

# Refining SemOpenAlex Concept Ontology: A Constraint-Aware Approach via Knowledge Graph Embeddings and SKOS Constraints

Özge Erten<sup>1,\*</sup>, Shervin Mehryar<sup>1</sup>, Bo Xiong<sup>2</sup>, Remzi Çelebi<sup>1</sup> and Christopher Brewster<sup>1,3</sup>

<sup>1</sup>*Institute of Data Science, Maastricht University, Paul-Henri Spaaklaan 1, 6229 GT, Maastricht, Netherlands*

<sup>2</sup>*Institute for Artificial Intelligence, University of Stuttgart, 70569, Stuttgart, Germany*

<sup>3</sup>*Data Science Group, TNO, Kampweg, Soesterberg, Netherlands*

## Abstract

The continuous growth in scientific publications has led to an increasing demand for efficient solutions in managing vast amounts of scholarly information. SemOpenAlex, an academic article Knowledge Graph (KG), aims to organize the scientific papers by tagging them with the concepts representing the relevant topics. The concepts are hierarchically organized using the Simple Knowledge Organization System (SKOS) vocabulary. However, this concept hierarchy contains noise resulting from Natural Language Processing (NLP) concept extraction. This paper proposes a link prediction based method to reduce the noise within SemOpenAlex concept hierarchy. The method utilizes informal SKOS consistency definitions to create negative triples which violate the definitions, combined with randomly generated negatives. The primary objective here is to integrate true negative samples into the knowledge graph embedding model during the learning process. This study contributes to refining SKOS-based KGs by enhancing the semantic quality of information within the KG.

## Keywords

Knowledge graph, Link prediction, SKOS vocabulary, negative sampling, custom negative sampling, constraint-aware embeddings, constraint-aware negative sampling, Knowledge graph embeddings

## 1. Introduction

According to the online citation index Web of Science, over 6 million articles were published between 2018 and 2022 [1]. The increasing number of publications, as shown in Figure 1, contribute to a rich pool of scholarly knowledge. Since scholarly knowledge is rapidly increasing and evolving; accessing to it is often insufficient [2]. For instance, from researchers' perspective, the volume of publications turns an efficient literature search and discovery of relevant articles and topics into a labor-intensive task. One popular approach to managing this knowledge involves the use of Knowledge Graphs (KGs). Beside representational capabilities, KGs offer the support of dividing topics and their sub-fields [3]. Recently, SemOpenAlex KG [4] was developed to organize and represent scientific publications and related information across diverse domains. This KG is structured around entities such as titles, authors, abstracts and article texts. It also includes a concept hierarchy that categorizes publications under different topics. This concept hierarchy was built using the Simple Knowledge Organization System (SKOS) standard – a common vocabulary representing controlled vocabularies or knowledge organization systems in a machine-readable way. Some popular knowledge bases facilitate it to organize their information. For example, AGROVOC[5], an agricultural thesaurus, uses SKOS to

---

SWAT4HCLS 2024: *The 15th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences, February 26–29, 2024, Leiden, The Netherlands*

\*Corresponding author.

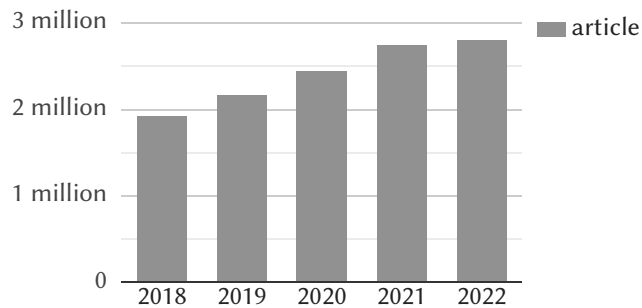
✉ o.erten@maastrichtuniversity.nl (Ö. Erten); shervin.mehryar@maastrichtuniversity.nl (S. Mehryar); bo.xiong@ipvs.uni-stuttgart.de (B. Xiong); remzi.celebi@maastrichtuniversity.nl (R. Çelebi); christopher.brewster@maastrichtuniversity.nl (C. Brewster)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

represent its semantic relations. GACS Core [6], an agricultural conceptual scheme developed to enhance the consistency of agricultural research, also uses SKOS.



**Figure 1:** The chart shows the published article counts on Web of Science from 2018 to 2022. There is a consistent increase in the number of articles published each year during this period [1].

