ESWC 2022

TOWARDS SUPERVISED BIOMEDICAL SEMANTIC SIMILARITY

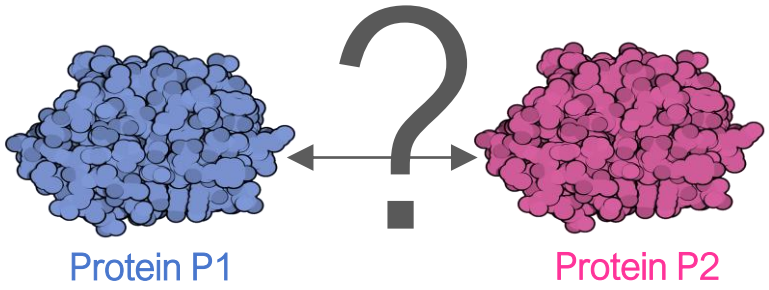
Rita T. Sousa, Sara Silva, Catia Pesquita
LASIGE, Faculdade de Ciências da Universidade de Lisboa

5th Workshop on Semantic Web solutions for large-scale biomedical data analytics
SeWeBMeDA 2022
May 29

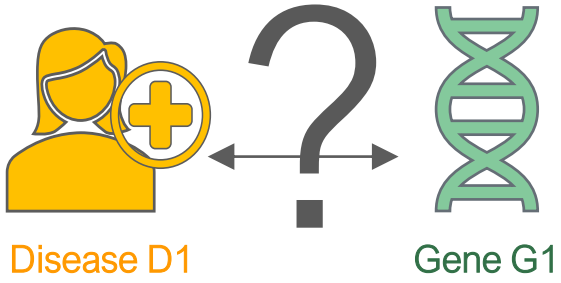
U LISBOA | Ciências ULisboa | LASIGE | FCT Fundação para a Ciência e a Tecnologia

MEASURING ENTITY SIMILARITY IN THE BIOMEDICAL DOMAIN IS FUNDAMENTAL

There are a wide variety of bioinformatics applications that benefit from using similarity.



