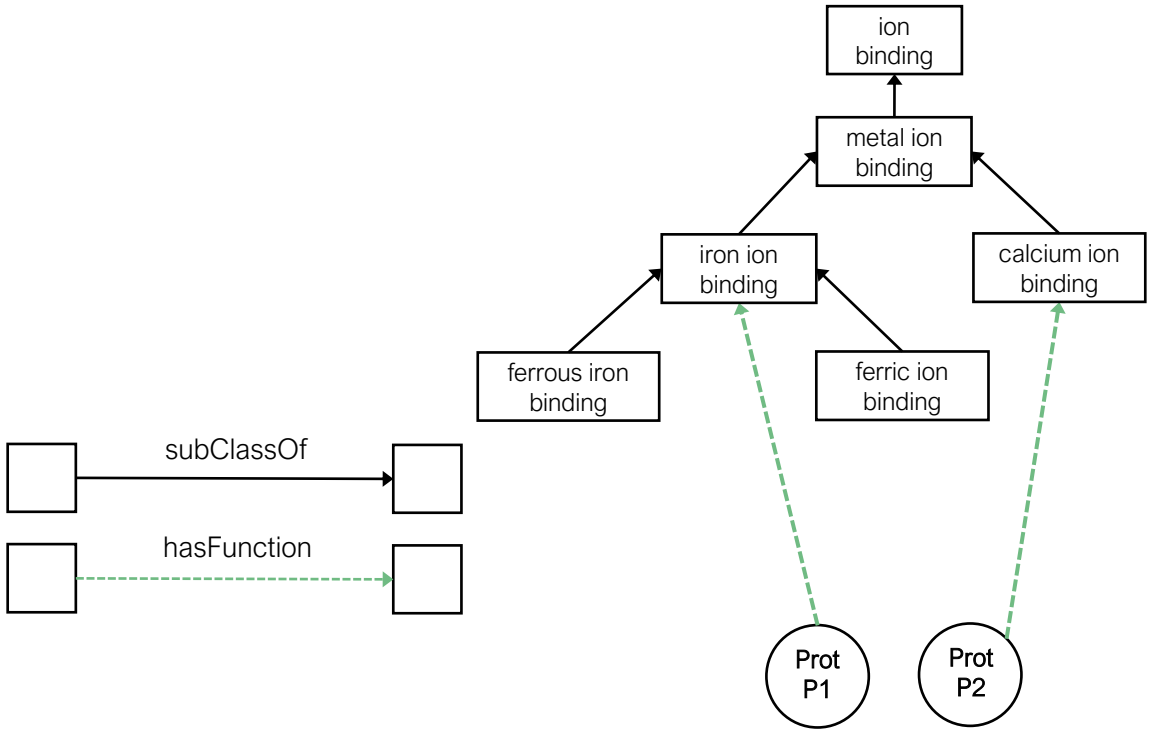
ESWC 2023

BENCHMARK DATASETS FOR BIOMEDICAL KNOWLEDGE GRAPHS WITH NEGATIVE STATEMENTS

Rita T. Sousa, Sara Silva, Catia Pesquita
LASIGE, Faculdade de Ciências da Universidade de Lisboa

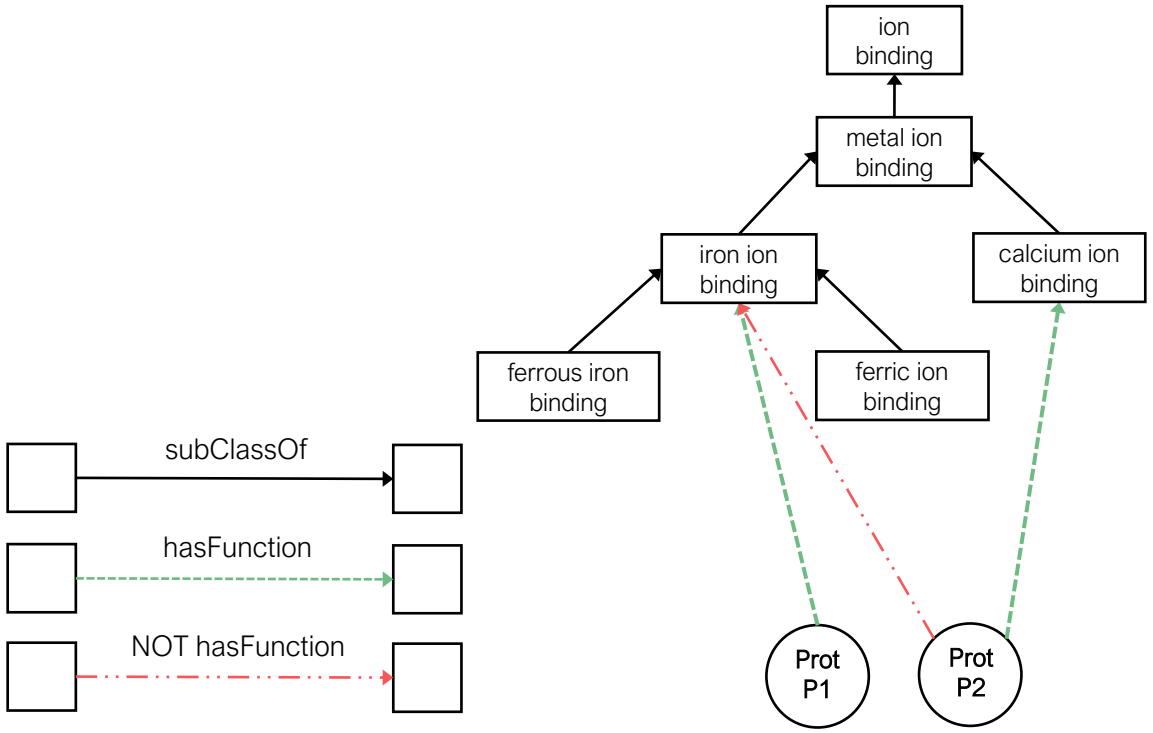
6th Workshop on SeWeBMeDA
29th of May 2023

MOST BIOMEDICAL KNOWLEDGE GRAPHS (KGS) USE ONTOLOGIES AS A BACKBONE TO DESCRIBE ENTITIES THROUGH ONTOLOGY-BASED ANNOTATION.



Biomedical entities can be described through positive statements that link them to an ontology class.

NEGATIVE STATEMENTS INDICATE THAT A BIOMEDICAL ENTITY IS NOT DESCRIBED BY AN ONTOLOGY CLASS CAN HELP COMPLETE SEMANTIC REPRESENTATION OF BIOMEDICAL ENTITIES.