However, the extraction and organization of SemOpenAlex concepts involve Natural Language Processing (NLP) techniques and heuristics, often introduce noise into the extracted knowledge. This leads to a loss of quality and reliability in the KG. Manual efforts in cleaning noise from KGs are time-consuming and laborious. In this research, we aim to improve the quality of SemOpenAlex concepts with an automated method [7]. Specifically, this work focuses on enhancing the data quality of SemOpenAlex by improving the quality of SKOS connections in the concept hierarchy. To achieve this, our approach first removes inconsistent relations in SemOpenAlex KG and subsequently predicts more accurate ones. We apply a customised negative sampling method within Knowledge Graph Embedding (KGE) techniques for link prediction. Namely, traditional KGEs commonly assume local completeness in the KG, employing a 'Close World' perspective when generating negative triples. This approach can lead to the generation of negative triples that are actually correct, as emphasized in Jain et al.(2021)'s study [8]. To address this limitation, we propose a constraint-aware negative sampling technique that leverages informal SKOS constraints for triple corruptions. Additionally, we provide evidence that the predicted SemOpenAlex concept connections align with established standards, by validating the predictions with the Unified Medical Language System (UMLS) ontology.

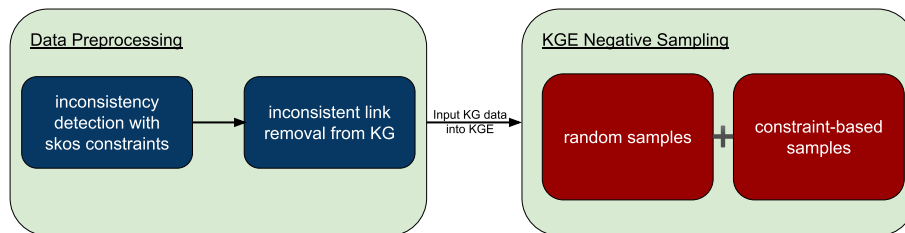
## 2. Related Work

In this context, Jain et al. (2021) [8] point out a weakness in common methods for learning from Knowledge Graphs (KGs). They argue that these methods assume KGs are locally complete, however this can not guarantee truly incorrect negative samples for the training. To address this, the authors introduce a method called ReasonKGE. ReasonKGE assesses whether the learning process aligns with the description logic used. If not, it identifies predictions that violate the logic and includes them, along with the similarly generated triples, as negative samples for the next training iteration. The authors experiments conducted indicate that ReasonKGE produces more accurate predictions compared to standard methods. On the other hand, Alam et al. (2020) [9] highlight a common challenge in KGE models related to the random generation of negative samples. Commonly, KGE models follow the Close World Assumption (CWA). They treat unseen triples in KGs as unknown and rely on existing data. To overcome this limitation, the authors introduce a negative sampling approach that considers triple affinity. Initially, the method calculates the distance of a candidate triple for corruption using cosine similarity, and compares it with the remaining triples in the KG. Subsequently, an affinity function incorporates this distance and the comparison to find the fitness of

each entity for the corruption task. As its output, the function produces a set of entities with their likelihood scores for corruption. Random selection of these entities forms a batch of negative samples from the candidate triple. Ultimately, the proposed method enhances the performance of KGE models, particularly in terms of time efficiency, for link prediction tasks as demonstrated in their experiments. Another work is conducted by Yao et al. (2022) [10] which aims to enhance the quality of negative samples in KGE models. Their approach takes into account triples with similar context in the KG, and generates more meaningful negative samples. For example, given the positive triple (Steve Jobs, FounderOf, Apple Inc.), a higher-quality negative sample would be (Jerry Yang, FounderOf, Apple Inc.) rather than a less contextually relevant one like (Yahoo, FounderOf, Apple Inc.). The method assesses the quality of a negative sample based on its closeness to the positive entity during training. After, it selects the most similar entity for corruption. The authors conducted experiments comparing their approach to state-of-the-art methods in link prediction tasks, and their method demonstrated better performance [8, 9, 10]. While previous works concentrate on general rules or KG structures when generating custom negative samples, our approach distinguishes itself by emphasizing semantic simplicity and employing specific rules. This approach improves the reliability of generating accurate negative samples.

