# Evolving meaning for supervised learning in complex biomedical domains using knowledge graphs

Rita T. Sousa[0000−0002−7241−8970]

LASIGE, Faculdade de Ciências, Universidade de Lisboa, Portugal
risousa@ciencias.ulisboa.pt

**Abstract.** Knowledge graphs represent an unparalleled opportunity for machine learning, given their ability to provide meaningful context to data through semantic representations. Knowledge graphs provide multiple perspectives over an entity, describing it using different properties or multiple portions of the graph. State-of-the-art semantic representations are static and take into consideration all semantic aspects, ignoring that some may be irrelevant to the downstream learning task. The goal of this Ph.D. project is to discover suitable semantic representations of knowledge graph entities that are adapted to specific supervised learning tasks. I will use Genetic Programming to evolve tailored semantic representations, and develop novel approaches that integrate them with different supervised learning techniques. These novel approaches will be anchored by a framework that integrates different semantic representation approaches and two representative learning approaches, Support Vector Machine and Graph Convolutional Neural Networks, and allows a comparative evaluation using benchmarks. The developed approaches will be applied to two bioinformatics tasks, prediction of protein interactions and gene-disease associations, where the impact of data size and complexity will be investigated.

**Keywords:** Knowledge Graph · Ontology · Semantic Similarity · Graph Embedding · Graph Kernel · Machine Learning · Genetic Programming · Protein-Protein Interaction Prediction · Gene-Disease Association Prediction.

## 1 Introduction

Semantic information is recognized as a valuable knowledge resource in supporting data mining tasks, since it associates meaning and context to raw data in a structured way. Although many data mining approaches are limited to what can be extracted directly from the data, understanding the meaning of data increases the performance of these approaches for knowledge discovery [20].

There are three main sources of meaning used to build semantic representations: (i) text corpora that can be used to produce vectorial representations based on distributional semantics [6,15,16]; (ii) handcrafted rules, generally designed

using expert knowledge or learning from real data [14]; (iii) knowledge graphs (KGs) (typically built by integrating ontologies and linked data) which provide a conceptualization of a domain based on a formal definition of its entities and their relations. In recent years, the explosion in complexity and heterogeneity of biomedical data has motivated a new paradigma, where millions of semantically-described biological entities are available as linked data, building a biomedical Semantic Web [20]. Given their ability to provide meaningful context to the data, KGs represent an unparalleled opportunity for machine learning [20]. A cornerstone challenge to this is how to represent or encode the semantic information contained in the graph structure so that it can be easily exploited by machine learning models (i.e., producing semantic representations). KG-based representations, such as graph kernels[13] and graph embeddings [4], are a solution to bridge the gap between KGs and the typical vector-based representations of entities used by most machine learning techniques. A less well-known alternative is to employ semantic similarity as a representation [24]. Different machine learning algorithms can then process these representations for a wide variety of downstream learning tasks.

A severe limitation of several approaches for machine learning using KGs is that the construction of semantic representations often ignores the learning task. Consider the prediction of protein-protein interactions. It is well established that semantic similarity kernels over protein KGs can support the prediction task. However, the prediction is more accurate if just a portion of the KG is used (in this case, the one concerning biological processes) rather than the whole KG [1]. Therefore, adjusting the semantic representation to the machine learning task can improve its performance, but achieving it in an automated fashion is an open challenge.

The research focus of the Ph.D. is addressing this issue by using Genetic Programming (GP) to learn suitable semantic representations of data objects extracted from KGs to support supervised learning tasks. Thus, the plan of this thesis proposal includes the development of GP-based methods for evolving semantic representations (graph kernels, graph embeddings, and semantic similarity). Also, the plan comprises the adaptation of existing machine learning algorithms to explore semantic representations. The developed approaches will be evaluated in bioinformatics applications, particularly the prediction of protein interaction and disease-associated genes.

## 2 State of the Art

The research associated with this Ph.D. thesis proposal builds on the state of the art and related work from two domains: KG-based semantic representations and machine learning algorithms.

### 2.1 KG-based semantic representations

State-of-the-art KG-based semantic representations include graph kernels and graph embeddings. Semantic similarity kernels can also be used as a semantic

representation by comparing entities based on the properties they share and their taxonomic relationships.