There is a difference between a positive and a negative regarding the implied inheritance of properties of the assigned class:

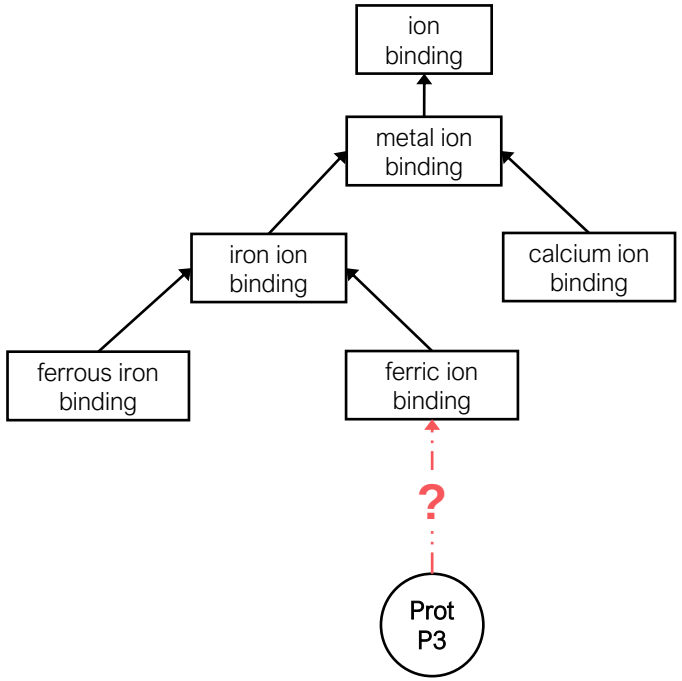
- A protein that performs *'iron ion binding'* also performs *'metal ion binding'*.
- A protein that does not perform *'iron ion binding'* also does not perform *'ferric ion binding'*, but there are no guarantees that it does not perform *'iron ion binding'*.

NEGATIVE STATEMENTS CAN BE INCORPORATED USING NEGATIVE OBJECT PROPERTY ASSERTIONS.

A negative object property assertion to state that the individual representing a biomedical entity is not connected by the object property expression to the individual representing an ontology class.

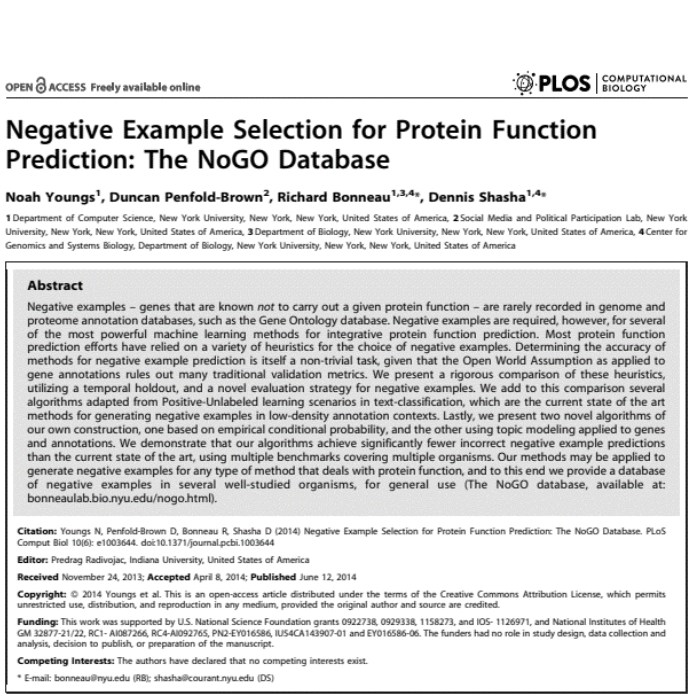
```
<owl:NamedIndividual rdf:about="http://purl.obolibrary.org/obo/GO_0048268">
    <rdf:type rdf:resource="http://purl.obolibrary.org/obo/GO_0048268"/>
</owl:NamedIndividual>
<rdf:Description>
    <rdf:type rdf:resource="http://www.w3.org/2002/07/owl#NegativePropertyAssertion"/>
    <owl:sourceIndividual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

THE LACK OF NEGATIVE STATEMENTS IS A PROBLEM SINCE BIOMEDICAL ONTOLOGY ANNOTATIONS RESIDE UNDER THE OPEN WORLD ASSUMPTION.



The lack of negative statements can lead to confusion regarding whether the absence of a positive statement is due to a lack of knowledge or the actual absence of the relationship.

SEVERAL METHODOLOGIES TACKLE THE PROBLEM OF THE LACK OF NEGATIVE STATEMENTS.



OPEN ACCESS Freely available online

PLOS | COMPUTATIONAL BIOLOGY

Negative Example Selection for Protein Function Prediction: The NoGO Database

Noah Youngs[1], Duncan Penfold-Brown[2], Richard Bonneau[1,3,4]*, Dennis Shasha[1,4]*

1 Department of Computer Science, New York University, New York, New York, United States of America, 2 Social Media and Political Participation Lab, New York University, New York, New York, United States of America, 3 Department of Biology, New York University, New York, New York, United States of America, 4 Center for Genomics and Systems Biology, Department of Biology, New York University, New York, New York, United States of America

Abstract

Negative examples – genes that are known *not* to carry out a given protein function – are rarely recorded in genome and proteome annotation databases, such as the Gene Ontology database. Negative examples are required, however, for several of the most powerful machine learning methods for integrative protein function prediction. Most protein function prediction efforts have relied on a variety of heuristics for the choice of negative examples. Determining the accuracy of methods for negative example prediction is itself a non-trivial task, given that the Open World Assumption as applied to gene annotations rules out many traditional validation metrics. We present a rigorous comparison of these heuristics, utilizing a temporal holdout, and a novel evaluation strategy for negative examples. We add to this comparison several algorithms adapted from Positive-Unlabeled learning scenarios in text-classification, which are the current state of the art methods for generating negative examples in low-density annotation contexts. Lastly, we present two novel algorithms of our own construction, one based on empirical conditional probability, and the other using topic modeling applied to genes and annotations. We demonstrate that our algorithms achieve significantly fewer incorrect negative example predictions than the current state of the art, using multiple benchmarks covering multiple organisms. Our methods may be applied to generate negative examples for any type of method that deals with protein function, and to this end we provide a database of negative examples in several well-studied organisms, for general use (The NoGO database, available at: bonneaulab.bio.nyu.edu/nogo.html).