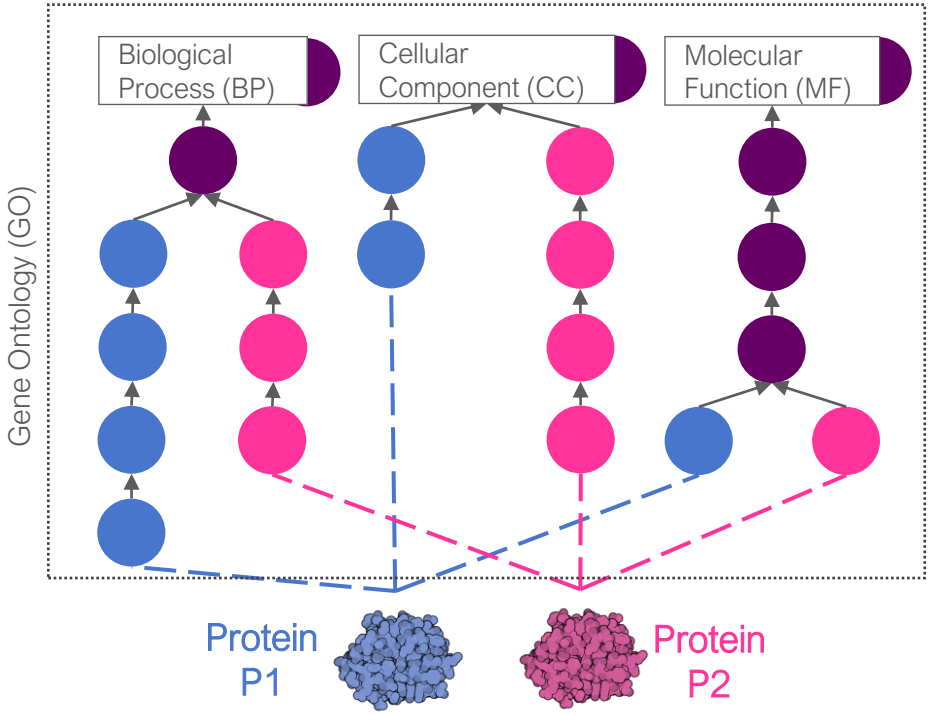
Protein-Protein Interaction Prediction



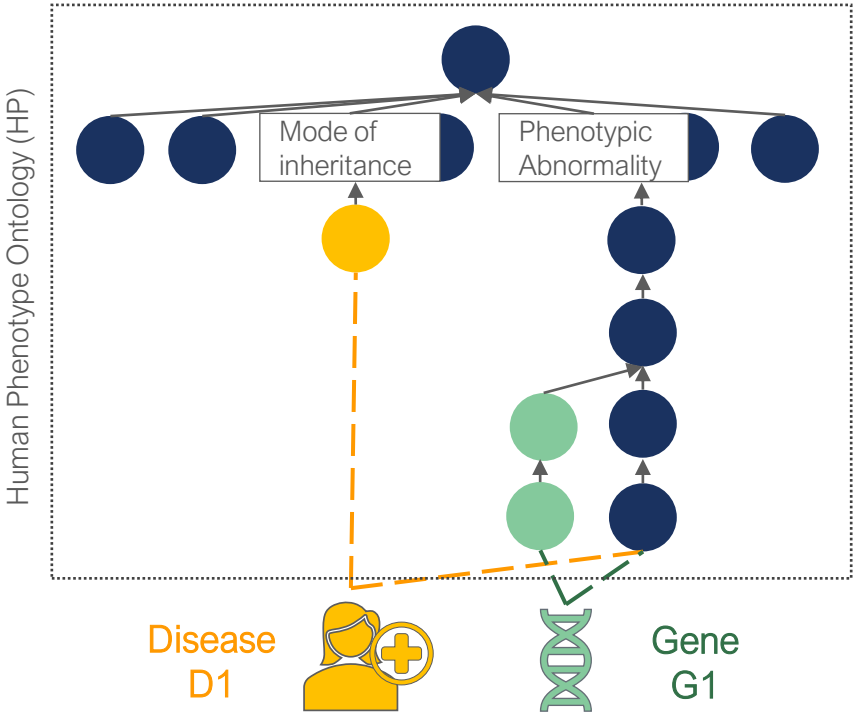
Disease-associated Genes Identification

BIOMEDICAL ONTOLOGIES AND KNOWLEDGE GRAPHS (KGs) CAN BE USED TO COMPUTE SEMANTIC SIMILARITY

Ontologies and Knowledge Graphs (KGs) provide the scaffolding for comparing biological entities at a higher level of complexity by comparing the ontology classes with which they are annotated.