**Graph Kernels** In the past years, many graph kernels [13] have been proposed and widely-used for solving classification tasks in graphs. Graph kernels are functions that measure the similarity between graphs. Most of the graph kernels are instances of convolution kernels. The main idea is to decompose structured objects into their sub-structures and define valid local kernels among them. The three major graph kernel families considered in this proposal are (i) graph kernels based on the distribution of limited-size subgraphs; (ii) graph kernels based on subtree pattern; (iii) graph kernels based on walks and paths.

**Graph Embeddings** An embedding maps each node to a lower-dimensional space in which its graph position and the structure of its local graph neighborhood is preserved as much as possible. There are a variety of methods for building KG embeddings [4]. While some focus on exploring solely the KG facts (like translational distance models or semantic matching models), others also include additional information, such as entity types, relation paths, axioms, and rules or textual information. More recently, path-based approaches have been proposed by transforming the KG into node sequences [19]. After representing a graph as a set of random walk paths sampled from it, natural language methods are then applied to the sampled paths for graph embedding.

**Semantic Similarity** Semantic similarity kernels [10] compare entities (ontology classes or KG entities) based on the taxonomic relations within the ontology graph. The majority of semantic similarity measures explore the properties of each class involved, typically relying on the information content of a class, a measure of how specific and informative a class is. In instance-based semantic similarity, each instance is annotated with a set of classes which are then processed using one of two approaches: pairwise, where pairwise comparisons between all classes annotating each instance are considered; groupwise, where set, vector, or graph-based measures are employed, avoiding the need for pairwise comparisons [17].

### 2.2 Machine Learning

In the context of this Ph.D., machine learning is used both to learn a suitable representation for a specific classification task and to train the classification models based on the representation. While GP is employed in the first task, the second task is more flexible and can, in principle, employ any machine learning algorithm able to handle vector or graph-based inputs. SVM and GCNNs were selected as representative approaches of these types of algorithms, but during the course of research, others may be investigated as well.

**Genetic Programming** GP is inspired by Darwinian evolution and Mendelian genetics and is a population-based search procedure that can evolve solutions to complex problems of different domains [18]. One of the major strengths of GP is its ability to explore large search spaces with a diverse population of free-form individuals and produce potentially readable white-box models, without compromising predictive ability. GP can be easily applied to supervised learning problems, with regression and classification being the most common types [8].

**Support Vector Machine** SVM is a kernel method that performs classification tasks by constructing hyperplanes in a multidimensional space that separate cases belonging to different classes. In the last decade, some approaches combining SVM with graph kernels have been proposed, which can be adapted to be used with KGs [20]. More recently, attention has shifted to approaches that use graph embeddings to learn vectorial representations that are then used with SVM [19].

**Graph Convolutional Neural Networks** GCNNs are powerful deep neural networks for graph structured data [9]. The "graph convolution" operation applies the same linear transformation to all the neighbors of a node, followed by mean pooling and nonlinearity. By stacking multiple graph convolution layers, GCNNs can learn node representations by using information from distant neighbors [3,7,5,12]. Very recently, relational GCNNs [22] were proposed as a generalization of GCNNs developed for dealing with highly multi-relational data, such as KGs, and were applied to link prediction and entity classification.

## 3 Problem Statement and Contributions

KGs are a recognized valuable source for background information in many data mining tasks, encoding semantics that describes entities in terms of several semantic aspects (Definition 1) [20].

Many of the existing KG-based approaches use KGs for generating semantic representations (Definition 2) which are used as features in various data mining tasks. These can be considered static semantic representations (Definition 3), since they take equally in consideration all semantic aspects, blind to the fact that some may be irrelevant to the downstream machine learning task, potentially introducing noise. In some applications, such as link prediction, the classification target is encoded in the KG, so this aspect is mitigated. But in applications where the classification target is not encoded in the KG, this is inevitable, since embeddings cannot be trained on the targets. Furthermore, in complex domains, KGs can be quite large and using the whole graph can be time-consuming and cumbersome and employing irrelevant features can negatively impact the performance of machine learning algorithms.