Citation: Youngs N, Penfold-Brown D, Bonneau R, Shasha D (2014) Negative Example Selection for Protein Function Prediction: The NoGO Database. PLoS Comput Biol 10(6): e1003644. doi:10.1371/journal.pcbi.1003644

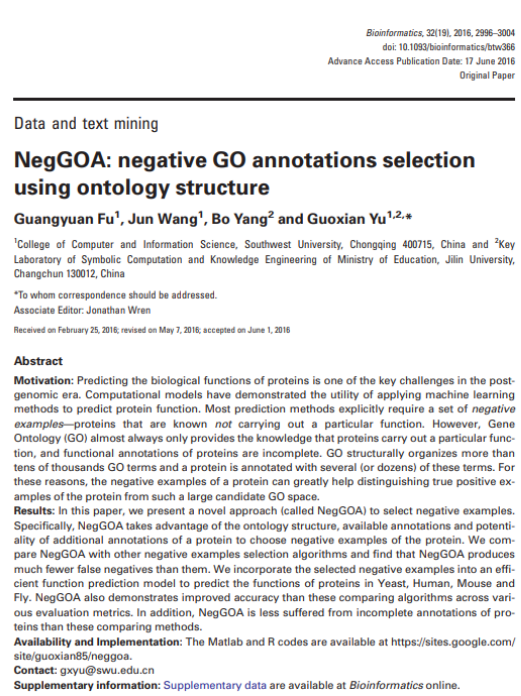
Editor: Predrag Radivojac, Indiana University, United States of America

Received November 24, 2013; **Accepted** April 8, 2014; **Published** June 12, 2014

Funding: This work was supported by U.S. National Science Foundation grants 0922738, 0929338, 1158273, and IOS- 1126971, and National Institutes of Health GM 32877-21/22, RC1- AI087266, RC4-AI092765, PN2-EY016586, IU54CA143907-01 and EY016586-06. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: bonneau@nyu.edu (RB); shasha@courant.nyu.edu (DS)



Bioinformatics, 32(19), 2016, 2996–3004
doi: 10.1093/bioinformatics/btw366
Advance Access Publication Date: 17 June 2016
Original Paper

OXFORD

Data and text mining

NegGOA: negative GO annotations selection using ontology structure

Guangyuan Fu[1], Jun Wang[1], Bo Yang[2] and Guoxian Yu[1,2,*]

[1]College of Computer and Information Science, Southwest University, Chongqing 400715, China and [2]Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on February 25, 2016; revised on May 7, 2016; accepted on June 1, 2016

Abstract

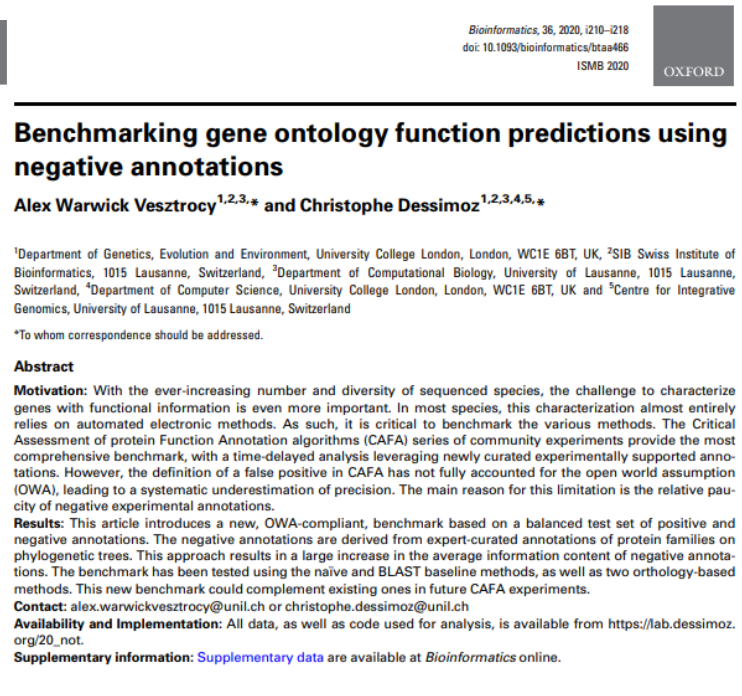
Motivation: Predicting the biological functions of proteins is one of the key challenges in the post-genomic era. Computational models have demonstrated the utility of applying machine learning methods to predict protein function. Most prediction methods explicitly require a set of *negative examples*—proteins that are known *not* carrying out a particular function. However, Gene Ontology (GO) almost always only provides the knowledge that proteins carry out a particular function, and functional annotations of proteins are incomplete. GO structurally organizes more than tens of thousands GO terms and a protein is annotated with several (or dozens) of these terms. For these reasons, the negative examples of a protein can greatly help distinguishing true positive examples of the protein from such a large candidate GO space.

Results: In this paper, we present a novel approach (called NegGOA) to select negative examples. Specifically, NegGOA takes advantage of the ontology structure, available annotations and potentiality of additional annotations of a protein to choose negative examples of the protein. We compare NegGOA with other negative examples selection algorithms and find that NegGOA produces much fewer false negatives than them. We incorporate the selected negative examples into an efficient function prediction model to predict the functions of proteins in Yeast, Human, Mouse and Fly. NegGOA also demonstrates improved accuracy than these comparing algorithms across various evaluation metrics. In addition, NegGOA is less suffered from incomplete annotations of proteins than these comparing methods.

Availability and Implementation: The Matlab and R codes are available at https://sites.google.com/site/guoxian85/neggoa.

Contact: gxyu@swu.edu.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.



Bioinformatics, 36, 2020, i210–i218
doi: 10.1093/bioinformatics/btaa466
ISMB 2020

OXFORD

Benchmarking gene ontology function predictions using negative annotations

Alex Warwick Vesztrocy[1,2,3,*] and Christophe Dessimoz[1,2,3,4,5,*]

[1]Department of Genetics, Evolution and Environment, University College London, London, WC1E 6BT, UK, [2]SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland, [3]Department of Computational Biology, University of Lausanne, 1015 Lausanne, Switzerland, [4]Department of Computer Science, University College London, London, WC1E 6BT, UK and [5]Centre for Integrative Genomics, University of Lausanne, 1015 Lausanne, Switzerland

*To whom correspondence should be addressed.

