ISWC 2023

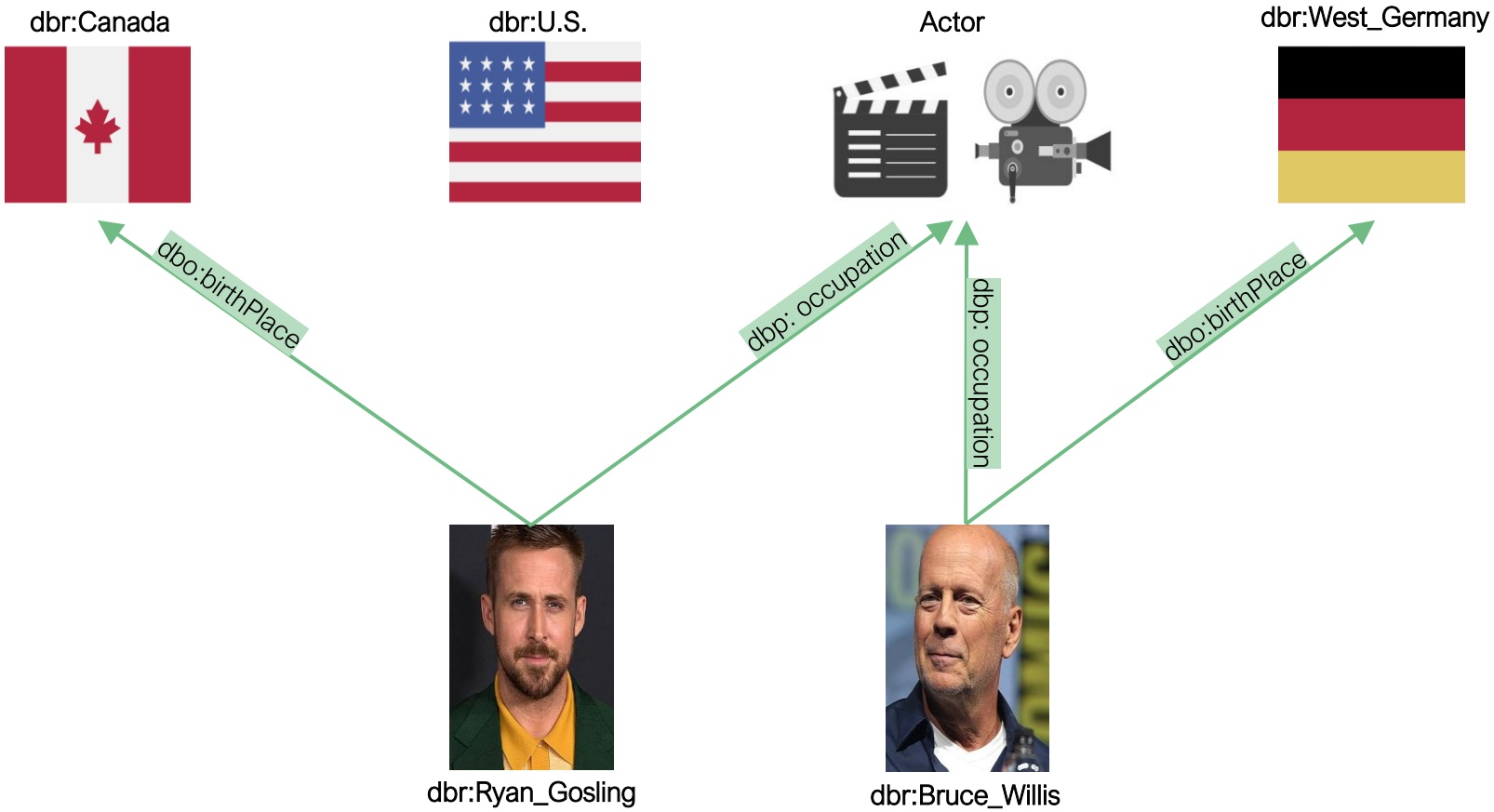
BIOMEDICAL KNOWLEDGE GRAPH EMBEDDINGS WITH NEGATIVE STATEMENTS

Rita T. Sousa[1], Sara Silva[1], Heiko Paulheim[2], Catia Pesquita[1]
[1]LASIGE, Faculdade de Ciências da Universidade de Lisboa, Portugal
[2]Data and Web Science Group, Universität Mannheim, Germany

