For future work, several directions can be explored to improve the proposed approach. First, in contrastive learning, for each triple in the KG, either the head or tail is paired with all KG entities. While effective, this approach is not scalable for very large KGs. Future work could investigate more efficient strategies, such as random sampling from the set of entities or selecting entities based on their embedding proximity, to substantially reduce computational cost. Second, further improvements could be achieved through systematic hyperparameter optimization, which may lead to additional performance gains. Finally, the current implementation trains the two models synchronously, applying the same number of initial training epochs to both the positive and negative KGE models. However, an asynchronous training strategy, allowing different numbers of epochs for each model, could be explored in the future.

## References

- [1] Hiba Arnaout, Tuan-Phong Nguyen, Simon Razniewski, and Gerhard Weikum. 2023. UnCommonSense in Action! Informative Negations for Commonsense Knowledge Bases. In *ACM International Conference on Web Search and Data Mining*. 1120–1123.
- [2] Hiba Arnaout, Simon Razniewski, Gerhard Weikum, and Jeff Z. Pan. 2021. Negative knowledge for open-world wikidata. In *The Web Conference*. 544–551.
- [3] Hiba Arnaout, Simon Razniewski, Gerhard Weikum, and Jeff Z. Pan. 2021. Wikinegata: a knowledge base with interesting negative statements. *Proceedings of the VLDB Endowment* 14, 12 (2021), 2807–2810.
- [4] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems* 26 (2013).
- [5] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.
- [6] Tadeusz Calinski and Jerzy Harabasz. 1974. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods* 3, 1 (1974), 1–27.
- [7] David L. Davies and Donald W. Bouldin. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2 (1979), 224–227.
- [8] Patrick Doherty, Witold Lukaszewicz, and Andrzej Szalas. 2000. Efficient reasoning using the local closed-world assumption. In *International Conference on Artificial Intelligence: Methodology, Systems, and Applications*. 49–58.
- [9] Luis Antonio Galárraga, Christina Teflioudi, Katja Hose, and Fabian Suchanek. 2013. AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In *International Conference on World Wide Web*. 413–422.
- [10] GO Consortium. 2021. The Gene Ontology resource: enriching a GOLD mine. *Nucleic Acids Research* 49, D1 (2021), D325–D334.
- [11] Qingyu Guo, Fuzhen Zhuang, Chuan Qin, Hengshu Zhu, Xing Xie, Hui Xiong, and Qing He. 2020. A survey on knowledge graph-based recommender systems. *IEEE Transactions on Knowledge and Data Engineering* 34, 8 (2020), 3549–3568.
- [12] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia d'Amato, Gerard de Melo, Claudio Gutierrez, Sabrina Kirrane, et al. 2021. Knowledge Graphs. *ACM Computing Surveys* 54, 4 (2021), 1–37.
- [13] Roger A. Horn. 1990. The hadamard product. In *Proceedings of Symposia in Applied Mathematics*, Vol. 40. 87–169.
- [14] Junjie Huang, Huawei Shen, Liang Hou, and Xueqi Cheng. 2019. Signed graph attention networks. In *International Conference on Artificial Neural Networks*. Springer, 566–577.
- [15] Junjie Huang, Ruobing Xie, Qi Cao, Huawei Shen, Shaoliang Zhang, Feng Xia, and Xueqi Cheng. 2023. Negative can be positive: Signed graph neural networks for recommendation. *Information processing and management* 60, 4 (2023), 103403.
- [16] Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. 2019. Knowledge graph embedding based question answering. In *ACM International Conference on Web Search and Data Mining*. 105–113.
- [17] Nicolas Hubert, Pierre Monnin, Armelle Brun, and Davy Monticcolo. 2023. Sem@K: Is my knowledge graph embedding model semantic-aware? *Semantic Web* 14, 6 (2023), 1273–1309.
- [18] Nicolas Hubert, Pierre Monnin, Armelle Brun, and Davy Monticcolo. 2024. Treat different negatives differently: Enriching loss functions with domain and range constraints for link prediction. In *Extended Semantic Web Conference*. 22–40.
