SPHOTA-NEXT: Ontology Enhanced Knowledge Injection for Biochemical Link Prediction

Bharath Chand[1,*], Thara Prabhakaran[1] and Sanju Tiwari[2]

[1]Department of Futures Studies, University of Kerala, Kerala, India
[2]Sharda University, Delhi-NCR, India

Abstract

Link prediction in biochemical knowledge graphs requires enriched contextual information to improve entity disambiguation and embedding quality. This work extends the Knowledge Graph Structure Prediction framework by integrating chemical ontology data from RDKit to augment domain-specific knowledge representations. The proposed approach enhances the K-BERT knowledge injection methodology by incorporating compound-level semantic information, mapping RDKit chemical properties and classifications as structured triples aligned with entity names for effective knowledge integration. Applied to the NuBBE DB dataset used in the Biochemical Knowledge Extraction (BiKE) task, the method combines compound name-based entity linking with enriched chemical property triples. Empirical evaluation using the NatUKE benchmark demonstrates that the augmented knowledge graph improves link prediction accuracy across multiple co-related properties. The study validates that domain-specific ontology enhancement strengthens K-BERT's ability to capture the semantics, advancing biochemical knowledge extraction and establishing a foundation for Ontology enhanced knowledge graph embedding strategies.

Keywords

Knowledge Injection, K-BERT, Knowledge Graph Embedding, Ontology

1. Introduction

Link prediction in biochemical knowledge graphs is challenging because scientific knowledge is often represented as sparse heterogeneous networks containing entities such as compounds, species, bioactivities, collection sites, and isolation methods. The NuBBE DB [1] data used in the BiKE challenge can be represented as a paper centric knowledge graph in which each publication node is connected to extracted biochemical properties. Such a representation enables knowledge extraction and graph completion tasks. However, it suffers from limited semantic connectivity between related entities.

Our previous SPHOTA framework [2] addressed this challenge by combining K-BERT [3] based knowledge injection with the EPHEN [4] embedding propagation framework. The approach enriched node representations using textual information extracted from scientific articles and propagated semantic information throughout the graph using regularization based embedding method. Although this hybrid approach achieved strong performance across several prediction tasks, the underlying knowledge graph remained structurally sparse. In particular, compounds rarely share direct connections with chemically related compounds, species lack taxonomic relationships, and bioactivities are represented without higher-level semantic organization.

To address this limitation, we propose SPHOTA-NEXT, an ontology-enhanced extension of the original framework. The central idea is to enrich graph entities with hierarchical knowledge derived from external domain resources. Compound entities are aligned with chemical classifications derived using RDKit [5], enabling the generation of ontology triples that capture chemical categories and hierarchical relationships. These additional triples are injected into K-BERT [3] at the entity level,

BiKE'26: Third International Biochemical Knowledge Extraction Challenge, May 10–May 14, 2026, co-located with the 23rd European Semantic Web Conference (ESWC), Dubrovnik, Croatia
*Corresponding author.
bharathchand@keralauniversity.ac.in (B. Chand); tharap@keralauniversity.ac.in (T. Prabhakaran); sanju.tiwari@sharda.ac.in (S. Tiwari)