**Definition 1.** *A **semantic aspect** represents a perspective of the representation of KG entities. It can correspond to a given set of property types or portions of the graph.*

**Definition 2.** *A **semantic representation** is a set of features describing a KG entity and obtained by processing the KG.*

**Definition 3.** *A **static semantic representation** is a set of features describing a KG entity that are obtained by processing the full KG without additional external input or tailoring to a specific task.*

The guiding hypothesis of this Ph.D. proposal is that GP can learn suitable semantic representations of data objects extracted from KGs optimized towards a specific supervised learning task and without needing to have the target encoded in the KG. Although an analysis of the related work showed that there are no known approaches that use GP to improve semantic representations, preliminary results for semantic similarity kernels [23] encourage this direction of investigation.

The developed approaches will be used to support classification tasks, taking as input a KG and a set of KG entity pairs. The models are trained using external information (not encoded in the KG) about the classification targets for each pair. Many important biomedical tasks can benefit from this work. The detection of biomedical relations between pairs of biological entities has received growing attention recently, with numerous biological and clinical applications including prediction of protein interactions, drug interactions, and gene-disease relationships. High quality predictions in these areas can help target biomedical research into more promising areas. For these reasons, this domain is the main evaluation target for the proposed approaches.

This proposal is organized around three research questions (RQ):

**RQ1** Which are the static semantic representations that are more suitable to support supervised learning over KGs?
**RQ2** How can GP be applied to adapt semantic representations, improving on the best solutions achieved by domain experts?
**RQ3** Are the improved semantic representations useful to bioinformatics applications?

Therefore, the expected contributions (C) of this research are:

**C1** A novel GP-based approach to learn suitable semantic representations for KG-based classification tasks;
**C2** A novel integration of semantic representations with GCNNs;
**C3** An evaluation framework to support the comparative evaluation of expert and machine learning-based semantic representations focusing on bioinformatics classification tasks;
**C4** Open source release of all produced software.

## 4 Research Methodology and Approach

The core of the research will be supported by a framework that integrates semantic representations and machine learning approaches, allowing the comparative

evaluation of existing semantic representation approaches for machine learning, as well as the development and evaluation of novel approaches to learn improved semantic representations using GP. The methodology adopted is organized around three tasks that articulate themselves to answer the RQs: construction of static semantic representation (T1), evolving semantic representations (T2), application of evolved semantic representations (T3).

**T1: Construction of static semantic representations** This task focuses on two goals: (1) building the framework; and (2) the comparative evaluation of existing approaches for semantic representation. The framework will be composed of two modules: the semantic representation module and the machine learning module. The semantic representation module will comprise three different approaches for semantic representation: graph embeddings, graph kernels, and semantic similarity. The machine learning module will encompass the two targeted supervised learning approaches: SVMs and GCNNs. A comparative evaluation of the three semantic representation approaches combined with the two machine learning algorithms will be conducted using a state-of-the-art benchmark suite of KGs for supervised learning tasks [21].

**T2: Evolving semantic representations** The main goal of this task is to address RQ2. To do so, it will focus on developing novel approaches based on GP that are able to learn which properties or portions of the graph are more relevant and how to combine them to produce adaptive semantic representations to address a given machine learning task. These novel approaches will focus on two targets: (1) evolving a combination of properties, from which a partial graph can be extracted to support semantic representation methods; (2) evolving a combination of subgraphs that can be employed by semantic representation methods. GP can be seen as a wrapper method, where the fitness function that guides evolution is based on the success of a given combination of semantic representation and machine learning algorithm in a specific task.

The notion behind using both SVMs and GCCNs as the machine learning approaches is to support a comparison between a more classical approach and a deep learning-based one, investigating the impact of dataset size and complexity, as well as the potential contributions of adaptive semantic representations versus static ones. Furthermore, it will also allow a comparison to employing directly GCCNs to learn the semantic representations.

**T3: Application of evolved semantic representations** In this task, the novel approaches will be integrated into the framework developed in T2, evaluated using the benchmark datasets, and applied to bioinformatics challenges.