Abstract

Motivation: With the ever-increasing number and diversity of sequenced species, the challenge to characterize genes with functional information is even more important. In most species, this characterization almost entirely relies on automated electronic methods. As such, it is critical to benchmark the various methods. The Critical Assessment of protein Function Annotation algorithms (CAFA) series of community experiments provide the most comprehensive benchmark, with a time-delayed analysis leveraging newly curated experimentally supported annotations. However, the definition of a false positive in CAFA has not fully accounted for the open world assumption (OWA), leading to a systematic underestimation of precision. The main reason for this limitation is the relative paucity of negative experimental annotations.

Results: This article introduces a new, OWA-compliant, benchmark based on a balanced test set of positive and negative annotations. The negative annotations are derived from expert-curated annotations of protein families on phylogenetic trees. This approach results in a large increase in the average information content of negative annotations. The benchmark has been tested using the naïve and BLAST baseline methods, as well as two orthology-based methods. This new benchmark could complement existing ones in future CAFA experiments.

Contact: alex.warwickvesztrocy@unil.ch or christophe.dessimoz@unil.ch

Availability and Implementation: All data, as well as code used for analysis, is available from https://lab.dessimoz.org/20_not.

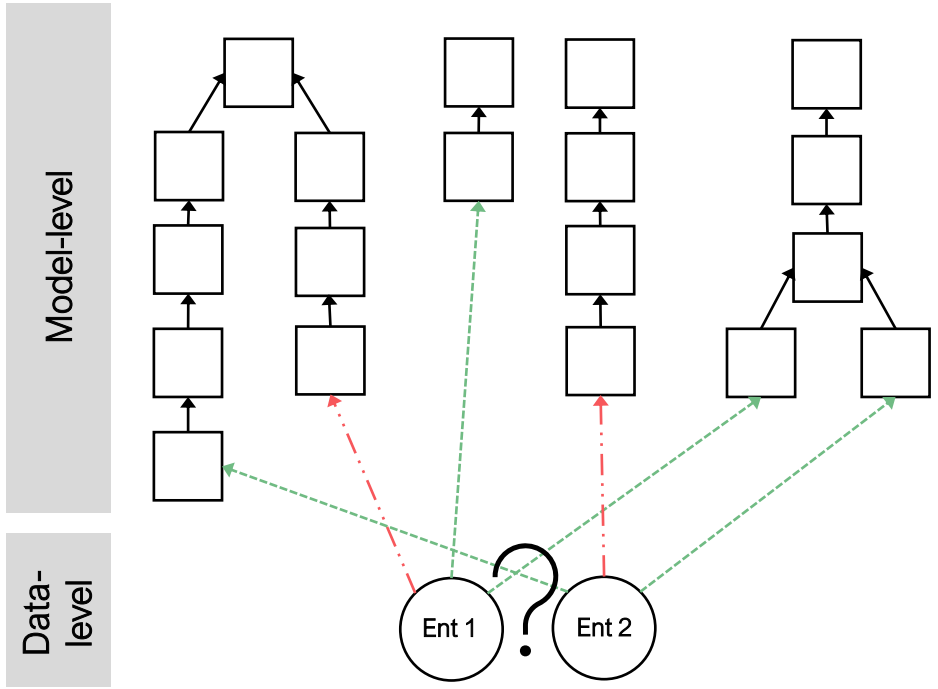
Supplementary information: Supplementary data are available at *Bioinformatics* online.

NO BENCHMARK DATASETS HAVE BEEN ESTABLISHED TO EVALUATE LEARNING TASKS OVER THE KGS WITH NEGATIVE STATEMENTS.

BUILDING BENCHMARK DATASETS

Collection of datasets that work over 2 enriched KGs for 3 relation prediction tasks:

- Protein-Protein Interaction (PPI) prediction;
- Gene-Disease Association (GDA) prediction;
- Disease prediction.



BUILDING BENCHMARK DATASETS

Methodology

Constructing KGs

Input:

- Ontology file in OWL format;
- Annotations file.

Output:

- KG in OWL format.

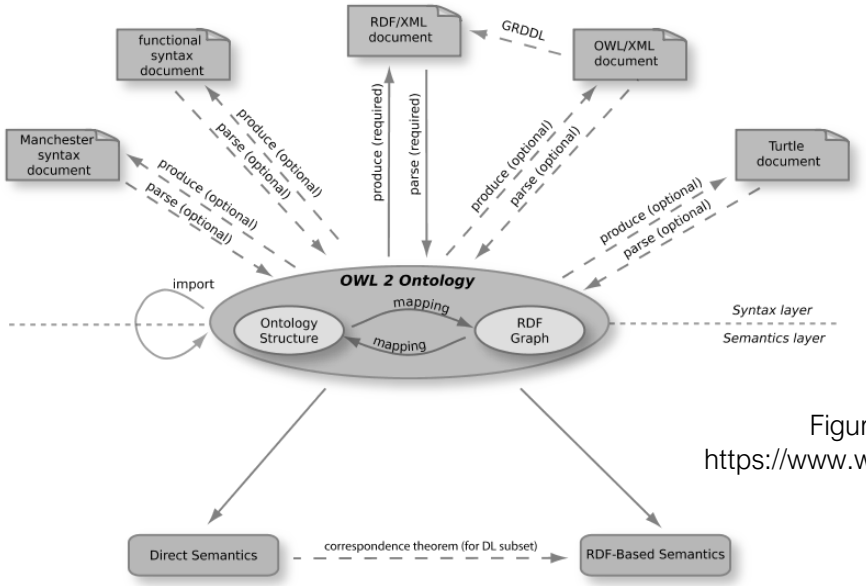
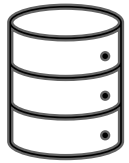


Figure extracted from https://www.w3.org/TR/owl2-overview/.

