Bio-KGvec2go: Serving up-to-date Dynamic Biomedical Knowledge Graph Embeddings[*]

Hamid Ahmad[1], Heiko Paulheim[2] and Rita T. Sousa[2,*]

[1]*University of Mannheim, Mannheim, Germany*

[2]*Data and Web Science Group, University of Mannheim, Mannheim, Germany*

Abstract

Knowledge graphs and ontologies represent entities and their relationships in a structured way, having gained significance in the development of modern AI applications. Integrating these semantic resources with machine learning models often relies on knowledge graph embedding models to transform graph data into numerical representations. Therefore, pre-trained models for popular knowledge graphs and ontologies are increasingly valuable, as they spare the need to retrain models for different tasks using the same data, thereby helping to democratize AI development and enabling sustainable computing.

In this paper, we present Bio-KGvec2go, an extension of the KGvec2go Web API, designed to generate and serve knowledge graph embeddings for widely used biomedical ontologies. Given the dynamic nature of these ontologies, Bio-KGvec2go also supports regular updates aligned with ontology version releases. By offering up-to-date embeddings with minimal computational effort required from users, Bio-KGvec2go facilitates efficient and timely biomedical research.

Keywords

Knowledge Graph Embeddings, Biomedical Ontologies, Gene Ontology, Human Phenotype Ontology

1. Introduction

Knowledge Graphs (KGs) contain factual knowledge about real-world entities and their relations in a fully machine-readable format [1]. Many modern KGs represent this information according to a formal definition of the domain knowledge given by an ontology. Ontologies are semantic models for a domain in which each entity is precisely defined, and the relationships between entities are parameterized or constrained [2]. In life sciences, the use of ontologies has gained prominence, with increasing importance in biomedical research [3]. Ontologies are applied across various areas of biology and medicine, ranging from gene function [4] to drug characterization [5]. Phenotype ontologies are also available for multiple species for the characterization of diseases [6]. Open repositories, such as BioPortal [7], provide access to hundreds of biomedical ontologies.

Given the richness of these semantic resources, they have been exploited in a wide variety of machine learning (ML) tasks, including entity classification, link prediction, graph classification, and relation prediction, among others. One of the challenges faced by approaches that combine artificial intelligence (AI) with KGs and ontologies is transforming graph data into a suitable representation that can be processed by ML algorithms. A current major trend is the use of KG embedding (KGE) methods, which transform entities and relationships in a KG into a lower-dimensional vector space while attempting to preserve the graph structure and, in some cases, semantic information [8]. KGEs are then fed as features for ML algorithms to support several applications, with particular success in the life sciences. From finding new treatments for existing drugs to diagnosing patients and identifying associations between diseases and genes, KGEs have been employed in a wide range of biomedical applications [9].

Therefore, pre-trained embeddings for popular biomedical ontologies are increasingly valuable, sparing the need to retrain the models for different tasks using the same data, and allowing greener

computing and sustainable AI development. However, such pre-trained models are not always readily available for the biomedical domain, and when available, they typically reflect a static snapshot of the KG at a specific point in time. This poses a challenge, given that knowledge is constantly evolving. Discoveries are published daily, rendering some facts obsolete and revealing new knowledge. As a result, the development of KGs and ontologies is dynamic [10]. Relying on embeddings generated on outdated versions risks overlooking critical insights and recent advances, limiting downstream performance.

This paper addresses the challenge of providing up-to-date embeddings for the two most successful biomedical ontologies - Gene Ontology [4] and Human Phenotype Ontology [6]. We present a framework capable of periodically collecting new KG versions, computing embeddings, and making them publicly available to support downstream research. These embeddings are provided via the Bio-KGvec2go platform, www.bio.kgvec2go.org, which is built upon KGVec2go [11]. KGvec2go provides a Web API that enables access to embeddings, computes entity similarity, and identifies related concepts based on input embeddings. While KGvec2go makes available RDF2Vec embeddings [12] for four general-purpose KGs (ALOD, DBpedia, WordNet, and Caligraph), it does not support biomedical KGs or reflect KG evolution. Bio-KGvec2go expands the range of KGs to encompass biomedical ontologies and other KGE models, but also recomputes the embeddings when new ontology versions are released.

By publicly providing regularly updated and accessible KGEs, we aim to facilitate ongoing research and democratize access to these resources. Even researchers without computational power to train KGE models can conduct analyses and investigations with the latest data representations. Furthermore, it facilitates the study of knowledge evolution, allowing researchers to explore how changes across KG versions impact the resulting embeddings and reflect shifts in domain knowledge over time.