## 5 Evaluation Plan

The proposed methodology will be evaluated in general-purpose benchmarks and the tasks of protein-protein interaction (PPI) prediction and gene-disease association prediction. Both tasks address RQ3.

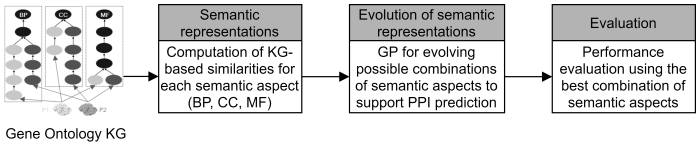


**Fig. 1.** evoKGsim methodology

**Evaluation in general-purpose benchmarks** The reference datasets that will be used are presented in [21]. This benchmark suite is comprised of 22 datasets that cover multiple domains (e.g., automotive, geology, common knowledge), range in size from 100 to 4.6 million instances, and support both classification and regression tasks.

**Application to protein-protein interaction prediction** A major challenge in systems biology is the accurate mapping of the interactome, i.e., the set of all PPI within a cell [1]. The developed approaches will be applied to the prediction of PPI, employing the Gene Ontology (GO), the most popular biomedical ontology, and several benchmark datasets [23].

**Application to gene-disease association prediction** The identification of genes responsible for human hereditary diseases can contribute to the improvement of medical care and the understanding of disease mechanisms [11]. In this task, the approaches developed and implemented will be used to predict disease-associated genes using datasets extracted from human disease databases, and KGs covering protein function, biomolecules, and metabolic pathways.

## 6 Preliminary Results

This PhD project is still in an early stage. So far, investigations have focused on similarity-based semantic representations using GP directly as a classifier. This novel approach, evoKGsim, uses GP to learn suitable combinations of semantic similarity aspects to support the classification of instances modeled as pairs of KG individuals. We evaluate its performance in PPI prediction using the GO as the KG, with its three semantic aspects, molecular function (MF), biological process (BP) and cellular component (CC), and a set of nine benchmark datasets[1]. evoKGsim currently supports two different approaches based on different semantic representations: taxonomic semantic similarity (evoKGsim-SS) calculated using $\text{ResnikMax}_{\text{Seco}}$ [17] and graph embedding similarity (evoKGsim-ES) calculated as cosine similarity over RDF2Vec embeddings [19]. The models returned by GP are the combinations of the similarity scores of the three GO aspects, evolved to support PPI prediction. An overview of the evoKGsim methodology for PPI prediction is shown in Figure 1.

We have used five static representations as baselines: the BP, CC and MF single aspects, and the average (Avg) and maximum (Max) of the single aspect

[1] These results have been partially published in [23].

**Table 1.** Median of WAFs with static representations and with evoKGsim using embeddings similarity (evoKGsim-ES) and semantic similarity (evoKGsim-SS). In bold, the best result for each dataset. The median WAF for each baseline is underlined when evoKGsim significantly outperforms the baseline ($p$-value $< 0.01$).