### 3. Methodology

This paper focuses on refining noisy SKOS relations in the SemOpenAlex scholarly KG. We formulate the problem as a KG completion task with the aim to improve the concept hierarchy. We apply a custom negative sampling method for KG completion models. This method is based on the World Wide Web Consortium (W3C) informal SKOS constraint definitions. We use Knowledge Graph Embedding (KGE) and Graph Neural Network (GNN) methods to measure the predictive accuracy of our approach.



**Figure 2:** Methodology. The method starts by identifying the triples which conflict with the aforementioned SKOS constraints. Since the triples causing an inconsistency can not be determined, we remove all triples involved in inconsistent cases. Then, this updated KG was used for training KGE. During training, true negative samples are introduced in the KG embedding models to predict missing SKOS relations in SemOpenAlex concept hierarchy.

Specifically, we experiment with baseline KGE models TransE [11, 12], DistMult [13] and QuatE [14] along with a GNN model. TransE is a translation type model which embeds vector representations for entities with their relation to the translation of the entities in low dimensional vector space. More specifically, the embedding of an tail entity is expected to be in close proximity to the embedding of a head entity, augmented by their relationship vector. For instance, in the triple (Steve Jobs, FounderOf, Apple Inc.), "Steve Jobs" is the head entity, "Apple Inc." is the tail entity, and they are connected by the "FounderOf" relation. In TransE, the location of "Steve Jobs" plus "FounderOf" is expected to be close to "Apple Inc." in the embedding space. RESCAL [15] models each relation as a matrix and models each triple as a three-way interaction between the head, relation and tail entity. However, RESCAL has quadratic number of relational parameters. DistMult [13]

simplifies RESCAL by restricting matrices representing relations as diagonal matrices. Due to the simple diagonal matrix relational modeling, DistMult cannot model symmetric relations. QuatE [14] embeds entities and relations as quaternions—4D hypercomplex numbers with one real component and three imaginary components. For each triple, QuatE first rotates the head quaternion number by performing a Hamilton Product with the corresponding relation quaternion. The plausibility of the triple is then measured by the quaternion inner product between the rotated head entity and the tail entity. QuatE has been proven to subsume DistMult and ComplEx. These methods produce negative samples by randomly mixing head and tail entities. Since they follow the Close World Assumption, there is a possibility for generating negative samples which are actually correct [8, 11, 13, 14].

We first applied SKOS inconsistency definitions, formulation 1 and formulation 2, to identify groups of triples that did not satisfy *Constraint-1* or *Constraint-2*. More specifically, the formulation 1, implies that there should not be entity  $a$  that is broader than entity  $b$ , and entity  $a$  that is related to entity  $b$ , as it would create an inconsistency. Likewise, the formulation 2 states that for all entities  $a$ ,  $b$ , and  $c$ , if  $a$  is broader than  $b$  and  $a$  is related to  $c$ , then it should not be the case that  $b$  is broader than  $c$  to maintain consistency [16].

$$\textbf{Constraint-1: } \forall a, b : \textit{skos : related}(a, b) \wedge \textit{skos : broader}(a, b) \rightarrow \perp \quad (1)$$

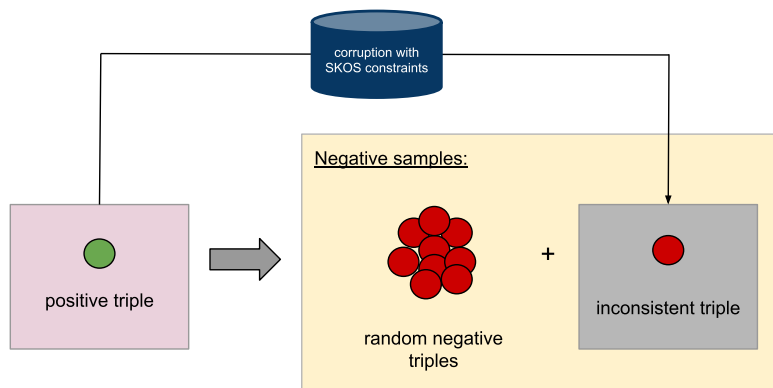
$$\textbf{Constraint-2: } \forall a, b, c : (\textit{skos : broader}(a, b) \wedge \textit{skos : related}(a, c)) \rightarrow \neg \textit{skos : broader}(b, c) \quad (2)$$

Second, we cleaned the KG from the links that conflict with Constraint-1 and Constraint-2 to KG completion step. For the link prediction, our method utilised KGE and GNN models. Briefly, KGE is a machine learning technique which is used for KG applications including KG completion and recommender systems. KGE representation reduces the complexity of the graph structure by mapping KG relations and entities to a low-dimensional vector space. Namely, the KGE model maps KG triples into the embedding space by taking into account positive and negative triples. During the training process, model should learn to score positive triples higher than negative ones. Through an iterative learning, the model improves its understanding of the semantic and structural patterns present in the KG. On the other hand, GNNs enhance entity-centric KGE models by incorporating neighboring information and graph structure during model training. The main goal of a GNN model is to embed entities while considering the encoding of their neighbors in the graph [8, 17, 18].