The remainder of this paper is structured as follows: we first describe the two biomedical ontologies explored, followed by an overview of the KGE models employed. Finally, we present the implementation details and functionalities of the Bio-KGvec2go platform.

2. Biomedical Ontologies

Currently, Bio-KGvec2go focuses on providing embeddings for two widely used biomedical ontologies.

Gene Ontology (GO) GO [4] defines a hierarchy of more than 40 000 classes that describe protein functions and their relationships. It can be represented as a graph where nodes are GO classes and edges define relationships between them (e.g., *is_a*, *part_of*, *regulates*), with the majority of *is_a* relations. Functions in GO are described across three domains: the biological processes, the molecular functions, and the cellular components. GO was initially proposed in 1998 by a consortium of researchers. Since then, it has been constantly reviewed, including the addition or deprecation of terms and reorganization of the relationship structure. Most revisions result from advances in biological knowledge or improvements in the precision of experimental technologies. Official GO versions are released monthly[1]. GO embeddings have been widely used in multiple applications, such as protein function prediction [13], protein interactions prediction [14, 15], and gene-disease associations discovery [16].

Human Phenotype Ontology (HP) HP [6] characterizes the phenotypic abnormalities in human hereditary diseases, covering key aspects such as the phenotypic abnormalities themselves, past medical history, mode of inheritance, clinical course, clinical modifiers, and frequency. The HP contains more than 18 000 classes represented in a directed acyclic graph, where each node represents a distinct phenotype, and all relationships are of the type *is_a*, establishing a hierarchy. HP was initially developed in 2008 at the Charité University Hospital in Berlin, and it has been continuously updated through a combination of expert curation, integration of new findings from the biomedical literature, and feedback from the global community of clinicians and researchers who use it. While HP does not follow a formal monthly release model like GO, new official versions are made available regularly (approximately every

---

[1]https://release.geneontology.org/

month to two months) through its GitHub repository[2]. HP embeddings have also been employed in critical biomedical tasks, including patient similarity computation [17], genotype-phenotype association prediction [18], and gene-disease association prediction [19].

3. Knowledge Graph Embedding Models

KGE methods map each node to a lower-dimensional space where the underlying KG structure and other semantic information are preserved as much as possible. Numerous KGE models have been proposed in the literature, contributing to a substantial body of work as highlighted in different surveys [20].

The KGE methods can be broadly categorized based on their underlying mechanisms for capturing graph structure and semantic information: translational distance models interpret relations as vector translations, semantic matching models focus on similarity scoring, geometric models exploit spatial structures to encode logical constraints, and random walks-based models use paths through the graph to capture long distance relationships. While translational distance models and semantic matching models focus on exploring the KG triples solely, random walks-based and geometric models also include additional information, namely the hierarchical information. To encompass a diverse range of knowledge representation approaches, this work employs six representative KGE models spanning the different categories:

- **TransE** [21] is the most representative translational distance model, treating relations as vector translations between entities. However, TransE struggles to handle one-to-many and many-to-many relations. Tackling this, **TransR** [22] introduces a space for each relation.
- Semantic matching approaches exploit similarity between entities and relations. **DistMult** [23] achieves this by employing a bilinear scoring function with diagonal relation matrices. **HolE** [24] uses circular correlation to create compositional representations while remaining scalable.
- **RDF2Vec** [12] is a random walk-based approach built upon two main steps: (i) producing random walks in the graph that are akin to a corpus of sentences; (2) using those sequences as input to a neural language model that learns a latent low-dimensional representation.
- **BoxE** [25] is a geometric approach that represents entities as points, and relations as a set of hyper-rectangles (boxes), capturing logical patterns such as hierarchy, symmetry, or intersection.

Regarding implementation, the PyKEEN package[3] is used to train TransE, TransR, DistMult, HolE, and BoxE, while pyRDF2Vec[4] is employed for RDF2Vec. To ensure a fair comparison, all models are trained with default hyperparameters, except for the number of epochs, set to 100, and the embedding dimension, set to 200.

4. Bio-KGvec2go

This paper presents a framework that collects new versions of biomedical ontologies, computes embeddings, and makes them publicly available through Bio-KGvec2go, www.bio.kgvec2go.org. The framework is designed to support an automated update mechanism that periodically downloads ontology releases from predefined URLs, computes checksums, and compares them with those of previously stored versions. If a change is detected, all embeddings are recomputed and made available.