- [19] Bhushan Kotnis and Vivi Nastase. 2017. Analysis of the impact of negative sampling on link prediction in knowledge graphs. *arXiv:1708.06816* (2017).
- [20] Denis Krompaß, Stephan Baier, and Volker Tresp. 2015. Type-constrained representation learning in knowledge graphs. In *International Semantic Web Conference*. Springer, 640–655.
- [21] Ziyang Liu, Chaokun Wang, Jingcao Xu, Cheng Wu, Kai Zheng, Yang Song, Na Mou, and Kun Gai. 2023. PANE-GNN: Unifying positive and negative edges in graph neural networks for recommendation. *arXiv:2306.04095* (2023).
- [22] Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. *arXiv:2109.03764* (2021).
- [23] Patrick E. McKight and Julius Najab. 2010. Kruskal-wallis test. *The corsini encyclopedia of psychology* (2010), 1–1.
- [24] Jan Portisch, Nicolas Heist, and Heiko Paulheim. 2022. Knowledge graph embedding for data mining vs. knowledge graph embedding for link prediction—two sides of the same coin? *Semantic Web* (2022), 1–24.
- [25] Petar Ristoski and Heiko Paulheim. 2016. RDF2Vec: RDF graph embeddings for data mining. In *International Semantic Web Conference*. Springer, 498–514.
- [26] Andrea Rossi, Denilson Barbosa, Donatella Firmani, Antonio Matinata, and Paolo Merialdo. 2021. Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data* 15, 2 (2021), 1–49.
- [27] Peter J. Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* 20 (1987), 53–65.
- [28] Ali Shaban and Heiko Paulheim. 2024. SnapE—training snapshot ensembles of link prediction models. In *International Semantic Web Conference*. Springer, 3–22.
- [29] Kartik Shenoy, Filip Ilievski, Daniel Garijo, Daniel Schwabe, and Pedro Szekely. 2022. A study of the quality of Wikidata. *Journal of Web Semantics* 72 (2022), 100679.
- [30] Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. *Advances in Neural Information Processing Systems* 26 (2013).
- [31] Rita T. Sousa, Sara Silva, Heiko Paulheim, and Catia Pesquita. 2023. Biomedical knowledge graph embeddings with negative statements. In *International Semantic Web Conference*. Springer, 428–446.
- [32] Rita T. Sousa, Sara Silva, and Catia Pesquita. 2020. Evolving knowledge graph similarity for supervised learning in complex biomedical domains. *BMC Bioinformatics* 21, 1 (2020), 1–19.
- [33] Rita T. Sousa, Sara Silva, and Catia Pesquita. 2023. Benchmark datasets for biomedical knowledge graphs with negative statements. In *Workshop on Semantic Web Solutions for Large-scale Biomedical Data Analytics co-located with Extended Semantic Web Conference*.
- [34] Damian et al. Szklarczyk. 2020. The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research* 49, D1 (11 2020), D605–D612.
- [35] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. 2016. Complex Embeddings for Simple Link Prediction. In *International Conference on International Conference on Machine Learning*, Vol. 48. 2071–2080.
- [36] Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledgebase. *Communications of the ACM* 57, 10 (2014), 78–85.
- [37] Quan Wang, Zhendong Mao, Bin Wang, and Li Guo. 2017. Knowledge graph embedding: A survey of approaches and applications. *IEEE Transactions on Knowledge and Data Engineering* 29, 12 (2017), 2724–2743.
- [38] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge Graph Embedding by Translating on Hyperplanes. In *AAAI Conference on Artificial Intelligence*. 1112–1119.
- [39] Alex Warwick Vesztrocy and Christophe Dessimoz. 2020. Benchmarking Gene Ontology function predictions using negative annotations. *Bioinformatics* 36, Supplement\_1 (07 2020), i210–i218.
- [40] Bishan Yang, Scott Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *International Conference on Learning Representations*.

## A Hyperparameters for Embedding Models

For our experiments, we adapt the OpenKE library<sup>4</sup>. All models (TransE, DistMult, ComplEx) are trained for 400 epochs with 100 batches, an entity negative sampling rate of 1, and a relation negative rate of 0. TransE uses SGD for optimization, while DistMult and ComplEx use Adagrad. All remaining hyperparameters were left at their default values provided by the OpenKE library.

<sup>4</sup><https://github.com/thunlp/OpenKE/tree/OpenKE-Tensorflow1.0>