| Dataset (# interactions) | | Static Semantic Representations | | | | | evoKGsim |
|---|---|---|---|---|---|---|---|
| | | BP | CC | MF | Avg | Max | |
| STRING-EC | ES | 0.729 | 0.806 | 0.716 | 0.815 | 0.813 | **0.824** |
| (2245) | SS | 0.667 | 0.821 | 0.644 | 0.815 | 0.827 | **0.861** |
| STRING-DM | ES | 0.809 | 0.871 | 0.761 | 0.872 | 0.882 | **0.891** |
| (550) | SS | 0.908 | 0.883 | 0.745 | 0.910 | **0.945** | 0.936 |
| BIND-SC | ES | 0.760 | 0.768 | 0.733 | 0.801 | 0.764 | **0.803** |
| (1366) | SS | 0.876 | 0.852 | 0.779 | 0.905 | 0.894 | **0.919** |
| DIP/MIPS-SC | ES | 0.787 | 0.761 | 0.723 | 0.801 | 0.773 | **0.811** |
| (13807) | SS | 0.841 | 0.793 | 0.701 | 0.832 | 0.834 | **0.847** |
| STRING-SC | ES | 0.778 | 0.758 | 0.695 | 0.796 | 0.768 | **0.806** |
| (30384) | SS | 0.826 | 0.788 | 0.677 | 0.830 | 0.824 | **0.844** |
| DIP-HS | ES | 0.698 | 0.577 | 0.632 | 0.643 | 0.659 | **0.705** |
| (2739) | SS | 0.877 | 0.818 | 0.755 | 0.876 | 0.859 | **0.894** |
| STRING-HS | ES | 0.766 | 0.712 | 0.679 | 0.756 | 0.743 | **0.782** |
| (6912) | SS | 0.853 | 0.763 | 0.722 | 0.851 | 0.814 | **0.873** |
| GRID/HPRD-unbal-HS | ES | 0.607 | 0.560 | 0.567 | 0.601 | 0.594 | **0.613** |
| (31320) | SS | 0.715 | 0.677 | 0.662 | 0.731 | 0.706 | **0.738** |
| GRID/HPRD-bal-HS | ES | 0.639 | 0.617 | 0.599 | **0.663** | 0.641 | **0.663** |
| (31349) | SS | 0.653 | 0.602 | 0.598 | 0.654 | 0.641 | **0.658** |
| | ES | 0.730 | 0.714 | 0.678 | 0.750 | 0.737 | 0.766 |
| *Average on all datasets* | SS | 0.802 | 0.777 | 0.698 | 0.823 | 0.816 | 0.841 |

scores. The static representations are employed as a simple similarity threshold-based classifier, where a semantic similarity score for a protein pair exceeding a certain threshold predicts a positive interaction. To select the threshold, we applied stratified 10-fold cross-validation, where the training set is used to select the best classification threshold, which is then applied to the test set. This emulates the best choice that a human expert could theoretically select.

Table 1 presents the results obtained when using the graph embeddings similarity representation (evoKGsim-ES) and the semantic similarity representation (evoKGsim-SS). For evaluating the quality of a predicted classification, we use the weighted average of F-measures (WAF) for stratified 10-fold cross-validation. Statistical significance of the results was determined using pairwise non-parametric Kruskal-Wallis tests [2] at $p < 0.01$. The results indicate that the performance of evoKGsim is always better than the static baselines, except against SS Max for STRING-DM (and against ES Avg for GRID/HPRD-bal-HS, with equal performance). These results are especially relevant when we recall that the baselines were built to emulate a domain expert using an optimal threshold selection. When comparing evolved semantic representations, evoKGsim-SS achieves a better performance than evoKGsim-ES in all datasets

except GRID/HPRD-bal-HS. Since the SS representation is limited to the taxonomic relations within the ontology, whereas ES takes into account all types of relations, the ES representations could, in principle, be more informative. However, they do not take into account the specificity of annotations, which can hinder their ability to estimate similarity more accurately.

These results are a first step towards answering our research questions and may be used as a starting point for extending to other semantic representations and classification problems.

## 7 Conclusions and Lessons Learned

This Ph.D. project aims to develop novel GP-based approaches that can learn suitable semantic representations based on KGs to support supervised learning. Until now, I have developed a methodology that employs GP to evolve similarity-based semantic representations for KGs. The work to follow includes integrating machine learning algorithms and extending the approach to other semantic representations. To do so, some challenges will need to be overcome. First, machine learning algorithms such as SVM and CNNs, are usually suitable for problems with a significant number of features. This needs to be considered to ensure a fair comparison between feature-rich representations such as embeddings and the simpler similarity-based representations. Second, for semantic representations like graph embeddings, the embeddings themselves can be evolved. This may prove challenging since existing GP implementations are unable to handle vectors as a single data item.

**Acknowledgements** I would like to thank my Ph.D. supervisors, Prof. Catia Pesquita and Prof. Sara Silva, for their valuable feedback and support in the realization of this work. This research has been supported by the Fundação para a Ciência e a Tecnologia through the LASIGE Research Unit, UIDB/00408/2020 and UIDP/00408/2020, the PhD grant SFRH/BD/145377/2019, and the projects DSAIPA/DS/0022/2018, PTDC/CCI-CIF/29877/2017, PTDC/CCI-INF/29168/2017, PTDC/EEI-ESS/4633/2014.