Regarding the platform Bio-KGvec2go, it is built upon the existing Web API, KGvec2go [11], to provide access to up-to-date embedding models trained on biomedical ontologies. KGvec2go is implemented in Python using Flask and can be deployed with the Apache HTTP Server. The KGvec2go API, www.kgvec2go.org, offers a RESTful service designed to operate efficiently on Internet-connected devices with limited CPU and RAM (e.g., smartphones). Currently, Bio-KGvec2go offers three functionalities,

---

[2]https://github.com/obophenotype/human-phenotype-ontology/releases
[3]https://pykeen.readthedocs.io/en/stable/
[4]https://pyrdf2vec.readthedocs.io/en/latest/

as shown in Figure 1: (i) downloading the embeddings for the various ontology versions, (ii) computing semantic similarity between two classes, and (iii) retrieving the top 10 most similar classes for any given ontology class.

All the code is available on GitHub[5]. Additionally, the trained KGE models are available on Zenodo[6], ensuring long-term preservation. The models are accompanied by metadata using the PROV standard [26], describing the input ontology, the KGE model used, and the corresponding hyperparameters.

Download   Users can select an embedding model and download a corresponding JSON file containing the vector representations for each ontology class, encoded as 200-dimensional floating-point arrays. As of now, Bio-KGvec2go hosts embeddings for six distinct versions of each biomedical ontology, with the first version dating back to 2023 and subsequent versions released approximately every six months. Besides the downstream use of the embeddings, this functionality enables researchers to compare embeddings across different ontology snapshots, supporting studies on ontology evolution.

Similarity   Users can access the semantic similarity between two ontology classes by selecting an embedding model and providing either class identifiers or textual labels (with automatic normalization of case and whitespace). Bio-KGvec2go retrieves the corresponding vectors of the two classes from the most up-to-date version and computes the cosine similarity. The resulting score, ranging from -1 to 1, indicates the degree of similarity, where 1 denotes perfect similarity and -1 reflects complete dissimilarity. This functionality is particularly useful for ontology curation and annotation.

Top Closest Concepts   Users can select an embedding model and specify the target class by the identifier or label (with automatic normalization) to obtain the top 10 most semantically similar ontology classes. Bio-KGvec2go retrieves the most up-to-date version of embeddings and computes cosine similarities between the input vector and all other class vectors, returning a ranked list. The output is presented as a detailed table listing each related class by its identifier and label, the similarity score, and a direct URL for further exploration. This functionality is well-suited for semantic search and identification of candidates for enrichment analyses.
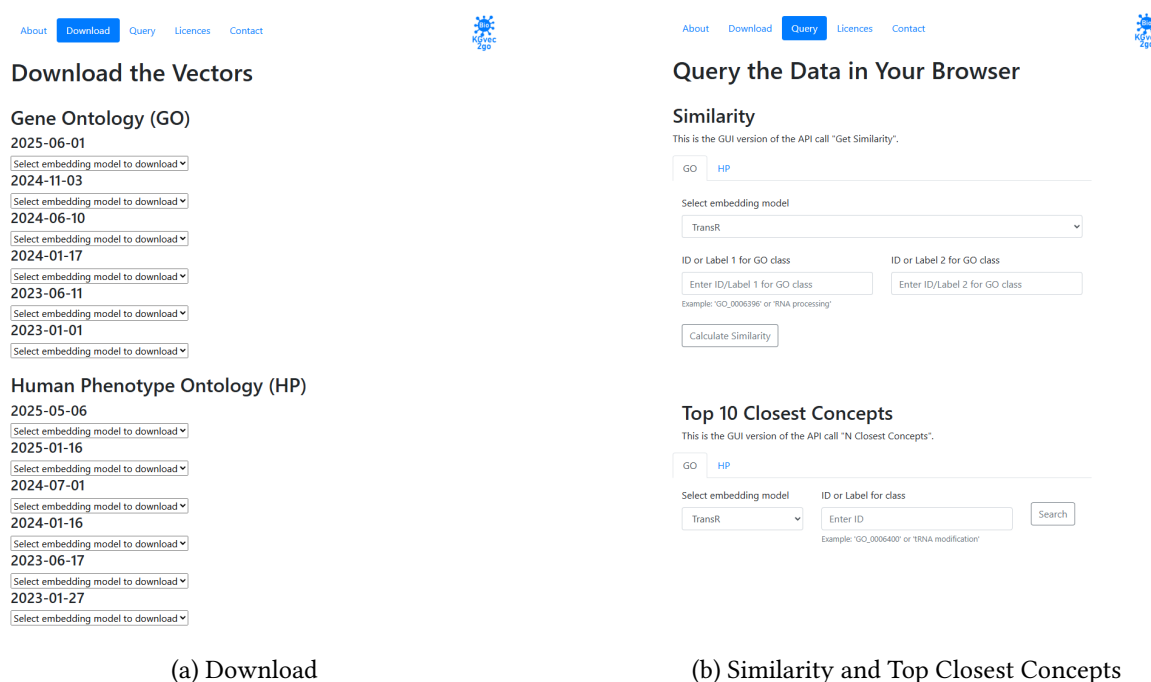