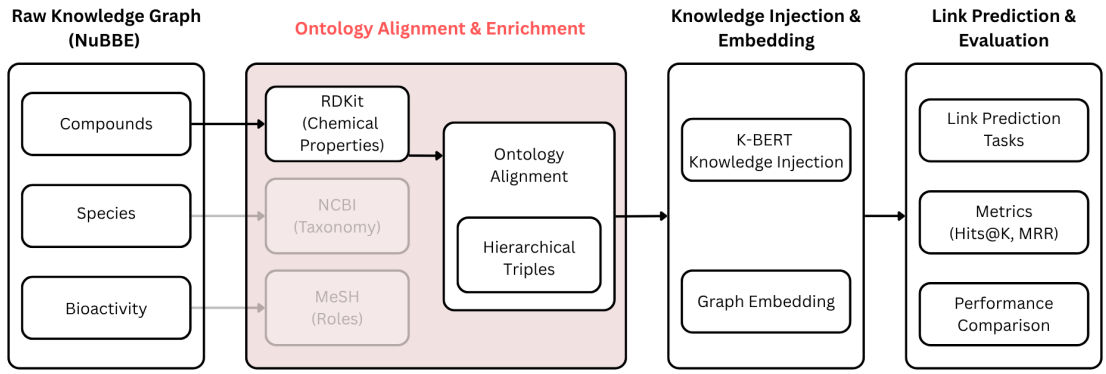

Figure 1: Architecture of the proposed SPHOTA-NEXT framework. The pipeline enriches the NuBBE DB knowledge graph with ontology-derived chemical triples, injects the enriched knowledge into K-BERT-based textual representations, and applies graph embedding propagation before cosine-similarity-based link prediction.

allowing the language model to exploit both textual context and ontology structure during embedding generation.

The motivation behind this enhancement is that while individual compounds may occur infrequently within the corpus, their broader chemical classes appear repeatedly across documents. Consequently, ontology hierarchies create shared semantic neighborhoods that enable better generalization during embedding, learning and link prediction. Experimental evaluation on NuBBE DB [1] with the NatUKE benchmark [6] demonstrates that ontology enhanced knowledge injection improves prediction performance for several biochemical extraction tasks, particularly bioactivity and isolation type prediction, highlighting the value of integrating domain ontologies into language-model driven knowledge graph completion.

2. Approach

The proposed SPHOTA-NEXT framework extends the original SPHOTA [2] architecture by introducing an ontology enrichment layer between the raw knowledge graph and the knowledge injection stage. Figure 1 presents the overall workflow, showing how biochemical entities are aligned with external chemical knowledge, transformed into ontology triples, injected into K-BERT [3], and propagated through the graph for link prediction.

2.1. Dataset

The experiments were conducted using NuBBE DB [1] as the underlying biochemical dataset, with evaluation following the NatUKE benchmark [6]. The knowledge graph contains publication nodes identified by DOI and associated property nodes including compound names, bioactivities, species, collection sites, and isolation types. Link prediction is evaluated by removing selected edges from the graph and measuring the ability of proposed method to recover the missing relationships through retrieval.

2.2. Motivation for Ontology Enhancement

Analysis of the existing benchmark setting reveals several structural limitations. The graph is largely paper-centric, with connections primarily occurring between publications and extracted properties. There are no explicit links between chemically related compounds, taxonomically related species, or semantically related bioactivities. Furthermore, approaches such as DeepWalk, Node2Vec, and Metapath2Vec rely heavily on graph topology and therefore have limited access to domain knowledge. Although EPHEN incorporates textual embeddings, any noise in the initial embeddings may propagate throughout the network.

To overcome these limitations, we enrich the graph with ontology-derived hierarchical relationships. The objective is to create additional semantic connections between entities that share higher-level conceptual categories.

2.3. Ontology Alignment and Enrichment

Compound entities are aligned with external chemical knowledge resources using RDKit [5] derived classifications. For each compound, ontology triples describing chemical categories and hierarchical relationships are generated. These triples are attached directly to the corresponding compound entities.

Unlike simple attribute enrichment, the proposed approach focuses on introducing hierarchy and class-level knowledge. For example, multiple compounds belonging to the same chemical family become connected through shared ontology concepts. This creates common semantic neighborhoods that are absent in the original graph and enables information sharing between otherwise disconnected entities.

The architecture is designed to support additional ontology resources such as NCBI Taxonomy for species and MeSH-based semantic categories for biological activities, thereby providing a general framework for ontology-driven graph enrichment.

2.4. Knowledge Injection and Embedding Learning

The enriched ontology triples are incorporated into textual representations using K-BERT [3]. During preprocessing, entity mentions are matched to ontology concepts and relevant triples are injected into the sentence structure through the K-BERT [3] sentence-tree mechanism. The visible matrix ensures that injected knowledge influences only the corresponding entity representation while preserving the original textual context.

The resulting ontology-aware embeddings are subsequently propagated through the graph using the EPHEN [4] regularization framework. This process integrates textual semantics, ontology knowledge, and graph structure into a unified embedding space.

2.5. Link Prediction

Following the embedding generation, link prediction is performed using a K-nearest-neighbor retrieval strategy based on cosine similarity. Candidate entities are ranked according to embedding similarity, and evaluation is using Hits@K and Mean Reciprocal Rank metrics.