BUILDING BENCHMARK DATASETS

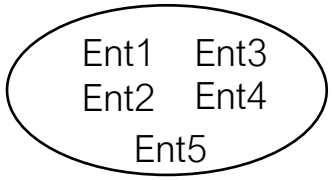
Methodology

Extracting entity pairs



Positive Pairs

Ent1	Ent2	1
Ent2	Ent3	1
Ent4	Ent5	1



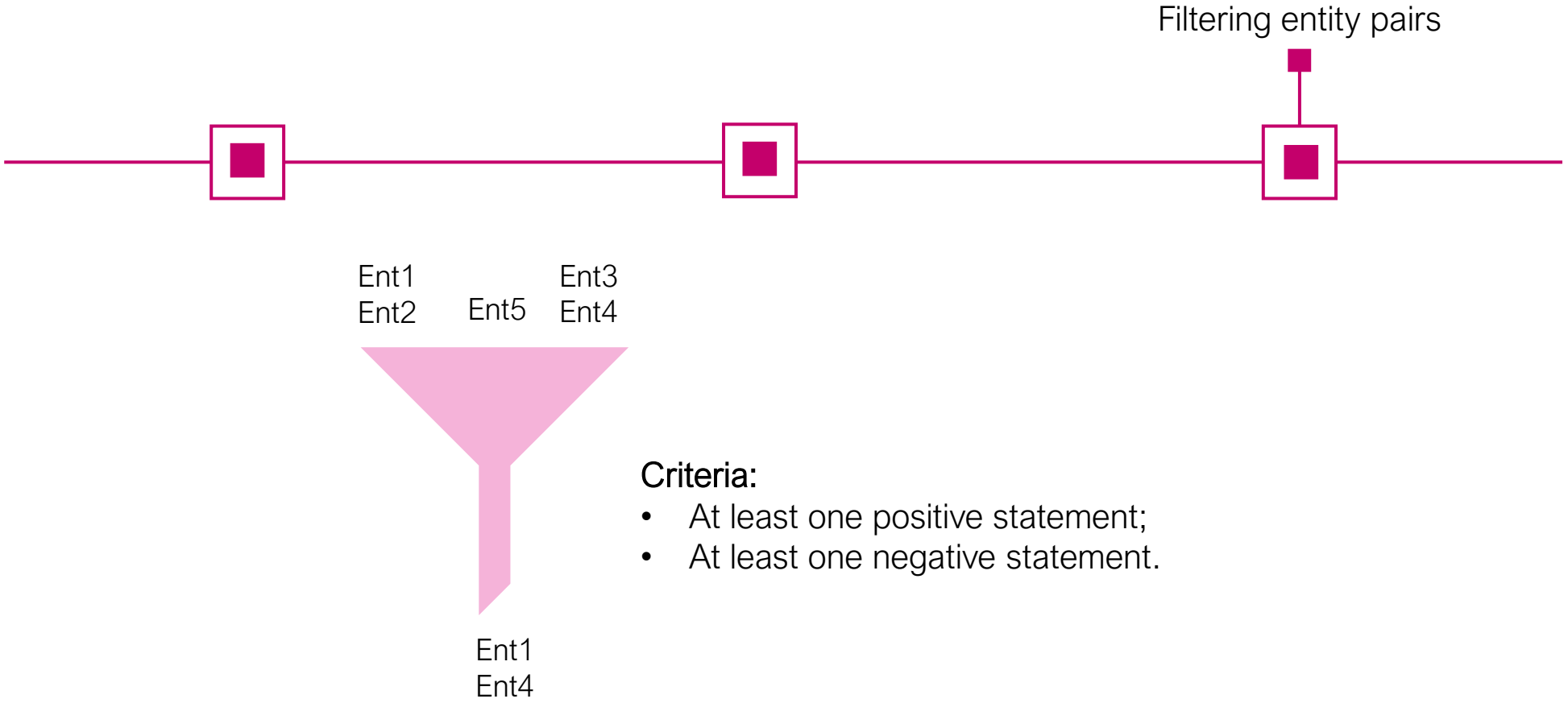
Negative Sampling

Negative Pairs

Ent1	Ent3	0
Ent5	Ent3	0
Ent2	Ent4	0

BUILDING BENCHMARK DATASETS

Methodology



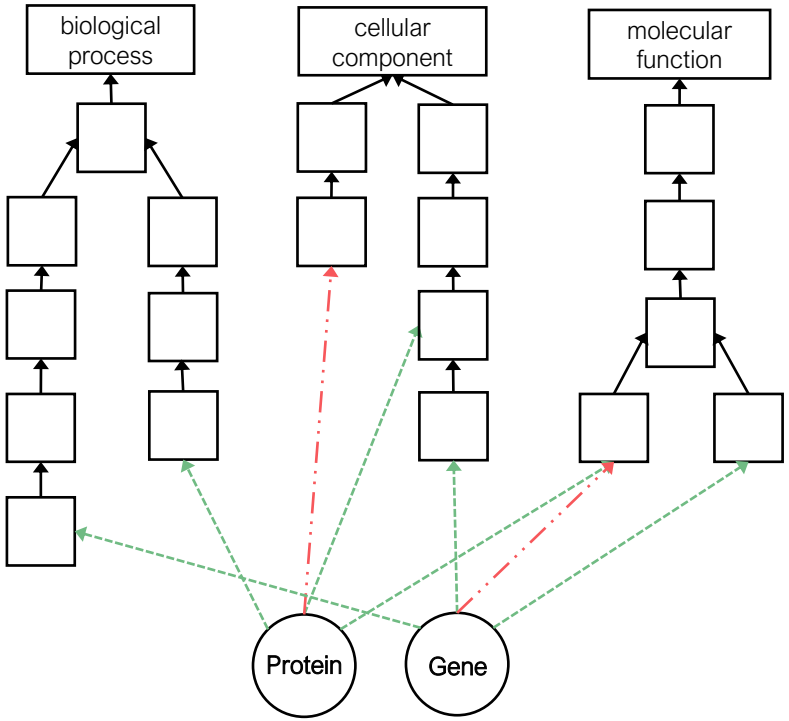
Criteria:

- At least one positive statement;
- At least one negative statement.

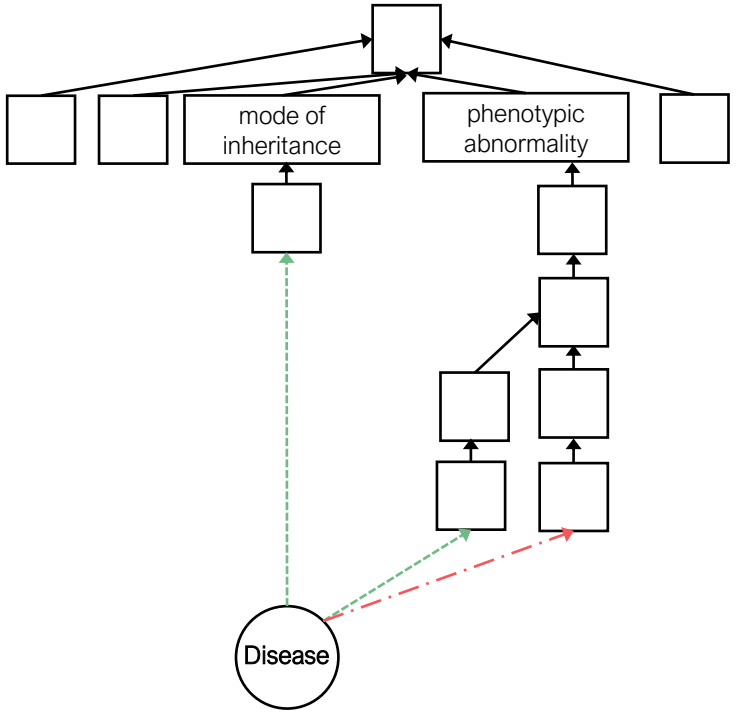
BUILDING BENCHMARK DATASETS

Biomedical knowledge graphs

Gene Ontology (GO) contains information about gene products functions.