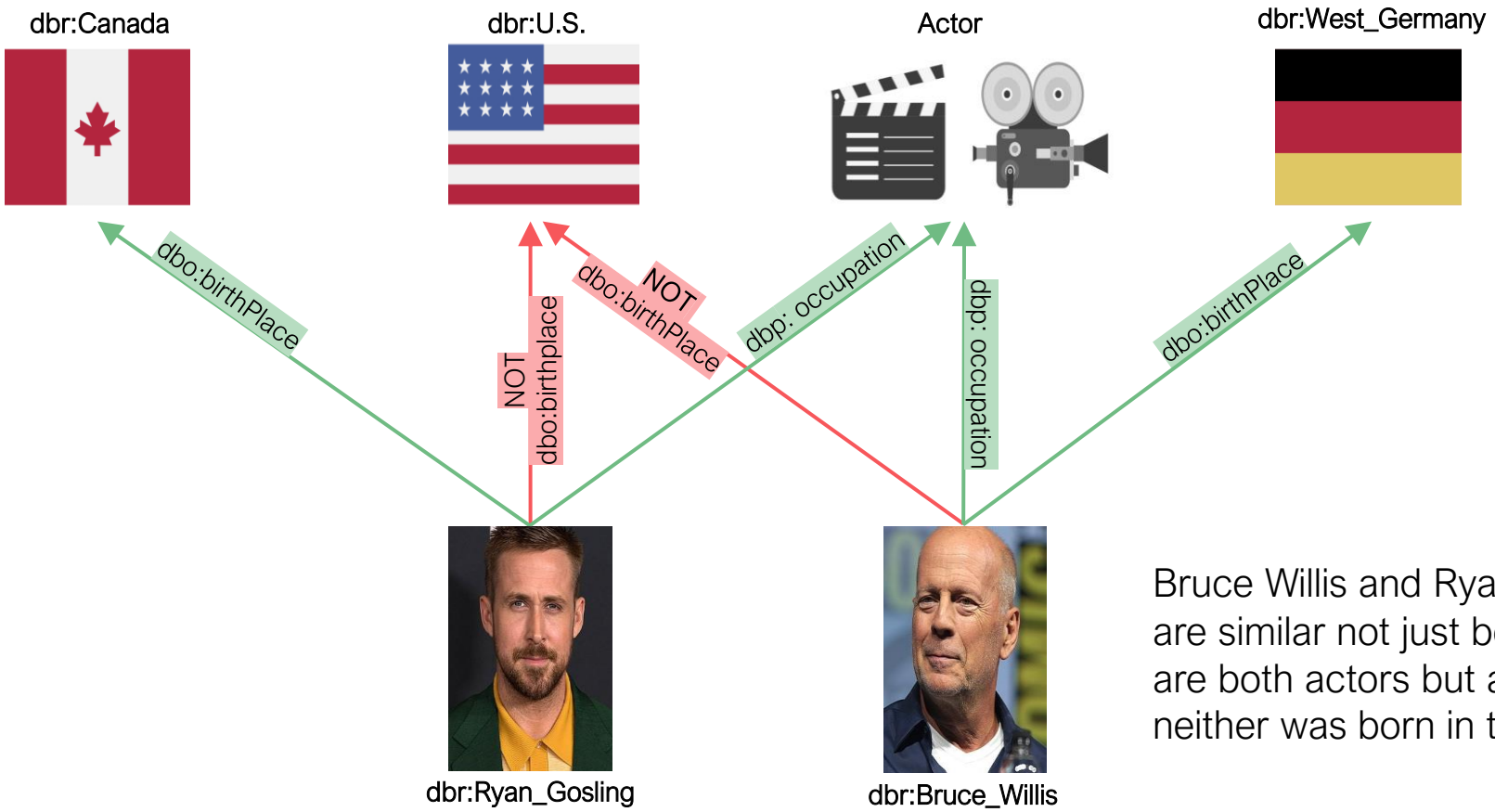
International Semantic Web Conference
6-10 November 2023

LASIGE driven by excellence · Ciências ULisboa · fct Fundação para a Ciência e a Tecnologia · UNIVERSITÄT MANNHEIM

The vast majority of knowledge graph (KG) relations are defined as positive statements.



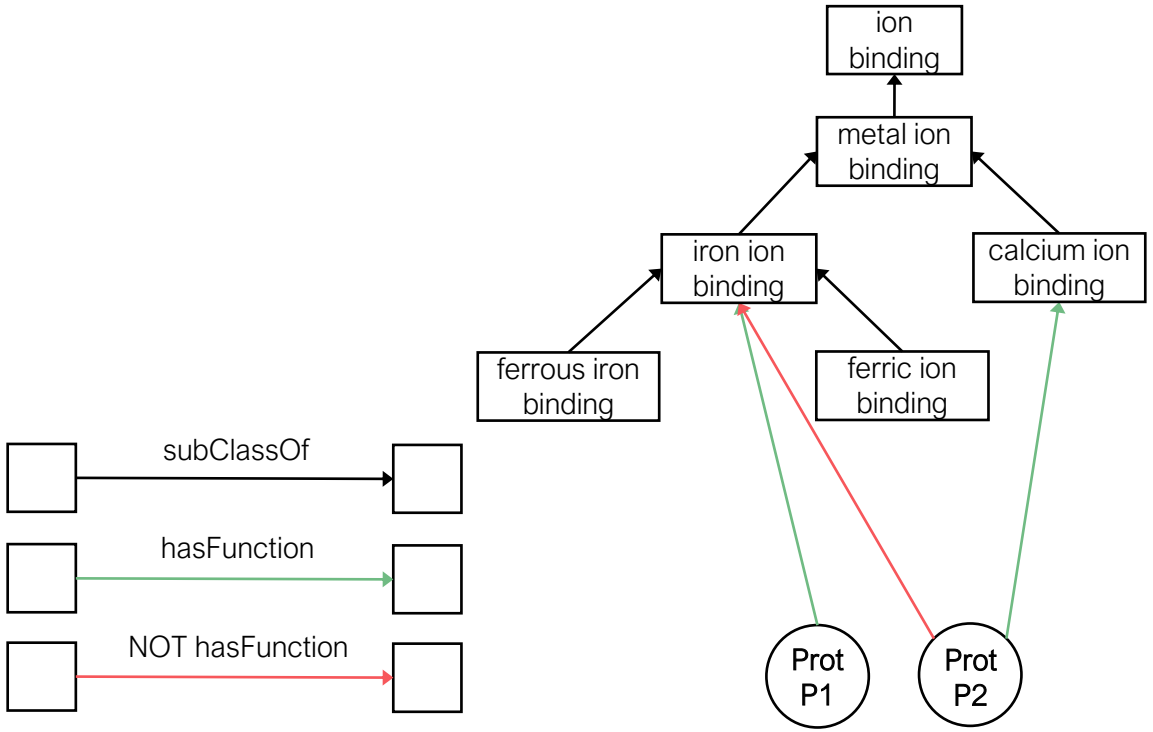
Explicitly considering negative statements improves the accuracy of representations.



Bruce Willis and Ryan Gosling are similar not just because they are both actors but also because neither was born in the U.S.