Third, we addressed the negative sampling issue by implementing a customized approach to improve the quality of negative samples. We achieved this by using SKOS inconsistency definitions which are described on the W3C Spec webpage<sup>1</sup>. Initially, we defined SKOS inconsistency definitions as first-order logic formulas for having structured inconsistency representations. Then, we generated corrupted triples by using these definitions on the KG triples during KGE model training as additional negative samples. Lastly, we explored how an increased quantity of noisy samples affects KGE link prediction accuracy on the SemOpenAlex concept hierarchy[16].

In the final step, we trained KGE models using custom constraint-aware negative sampling. The primary goal is to demonstrate that this method produces predictions that refine previously identified erroneous connections. Since we initially remove all triples contributing to failed groups, it becomes

<sup>1</sup><https://www.w3.org/TR/swbp-skos-core-spec/>



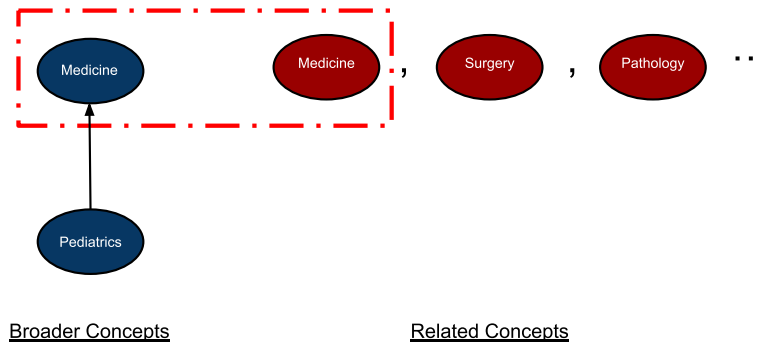
**Figure 3:** Negative sample creation. Inconsistent cases occur if there is a set of triples that violate the constraints. We separate these triples into two categories to establish which of the set cause the inconsistency: Valid triples adhere to the rule, while inconsistency-causing triples defy it. This distinction can be used in negative sampling, where the valid part is regarded as a positive sample and the inconsistent part as a negative one.

crucial to understand which triples are causing conflicts. To achieve this, we leveraged the Unified Medical Language System (UMLS) ontology. Specifically, we contextually matched Medicine and its sub-concepts in SemOpenAlex with UMLS concepts to create a proofing test set. This allows us to distinguish which triples align with SKOS constraints and which ones introduce noise. We assume the presence of a parent-child relation in UMLS proves their hierarchical relationship, so we add such triples into the test set.

## 4. Experimental Setup

To generate a smaller and more manageable dataset from SemOpenAlex, we extracted concepts related to Medicine and its sub-concepts, as well as their relations using `skos:related` and `skos:broader` relations. Initially, the dataset contained 345,119 triples with 51,885 different concepts. Following that, we identified and removed 1,456 conflicting triples for Constraint-1 and 1,044 triples for Constraint-2 from the dataset. Figure 4 illustrates an example inconsistent group of triples. The inconsistencies are used as negative samples as previously detailed out. For each training triplet, we perform experiments with no negative sampling, a negative sample drawn at random, or using the proposed negative sampling according to inconsistency constraints. By introducing these semantic negatives during training, we aim to improve the model’s understanding about the dataset.

In order to further test our methodology, we utilize the conflicting triples as a proof set, by focusing on the `skos:broader` relation and extracting entities that have a matching UMLS concept. For instance as shown in Figure 4, the SemOpenAlex concept Pediatrics (C187212893) was part of a conflict group that matched with the UMLS Pediatrics (C0030755) concept. After identifying the match, we added (Pediatrics, `skos:broader`, Medicine) triple into the test set since this UMLS pair can be a proof of hierarchical connection between Pediatrics and Medicine concepts. We extended this matching process for all Constraint-1 and Constraint-2 conflicts to obtain 204 triples for Constraint-1 and 203 triples for Constraint-2, ultimately used as a test set. Table 1 provides the details on how the data is set up for training, validation, and testing the models.