(a) Download

(b) Similarity and Top Closest Concepts

Figure 1: Functionalities of Bio-KGvec2go.

---

[5]https://github.com/ritatsousa/biokgvec2go
[6]https://zenodo.org/records/15865665

5. Use Cases

Bio-KGvec2go has been designed as a user-friendly platform, and it can support a broad spectrum of biomedical research. One key use case is in ontology-based ML approaches, where the embeddings can serve as input features for predictive models across diverse biomedical tasks. These approaches have become increasingly popular, as they allow the integration of structured biological knowledge. For example, GO embeddings have been used to predict the function of uncharacterized proteins or to identify which proteins are likely to interact with each other. These interactions are crucial for many functions in biology and are highly relevant to disease states. Similarly, HP embeddings have been employed to uncover new associations between genes and diseases, improving the understanding of disease mechanisms. HP embeddings have also been used to improve disease diagnosis by comparing a patient's phenotypic profile to known disease profiles represented in the embedding space. Since discovering protein interactions or gene-disease associations through laboratory experiments is expensive and time-consuming, ML-based approaches help to generate candidate pairs, narrowing the search space for lab validation and substantially reducing both the time and cost of experimental research.

Another important application lies in ontology development, curation, and semantic annotation. Both GO and HP are manually curated by domain experts, many of whom have limited computational experience and therefore benefit from accessible tools. The *similarity* and *top closest concepts* functionalities provided by Bio-KGvec2go are particularly useful for these tasks. For instance, when annotating a gene with a specific function, researchers can use the tool to find the most semantically similar GO terms, ensuring more accurate and consistent annotations. Beyond annotation, this tool can help identify gaps or inconsistencies in the ontology itself.

6. Conclusion

This paper presents Bio-KGvec2go, a FAIR resource designed to provide up-to-date pre-trained biomedical embeddings. Bio-KGvec2go extends the original KGvec2go API by incorporating multiple KGE models beyond RDF2Vec and by supporting several versions of the same biomedical KGs. By democratizing access to these embeddings, Bio-KGvec2go enables researchers to accelerate experimentation and improve performance across various tasks, including disease prediction, identification of gene-disease associations, and drug discovery. Moreover, by facilitating the reuse of pre-trained embeddings, it contributes to reducing the carbon footprint.

As future work, we plan to expand Bio-KGvec2go to support additional biomedical KGs and embedding models. We also aim to improve the similarity and top closest concepts search functionalities by introducing features such as autocomplete for concept labels and tolerance to minor typos, ensuring that users can retrieve relevant concepts even if the input is not an exact match.

Acknowledgments

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly for grammar checks, paraphrasing, and rewording. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. e. a. Kirrane, Knowledge graphs, ACM Computing Surveys 54 (2021) 1–37.

[2] S. Staab, R. Studer, Handbook on Ontologies, International Handbooks on Information Systems, Springer, 2010.

[3] R. Hoehndorf, M. Dumontier, G. V. Gkoutos, Evaluation of research in biomedical ontologies, Briefings in Bioinformatics 14 (2013) 696–712.

[4] S. A. Aleksander, J. Balhoff, S. Carbon, J. M. Cherry, H. J. Drabkin, D. Ebert, M. Feuermann, et al., The Gene Ontology Knowledgebase in 2023, Genetics 224 (2023) iyad031.

[5] K. Degtyarenko, P. De Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, M. Ashburner, ChEBI: a database and ontology for chemical entities of biological interest, Nucleic Acids Research 36 (2007) D344–D350.