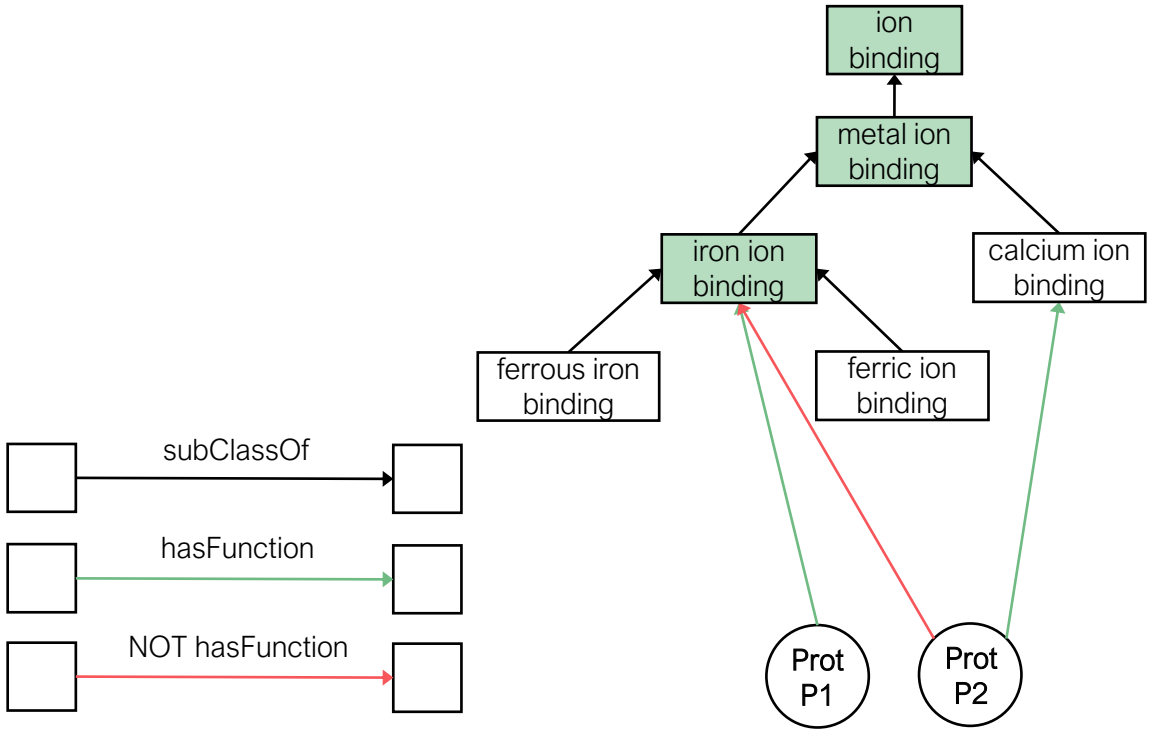
However, little attention has been given to the exploration of negative statements by KG embedding approaches.

In ontology-rich KGs, there is a difference between positive and negative statements regarding the implied inheritance of properties of the assigned class.



Differences in the inheritance of properties exhibited by the superclasses or subclasses of the assigned class.

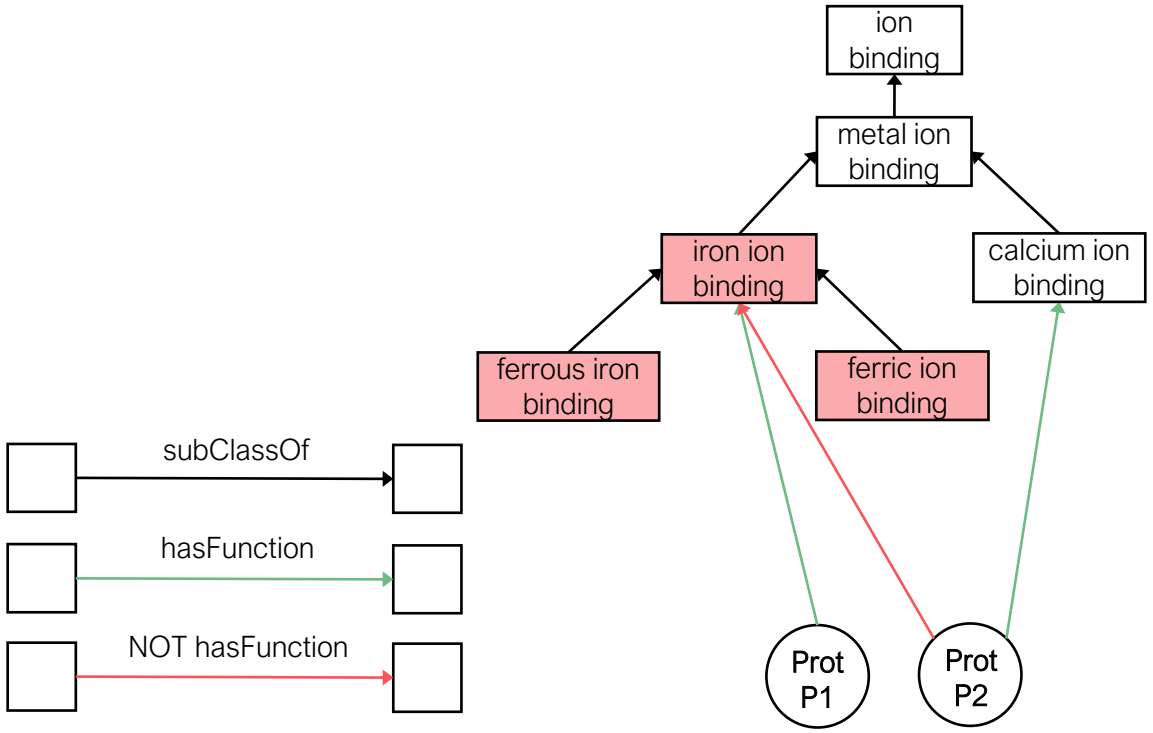
In ontology-rich KGs, there is a difference between positive and negative statements regarding the implied inheritance of properties of the assigned class.



Differences in the inheritance of properties exhibited by the superclasses or subclasses of the assigned class.

- A protein that performs *'iron ion binding'* also performs *'metal ion binding'*.

