IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Supervised Biomedical Semantic Similarity

RITA T. SOUSA[1], SARA SILVA[1], and CATIA PESQUITA[1]
[1]LASIGE, Faculdade de Ciências da Universidade de Lisboa, Portugal
Corresponding author: Rita T. Sousa (e-mail: risousa@ciencias.ulisboa.pt).

ABSTRACT Semantic similarity between concepts in knowledge graphs is essential for several bioinformatics applications, including the prediction of protein-protein interactions and the discovery of associations between diseases and genes. Although knowledge graphs describe entities in terms of several perspectives (or semantic aspects), state-of-the-art semantic similarity measures are general-purpose. This can represent a challenge since different use cases for the application of semantic similarity may need different similarity perspectives and ultimately depend on expert knowledge for manual fine-tuning.
We present a new approach that uses supervised machine learning to tailor aspect-oriented semantic similarity measures to fit a particular view on biological similarity or relatedness. We implement and evaluate it using different combinations of representative semantic similarity measures and machine learning methods with four biological similarity views: protein-protein interaction, protein function similarity, protein sequence similarity and phenotype-based gene similarity.
The results demonstrate that our approach outperforms non-supervised methods, producing semantic similarity models that fit different biological perspectives significantly better than the commonly used manual combinations of semantic aspects.

INDEX TERMS Semantic Similarity; Ontology; Knowledge Graph; Supervised Learning;

I. INTRODUCTION

THE life sciences field has increasingly taken advantage of ontologies to tackle the challenges of managing and analyzing the growing volumes of biomedical data. In the computer science context, ontologies are artifacts that express knowledge about a domain in a shareable and computationally accessible form [15]. To enable such a description, ontologies consist of classes that describe types of entities in a domain and relationships between the classes as well as restrictions, rules, and axioms. The ontology data model can be applied to a set of individual entities to create a knowledge graph (KG) [8], where the nodes represent ontology classes and real-world entities, and edges are employed in defining ontology classes' relations and semantic annotations (i.e., the assignment of a real-world entity to an ontology class that describes it [18]).

In the life sciences, we have witnessed in the last decade not only an increase in the number and size of available ontologies, with over 800 biomedical ontologies in BioPortal [43] but also in their relevance in biomedical data management and research [15]. Ontologies are also increasingly used to support data analysis and mining. One of the fundamental tasks in this area is measuring the similarity between entities described in an ontology, i.e., semantic similarity [28]. A semantic similarity measure can be defined as a function that estimates the closeness in meaning between two entities. Ontologies allow the description of complex biological phenomena that are not easily captured in mathematical form. As such, they provide the scaffolding for comparing biological entities at a higher level of complexity by comparing the ontology classes with which they are annotated. There are a wide variety of bioinformatics applications that benefit from using semantic similarity over biomedical ontologies, namely protein-protein interaction (PPI) prediction [7], [46], disease-associated genes identification [3], [14], and drug-drug interaction prediction [1], [19].

The specificity of these data mining tasks contrasts with the broad domains covered by many biomedical ontologies. Large and successful biomedical ontologies often afford multiple perspectives over the entities it describes, i.e., semantic aspects. A semantic aspect represents a perspective of the representation of KG entities and can correspond to a given set of property types or portions of the graph. For instance, the Gene Ontology (GO) [38] describes protein function

according to three semantic aspects: the *molecular functions* they perform, the *biological processes* they intervene in and the *cellular components* where they are active. Moreover, it can also be the case that multiple ontologies describe the same real-world entities, each covering a different semantic aspect.

Depending on our viewpoint of the domain or the analytical task for which we want to use semantic similarity, some semantic aspects may be irrelevant to a specific definition of similarity. Consider the following example of comparing proteins according to their function. From a biochemist's point of view, two proteins playing the same molecular functions are very similar. However, these proteins can be very different from a physiological perspective if they participate in different biological processes at the whole-organism level. Therefore, depending on our goal, different semantic aspects should be considered in similarity computation. Selecting which semantic aspects to use and how they should be taken into account usually falls to the domain expert, rendering semantic similarity applications dependent on fine-tuning. This brings us to the challenge of tailoring semantic similarity measures (SSMs) to fit a specific application and biological perspective on similarity.

This work presents a novel approach that integrates semantic similarity and supervised learning methods to learn semantic similarity models tailored to capture particular biological similarity views better, producing a supervised similarity. Since no gold standard exists for the similarity between complex biomedical entities, we take advantage of objective similarities to train the models and evaluate them [6]. These objective similarities rely on objective representations of entities (e.g., gene sequence, domains) and calculate similarity using mathematical expressions or other algorithms (e.g., BLAST-based similarity for sequences). Although these objective similarities do not provide the broad spectrum comparison that semantic similarity supports, they are known to relate to relevant characteristics of the underlying entities. The results achieved on the benchmark datasets demonstrate our approach's ability to significantly improve the estimation of similarity between biomedical entities.

Our main contributions are the following:

- We propose a novel approach that considers the different KG semantic aspects used to describe entities and relies on ML to learn a supervised semantic similarity that fits an objective similarity.
- We design a comparative evaluation that includes five KG-based similarity measures based on embeddings or taxonomic semantic similarity and eight ML methods.
- We report extensive experimental results demonstrating that our approach can produce a supervised semantic similarity that outperforms static semantic similarity for 21 benchmark biomedical datasets.

II. RELATED WORK

An SSM can be defined as a function that estimates the closeness in meaning between two entities. Several SSMs have been proposed, with most measures falling in the category of taxonomic semantic similarity (also referred to as ontology-based semantic similarity, or only semantic similarity) [12]. However, KG embeddings, a more recent research direction, can also be used to compute semantic similarity [20], [34], [35].

Taxonomic semantic similarity compares entities based on the taxonomic relations within the ontology graph [28]. Taxonomic SSMs are generally designed by an expert based on assumptions about how an ontology is used and what should constitute a similarity. They extensively use the taxonomical aspect of an ontology, comparing classes based on subclass/superclass relations. Taxonomic SSMs can be distinguished based on the entities they intend to compare since we can measure the similarity between either ontology classes or real-world entities (annotated with a set of classes). In the case of GO, semantic similarity can be calculated for two ontology classes, for instance, calculating the similarity between two GO classes (e.g., the GO term *protein metabolic process* and the GO term *protein stabilization*); or between two entities each annotated with a set of classes, for instance calculating the similarity between two proteins. Each protein can be annotated with several GO classes, so to assess the similarity between proteins, it is necessary to compare sets of classes rather than single classes.

For class-based semantic similarity, edge-based measures rely on algorithms designed for graph analysis [24], [29]. However, the majority of methods explore the properties of each class involved, typically relying on the information content (IC) of a class, a measure of how informative (or, in other words, specific) a class is, and then using it to measure the shared meaning between two classes. IC can be calculated using external data, for instance, the frequency of annotations of entities in a corpus [30], or based on intrinsic properties, such as the ontology's structure [33]. In entity-based semantic similarity, each instance is described with a set of classes which are then processed using one of two approaches: pairwise or groupwise. In pairwise approaches, the semantic similarity is calculated between classes in one set and classes in the other (using class-based measures). In groupwise approaches, the measures can directly compare the sets of classes according to information defined in the ontology, circumventing the need for pairwise comparisons [26], [39]. Purely set-based and vector-based approaches are rare. In vector-based approaches, the sets are compared through their vector representations, with each term corresponding to a dimension.