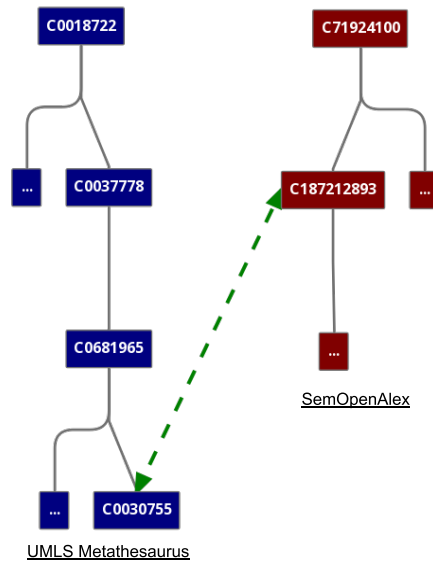
**Figure 4:** Broader and Related concepts of Pediatrics. Here the Pediatrics concept is linked with a skos:broader relation to Medicine concept. Also, it establishes a skos:related connection to the same concept. As per Constraint-1, these two triples form part of an inconsistency group.

**Table 1**  
Summary of the experiment data

<i>Data</i>	<i>Triple Count</i>
Conflicting triples Constraint-1	1456
Conflicting triples Constraint-2	1044
UMLS-matched Constraint-1 triples	204
UMLS-matched Constraint-2 triples	203
Train Batch Size (9:1 ratio)	10000
Validation Batch Size (9:1 ratio)	1000
Proving test set (Constraint-1 triples)	2162
Proving test set (Constraint-2 triples)	2040

We considered three different scenarios under which the negative sampling’s effect is measured. In the first scenario, no negative sampling is used in order to establish the capabilities of each model to learn embeddings solely based on the positive samples. In the second scenario, a random negative sampling component is added in order to establish a baseline for comparison. In the third scenario, the proposed constraint-aware method is evaluated against the baseline methods. To compensate for the imbalance in the number of negative samples, we further bootstrap this case at a ratio of 10% as follows [19]. In every tenth iteration during training, we enforce a full batch to include only constraint-based negative samples as our mixing strategy [20].

The embedding methods are trained for 100 epochs without early stopping and embedding dimensions of size 30 for TransE and DistMul, and size 50 for QuatE and GNN for best performance as reported in the original publication. The parameters are set empirically and according to the settings in [21] and [22]. During the training process, we use an adaptive stochastic optimizer with a batch size of 10,000 for training and hold out a validation set of size 1,000 (9-to-1 ratio), as in Table 1. The learning rate is set to 0.005 for all methods. The QuatE loss function is regularized with  $\lambda = 0.05$ . The GNN uses two convolution layers with an intermediate ReLu activation function. The code is accessible on our Github: <https://github.com/ozzygen/predict-KGE-SKOS>.



**Figure 5:** UMLS and SemOpenAlex ontology matching. To identify matches, Levenshtein distance with the threshold of  $> 92$  is used. For instance, the UMLS Metathesaurus identifies the concept *Pediatrics* by the code *C0030755*, while the corresponding identifier for the *Pediatrics* concept is *C187212893* in SemOpenAlex. This information helps us in pinpointing the correct triples in inconsistent cases to find source of the inconsistency.

## 5. Result and Discussion

We conduct experiments with TransE, DistMult, QuatE KGE models and a GNN to evaluate the efficacy of the proposed negative sampling on link prediction accuracy. Link prediction is one of the key components in knowledge completion and therefore the refinement of SemOpenAlex concept hierarchy. We train the models via the training process described above, followed by a 10-fold validation. The trained and validated models are then assessed on the UMLS-verified triples. The validation set results are shown in Table 2 and the test set results are shown in Table 3. We compare the performance in each case according to Hits@K, Mean Rank (MR), and Mean Reciprocal Rank (MRR). The Hits@K metric measures the precision of the model, while the MR and MRR are reported for generalizability purposes.