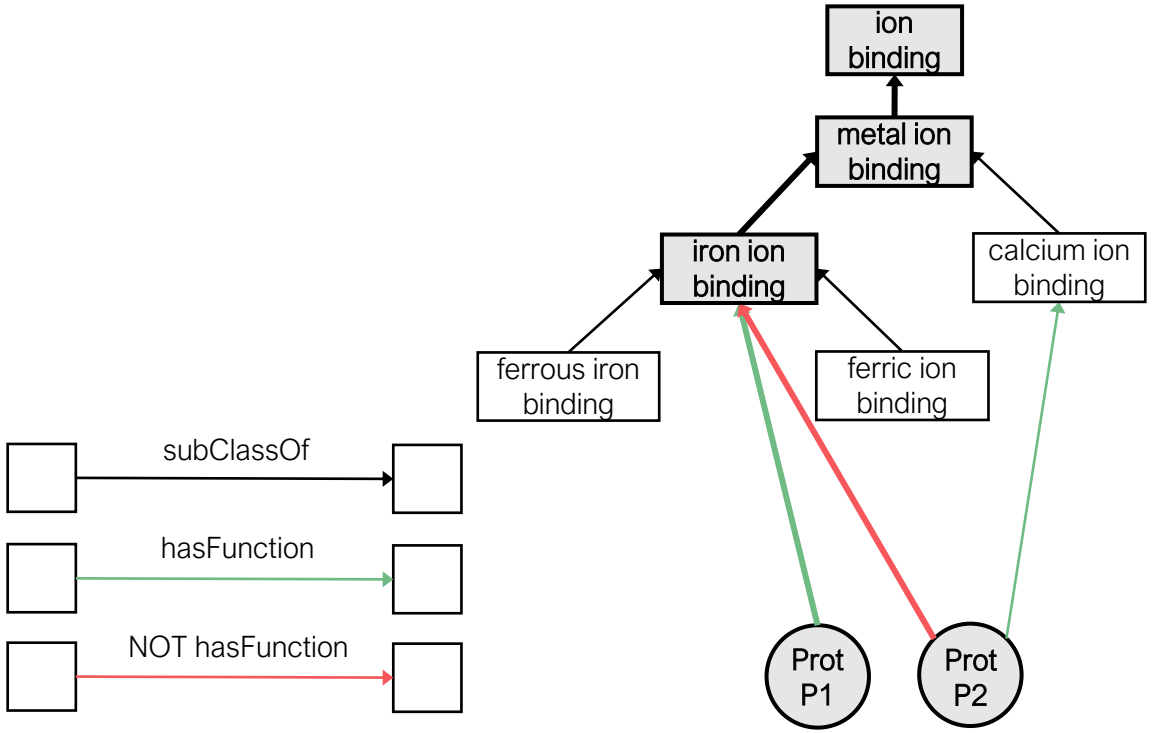
In ontology-rich KGs, there is a difference between positive and negative statements regarding the implied inheritance of properties of the assigned class.



Differences in the inheritance of properties exhibited by the superclasses or subclasses of the assigned class.

- A protein that performs *'iron ion binding'* also performs *'metal ion binding'*.
- A protein that does not perform *'iron ion binding'* also does not perform *'ferric ion binding'*, but there are no guarantees that it does not perform *'metal ion binding'*.

Since ontologies typically declare subclass axioms, the reverse inheritance of negative statements are not adequately explored by walk-based KG embedding methods.



Classical Random Walks:

Prot P1 > hasFunction > iron ion binding > subClassOf > metal ion binding > subClassOf > ion binding

Prot P2 > NOT hasFunction > iron ion binding > subClassOf > metal ion binding > subClassOf > ion binding

Challenges

(i) How can KG embedding methods distinguish between negative and positive statements?

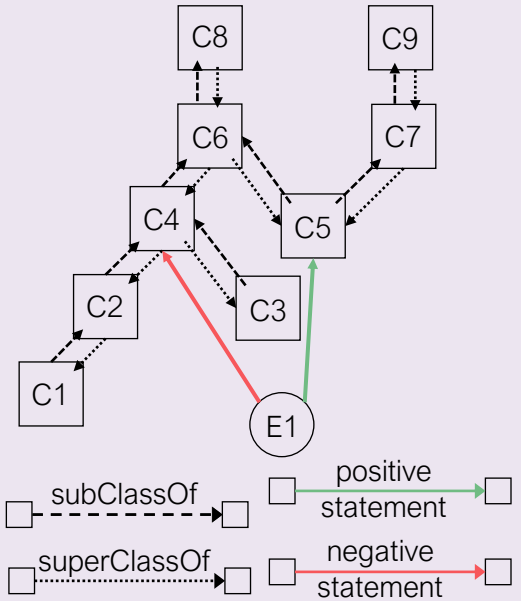
(ii) How can the reverse inheritance implied by negative statements be adequately explored by walk-based KG embedding methods?

TrueWalks

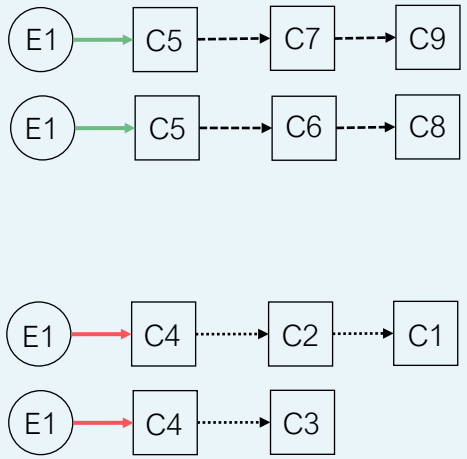
Novel method to generate random walks on ontology-rich KGs that can distinguish between positive and negative statements and consider the semantic implications of negation in KGs.

GitHub

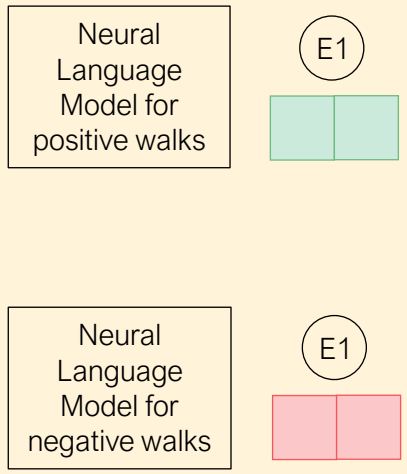
Creation of the RDF graph



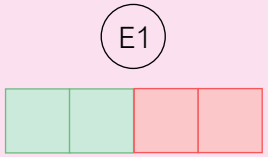
Random walk generation



Language model



Final representation



LASIGE driven by excellence · Ciências ULisboa · fct Fundação para a Ciência e a Tecnologia · UNIVERSITÄT MANNHEIM

TrueWalks

Creation of the RDF graph	Random walk generation	Language model	Final representation

The first step is the conversion of an ontology-rich KG into an RDF graph according to the OWL to RDF Graph Mapping guidelines.