Regarding embedding semantic similarity, an embedding is a vector representation that maps each node to a lower-dimensional space. The structure of its local graph neighborhood and its graph position is preserved as much as possible. Several methods for building KG embeddings have been proposed [5]. While some focus on exploring the graph facts solely (like translational distance models [4], [42] or semantic matching [41], [44]), others also include additional information, such as entity types, relation paths, axioms

and rules, or textual information. More recently, path-based approaches, such as RDF2Vec [31] and OPA2Vec [35], have been proposed by transforming the ontology graph into node sequences. For these approaches, a KG is represented as a set of random walk paths sampled from it, and then natural language methods are applied to the sampled paths for KG embedding. After employing KG embedding methods, each entity is represented by a vector. It is then possible to compute the KG embedding similarity between two entities by computing the distance of their corresponding vectors in Euclidean space. In the GO case, the embedding methods represent proteins or GO classes in a low-dimensional space such that similar nodes in the ontology graph correspond to close points.

More recently, approaches that combine taxonomic semantic similarity with ML have been proposed. GARUM [40] is based on a supervised regression algorithm that receives several similarity measures of hierarchy, neighborhood, shared information, and attributes, and then predicts a final similarity score. In evoKGsim [36], we have used genetic programming over aspect-oriented semantic similarities to predict PPIs. However, most of the work combining ontologies and ML is focused on embeddings. Kulmanov *et al.* [21] provide an overview of methods incorporating SSMs and ontology embeddings into ML methods.

III. METHODS

We have developed a novel approach[1] [37] to learn the similarity between entities represented in KGs (Definition 1) optimized towards a specific objective similarity. This tailoring is achieved by considering the similarities for different semantic aspects (Definition 2), as opposed to the static SSMs (Definition 5).

Definition 1: A **KG** is a graph $KG = (V, E)$ where V is the set of vertices that represent either ontology classes V_c or individuals V_i, and E is the set of edges that are established between vertices, representing either ontology-level axioms, such as subclass statements or property restrictions and the assignment of an individual to a class through type declarations.

Definition 2: A **semantic aspect** is a subgraph extracted from the full KG, $KG_{SA} = (V_c', V_i', E')$ rooted in class a, where each vertex $v_c' \in V_c'$ is a subclass (directly or through inference) of a, each vertex $v_i' \in V_i'$ is an individual of a class in V_c', and where each $e' \in E'$ corresponds to an edge between elements of $V_c' \cup V_i'$.

Definition 3: A **semantic similarity** is a function that compares two individuals based on their representations in the KG and returns a numerical score that reflects the closeness in meaning between the individuals.

Definition 4: An **objective similarity** is a similarity metric that compares two individuals based on an objective representation of a specific property (e.g. two proteins represented by their amino acid sequences can be compared through their sequence similarity score.)

Definition 5: A **static semantic similarity** is a semantic similarity that does not consider additional external input or tailoring to a specific objective similarity.

Figure 1 shows an overview of the approach. The first step involves identifying the semantic aspects describing the KG entities. Our approach takes as pre-defined semantic aspects the subgraphs when the KGs have multiple roots (such as GO) or the subgraphs rooted in the classes at a distance of one from the KG root class. As an alternative, semantic aspects can be manually defined. The next step is representing each instance (i.e., a pair of KG entities) according to static KG-based similarities computed for each semantic aspect. The third step in our approach is to select the objective similarity for which we want to tailor the similarity. The last step is employing an ML method to learn a supervised semantic similarity. The ML algorithms are used for regression where the expected outputs are the objective similarity values. The models returned in the second step are then the combinations of the similarity scores of the three GO aspects.

In addition to the three GO aspects, the similarity is also calculated for the HP phenotypic abnormality subgraph for the gene dataset. Therefore, instead of three semantic aspects, we consider four semantic aspects. However, the general approach is independent of the semantic aspects, the specific implementation of KG-based similarity and the ML algorithm employed in regression.

A. DATA

Our approach takes as input an ontology file, an instance annotation file and a list of instance pairs with objective similarity values. We evaluate our approach using benchmark datasets and two different KGs.

1) Benchmark Datasets

The 21 benchmark datasets are presented in Cardoso *et al.* [6] and are available online[2] (dated June 2020). These datasets explore four objective similarities based on protein and gene properties. This resulted in one gene dataset and 16 protein datasets, divided by species, level of annotation completion and objective similarity, and four additional datasets, combining all species' protein pairs in the same objective similarity group. Datasets range from 264 individual proteins and 428 pairs to 27 thousand proteins and 158 thousand pairs.

The protein datasets are constituted of proteins identified by their UniProt Accession Numbers and annotated with GO classes. The number of proteins and pairs for each protein dataset is supplied in Table S1 of the Supplementary File. The gene dataset has 2026 distinct human genes identified by their Entrez Gene Code and 12000 gene pairs. Each gene is annotated with GO classes and HP classes.

In the PFAM datasets, two objective protein similarities based on their biological properties were employed: sequence

[1] https://github.com/liseda-lab/Supervised-SS

[2] https://github.com/liseda-lab/kgsim-benchmark

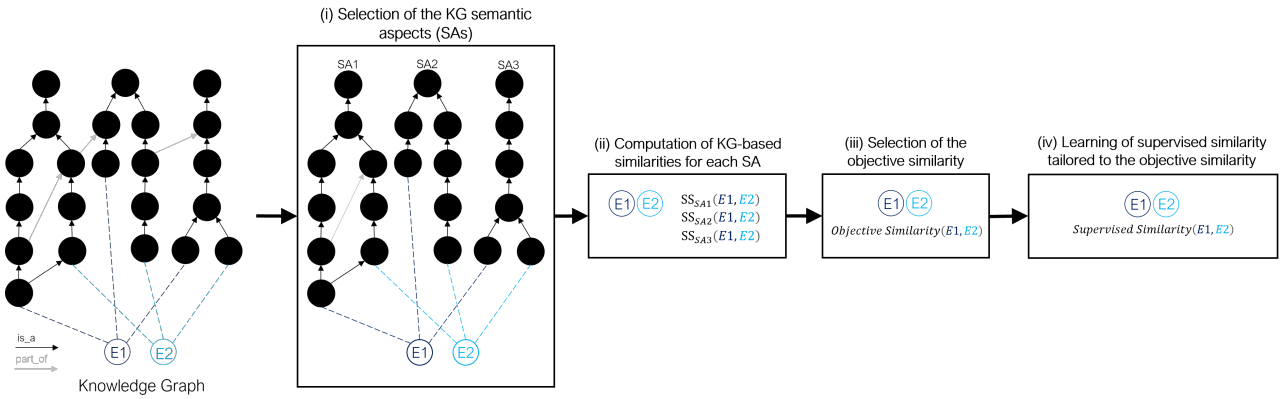

FIGURE 1: Overview of the proposed approach.

similarity and PFAM similarity. In PPI protein datasets, two objective similarities were also employed: sequence similarity and PPI similarity. Concerning the gene benchmark dataset, the objective similarity is based on phenotypic series.