From Table 2, it can be observed that including negative sampling given the same number of training epochs improves the performance. The results indicate that including negative samples (random or constraint-based) improves the training accuracy across the board. For instance, for either TransE or DistMult, hits@10 is consistently above 0.70 using either constraint sets. The improvement is more palpable in the case of TransE and Hits@10 metric, where an increase of around 0.60 points is observed for the first set of SKOS constraints and an increase 0.73 for the second set of SKOS constraints, over the baseline. This observation highlights the inherent difference in the constraints used and the effect this phenomenon can have on the ‘translation’ based methods such as TransE. Unlike TranE, the baseline Hits@10 results of DistMult already are competitive at 0.70 and 0.74 due to its algorithmic benefits. Nevertheless, still some improvements can be achieved using the proposed negative sampling. The improvements over the base line are 0.16 points in the case of the first set of SKOS constraints and 0.03 points in the case of the second set of SKOS constraints. It scores 0.08 points better using the second set of SKOS constraints, while the first set of constraints appears to have a neutral effect. Similar to TransE, the choice of negative samples and whether the first or second constraint set is used, conclusively make a difference. The GNN and QuatE are widely adopted methods in knowledge embedding and reasoning tasks due to their expressive power. As expected, the baseline validation performance without negative sampling using either method is

remarkable, with Hits@10 results of 0.78 and 0.8 from the GNN as well as 0.9 and 0.87 from QuatE, under the first set of SKOS constraints and the second set of SKOS constraints respectively. The slight difference observed in performance under either constrain set is due to the structural differences in the resulting knowledge graphs after applying the cleaning step described in the previous section.

Under Constraint 1 refinement criteria, the GNN’s performance improves generally with the addition of negative samplings, however random sampling evidently has as good a performance as the proposed one. Under Constraint 2 however, the proposed approach gains an advantage of 0.02 points in Hits@10. The reason the effect of the proposed negative sampling is mitigated in the first case is deemed to be due to the fact that more entities are involved in the second case which can have an impact based on the methodology of GNN which is a graph-based algorithm and therefore locality information and the number of neighbouring nodes (two versus three) can play an important role. Under the second SKOS constraint with three entities involved, an improved Hits@1 and Hits@10 of 0.71 and 0.82 points are achieved. Among all validated methods, QuatE together with the proposed sampling method achieves the best performance with a Hits@10 of 1.00 and a remarkable MRR of 0.85 Under either SKOS constraint, using negative sampling improves performance. The addition of constraint-based sampling, specially in the second case, achieves an improvement of 0.02 points from 0.94 to 0.98 in Hits@10 and from 0.56 to 0.69 in Hits@1. The MRR scores accordingly improve as well.

Lastly, we evaluate our constraint-aware method using a test set for each KGE method and report the results in Table 3. The test set is created according to the UMLS hierarchy such that the skos:broader predicate is valid. Accordingly, this test set evaluates the performance for link prediction in each case based on the quality of the embeddings and the underlying refined hierarchy. It can be seen that consistent with the validation results, QuatE and GNN continue to achieve the best performance under both constraint criteria. The GNN’s high performance with respect to Hits@1 and Hits@10 which are a measure of prediction accuracy, reflect the ability of GNNs in capturing the underlying structure of the ontology. QuatE benefits from more degrees of freedom and in both constrain settings remains the best performing KGE model after refinement through our proposed approach. These results corroborate the strength of using verified triples from a well maintained medical ontology.

**Table 2**

Performance comparison using TransE, DistMult, GNN, and QuatE. Reported metrics are Hits@1, Hits@10, MR, and MRR. The results compare the two embedding methods under SKOS constraint 1 and constraint 2 as discussed before. We further investigate for each method, three scenarios: without negative sampling (w/o n.s.), with random negative sampling (w/ r.n.s.), and the proposed negative sampling method.