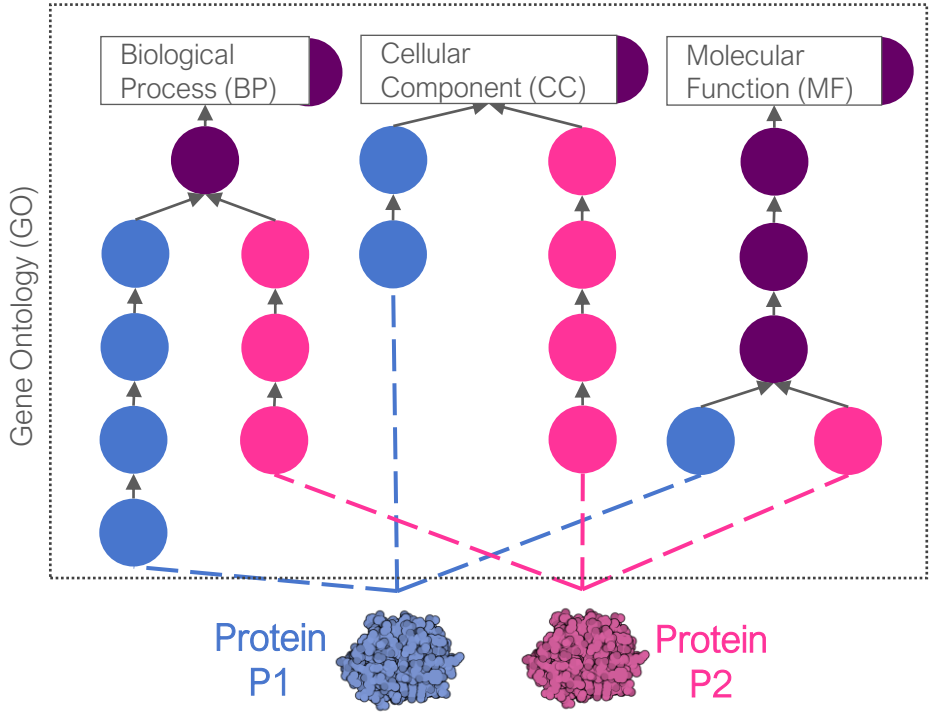
Protein-Protein Interaction Prediction



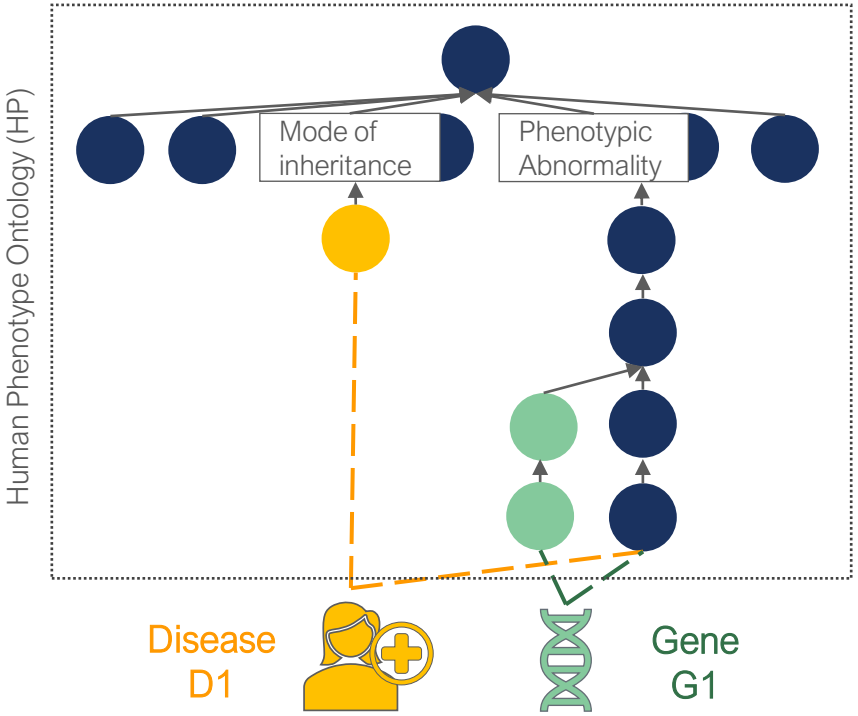
Disease-associated Genes Identification

TAXONOMIC SEMANTIC SIMILARITY MEASURES (SSMs) ARE GENERALLY DESIGNED BY AN EXPERT

$$\mathrm{SS}(\mathrm{E1}, E2) = \frac{A(\mathrm{E1}) \cap A(E2)}{\mathrm{A(E1)} \cup \mathrm{A(E2)}}$$