[6] S. Köhler, M. Gargano, N. Matentzoglu, L. C. Carmody, D. Lewis-Smith, N. A. Vasilevsky, D. Danis, G. Balagura, G. Baynam, A. M. Brower, et al., The Human Phenotype Ontology in 2021, Nucleic Acids Research 49 (2021) D1207–D1217.

[7] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, M. A. Musen, BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications, Nucleic Acids Research 39 (2011) W541–W545.

[8] Q. Wang, Z. Mao, B. Wang, L. Guo, Knowledge graph embedding: A survey of approaches and applications, IEEE Transactions on Knowledge and Data Engineering 29 (2017) 2724–2743.

[9] S. K. Mohamed, A. Nounu, V. Nováček, Biological applications of knowledge graph embedding models, Briefings in Bioinformatics 22 (2021) 1679–1693.

[10] G. Flouris, D. Manakanatas, H. Kondylakis, D. Plexousakis, G. Antoniou, Ontology change: classification and survey, The Knowledge Engineering Review 23 (2008) 117–152.

[11] J. Portisch, M. Hladik, H. Paulheim, KGvec2go–Knowledge Graph Embeddings as a Service, in: Language Resources and Evaluation Conference, 2020, pp. 5641–5647.

[12] P. Ristoski, H. Paulheim, RDF2Vec: RDF graph embeddings for data mining, in: International Semantic Web Conference, 2016, pp. 498–514.

[13] X. Zhong, J. C. Rajapakse, Graph embeddings on gene ontology annotations for protein–protein interaction prediction, BMC Bioinformatics 21 (2020) 560.

[14] K.-H. Chen, T.-F. Wang, Y.-J. Hu, Protein-protein interaction prediction using a hybrid feature representation and a stacked generalization scheme, BMC Bioinformatics 20 (2019) 308.

[15] Ieremie, Ioan and Ewing, Rob M and Niranjan, Mahesan, TransformerGO: predicting protein–protein interactions by modelling the attention between sets of gene ontology terms, Bioinformatics 38 (2022) 2269–2277.

[16] S. Nunes, R. T. Sousa, C. Pesquita, Multi-domain knowledge graph embeddings for gene-disease association prediction, Journal of Biomedical Semantics 14 (2023) 11.

[17] F. Shen, S. Peng, Y. Fan, A. Wen, S. Liu, Y. Wang, L. Wang, H. Liu, HPO2Vec+: Leveraging heterogeneous knowledge resources to enrich node embeddings for the human phenotype ontology, Journal of Biomedical Informatics 96 (2019) 103246.

[18] R. Patel, Y. Guo, A. Alhudhaif, F. Alenezi, S. A. Althubiti, K. Polat, Graph-based link prediction between human phenotypes and genes, Mathematical Problems in Engineering 2022 (2022) 7111647.

[19] S. Mukherjee, J. D. Cogan, J. H. Newman, J. A. Phillips, R. Hamid, J. Meiler, J. A. Capra, Identifying digenic disease genes via machine learning in the Undiagnosed Diseases Network, The American Journal of Human Genetics 108 (2021) 1946–1963.

[20] J. Cao, J. Fang, Z. Meng, S. Liang, Knowledge graph embedding: A survey from the perspective of representation spaces, ACM Computing Surveys 56 (2024) 1–42.

[21] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, O. Yakhnenko, Translating Embeddings for Modeling Multi-relational Data, in: Advances in Neural Information Processing Systems 26, Curran Associates, Inc., 2013.

[22] Y. Lin, Z. Liu, M. Sun, Y. Liu, X. Zhu, Learning Entity and Relation Embeddings for Knowledge Graph Completion, in: AAAI Conference on Artificial Intelligence, 2015.

[23] B. Yang, S. W.-t. Yih, X. He, J. Gao, L. Deng, Embedding Entities and Relations for Learning and

Inference in Knowledge Bases, in: International Conference on Learning Representations, 2015.

[24] M. Nickel, L. Rosasco, T. Poggio, Holographic Embeddings of Knowledge Graphs, in: AAAI Conference on Artificial Intelligence, AAAI Press, Washington DC, USA, 2016.

[25] R. Abboud, I. Ceylan, T. Lukasiewicz, T. Salvatori, Boxe: A box embedding model for knowledge base completion, Advances in Neural Information Processing Systems 33 (2020) 9649–9661.

[26] L. Moreau, P. Groth, Provenance: an introduction to PROV, Springer Nature, 2022.