- ***Sequence similarity*** (Sim_{seq}) measures the relationship between two sequences, and it establishes the likelihood for sequence homology. We infer homology (i.e., common evolutionary ancestry) when two sequences share more similarity than would be expected by chance. A sequence similarity value is aimed to approximate the evolutionary distance between proteins.
- ***PFAM similarity*** (Sim_{PFAM}) is computed by comparing the functional regions (commonly termed domains) that exist in each protein sequence. Protein functional domains were extracted from the PFAM [9]. Since protein domains typically correspond to functional sites of a protein, determining the similarity between domains can help to define protein function.
- ***Protein-protein interaction similarity*** (Sim_{PPI}) has a binary representation: 1 if the proteins interact, 0 otherwise. Two proteins are considered to be similar if they interact. PPIs are responsible for many critical functions in biology and are highly relevant to disease states.
- ***Phenotypic series similarity*** (Sim_{PS}) is based on OMIM's Phenotypic Series [2], which are groups of identical or similar phenotypes and their associated genes. Phenotypic similarity reflects the similarity between genes and can help to find biological modules of functionally related genes.

2) Gene Ontology Knowledge Graph

GO [38] is the most widely used biological ontology. It defines the universe of classes, also called "GO terms", associated with gene product (proteins or RNA) functions and how these functions are related to each other concerning three aspects: (i) molecular function (MF), the activities that occur at the molecular level performed by the gene product; (ii) biological process (BP), the larger process in which the gene product is active; (iii) cellular component (CC), the cellular compartments in which the gene product performs a function. Figure 2 shows a small fraction of the GO and annotated proteins.

We built the GO KG with explicit type declarations that link proteins to the GO classes describing them according to their GO annotations. Therefore, the nodes of the GO KG represent proteins or GO classes, while edges represent relationships between the GO classes or links between proteins annotated with GO classes. In this work, the GO KG, with its three semantic aspects (BP, CC and MF), is used to compute the similarity between two proteins for the protein datasets and two genes for the gene dataset.

3) Human Phenotype Knowledge Graph

The HP [22] contains terms describing phenotypic abnormalities found in human hereditary diseases. The HP is organized as independent subontologies that cover different categories: "Phenotypic abnormality", "Mode of inheritance", "Clinical course", "Clinical modifier" and "Frequency". Since the subontology "Phenotypic Abnormality" is the ontology branch that describes the phenotypes associated with the gene, the HP KG comprises this subontology and HP annotations. An HP annotation associates a specific gene with a particular HP class.

In the HP KG, the nodes are HP classes or genes. The edges represent ontology relations or links between genes and HP classes via their annotations. Figure 3 shows an example subgraph of the HP KG. In this work, the HP KG is used to compute the semantic similarity between two genes based on the phenotypes that describe them.

B. STATIC SIMILARITY COMPUTATION

The following subsections present the specific details of the five different KG-based SSMs: two based on taxonomic similarity and three based on embeddings.

1) Taxonomic Semantic Similarity

We employ two state-of-the-art measures, derived by combining one IC approach (IC_{Seco}) with one of two set similarity

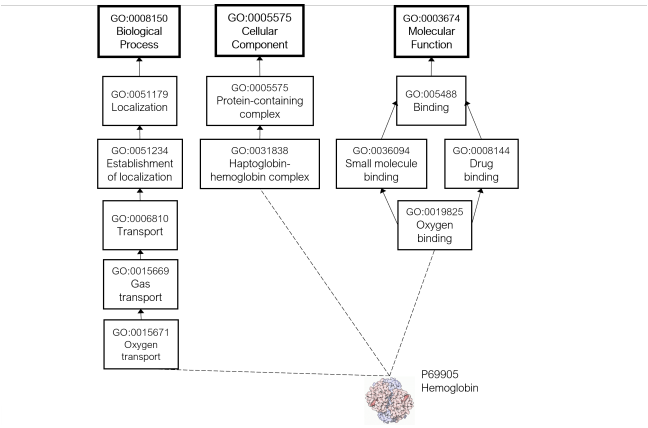

FIGURE 2: Example of a protein represented under three semantic aspects of the Gene Ontology: biological process, molecular function and cellular component. For simplicity, only a small portion of each semantic aspect subgraph is shown.

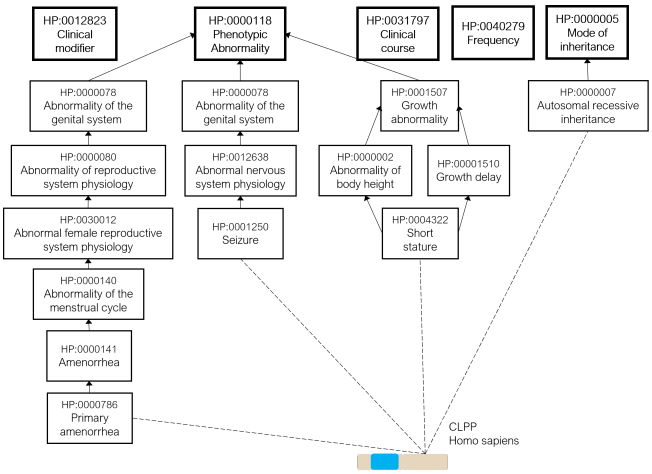

FIGURE 3: Example of a disease represented under two out of five semantic aspects of the Human Phenotype Ontology: phenotypic abnormality and mode of inheritance. For simplicity, only a small portion of each semantic aspect subgraph is shown.

measures (ResnikBMA, SimGIC), using the Semantic Measures Library 0.9.1 [13]. These were selected by their high performance in the biomedical domain [25].

IC_{Seco} is a structure-based approach proposed by Seco *et al.* [33] based on the number of direct and indirect descendants that measures how informative (or, in other words, specific) a class is. It is given by

$$\text{IC}_{\text{Seco}}(t) = 1 - \frac{\log\left[\text{Ndescendants}(t) + 1\right]}{\log\left[\text{Nnodes}\right]} \tag{1}$$

where $\text{Ndescendants}(t)$ is the number of indirect and direct descendants from term t (including term t), and Nnodes is the total number of concepts in the ontology.

ResnikBMA is a pairwise approach based on the class-based measure proposed by Resnik [30] in which the similarity between two classes corresponds to the IC of their most informative common ancestor. In this pairwise approach, the semantic similarity between two instances is calculated between classes in one set and classes in the other

$$\text{ResnikBMA}(e_1, e_2) = \frac{\sum_{t_1 \in \text{S}(e_1)} \text{sim}(t_1, t_2)}{2|\text{S}(e_1)|} + \frac{\sum_{t_2 \in \text{S}(e_2)} \text{sim}(t_1, t_2)}{2|\text{S}(e_2)|} \tag{2}$$

where $\text{S}(e_i)$ is the set of annotations for entity e_i and $\text{sim}(t_1, t_2)$ is the semantic similarity between class t_1 and class t_2 and is defined as:

$$\text{sim}(c_1, c_2) = \max\left\{\text{IC}(c) : c \in \{\text{A}(c_1) \cap \text{A}(c_2)\}\right\} \tag{3}$$

where $\text{A}(c_i)$ is the set of ancestors of c_i.