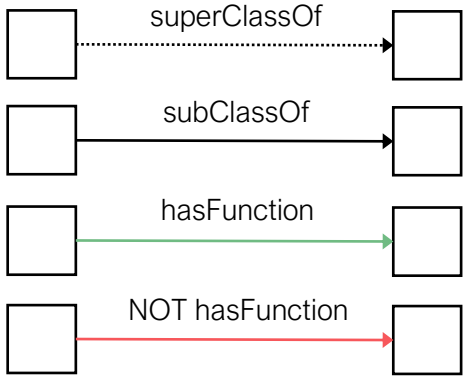
Negative statements are incorporated in the KG using **negative object property assertions** stating that the individual representing a biomedical entity is not connected by the object property expression to the individual representing an ontology class.

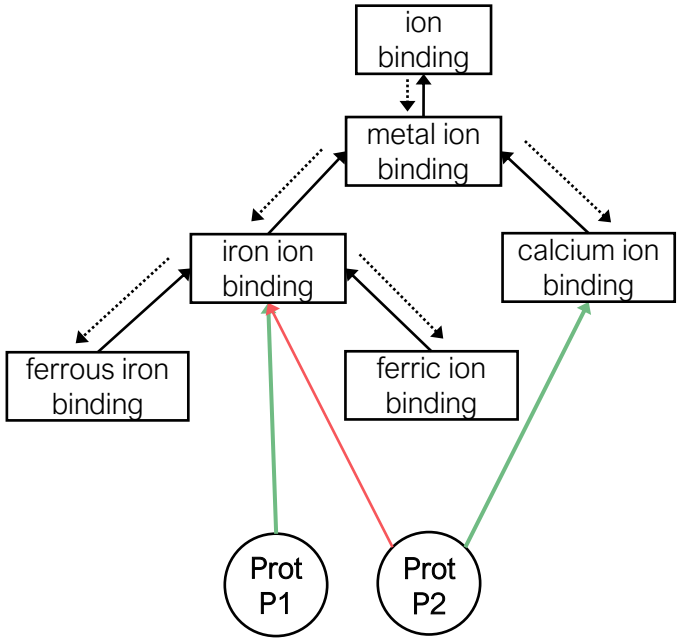
```
<owl:NamedIndividual rdf:about="http://purl.obolibrary.org/obo/GO_0048268">
    <rdf:type rdf:resource="http://purl.obolibrary.org/obo/GO_0048268"/>
</owl:NamedIndividual>
<rdf:Description>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#NegativePropertyAssertion"/>
    <owl:sourceIndividual rdf:resource="http://purl.obolibrary.org/obo/GO_0048268"/>
    <owl:assertionProperty
rdf:resource="http://purl.obolibrary.org/obo#has_function"/>
    <owl:targetIndividual rdf:resource="https://www.uniprot.org/uniprotkb/Q9BY11"/>
</rdf:Description>
```

TrueWalks

Creation of the RDF graph	Random walk generation	Language model	Final representation

Biased walks: a positive statement implies paths using subclass edges, whereas a negative statement uses superclass edges.

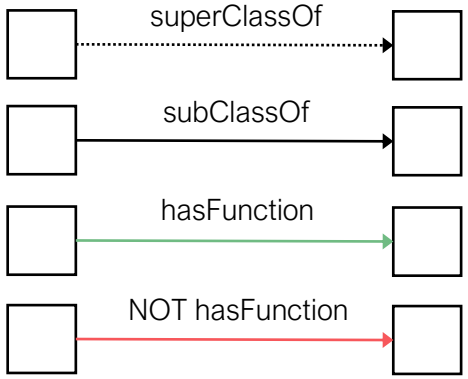


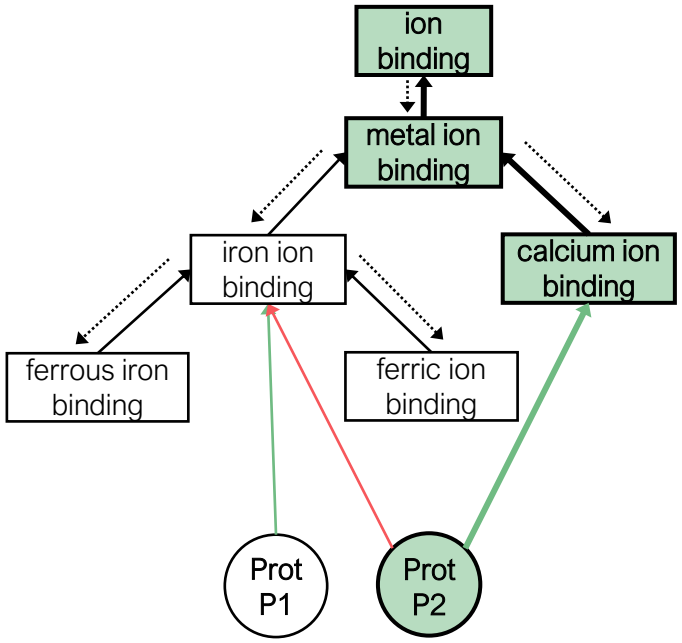


TrueWalks

Creation of the RDF graph	Random walk generation	Language model	Final representation

Biased walks: a positive statement implies paths using subclass edges, whereas a negative statement uses superclass edges.