Method	<i>Skos Constraint 1</i>				<i>Skos Constraint 2</i>			
	Hits@1	Hits@10	MR	MRR	Hits@1	Hits@10	MR	MRR
TransE (w/o n.s.)	0.04	0.26	44.7	0.13	0.14	0.14	46.1	0.07
TransE (w/ r.n.s.)	<u>0.52</u>	<u>0.86</u>	<u>6.55</u>	<u>0.65</u>	<u>0.53</u>	<u>0.88</u>	<b>6.01</b>	<u>0.66</u>
TransE (proposed)	<b>0.55</b>	<b>0.88</b>	<b>5.89</b>	<b>0.68</b>	<b>0.55</b>	<b>0.88</b>	<u>6.08</u>	<b>0.67</b>
DistMult (w/o n.s.)	0.03	0.70	18.3	0.05	0.06	0.74	17.2	0.06
DistMult (w/ r.n.s.)	<u>0.08</u>	<u>0.86</u>	<u>12.9</u>	<u>0.08</u>	0.03	<u>0.77</u>	<u>16.3</u>	<u>0.06</u>
DistMult (proposed)	<b>0.10</b>	<b>0.86</b>	<b>12.4</b>	<b>0.08</b>	<b>0.06</b>	<b>0.85</b>	<b>13.6</b>	<b>0.07</b>
GNN (w/o n.s.)	<u>0.65</u>	<u>0.78</u>	<u>12.5</u>	<u>0.08</u>	0.66	<u>0.80</u>	12.4	<u>0.08</u>
GNN (w/ r.n.s.)	<b>0.73</b>	<b>0.86</b>	<b>8.57</b>	<b>0.12</b>	<u>0.68</u>	<u>0.80</u>	12.5	<u>0.08</u>
GNN (proposed)	0.64	0.76	14.4	0.07	<b>0.71</b>	<b>0.82</b>	<b>11.2</b>	<b>0.09</b>
QuatE (w/o n.s.)	0.60	0.90	6.90	0.71	0.56	0.87	6.78	0.67
QuatE (w/ r.n.s.)	<b>0.78</b>	<u>1.00</u>	<u>1.54</u>	<b>0.87</b>	<u>0.56</u>	<u>0.94</u>	<u>2.51</u>	<u>0.71</u>
QuatE (proposed)	<u>0.76</u>	<b>1.00</b>	<b>1.64</b>	<u>0.85</u>	<b>0.69</b>	<b>0.98</b>	<b>1.98</b>	<b>0.80</b>



**Table 3**

Performance comparison on UMLS-verified test set, 2162 triples for Constraint 1 and 2040 triples for Constraint 2. The test sets are created by verifying the triple candidates against the UMLS concept hierarchy and including those with a matching skos:broader predicate. Reported metrics are Hits@1, Hits@10, MR, and MRR using the proposed negative sampling method.

Method	<i>Skos Constraint 1</i>				<i>Skos Constraint 2</i>			
	Hits@1	Hits@10	MR	MRR	Hits@1	Hits@10	MR	MRR
TransE	0.13	0.46	15.86	0.24	0.05	0.10	30.6	0.10
DistMult	0.20	0.43	24.36	0.04	0.25	0.45	23.5	0.04
GNN	<b>0.93</b>	<u>0.96</u>	<u>2.38</u>	0.42	<b>0.87</b>	<u>0.92</u>	<b>4.17</b>	<u>0.24</u>
QuatE	<u>0.85</u>	<b>1.00</b>	<b>1.22</b>	<b>0.92</b>	<u>0.83</u>	<b>1.00</b>	<u>1.26</u>	<b>0.91</b>

## 6. Conclusion

This study assesses the impact of SKOS constraints, informally defined, on the predictive performance of KGE models. Our experiments reveal that incorporating logically inferred negative samples during model training enhances learning by leveraging logical formulations derived from SKOS website’s text definitions. By injecting two types of negative samples into QuatE embedding model, the proposed method achieves Hits@1 of 0.85, Hits@10 of 1.00, MR of 1.22, and MRR of 0.92 in one scenario, and Hits@1 of 0.83, Hits@10 of 1.00, MR of 1.26, and MRR of 0.91 in the other scenario depending on the selected SKOS constraints, which we have verified against the well-known UMLS medical ontology. We plan to extend this research as a future work by covering the remaining concepts and their sub-classes in the SemOpenAlex concept hierarchy. We will also explore the incorporation of ontological axioms into the KGE model learning process.