SimGIC is a groupwise approach where the sets of classes are directly compared according to information defined in the ontology, circumventing the need for pairwise comparisons. It was proposed by Pesquita *et al.* [26] and is based on a Jaccard index in which each term is weighted by its IC

$$\text{SimGIC}(e_1, e_2) = \frac{\sum_{t \in \{\text{S}(e_1) \cap \text{S}(e_2)\}} \text{IC}(t)}{\sum_{t \in \{\text{S}(e_1) \cup \text{S}(e_2)\}} \text{IC}(t)} \tag{4}$$

where $\text{S}(e_i)$ is the set of annotations (direct and inherited) for entity e_i.

2) Knowledge Graph Embedding Similarity

We apply three KG embedding approaches, namely RDF2Vec, TransE, and distMult, using an RDF2Vec python implementation[3] and the OpenKE library[4]. These approaches were selected because they represent the main types of KG embedding techniques. *RDF2Vec* [31] is a path-based approach adapted to RDF graphs, that employs neural language models over random walks on the graph. *TransE* [4] is the most representative translational distance embedding approach that exploits distance-based scoring functions. *distMult* [44] is a semantic matching approach that exploits similarity-based scoring functions.

We generate protein or gene KG embeddings for each semantic aspect using these approaches (parameters for each embedding method are supplied in the Supplementary File), and then, to compute the KG embeddings similarities, we employ cosine similarity between the vectors representing each entity in the pair.

C. SUPERVISED SIMILARITY COMPUTATION

Our approach combines the semantic similarities computed for each semantic aspect and returns a supervised similarity (see Algorithm 1). A supervised regression algorithm computes the aggregation function. Therefore, each regressor receives the similarity values for each semantic aspect as

[3]https://github.com/IBCNServices/pyRDF2Vec
[4]https://github.com/thunlp/OpenKE/tree/OpenKE-Tensorflow1.0

input features (independent variables) and an objective similarity value as the expected output (dependent variable) and returns an aggregated similarity score as the predicted output. We evaluate eight representative classes of ML models to train regressors using scikit-learn 21.3 [23] library: linear regression (LR), bayesian ridge (BR), K-nearest neighbor (KNN), genetic programming (GP), decision tree (DT), random forest (RF), XGBoost (XGB), and multi-layer perception (MLP). Details and parameters are available in the Supplementary File.

Algorithm 1 Supervised Semantic Similarity

```
ssm ← semantic_similarity_measure
regressor ← regression_algorithm
function GET_SS(ent_pairs, kg)
    ss ← ∅    ▷ dictionary to hold ss scores for each pair
    root ← GET_ROOT(kg)
    semantic_aspects ← GET_SA(root)
    for s in semantic_aspects do
        sg ← GET_SG(s)          ▷ subgraph rooted in s
        for e1, e2 in ent_pairs do
            ss[e1, e2].append(COMPUTE_SS(e1,e2,sg,ssm))
    return ss
function GET_SUP_SS(ent_pairs, objective_sim, kg)
    ss_scores ← GET_SS(ent_pairs, kg)
    X_train, X_test ← SPLIT_TRAIN(ss_scores)
    y_train, y_test ← SPLIT_TRAIN(objective_sim)
    model ← CREATE_MODEL(regressor)
    model.fit(X_train, y_train)
    sup_ss_scores ← model.predict(X_test)
    return sup_ss_scores
```

IV. RESULTS AND DISCUSSION

The focus of our evaluation approach is to assess the ability of our approach to improve semantic similarity computations, avoiding the need for expert knowledge. For each combination of an SSM with an ML algorithm, we compute the Pearson's correlation coefficient between the obtained supervised similarity (predicted values) and the respective objective similarities (expected values). For cross-validation, each dataset is split into ten folds. The same ten folds are used throughout all the experiments. For each fold, we take that fold as the test set and the remaining nine folds as the training set. Each ML algorithm learns on the training set and outputs its predictions for the test set, where the Pearson correlation coefficient is calculated. The results we report are the median and the interquartile range (IQR) of the ten Pearson correlation coefficients calculated on the ten folds. We compute the static similarity for each semantic aspect and use, as baselines, the single aspect similarities and two well-known strategies for combining the single aspect scores, the average and maximum. By comparing these baselines to the supervised approaches, we aim to investigate the ability of ML methods to learn combinations of semantic aspects that improve the calculation of similarity.

Table 1 compares the results obtained using static similarity and supervised similarity for sequence, PFAM, PPI and phenotypic series similarities. The static similarity was obtained using different SSMs, and then the Pearson correlation coefficient was computed for each objective similarity. Regarding supervised similarity, the median and IQR of Pearson correlation values were calculated for the proposed approach using an SSM with an ensemble method (XGB or RF) for each objective similarity, the combinations previously shown to produce the best results. For the sake of brevity, Table 1 only shows the results for the protein datasets with one level of annotation combining all species' protein pairs in the same objective similarity group. However, Tables S5-S8 of the Supplementary File provide the results for the remaining protein datasets, SSMs and ML algorithms and show that the combination of SSM-ML that increases performance is always composed of a taxonomic SSM and an ensemble method.

A. STATIC SIMILARITY

The behavior of the five similarity-based semantic measures employed is, for most datasets, consistent. Comparing the two taxonomic semantic similarity approaches, we verify that, in most cases, the maximum correlation is achieved when the ResnikBMA approach is used. Regarding the KG embedding approaches, TransE has performed worse than the other embedding methods. Therefore, the results obtained with TransE were excluded from Table 1 but are shown in the Supplementary File. distMult, a semantic matching method, is the second-best class of embeddings. Finally, RDF2Vec achieves the maximum correlation in the majority of datasets.

The differences between KG embedding approaches are not unexpected since the methods that put more emphasis on local neighborhoods, such as translational distance approaches, are less suitable since they fail to capture longer-distance relations. This is relevant when most of the information to be processed is represented in the ontology portion of the KG, where taxonomic relations play an essential role. RDF2Vec, a path-based approach, can capture taxonomic (longer-distance) relations, which translates into a broader representation of the entities, achieving better results than the other embedding methods in most experiments.

When comparing the two types of semantic similarity, taxonomic similarity performs well across many evaluations and, in most datasets, performs better than embedding similarity. The initial assumption was that embedding similarity could outperform taxonomic similarity since semantic similarity is limited to the taxonomic relations within the ontology. In contrast, embeddings consider all types of relations, and therefore, the embedding representations could be more informative in principle. However, the ability of taxonomic similarity to take into account class specificity may give it the advantage over embedding similarity to estimate similarity more accurately. Besides, taxonomic similarity measures are usually hand-crafted, providing human-interpretable results

TABLE 1: Pearson correlation coefficient between the objective similarity and different SSMs for the baselines and the median and IQR of Pearson correlation coefficient between the objective similarity and supervised similarity obtained using XGB or RF. In bold, the best result for each PPI_ALL1 dataset-SSM.