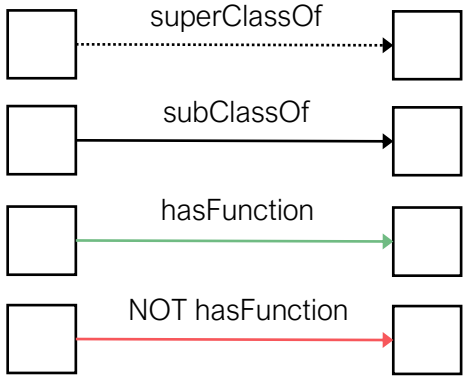


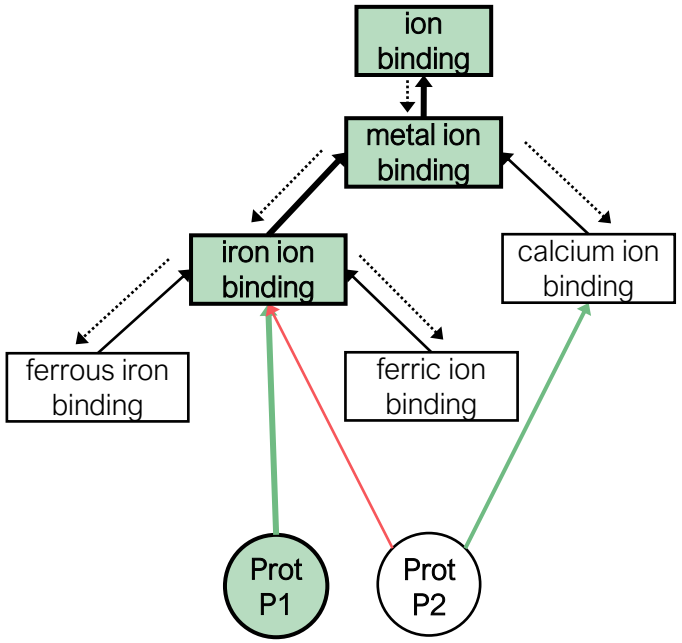
Prot P2 > hasFunction > calcium ion binding > subClassOf > metal ion binding > subClassOf > ion binding

TrueWalks

Creation of the RDF graph	Random walk generation	Language model	Final representation

Biased walks: a positive statement implies paths using subclass edges, whereas a negative statement uses superclass edges.





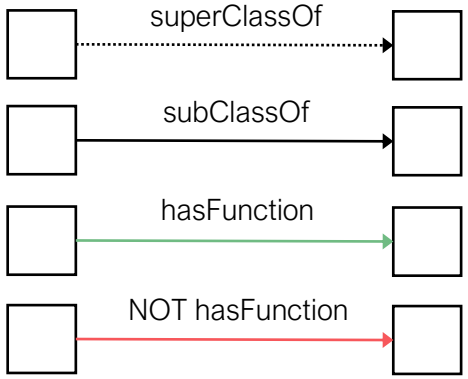
Prot P2 > hasFunction > calcium ion binding > subClassOf > metal ion binding > subClassOf > ion binding

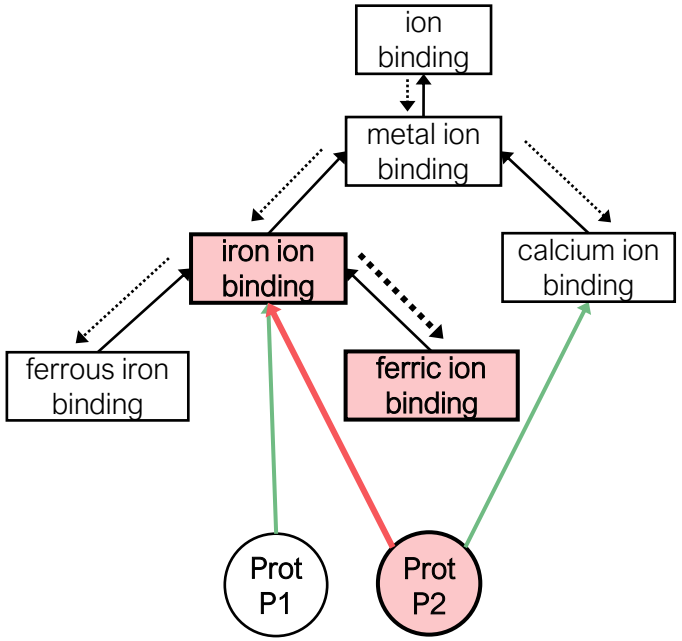
Prot P1 > hasFunction > iron ion binding > subClassOf > metal ion binding > subClassOf > ion binding

TrueWalks

Creation of the RDF graph	Random walk generation	Language model	Final representation

Biased walks: a positive statement implies paths using subclass edges, whereas a negative statement uses superclass edges.





Prot P1 > hasFunction > iron ion binding > subClassOf > metal ion binding > subClassOf > ion binding

Prot P2 > hasFunction > calcium ion binding > subClassOf > metal ion binding > subClassOf > ion binding

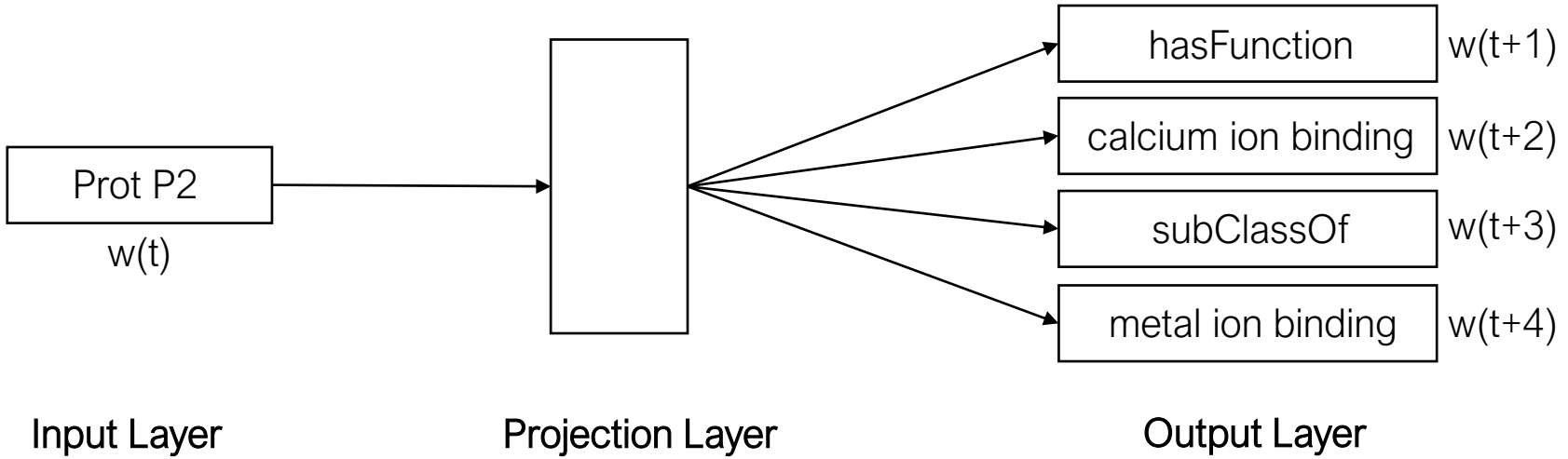
Prot P2 > NOT hasFunction > iron ion binding > superClassOf > ferric ion binding