3. Evaluation

The evaluation follows the official BiKE challenge protocol defined by the NatUKE benchmark [6]. In this setting, the knowledge graph is intentionally made incomplete during testing by removing known links between selected publication nodes and their ground-truth property nodes. The task is then to recover these hidden links from the remaining graph, textual context, and entity representations. This protocol reflects the practical knowledge extraction scenario in which a biochemical publication may be only partially annotated and the system must infer missing information such as compound names, bioactivities, species, collection sites, or isolation types.

For each test case, the embedding of the query node is compared with candidate nodes of the corresponding target property type. Candidates are ranked using cosine similarity, and the correct answer is considered successfully recovered when it appears among the top ranked predictions. The evaluation therefore measures knowledge graph completion ability rather than only local classification performance. We report `hits@k`, where k follows the official NatUKE setting for each property: 50 for compound name, 5 for bioactivity, 50 for species, 20 for collection site, and 1 for isolation type. These different values of k account for the varying candidate-space sizes and difficulty levels of the five extraction subtasks.

Table 1 compares the proposed ontology-enhanced method with the benchmark baselines and related strategies. DeepWalk and Node2Vec learn node representations from random walk neighborhoods, while Metapath2Vec extends this idea to heterogeneous networks by following type-aware paths. EPHEN incorporates contextual text embeddings and propagates them through graph regularization. The proposed method builds on this direction by combining K-BERT-style knowledge injection [3, 7, 8] with ontology-derived chemical triples before embedding-based retrieval.

The results show that ontology enhancement provides clear gains for several extraction tasks. The strongest improvements are observed for bioactivity, collection site and isolation type prediction, where the proposed method achieves the best scores. Collection site prediction indicates that enriched biochemical context can support the recovery of properties through propagated graph semantics. Species prediction is competitive, where Metapath2Vec remains stronger. Compound-name prediction remains difficult because exact compound recovery involves a large and sparse candidate set; nevertheless, the substantial gains observed across bioactivity, collection site and isolation type demonstrate the effectiveness of the proposed method.

Overall, the evaluation suggests that ontology-guided enrichment strengthens link prediction by providing semantic structure. The additional triples do not merely add attributes to isolated compounds; they connect rare entities through shared chemical categories, improving the embedding space used for nearest-neighbor retrieval. The open repository containing instructions for verifying the claimed results is available at https://github.com/bharathchand10/BiKE-SPHOTA-NEXT.

Table 1
Results table for extracting: compound name (C), bioactivity (B), specie (S), collection site (L), and isolation type (T) in NaTUKE's third evaluation stage. The best results for each extraction are bold.

Model	Property				
	C ($k = 50$)	**B** ($k = 5$)	**S** ($k = 50$)	**L** ($k = 20$)	**T** ($k = 1$)
DeepWalk	0.00	0.10	0.27	0.38	0.14
Node2Vec	0.00	0.03	0.25	0.36	0.05
Metapath2vec	**0.09**	0.13	**0.42**	0.42	0.19
EPHEN	0.03	0.60	0.29	0.55	0.75
Ontology Enhancement	0.00	**0.79**	0.35	**0.79**	**0.96**

4. Related Work

Several studies have sought to advance the automatic extraction of biochemical knowledge from scientific literature, particularly within the framework of the NatUKE benchmark. One notable approach, 'Leveraging ChatGPT API for Enhanced Data Preprocessing in NatUKE' [9], incorporates OpenAI's ChatGPT into the preprocessing workflow to derive structured metadata from research articles.

Another study 'Improving Natural Product Automatic Extraction With Named Entity Recognition' [10], augments the conventional NatUKE pipeline through the integration of Named Entity Recognition (NER) techniques. Rather than relying on simple token-based segmentation, the method identifies and extracts complete sentences containing relevant entities.

The work 'Enhancing Biochemical Extraction with BFS-driven Knowledge Graph Embedding ap-

proach' [11] presents a hybrid framework that combines graph traversal with representation learning. In this approach, Breadth-First Search (BFS) is employed to generate paths through a knowledge graph, which are subsequently treated as textual sequences for training Word2Vec embeddings.