## References

- [1] W. of Science, Web of science <https://www.webofscience.com/>, 2023. URL: <https://www.webofscience.com/>.
- [2] P. A. Bonatti, S. Decker, A. Polleres, V. Presutti, Knowledge graphs: New directions for knowledge representation on the semantic web (dagstuhl seminar 18371), in: Dagstuhl reports, volume 8, Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2019.
- [3] M. Dalle Lucca Tosi, J. C. dos Reis, Understanding the evolution of a scientific field by clustering and visualizing knowledge graphs, *Journal of Information Science* 48 (2022) 71–89.
- [4] SemOpenAlex, Semopenalex ontology <https://semopenalex.org/>, 2023. URL: <https://semopenalex.org/>.
- [5] C. Caracciolo, A. Stellato, A. Morshed, G. Johannsen, S. Rajbhandari, Y. Jaques, J. Keizer, The agrovoc linked dataset, *Semantic Web* 4 (2013) 341–348.
- [6] T. Baker, B. Whitehead, R. Musker, J. Keizer, Global agricultural concept space: lightweight semantics for pragmatic interoperability, *npj Science of Food* 3 (2019) 16.
- [7] F. Musa Aliyu, A. Ojo, Towards building a knowledge graph with open data—a roadmap, in: *e-Infrastructure and e-Services for Developing Countries: 9th International Conference, AFRICOMM 2017, Lagos, Nigeria, December 11-12, 2017, Proceedings* 9, Springer, 2018, pp. 157–162.
- [8] N. Jain, T.-K. Tran, M. H. Gad-Elrab, D. Stepanova, Improving knowledge graph embeddings with ontological reasoning, in: *International Semantic Web Conference*, Springer, 2021, pp. 410–426.

- [9] M. M. Alam, H. Jabeen, M. Ali, K. Mohiuddin, J. Lehmann, Affinity dependent negative sampling for knowledge graph embeddings., in: DL4KG@ ESWC, 2020.
- [10] N. Yao, Q. Liu, X. Li, Y. Yang, Q. Bai, Entity similarity-based negative sampling for knowledge graph embedding, in: Pacific Rim International Conference on Artificial Intelligence, Springer, 2022, pp. 73–87.
- [11] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating embeddings for modeling multi-relational data, *Advances in neural information processing systems* 26 (2013).
- [12] S. Mehryar, R. Celebi, Improving transitive embeddings in neural reasoning tasks via knowledge-based policy networks, in: *CEUR Workshop Proceedings*, volume 3337 of *CEUR Workshop Proceedings*, 2022, pp. 16–27.
- [13] B. Yang, W.-t. Yih, X. He, J. Gao, L. Deng, Embedding entities and relations for learning and inference in knowledge bases, *arXiv preprint arXiv:1412.6575* (2014).
- [14] S. Zhang, Y. Tay, L. Yao, Q. Liu, Quaternion knowledge graph embeddings, *Advances in neural information processing systems* 32 (2019).
- [15] M. Nickel, V. Tresp, H.-P. Kriegel, et al., A three-way model for collective learning on multi-relational data., in: *Icml*, volume 11, 2011, pp. 3104482–3104584.
- [16] A. Isaac, E. Summers, Skos simple knowledge organization system primer, Working Group Note, W3C (2009).
- [17] S. Choudhary, T. Luthra, A. Mittal, R. Singh, A survey of knowledge graph embedding and their applications, *arXiv preprint arXiv:2107.07842* (2021).
- [18] Z. Ye, Y. J. Kumar, G. O. Sing, F. Song, J. Wang, A comprehensive survey of graph neural networks for knowledge graphs, *IEEE Access* 10 (2022) 75729–75741.
- [19] Y. Wang, B. Hu, S. Yang, M. Zhu, Z. Zhang, Q. Zhang, J. Zhou, G. Ye, H. He, Not All Negatives Are Worth Attending to: Meta-Bootstrapping Negative Sampling Framework for Link Prediction, 2023. URL: <http://arxiv.org/abs/2312.04815>.
- [20] Y. Kalantidis, M. B. Sariyildiz, N. Pion, P. Weinzaepfel, D. Larlus, Hard Negative Mixing for Contrastive Learning, in: *Advances in Neural Information Processing Systems*, volume 33, Curran Associates, Inc., 2020, pp. 21798–21809. URL: <https://proceedings.neurips.cc/paper/2020/hash/f7cade80b7cc92b991cf4d2806d6bd78-Abstract.html>.
- [21] S. Mehryar, R. Celebi, Semantic Annotation of Tabular Data for Machine-to-Machine Interoperability via Neuro-Symbolic Anchoring, in: *CEUR Workshop Proceedings*, volume 3557, Rheinisch-Westfaelische Technische Hochschule Aachen\* Lehrstuhl Informatik V, 2023, pp. 61–71. URL: <https://ceur-ws.org/Vol-3557/paper5.pdf>.
- [22] J. Loesch, M. Dumontier, R. Celebi, Enhanced GAT: Expanding Receptive Field with Meta Path-Guided RDF Rules for Two-Hop Connectivity (2023). URL: <https://ceur-ws.org/Vol-3592/paper8.pdf>.