TrueWalks

Creation of the RDF graph	Random walk generation	Language model	Final representation

Positive and negative walks are fed to neural language models to learn a dual latent representation. Two alternative neural language models are employed: skip-gram (TrueWalks) and structured skip-gram model (TrueWalksOA).

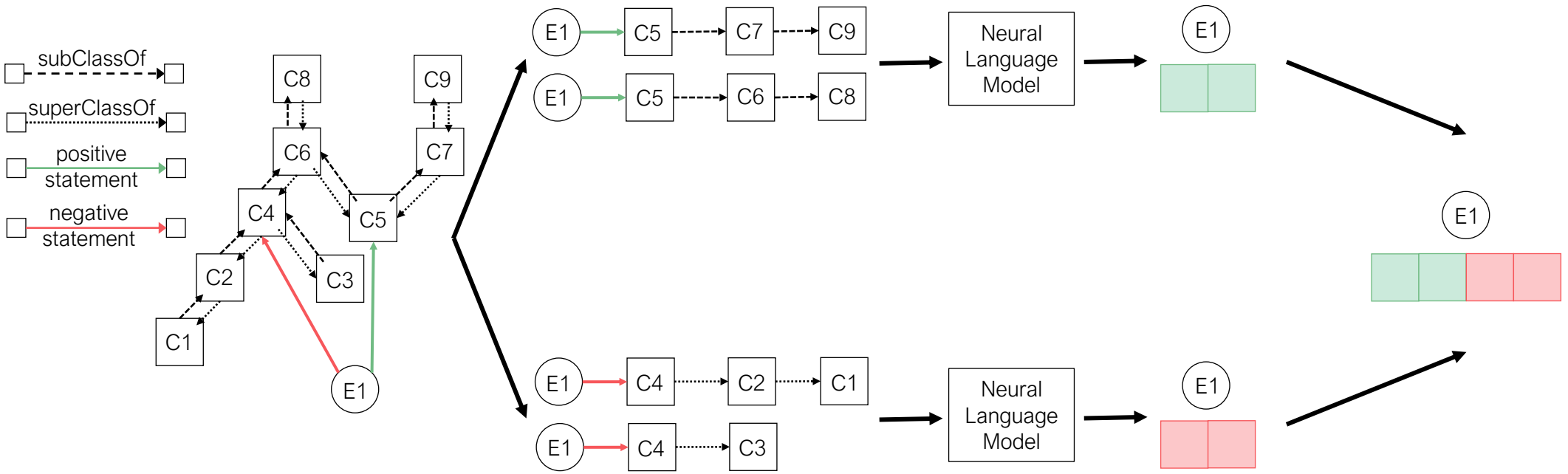
Corpus: Prot P2 hasFunction calcium ion binding subClassOf metal ion binding subClassOf ion binding



TrueWalks

Creation of the RDF graph	Random walk generation	Language model	Final representation

The two representations of each entity are combined using concatenation and produce a final representation.



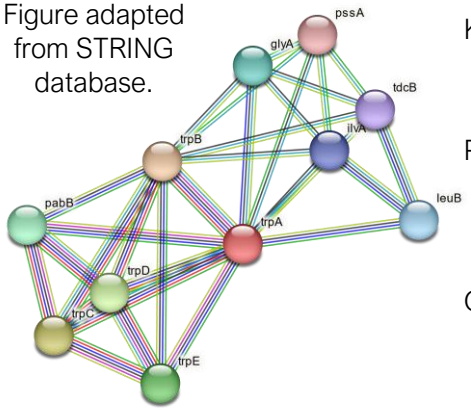
Data

TrueWalks is evaluated on two biomedical relation prediction tasks:

Protein-Protein Interaction (PPI) Prediction

- Target relations from STRING.
- Gene Ontology (GO) KG enriched with negative statements are used to describe proteins.

Figure adapted from STRING database.



Known interactions:
- from curated databases
- experimentally determined

Predicted interactions:
- gene neighborhood
- gene fusions
- gene co-occurrence

Others:
- text mining
- co-expression
- protein homology

Gene-Disease Association (GDA) Prediction

- Target relations from DisGeNET.
- Gene Ontology (GO) KG enriched with negative statements describe genes and Human Phenotype Ontology (HP) KG describe diseases.