Objective Similarity	Dataset	SSM	Static						Supervised			
			HP	BP	CC	MF	Avg	Max	XGB		RF	
									Median	IQR	Median	IQR
Sim_{seq}	PFAM	ResnikBMA		0.528	0.373	0.291	0.481	0.399	**0.803**	0.013	0.746	0.015
		SimGIC		0.552	0.406	0.415	0.547	0.406	**0.640**	0.033	0.589	0.004
		RDF2Vec		0.540	0.437	0.419	0.544	0.457	**0.657**	0.014	0.610	0.014
		disMult		0.398	0.236	0.322	0.467	0.429	0.570	0.009	**0.577**	0.009
	PPI	ResnikBMA		0.258	0.222	0.326	0.323	0.250	**0.774**	0.024	0.770	0.028
		SimGIC		0.317	0.257	0.370	0.380	0.280	**0.580**	0.030	0.542	0.029
		RDF2Vec		0.274	0.237	0.297	0.316	0.268	**0.735**	0.039	0.732	0.046
		disMult		0.277	0.239	0.202	0.369	0.310	**0.716**	0.047	**0.716**	0.045
Sim_{PFAM}	PFAM	ResnikBMA		0.448	0.370	0.456	0.525	0.500	**0.669**	0.008	0.638	0.005
		SimGIC		0.494	0.451	0.591	0.621	0.604	0.680	0.015	**0.691**	0.003
		RDF2Vec		0.524	0.466	0.619	0.627	0.623	0.661	0.007	**0.666**	0.009
		disMult		0.414	0.254	0.388	0.516	0.457	**0.527**	0.007	0.523	0.007
Sim_{PPI}	PPI	ResnikBMA		0.545	0.486	0.353	0.569	0.545	**0.634**	0.015	0.634	0.015
		SimGIC		0.486	0.428	0.318	0.496	0.458	0.584	0.008	**0.585**	0.007
		RDF2Vec		0.457	0.404	0.353	0.477	0.440	**0.510**	0.010	0.505	0.010
		disMult		0.452	0.244	0.015	0.388	0.396	**0.504**	0.014	**0.504**	0.014
Sim_{PS}	Gene	ResnikBMA	0.601	0.210	0.142	0.055	0.413	0.552	**0.648**	0.022	0.648	0.023
		SimGIC	0.489	0.205	0.158	0.095	0.399	0.429	**0.630**	0.011	0.629	0.013
		RDF2Vec	0.526	0.230	0.182	0.123	0.396	0.351	0.563	0.014	**0.564**	0.010
		disMult	0.015	0.184	0.105	0.041	0.179	0.182	0.172	0.018	**0.212**	0.052

for further analysis. On the contrary, embedding methods describe an entity as a numerical vector and, most of the time, are not interpretable since it is not possible to obtain an explanation for the results.

It is also important to point out the differences between semantic aspects. These differences depend on the objective similarity we are considering. For the sequence similarity, the differences between semantic aspects are not relevant, and no semantic aspect is clearly superior to others. Previous works [16] already suggested the absence of a strong correlation between sequence and semantic similarities since there are many proteins with low sequence similarity and high semantic similarity. Concerning the PPI similarity, proteins interacting in a cell are expected to participate in similar cellular locations and processes. As expected, the results indicate that using only the semantic similarity for MF provides worse results than the other single aspects. In opposition, we verify that the MF is a relevant semantic aspect for the PFAM similarity. The more functional (or PFAM) domains two proteins share, the more likely it will be to share molecular functions since these domains are usually responsible for assigning functions to proteins. For the gene dataset, the HP semantic aspect achieves better results than the GO semantic aspects. These results were also expected since the more phenotypic series two genes are associated with, the more likely they share HP classes. Regarding static combination approaches, in most cases, they achieve better results than the single aspects, with the average combination outperforming the maximum.

B. SUPERVISED SIMILARITY

The objective similarities reflecting different biological features allow us to use ML algorithms to learn a supervised similarity towards a domain viewpoint. We employ eight representative ML methods, including classical, ensemble, and neural network-based methods. The heat maps depicting the median Pearson correlation coefficient between the objective similarities and supervised similarity obtained with different ML methods and SSMs for each objective similarity are supplied in the Supplementary File and facilitate the comparison of ML algorithms.

Analysing the eight employed ML methods, the results show that the regression models obtained by DT are globally lower compared to the other ML algorithms. DT is one of the most commonly used approaches for supervised learning. However, since it is based on recursive binary splitting, DT may not be suitable for the current regression problem of finding the best combination of semantic aspects. LR and BR also show lower correlations in many cases. LR and BR assume a linear relationship between the independent and dependent variables, which is not valid for many cases. This characteristic may explain why these ML methods could not learn suitable combinations of semantic aspects. While KNN, GP, and MLP achieve comparable results, ensemble methods, like XGB and RF, achieve higher results in most experiments. This is not unexpected since ensemble methods combine the decisions from multiple models to improve the overall performance. These methods have been successfully applied to different domains [32].

The results indicate that taxonomic semantic similarity is a more suitable similarity-based semantic representation for learning. Although the static similarity results have already demonstrated that taxonomic semantic similarity achieves higher correlations than KG embedding similarity, these differences are more evident when we apply ML methods. Interestingly, statistical tests (see the Supplementary File) show that significant performance differences are more common when comparing SSMs rather than ML methods. Therefore, it is not straightforward to identify the best combination of SSM with an ML algorithm that will work for all datasets and use cases. Nevertheless, the results support that combining a taxonomic SSM (ResnikBMA or SimGIC) with an ensemble method (RF or XGB) is a safe choice.

C. STATIC VERSUS SUPERVISED SIMILARITY

The results in Table 1 show that whatever the ensemble method and taxonomic SSM, supervised similarity consistently achieves higher correlation values than static similarity. Improvements over the single aspect similarities are consistent for all datasets and also clear when considering the combination baselines. However, there are some differences between the objective similarities. For sequence similarity, it is known that the relationship between sequence similarity and semantic similarity is non-linear [27], so improvements over the best static similarity are very pronounced (up to 58% for PPI_ALL1). Regarding PFAM similarity, supervised similarity outperforms both single aspects and static combinations (average and maximum), although the improvements are more relevant for single aspects. Concerning PPI similarity, improvements over the single aspect baselines are, as expected, more pronounced for the MF baseline (between 44 and 47%). The differences between static and supervised similarity are much more accentuated in the gene dataset for the GO single aspects.

It is important to note that, although interpretable models achieve lower performance values than black-box models in most cases, as shown in heat maps (Figures S1-S5 of the Supplementary File), the supervised similarity obtained using LR and GP can still improve over the baselines. Furthermore, we verify that also for embedding similarity, our approach can learn a combination of semantic aspects that outperforms the best static similarity.

To better compare the static similarity and our supervised similarity, we also generated violin plots. Figure S6 of the Supplementary File shows, for each dataset, three violins: the distribution of the objective similarity values; the distribution of supervised similarity obtained using one of the best SSM-ML method combinations, ResnikBMA coupled with XGBoost; and the distribution of the static similarity using the average of the single semantic aspects similarities computed with the best overall measure, ResnikBMA. For the sequence similarity, the distribution and the median for the objective and supervised similarity values are very similar but differ entirely from the static similarity. Regarding PPI and PFAM and phenotypic series similarities, the supervised similarity distribution has a broader range of values than the static similarity, which is closer to the objective similarity distributions. In the PPI similarity, the shape of the distribution of the supervised similarity is also closer to the objective similarity, with two wider areas closer to zero and to one. These results confirm our approach finds semantic aspect combinations that capture a given similarity perspective.