Human Phenotype Ontology (HP) describes phenotypes and human diseases.



	GO	HP
Classes	50918	17060
Literals and blank nodes	532373	442246
Edges	1425102	1082859

BUILDING BENCHMARK DATASETS

PPI Prediction

Predicting PPIs is a fundamental task for understanding biological systems.

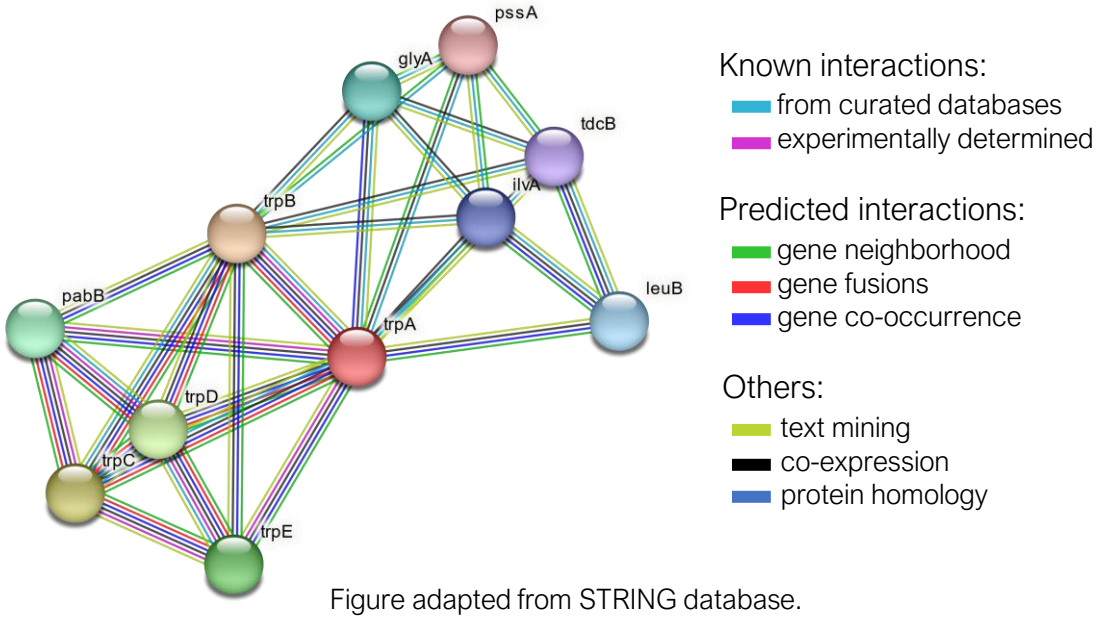


Figure adapted from STRING database.

Positive pairs extracted from STRING:

- curated or experimentally determined interactions with a confidence score > 0.950;
- proteins with at least one positive and one negative statement for a GO class.

Instances	440
Positive Pairs	1024
Negative Pairs	1024
Positive Statements	7364
Negative Statements	8579

BUILDING BENCHMARK DATASETS

GDA Prediction

Knowing GDAs is crucial to understanding the disease mechanisms and identifying potential therapeutic targets.

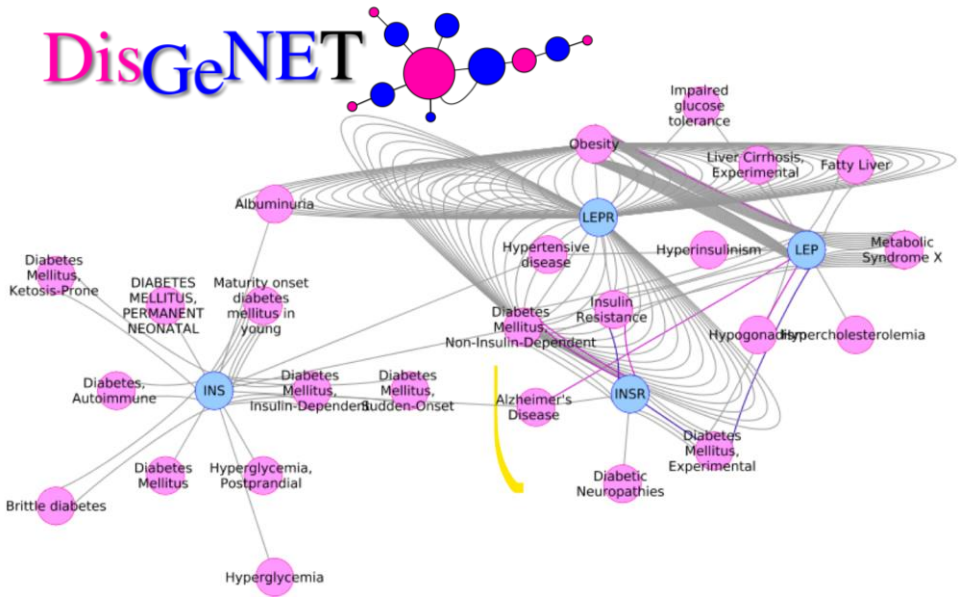


Figure extracted from DisGeNET-app.