## References

1. Bandyopadhyay, S., Mallick, K.: A new feature vector based on gene ontology terms for protein-protein interaction prediction. IEEE/ACM Trans. Comput. Biol. Bioinformatics **14**(4), 762–770 (2017)
2. Breslow, N.: A generalized kruskal-wallis test for comparing k samples subject to unequal patterns of censorship. Biometrika **57**(3), 579–594 (1970)
3. Bruna Estrach, J., Zaremba, W., Szlam, A., LeCun, Y.: Spectral networks and deep locally connected networks on graphs. In: 2nd Int. Conf. on Learning Representations (2014)
4. Cai, H., Zheng, V.W., Chang, K.C.: A comprehensive survey of graph embedding: Problems, techniques, and applications. IEEE Trans. on Knowl. and Data Eng. **30**(9), 1616–1637 (2018)

5. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: Proc. of the 30th Int. Conf. on Neural Information Processing Systems. p. 3844–3852 (2016)
6. Dumais, S.T.: Latent semantic analysis. Annual Review of Information Science and Technology **38**(1), 188–230 (2004)
7. Duvenaud, D., Maclaurin, D., Aguilera-Iparraguirre, J., Gómez-Bombarelli, R., Hirzel, T., Aspuru-Guzik, A., Adams, R.P.: Convolutional networks on graphs for learning molecular fingerprints. In: Proc. of the 28th Int. Conf. on Neural Information Processing Systems. p. 2224–2232 (2015)
8. Gandomi, A.H., Alavi, A.H., Ryan, C.: Handbook of Genetic Programming Applications. Springer International Publishing, Cham, Switzerland, 1st edn. (2015)
9. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al.: Recent advances in convolutional neural networks. Pattern Recogn. **77**(C), 354–377 (2018)
10. Harispe, S., Ranwez, S., Janaqi, S., Montmain, J.: Semantic Similarity from Natural Language and Ontology Analysis. Morgan & Claypool Publishers (2015)
11. Jimenez-Sanchez, G., Childs, B., Valle, D.: Human disease genes. Nature **409**(6822), 853–855 (2001)
12. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. CoRR **abs/1609.02907** (2016)
13. Kriege, N.M., Johansson, F.D., Morris, C.: A survey on graph kernels. Applied Network Science **5**(6) (2020)
14. Liu, H., Gegov, A., Cocea, M.: Rule-based systems: a granular computing perspective. Granular Computing **1**(4), 259–274 (2016)
15. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
16. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing. pp. 1532–1543 (2014)
17. Pesquita, C., Faria, D., Falcao, A.O., Lord, P., Couto, F.M.: Semantic similarity in biomedical ontologies. PLoS Computational Biology **5**(7), e1000443 (2009)
18. Poli, R., Langdon, W.B., McPhee, N.F., Koza, J.R.: A field guide to genetic programming. Published via `http://lulu.com` and freely available at `http://www.gp-field-guide.org.uk` (2008)
19. Ristoski, P., Paulheim, H.: RDF2Vec: RDF graph embeddings for data mining. In: The Semantic Web. pp. 498–514 (2016)
20. Ristoski, P., Paulheim, H.: Semantic web in data mining and knowledge discovery. Web Semantic **36**(C), 1–22 (2016)
21. Ristoski, P., de Vries, G.K.D., Paulheim, H.: A collection of benchmark datasets for systematic evaluations of machine learning on the semantic web. In: The Semantic Web. pp. 186–194. Springer International Publishing, Cham, Switzerland (2016)
22. Schlichtkrull, M., Kipf, T.N., Bloem, P., van den Berg, R., Titov, I., Welling, M.: Modeling relational data with graph convolutional networks. In: The Semantic Web. pp. 593–607 (2018)
23. Sousa, R.T., Silva, S., Pesquita, C.: Evolving knowledge graph similarity for supervised learning in complex biomedical domains. BMC Bioinformatics **21**(1), 6 (2020)
24. Zhu, G., Iglesias, C.A.: Computing semantic similarity of concepts in knowledge graphs. IEEE Trans. on Knowl. and Data Eng. **29**(1), 72–85 (2017)