SUPERVISED SIMILARITY INTERPRETABILITY

Although static SSMs, such as taxonomic SSMs, are handcrafted and interpretable, supervised learning can lead to losing this valuable characteristic. Therefore, it is interesting to compare ML algorithms not only in terms of performance but also in terms of interpretability. The models obtained by KNN, BR, MLP and ensemble methods are more challenging to interpret, although some methods for explaining black-box models have been proposed [10]. In opposition, the LR models predict the target as a weighted sum of the feature inputs. These linear equations have an easy-to-understand interpretation. Table 2 shows, for each objective similarity, an LR model obtained in one of the folds.

The solutions obtained by DT and GP are also, in principle, interpretable. However, in both cases, trees may grow to be very complex while learning complicated datasets, which can raise some difficulty in interpreting the solutions. Figure 4 shows, for each objective similarity, a GP model obtained in one of the folds. To allow a better understanding, these models were simplified to remove redundant and inviable code. Although the frequency in which a given variable appears in a GP model does not necessarily measure its importance for the predictions, the GP model analysis can indicate which semantic aspects are most relevant for each objective similarity. The obtained DT models are not shown since they are very large with multiple levels deep, which decreases their interpretability and visualization.

USING SUPERVISED SIMILARITY FOR PROTEIN-PROTEIN INTERACTION PREDICTION

The supervised similarity tailored to relevant biological similarities can be transferred to predictive tasks such as the PPI prediction. In several works, the prediction of PPI is formulated as a classification problem where a similarity score for a protein pair exceeding a certain threshold indicates a positive interaction [11], [17], [45]. Therefore, we used our supervised semantic similarity tailored to the PPI objective similarity to predict whether two proteins interact and compared it with supervised similarity tailored to the sequence similarity. Figure 5 compares supervised similarity with Precision-Recall curves evaluated using the best overall SSM, ResnikBMA, coupled with two ML methods, RF and XGB. The chart shows that the supervised similarity tailored to PPI obtained with XGB generally achieves the best AUC results. In contrast, the supervised similarity tailored to sequence similarity achieves the worst results. The difference between using the inappropriate supervised similarity and the suitable one is dramatic: between 0.15 and 0.20 for XGB

TABLE 2: Linear Regression models.

Objective Similarity	Model
Sim_{seq}	$0.1450\ \text{SS}_{\text{BP}} + 0.0693\ \text{SS}_{\text{CC}} + 0.0916\ \text{SS}_{\text{MF}} - 0.0058$
Sim_{PFAM}	$0.1943\ \text{SS}_{\text{BP}} + 0.1861\ \text{SS}_{\text{CC}} + 0.4344\ \text{SS}_{\text{MF}} + 0.1211$
Sim_{PPI}	$0.6001\ \text{SS}_{\text{BP}} + 0.6864\ \text{SS}_{\text{CC}} + 0.0132\ \text{SS}_{\text{MF}} + 0.1701$
Sim_{PS}	$0.8406\ \text{SS}_{\text{HP}} + 0.2063\ \text{SS}_{\text{BP}} + 0.1282\ \text{SS}_{\text{CC}} + 0.0004\ \text{SS}_{\text{MF}} + 0.2261$

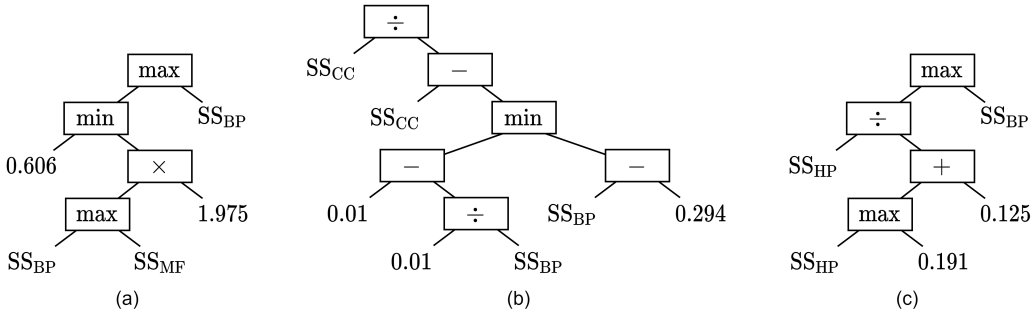

FIGURE 4: Parse trees representing GP models are shown for: (a) Sim_{PFAM}; (b) Sim_{PPI}; (c) Sim_{PS}.

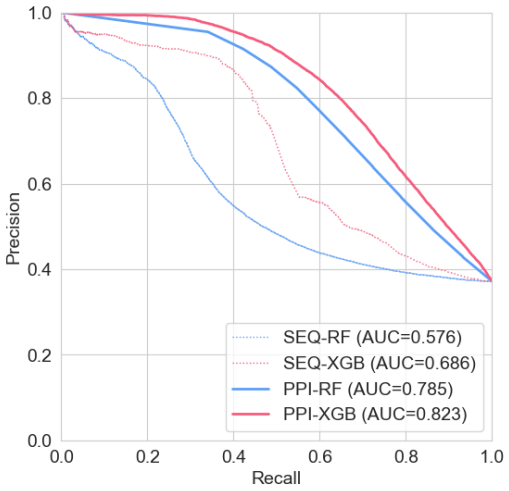

FIGURE 5: Precision-Recall curves and area under the curve (AUC) obtained for the PPI_ALL1 dataset using static similarity (Avg and Max), supervised similarity tailored to PPI similarity (PPI-RF and PPI-XGB), and supervised similarity tailored to sequence similarity (SEQ-RF and SEQ-XGB).

and RF, respectively. These results support the importance of calculating a similarity appropriate for our purpose.

The Precision-Recall curves for the remaining datasets are supplied in the Figure S7 of the Supplementary File. Comparing the charts for different datasets, we observe that the supervised similarity tailored to the PPI similarity obtained with XGB generally achieves the best AUC results.

V. CONCLUSION

Measuring the similarity between two gene products is fundamental to biomedical informatics research. Biomedical ontologies and KGs provide meaningful context to data and support the comparison of biomedical entities through semantic similarity. Many KGs afford different perspectives on the data. However, existing SSMs are general-purpose and typically depend on expert knowledge to select and combine the relevant KG semantic aspects for each use case. Tailoring semantic similarity to a viewpoint of the domain or a particular use case in an automated fashion had not yet been tackled.

We have developed a novel approach that considers the different KG semantic aspects used to describe entities and relies on ML to learn a supervised semantic similarity to fit an objective biological similarity. It captures a specific biological similarity view without needing domain experts to fine-tune it.

To evaluate the effectiveness of our approach, we used 21 benchmark datasets, categorized by species, annotation completeness level, knowledge graphs (KGs) used, and objective similarity measures employed. The objective similarities correspond to widely employed biological similarity metrics - PPI similarity, protein function similarity, protein sequence similarity and phenotype-based gene similarity - and were used to train and test the supervised models. The results show that our supervised similarity model achieves significant improvements over classical taxonomic SSMs as well as the more recently proposed KG embedding-based measures. Furthermore, it can find better semantic aspect combination functions than static combinations emulating expert knowledge. Finally, we demonstrate that tailoring an SSM to the appropriate use case has a marked influence on

predictive performance based on SSM, as evidenced by our case study on PPI prediction.