Positive pairs extracted from DisGeNET:

- the source did not rely on the databases Uniprot, OMIM, or Orphanet to avoid data leakage;
- genes and diseases with at least one positive and one negative statement for a GO class and HP class, respectively;

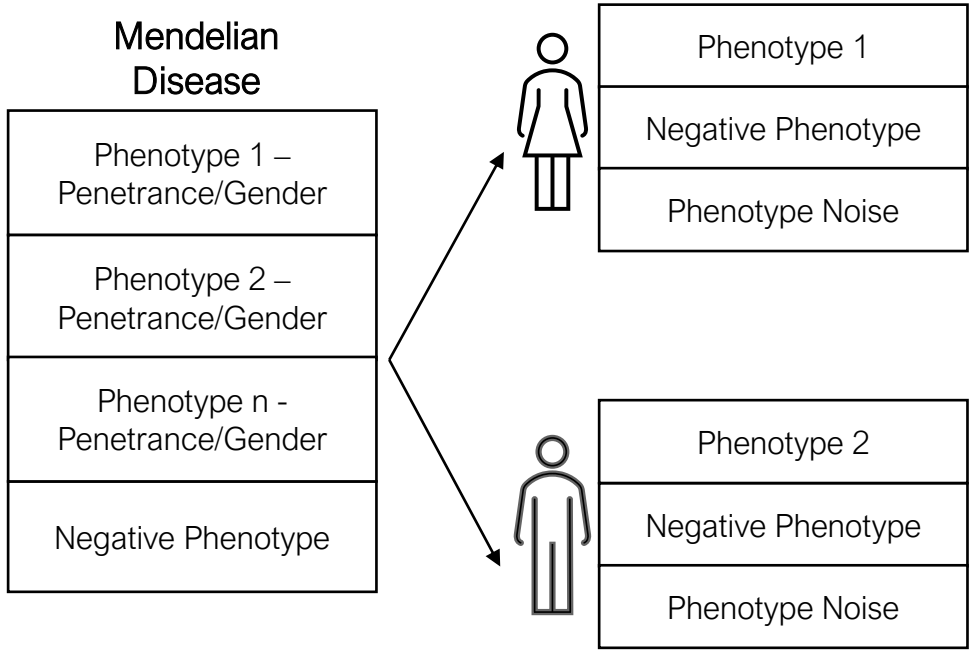
Instances	174 + 107
Positive Pairs	107
Negative Pairs	107
Positive Statements	14828
Negative Statements	9191

BUILDING BENCHMARK DATASETS

Disease Prediction

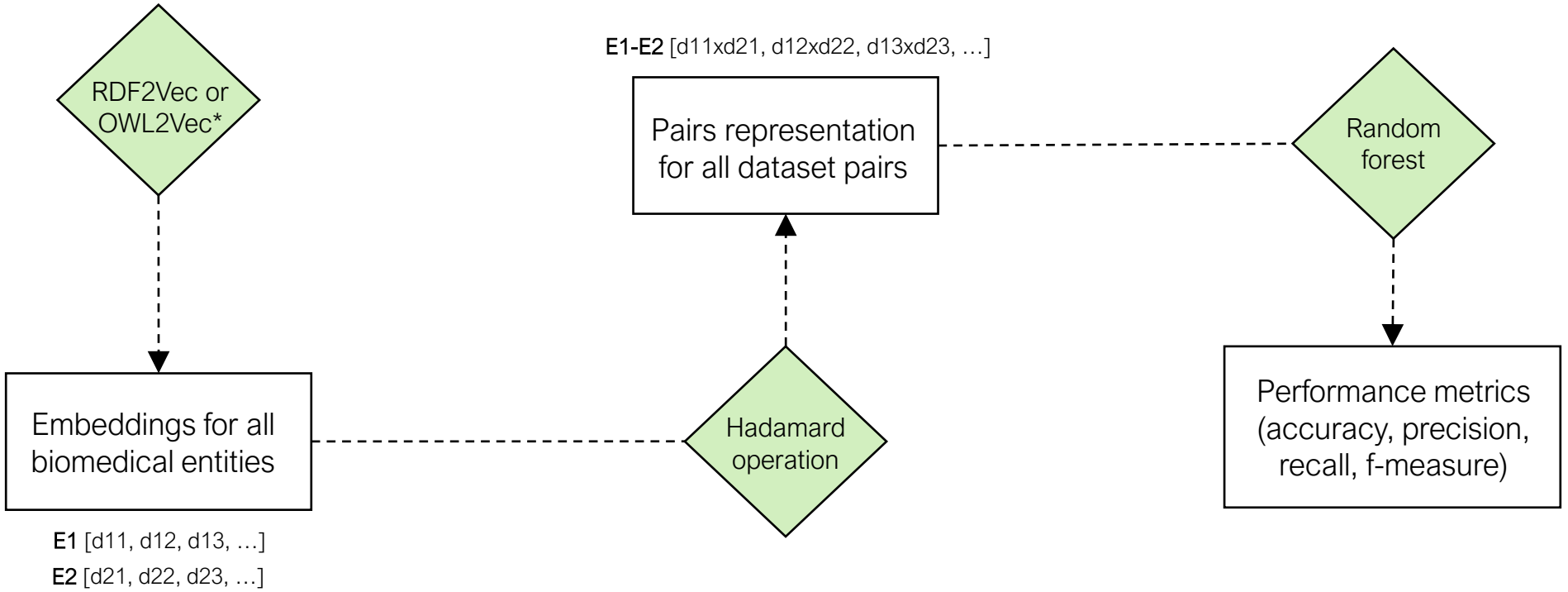
- 33 mendelian diseases where the penetrance of each phenotype is known.
- For each disease, 20 synthetic patients are created.
- 1000 diseases are randomly chosen.
- Random annotations can be added to patients to emulate a more realistic situation.

Instances	1033 + 660
Positive Pairs	660
Negative Pairs	681120
Positive Statements	38130
Negative Statements	179



VALIDATING BENCHMARK DATASETS

KG embedding methods have been successfully employed in biomedical relation prediction tasks.