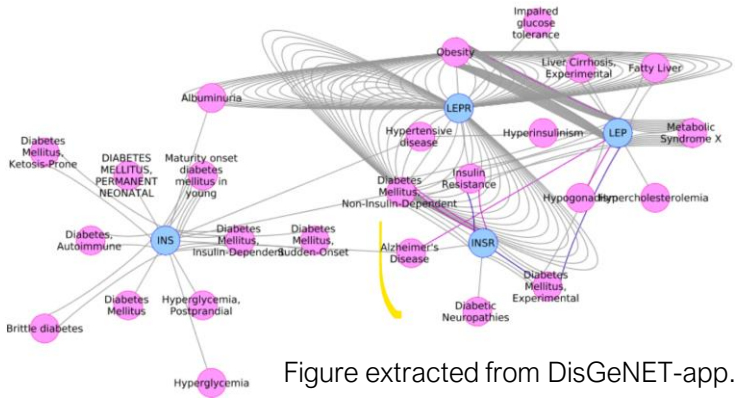


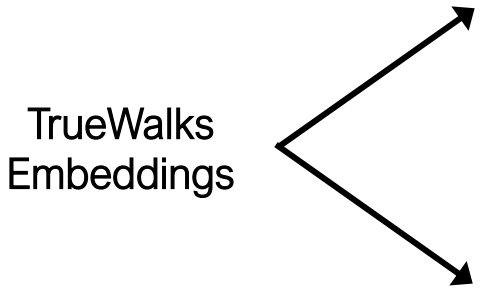
Figure extracted from DisGeNET-app.

	GO_{PPI}	GO_{GDA}	HP_{GDA}
Classes	50918	50918	17060
Instances	440	755	162
Positive Statements	7364	10631	4197
Negative Statements	8579	8966	225

Sousa, R. T., Silva, S., and Pesquita, C. (2023). Benchmark datasets for biomedical knowledge graphs with negative statements. In Workshop on Semantic Web solutions for large-scale biomedical data analytics (SemWebMeda) at ESWC 2023.

Experimental Setup

PPI and GDA prediction are relation prediction tasks. For PPI prediction, TrueWalks embeddings are employed as features for experiments (i) and (ii). For GDA prediction, since embeddings for genes and diseases are learned over two different KGs, TrueWalks embeddings are employed as features only for experiment (i).



(i) Relation prediction using Machine Learning:
KG embeddings of each entity in the pair are the features for a binary classifier.

(ii) Relation prediction using Semantic Similarity:
Comparison of the KG embeddings of each entity in the pair directly through a similarity metric.

Relation Prediction using Random Forest

	Method	PPI Prediction			GDA Prediction		
		P	R	F-score	P	R	F-score
P	TransE	0.553	0.546	0.554	0.533	0.538	0.531
	TransH	0.566	0.562	0.566	0.556	0.563	0.548
	TransR	0.620	0.607	0.616	0.594	0.600	0.592
	ComplEx	0.680	0.659	0.679	0.597	0.625	0.598
	distMult	0.765	0.737	0.754	0.585	0.600	0.575
	DeepWalk	0.813	**0.836**	0.822	0.618	**0.646**	0.629
	Node2vec	0.826	0.741	0.794	0.643	0.616	0.644
	metapath2vec	0.562	0.563	0.561	0.554	0.531	0.549
	OWL2Vec*	0.833	0.806	0.823	0.652	0.656	0.646
	RDF2Vec	0.831	0.826	0.828	0.623	0.625	0.615
P + N	TrueWalks	**0.870**	0.817	0.846	**0.667**	0.625	**0.661**
	TrueWalksOA	0.868	**0.836**	**0.858**	0.661	0.616	0.654

Table 1: Median scores using Monte Carlo 30 CV for both PPI and GDA prediction. P stands for KG which contains only positive statements. P+N refers to the KG where, in addition to the positive statements, negative statements were added with a new relationship.

- Vector representations are combined using the Hadamard operator and are then fed into a Random Forest algorithm.
- Negative statements produce more accurate representations of entities, which allow a better distinction of true positives from false positives.

Relation Prediction using Random Forest

	Method	PPI Prediction			GDA Prediction		
		P	R	F-score	P	R	F-score
P+N	TransE	0.584	0.582	0.585	0.597	0.585	0.586
	TransH	0.573	0.572	0.570	0.563	0.554	0.554
	TransR	0.722	0.678	0.704	0.633	0.625	0.630
	ComplEx	0.750	0.720	0.740	0.549	0.545	0.545
	distMult	0.813	0.740	0.784	0.530	0.523	0.534
	DeepWalk	0.843	0.834	0.841	0.615	0.646	0.630
	Node2vec	0.847	0.734	0.798	0.614	0.594	0.621
	metapath2vec	0.557	0.569	0.558	0.527	0.531	0.522
	OWL2Vec*	0.860	0.812	0.840	0.654	0.600	0.645
	RDF2Vec	0.847	**0.844**	0.845	0.625	**0.661**	0.630
	TrueWalks	**0.870**	0.817	0.846	**0.667**	0.625	**0.661**
	TrueWalksOA	0.868	0.836	**0.858**	0.661	0.616	0.654

Table 2: Median scores using Monte Carlo 30 CV for both PPI and GDA prediction. P+N refers to the KG which, in addition to the positive statements, negative statements were added with a new relationship.

- The added information given by negative statements generally improves the performance of most KG embedding methods.
- TrueWalks improve on precision and F-measure for both tasks when compared with the state-of-the-art methods.

Relation Prediction using Semantic Similarity

The semantic similarity is computed as the cosine similarity between the vectors of each entity in a pair.

	Method	Hits@10	Hits@100	MeanRank	AUC
P	DeepWalk	0.125	0.380	35.406	0.847
	Node2vec	0.163	0.375	37.275	0.827
	OWL2Vec*	0.152	0.386	33.192	0.860
	RDF2Vec	0.133	0.391	32.419	0.870
P+N	DeepWalk	0.148	0.383	35.365	0.849
	Node2vec	**0.166**	0.389	34.305	0.840
	OWL2Vec*	0.160	0.397	32.234	0.869
	RDF2Vec	0.155	0.401	30.281	0.879
	TrueWalks	0.161	0.392	32.089	0.869
	TrueWalksOA	**0.166**	**0.407**	**28.128**	**0.889**

Table 3: Performance for PPI prediction using cosine similarity.

Negative statements should not be ignored

- TrueWalks demonstrates the potential of designing artificial intelligence approaches that explore negative statements.
- TrueWalks can be generalized to other biomedical applications where negative statements play a decisive role, such as predicting disease-related phenotypes or performing differential diagnosis.

GitHub

https://github.com/liseda-lab/TrueWalks

risousa@ciencias.ulisboa.pt

@RitaTorresSousa