We evaluated both interpretable and black-box machine learning algorithms and compared their performance and interpretability. While the black-box models produced predictions with higher accuracy in our experiments, the supervised similarity obtained using LR and GP still showed improvement over the baseline models and allowed for an insightful analysis. This highlights the need to explore the trade-off between performance and interpretability.

Our approach is independent of the SSM and the chosen ML method. Until now, we have combined eight representative classes of ML models with five SSMs that consider semantic and structural information. Recently, embedding methods, such as OPA2Vec [35], that also consider lexical information, can be implemented and incorporated into our methodology.

Although we have applied supervised ML algorithms to tailor semantic similarity to different similarity objectives, the proposed approach is versatile and can also be applied to tailor semantic similarity to a specific learning task. Consequently, there are multiple real-world tasks where KG-based similarity is a suitable instance representation that can benefit from this work. Future work should evaluate the impact of supervised similarity in tasks such as drug-target interactions or gene-disease associations.

ACKNOWLEDGMENT

C. P., S. S., R. T. S. are funded by the FCT through LASIGE Research Unit (ref. UIDB/00408/2020 and ref. UIDP/00408/2020), and the FCT PhD grant (ref. SFRH/BD/145377/2019). It was also partially supported by the KATY project, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017453, and by HfPT: Health from Portugal under the Portuguese Plano de Recuperação e Resiliência.

REFERENCES

[1] Ibrahim Abdelaziz, Achille Fokoue, Oktie Hassanzadeh, Ping Zhang, and Mohammad Sadoghi. Large-scale structural and textual similarity-based mining of knowledge graph to predict drug–drug interactions. Journal of Web Semantics, 44:104 – 117, 2017.

[2] Joanna S Amberger, Carol A Bocchini, François Schiettecatte, Alan F Scott, and Ada Hamosh. OMIM.org: Online mendelian inheritance in man (omim®), an online catalog of human genes and genetic disorders. Nucleic Acids Research, 43(Database issue):D789—98, January 2015.

[3] Muhammad Asif, Hugo F. M. C. M. Martiniano, Astrid M. Vicente, and Francisco M. Couto. Identifying disease genes using machine learning and gene functional similarities, assessed through gene ontology. PLOS ONE, 13(12):1–15, 12 2018.

[4] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Durán, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In Proceedings of NIPS 2013, page 2787–2795, Red Hook, NY, USA, 2013. Curran Associates Inc.

[5] H. Cai, V. W. Zheng, and K. C. Chang. A comprehensive survey of graph embedding: Problems, techniques, and applications. IEEE Transactions on Knowledge and Data Engineering, 30(9):1616–1637, Sep. 2018.

[6] Carlota Cardoso, Rita T. Sousa, Sebastian Köhler, and Catia Pesquita. A collection of benchmark data sets for knowledge graph-based similarity in the biomedical domain. In Andreas Harth, Valentina Presutti, Raphaël Troncy, Maribel Acosta, Axel Polleres, Javier D. Fernández, Josiane Xavier Parreira, Olaf Hartig, Katja Hose, and Michael Cochez, editors, Proceedings of of ESWC 2020, pages 50–55, Cham, 2020. Springer.

[7] Kuan-Hsi Chen, Tsai-Feng Wang, and Yuh-Jyh Hu. Protein-protein interaction prediction using a hybrid feature representation and a stacked generalization scheme. BMC Bioinformatics, 20(1):308, 2019.

[8] Lisa Ehrlinger and Wolfram Wöß. Towards a definition of knowledge graphs. SEMANTiCS, 48:1–4, 2016.

[9] Sara El-Gebali, Jaina Mistry, Alex Bateman, Sean R Eddy, Aurélien Luciani, Simon C Potter, Matloob Qureshi, Lorna J Richardson, Gustavo A Salazar, Alfredo Smart, Erik L L Sonnhammer, Layla Hirsh, Lisanna Paladin, Damiano Piovesan, Silvio C E Tosatto, and Robert D Finn. The Pfam protein families database in 2019. Nucleic Acids Research, 47(D1):D427–D432, 10 2018.

[10] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. ACM Comput. Surv., 51(5), August 2018.

[11] Xiang Guo, Rongxiang Liu, Craig D. Shriver, Hai Hu, and Michael N. Liebman. Assessing semantic similarity measures for the characterization of human regulatory pathways. Bioinformatics, 22(8):967–973, 02 2006.

[12] Sebastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. Semantic Similarity from Natural Language and Ontology Analysis. Morgan Claypool Publishers, Williston, VT, USA, 2015.

[13] Sébastien Harispe, Sylvie Ranwez, Stefan Janaqi, and Jacky Montmain. The semantic measures library and toolkit: fast computation of semantic similarity and relatedness using biomedical ontologies. Bioinformatics, 30(5):740–742, 10 2013.

[14] Robert Hoehndorf, Paul N Schofield, and Georgios V Gkoutos. PhenomeNET: a whole-phenome approach to disease gene discovery. Nucleic Acids Research, 39(18):e119, 2011.

[15] Robert Hoehndorf, Paul N Schofield, and Georgios V Gkoutos. The role of ontologies in biological and biomedical research: a functional perspective. Briefings in bioinformatics, 16(6):1069–1080, 2015.

[16] N. Ikram, M. A. Qadir, and M. T. Afzal. Investigating correlation between protein sequence similarity and semantic similarity using gene ontology annotations. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 15(3):905–912, 2018.

[17] Shobhit Jain and Gary D Bader. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. BMC Bioinformatics, 11(1):562, 2010.

[18] Jelena Jovanović and Ebrahim Bagheri. Semantic annotation in biomedicine: the current landscape. Journal of biomedical semantics, 8(1):1–18, 2017.

[19] Andrej Kastrin, Polonca Ferk, and Brane Leskošek. Predicting potential drug-drug interactions on topological and semantic similarity features using statistical learning. PLOS ONE, 13(5):1–23, 05 2018.

[20] Maxat Kulmanov, Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf. Semantic similarity and machine learning with ontologies. Briefings in Bioinformatics, 10 2020. bbaa199.

[21] Maxat Kulmanov, Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf. Semantic similarity and machine learning with ontologies. Briefings in bioinformatics, page bbaa199, 2020.

[22] Sebastian Köhler, Leigh Carmody, Nicole Vasilevsky, Julius O B Jacobsen, Daniel Danis, Jean-Philippe Gourdine, Michael Gargano, Nomi L Harris, Nicolas Matentzoglu, Julie A McMurry, David Osumi-Sutherland, Valentina Cipriani, James P Balhoff, Tom Conlin, Hannah Blau, Gareth Baynam, Richard Palmer, Dylan Gratian, Hugh Dawkins, Michael Segal, Anna C Jansen, Ahmed Muaz, Willie H Chang, Jenna Bergerson, Stanley J F Laulederkind, Zafer Yüksel, Sergi Beltran, Alexandra F Freeman, Panagiotis I Sergouniotis, Daniel Durkin, Andrea L Storm, Marc Hanauer, Michael Brudno, Susan M Bello, Murat Sincan, Kayli Rageth, Matthew T Wheeler, Renske Oegema, Halima Lourghi, Maria G Della Rocca, Rachel Thompson, Francisco Castellanos, James Priest, Charlotte Cunningham-Rundles, Ayushi Hegde, Ruth C Lovering, Catherine Hajek, Annie Olry, Luigi Notarangelo, Morgan Similuk, Xingmin A Zhang, David Gómez-Andrés, Hanns Lochmüller, Hélène Dollfus, Sergio Rosenzweig, Shruti Marwaha, Ana Rath, Kathleen Sullivan, Cynthia Smith, Joshua D Milner, Dorothée Leroux, Cornelius F Boerkoel, Amy Klion, Melody C Carter, Tudor Groza, Damian Smedley, Melissa A Haendel, Chris Mungall, and Peter N Robinson. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. Nucleic Acids Research, 47(D1):D1018–D1027, 11 2018.