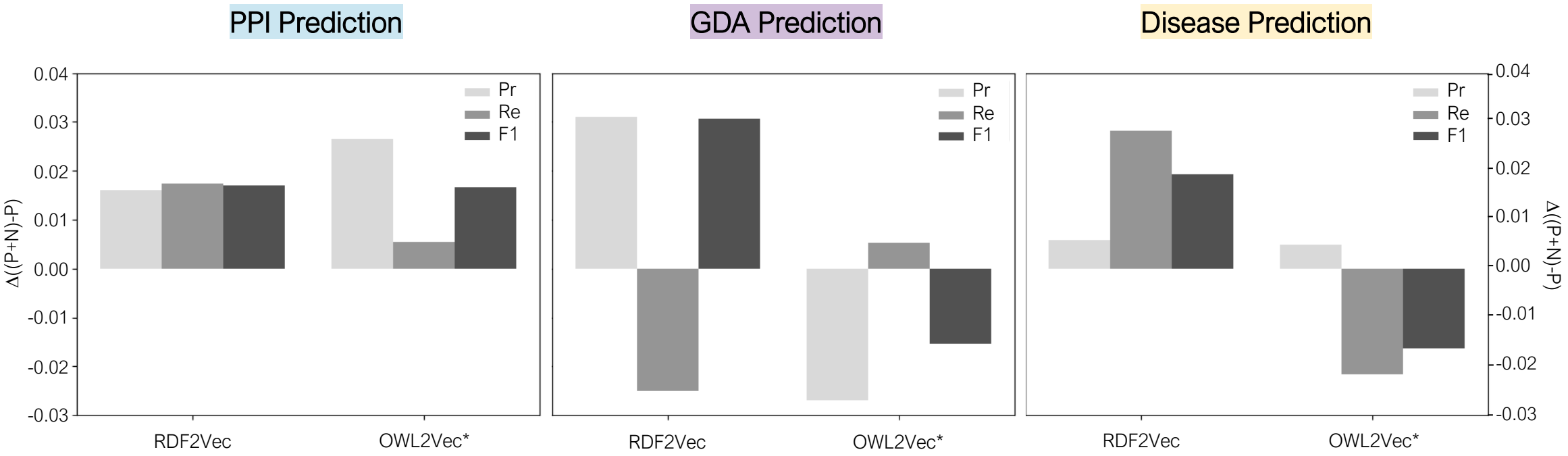


VALIDATING BENCHMARK DATASETS

Median precision, recall and weighted average F-measure (Pr, Re, and F1) for PPI, GDA, and disease prediction using only positive statements (Pos) or positive and negative statements (Pos+Neg).

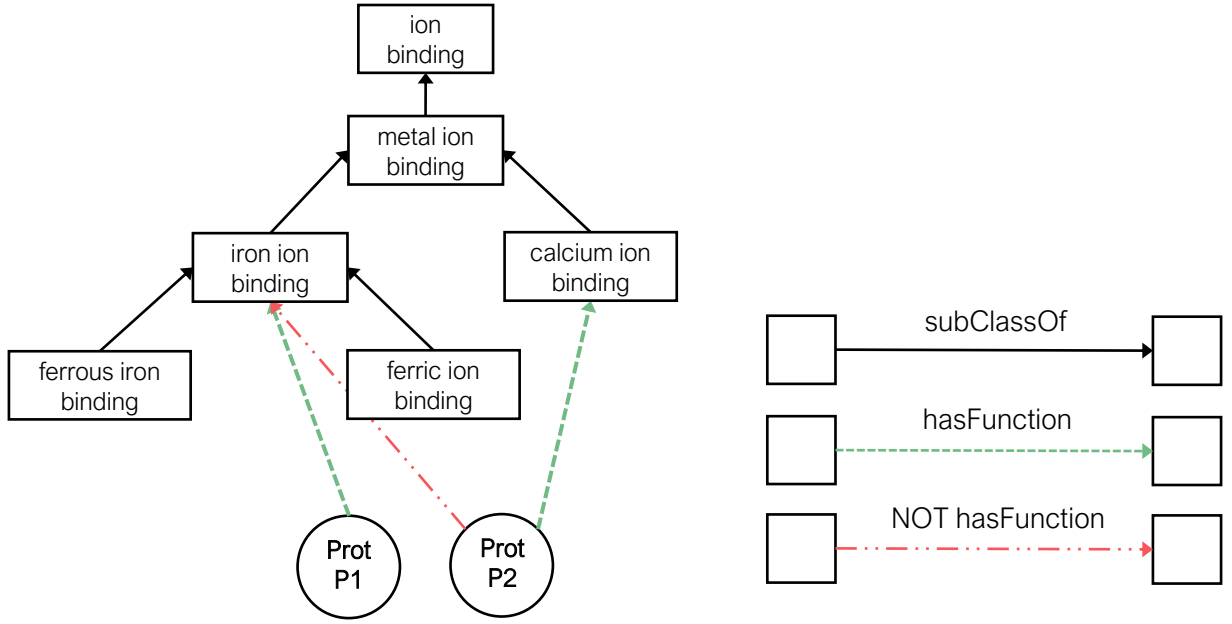
	Method	PPI Prediction			GDA Prediction			Disease Prediction		
		P	R	F-score	P	R	F-score	P	R	F-Score
P	OWL2Vec*	0.833	0.806	0.823	0.652	0.656	**0.646**	0.975	0.584	0.730
P	RDF2Vec	0.831	0.826	0.828	0.623	0.625	0.615	0.994	0.742	0.850
P+N	OWL2Vec*	**0.860**	0.812	0.840	0.625	**0.661**	0.630	0.980	0.563	0.713
P+N	RDF2Vec	0.847	**0.844**	**0.845**	**0.654**	0.600	0.645	**1.000**	**0.771**	**0.870**

VALIDATING BENCHMARK DATASETS



Barplots showing the differences in precision, recall, weighted average F-measure (Pr, Re, and F1) between using positive and negative statements or only positive statements.

CAN THE NEGATIVE STATEMENTS BE ADEQUATELY EXPLORED BY THESE KG EMBEDDING METHODS?



Classical Walks

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 > subClassOf > metal ion binding > subClassOf > ion binding

USING BENCHMARK DATASETS

zenodo

https://doi.org/10.5281/zenodo.7709195

For each dataset, we provide access to:

- TSV file containing pairs of entities and information about whether a relationship exists between them or not;
- OWL file(s) containing the KG used to describe the biomedical entities that appear in the TSV file.

CLOSING REMARKS

https://doi.org/10.5281/zenodo.7709195

- Benchmark datasets are essential for evaluating and comparing the performance of different approaches that work over KGs.
- We present a collection of datasets for 3 very relevant biomedical relation prediction tasks.
- The datasets are validated using two popular KG embeddings. The results highlight the importance of negative statements to create more accurate representations.
- The datasets open the opportunity for the emergence of new embedding methods that consider negative statements and their semantic implications.

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

THANK YOU FOR YOUR ATTENTION.

zenodo

https://doi.org/10.5281/zenodo.7709195

risousa@ciencias.ulisboa.pt

@RitaTorresSousa

This work was 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 (No 101017453), and by HfPT: Health from Portugal under the Portuguese Plano de Recuperação e Resiliência.

LASIGE driven by excellence

Ciências ULisboa

fct Fundação para a Ciência e a Tecnologia