The work EPHEN [4] serves as the foundational framework for the modifications introduced in our downstream implementation. EPHEN is specifically designed to unify semantic information derived from textual content with the structural characteristics of heterogeneous networks. The framework initially generates vector representations for text-associated nodes using a pretrained language model such as SBERT. These semantic embeddings are subsequently propagated to nodes lacking textual descriptions through a graph-regularization mechanism. Extending these developments, our previous work, SPHOTA[2], explored the integration of K-BERT [3] instead of SBERT, and it demonstrated the potential of this approach for supporting more accurate biochemical knowledge discovery. In this present work, SPHOTA is enhanced with Ontology details.

5. Conclusion & Future Works

The experimental results demonstrate that ontology guided enrichment improves the effectiveness of knowledge injection for biochemical link prediction. By introducing hierarchical chemical classification triples, the proposed SPHOTA-NEXT framework creates shared semantic neighborhoods that are absent in the original Knowledge Graph. The improvement is particularly evident for bioactivity, collection site and isolation type prediction, indicating that ontology structure provides useful contextual signals beyond those available from textual content alone. Future work will focus on expanding ontology coverage through NCBI Taxonomy and MeSH integration, exploring multi-ontology alignment strategies, and investigating ontology-aware embedding objectives capable of explicitly exploiting hierarchical relationships during representation learning.

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT and Claude in order to: Grammar and spelling check, Paraphrase and reword. The author(s) have not employed any Generative AI tools to Generate images. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] A. C. Pilon, M. Valli, A. C. Dametto, M. E. F. Pinto, R. T. Freire, I. Castro-Gamboa, A. D. Andricopulo, V. S. Bolzani, Nubbedb: an updated database to uncover chemical and biological information from brazilian biodiversity, Scientific Reports 7 (2017) 7215.

[2] B. Chand, S. Tiwari, Sphota: Knowledge graph structure prediction with a hybrid orientation of textual alignment using k-bert, Proceedings of the Second Biochemical Knowledge Extraction (BiKE) challenge co-located with Text2KG at ESWC (2025).

[3] W. Liu, P. Zhou, Z. Zhao, Z. Wang, H. Deng, Q. Ju, P. Wang, K-bert: Enabling language representation with knowledge graph, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 2901–2908.

[4] P. do Carmo, R. Marcacini, Embedding propagation over heterogeneous event networks for link prediction, in: 2021 IEEE International Conference on Big Data (Big Data), IEEE, 2021, pp. 4812–4821.

[5] G. Landrum, et al., Rdkit documentation, Release 1 (2013) 4.

[6] P. V. Do Carmo, E. Marx, R. Marcacini, M. Valli, J. V. S. e Silva, A. Pilon, Natuke: A benchmark for natural product knowledge extraction from academic literature, in: 2023 IEEE 17th International Conference on Semantic Computing (ICSC), IEEE, 2023, pp. 199–203.

[7] A. Cadeddu, A. Chessa, V. De Leo, G. Fenu, E. Motta, F. Osborne, D. R. Recupero, A. Salatino, L. Secchi, A comparative analysis of knowledge injection strategies for large language models in the scholarly domain, Engineering Applications of Artificial Intelligence 133 (2024) 108166.
[8] W. Xu, M. Fang, L. Yang, H. Jiang, G. Liang, C. Zuo, Enabling language representation with knowledge graph and structured semantic information, in: 2021 International Conference on Computer Communication and Artificial Intelligence (CCAI), IEEE, 2021, pp. 91–96.
[9] P. Fröhlich, J. Gwozdz, M. Jooß, Leveraging chatgpt api for enhanced data preprocessing in natuke., in: TEXT2KG/BiKE@ ESWC, 2023, pp. 244–255.
[10] S. Schmidt-Dichte, I. J. Mócsy, Improving natural product automatic extraction with named entity recognition., in: TEXT2KG/BiKE@ ESWC, 2023, pp. 226–234.
[11] B. Zope, S. Mishra, S. Tiwari, Enhancing biochemical extraction with bfs-driven knowledge graph embedding approach., in: TEXT2KG/BiKE@ ESWC, 2023, pp. 235–243.