[23] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer,

Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12:2825–2830, November 2011.

[24] Viktor Pekar and Steffen Staab. Taxonomy learning: Factoring the structure of a taxonomy into a semantic classification decision. In Proceedings of the 19th International Conference on Computational Linguistics - Volume 1, COLING '02, page 1–7, USA, 2002. Association for Computational Linguistics.

[25] Catia Pesquita. Semantic similarity in the gene ontology. In Christophe Dessimoz and Nives Škunca, editors, The Gene Ontology Handbook, pages 161–173. Springer New York, New York, NY, 2017.

[26] Catia Pesquita, Daniel Faria, Hugo Bastos, André Falcao, and Francisco Couto. Evaluating GO-based semantic similarity measures. In Proceedings of the 10th Annual Bio-Ontologies Meeting, pages 37–40, Vienna, Austria, July 2007.

[27] Catia Pesquita, Daniel Faria, Hugo Bastos, António E N Ferreira, André O Falcão, and Francisco M Couto. Metrics for GO based protein semantic similarity: a systematic evaluation. BMC bioinformatics, 9 Suppl 5:S4, 2008.

[28] Catia Pesquita, Daniel Faria, André O. Falcão, Phillip Lord, and Francisco M. Couto. Semantic similarity in biomedical ontologies. PLOS Computational Biology, 5(7):1–12, 07 2009.

[29] R. Rada, H. Mili, E. Bicknell, and M. Blettner. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man, and Cybernetics, 19(1):17–30, Jan 1989.

[30] Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 1, IJCAI'95, page 448–453, San Francisco, CA, USA, 1995. Morgan Kaufmann Publishers Inc.

[31] Petar Ristoski and Heiko Paulheim. RDF2Vec: RDF graph embeddings for data mining. In Proceedings of ISWC 2016, pages 498–514, Cham, 2016. Springer.

[32] Omer Sagi and Lior Rokach. Ensemble learning: A survey. WIREs Data Mining and Knowledge Discovery, 8(4):e1249, 2018.

[33] Nuno Seco, Tony Veale, and Jer Hayes. An intrinsic information content metric for semantic similarity in wordnet. In Proceedings of the 16th European Conference on Artificial Intelligence, ECAI'04, page 1089–1090, NLD, 2004. IOS Press.

[34] Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf. Onto2vec: joint vector-based representation of biological entities and their ontology-based annotations. Bioinformatics, 34(13):i52—i60, July 2018.

[35] Fatima Zohra Smaili, Xin Gao, and Robert Hoehndorf. OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction. Bioinformatics, 35(12):2133–2140, 11 2018.

[36] Rita T Sousa, Sara Silva, and Catia Pesquita. Evolving knowledge graph similarity for supervised learning in complex biomedical domains. BMC Bioinformatics, 21(1):6, January 2020.

[37] Rita T Sousa, Sara Silva, and Catia Pesquita. The supervised semantic similarity toolkit. In European Semantic Web Conference, pages 42–46. Springer, 2022.

[38] The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Research, 47(D1):D330–D338, 11 2018.

[39] Ignacio Traverso, Maria-Esther Vidal, Benedikt Kämpgen, and York Sure-Vetter. GADES: A graph-based semantic similarity measure. In Proceedings of SEMANTiCS 2016, pages 101–104, New York, NY, USA, 2016. Association for Computing Machinery.

[40] Ignacio Traverso-Ribón and Maria-Esther Vidal. Garum: A semantic similarity measure based on machine learning and entity characteristics. In Database and Expert Systems Applications - Volume 11029, pages 169–183, Cham, 2018. Springer.

[41] Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In Proceedings of ICML 2016, page 2071–2080, New York, NY, USA, 2016. JMLR.org.

[42] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In Proceedings of the 28th AAAI Conference on Artificial Intelligence, volume 28, page 1112–1119, Québec City, Québec, Canada, 2014. AAAI Press.

[43] Patricia L Whetzel, Natalya F Noy, Nigam H Shah, Paul R Alexander, Csongor Nyulas, Tania Tudorache, and Mark 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(suppl_2):W541–W545, 2011.

[44] Bishan Yang, Wen tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases, 2015.

[45] Jiongmin Zhang, Ke Jia, Jinmeng Jia, and Ying Qian. An improved approach to infer protein-protein interaction based on a hierarchical vector space model. BMC bioinformatics, 19(1):1–14, 2018.

[46] Shu-Bo Zhang and Qiang-Rong Tang. Protein–protein interaction inference based on semantic similarity of gene ontology terms. Journal of Theoretical Biology, 401:30–37, 2016.

RITA T. SOUSA received an M.S. degree in Bioinformatics and Computational Biology from the University of Lisbon. She is currently pursuing a Ph.D. degree in Informatics at LASIGE, in the Faculty of Sciences of the University of Lisbon. Her research interests include knowledge graphs, machine learning and biomedical applications.

SARA SILVA is a Principal Investigator at the Faculty of Sciences of the University of Lisbon, and a member of the LASIGE research center. She is the author of more than 100 peer-reviewed publications, including the book Lectures On Intelligent Systems. Her research interests are mainly in machine learning with a strong emphasis on genetic programming, where she has contributed with several new methods, and applied them in projects of different domains such as remote sensing, biomedicine, and radiomics, among others. She is an Associate Editor of the GPEM, SWEVO, and ACM TELO journals and has held Program Chair, Track Chair, Editor-in-Chief and General Chair roles at prestigious genetic and evolutionary computation conferences including EuroGP and GECCO. In 2018, she received the EvoStar Award for Outstanding Contribution to Evolutionary Computation in Europe. She is the creator and developer of GPLAB—A Genetic Programming Toolbox for MATLAB, and co-creator of GSGP—A Geometric Semantic Genetic Programming Library.

CATIA PESQUITA is an Associate Professor at the Faculty of Sciences of the University of Lisbon, where she leads the Research Line of Excellence in Health and Biomedical Informatics at LASIGE. Her research work is dedicated to knowledge engineering and data mining, particularly in the biomedical and clinical domains, supported by her multidisciplinary background. She has made significant contributions in data analytics and integration with ontologies and knowledge graphs, producing over 60 peer-reviewed publications in high impact venues including PLoS Computational Biology, BMC Bioinformatics, Journal of Biomedical Semantics, the International Semantic Web Conference and the Extended Semantic Web Conference. She has led and collaborated in multiple national and international research projects. Her research team and collaborators develop AML, an award-winning software for ontology matching.