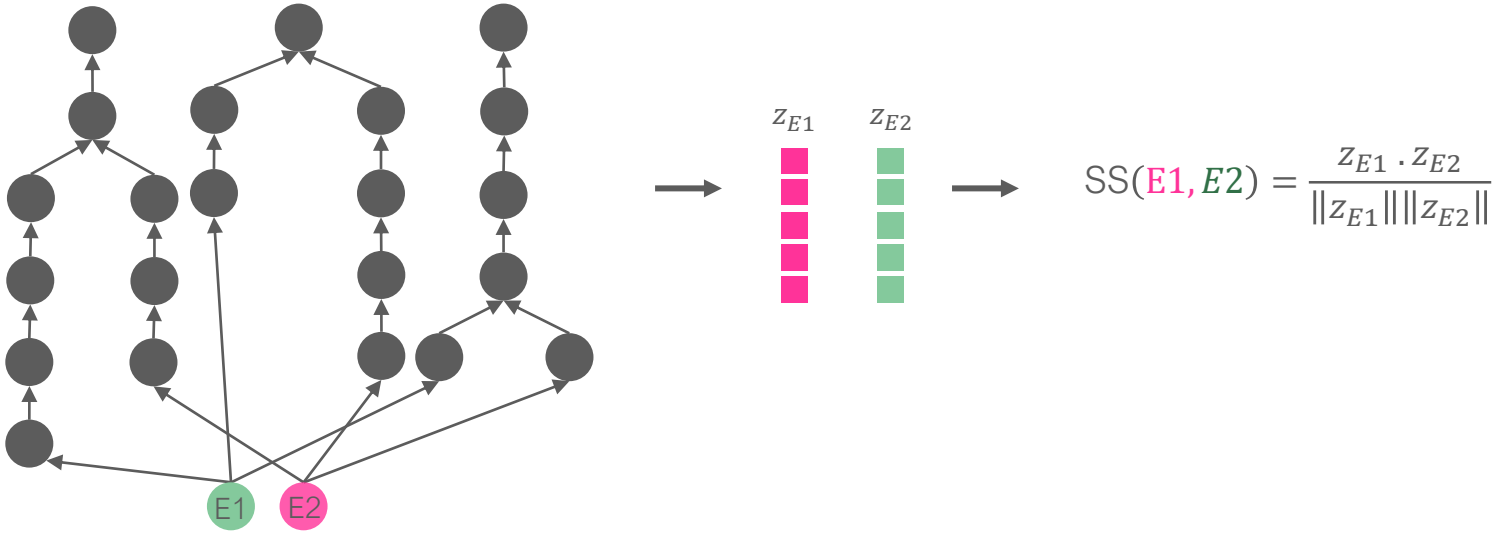
$$SS(\mathrm{P1}, P2) = {}^{7}/_{22}$$



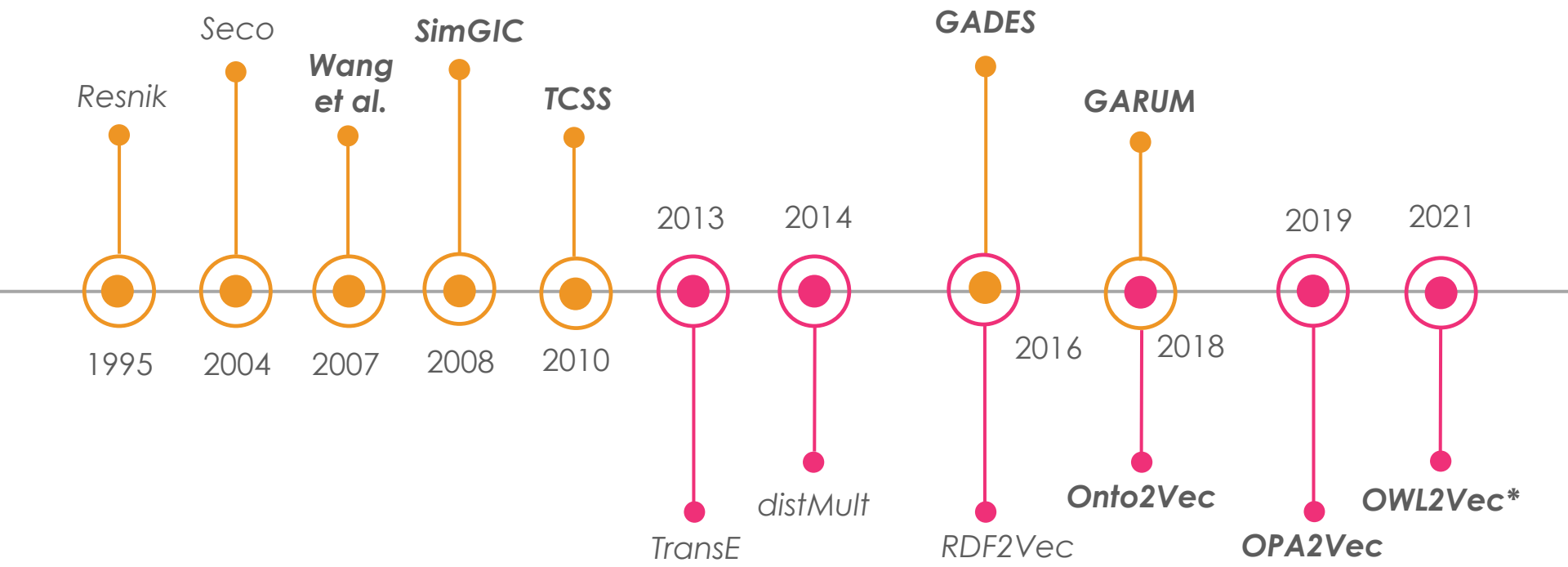
$$SS(\mathrm{D1}, G1) = {}^{10}/_{13}$$

KG EMBEDDINGS CAN ALSO BE USED TO COMPUTE SEMANTIC SIMILARITY THROUGH VECTOR SIMILARITY

KG embedding methods map each node to a lower-dimensional space in which its graph position and the structure of its local graph neighborhood are preserved.

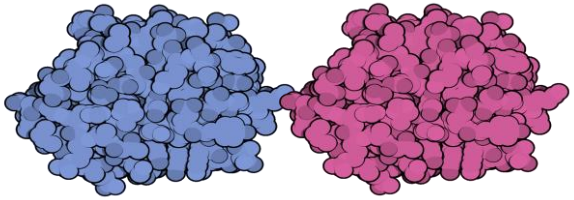


SEVERAL SSMs HAVE BEEN PROPOSED OVER THE YEARS AND APPLIED IN THE BIOMEDICAL DOMAIN



DIFFERENT USE CASES MAY REQUIRE DIFFERENT SIMILARITY PERSPECTIVES

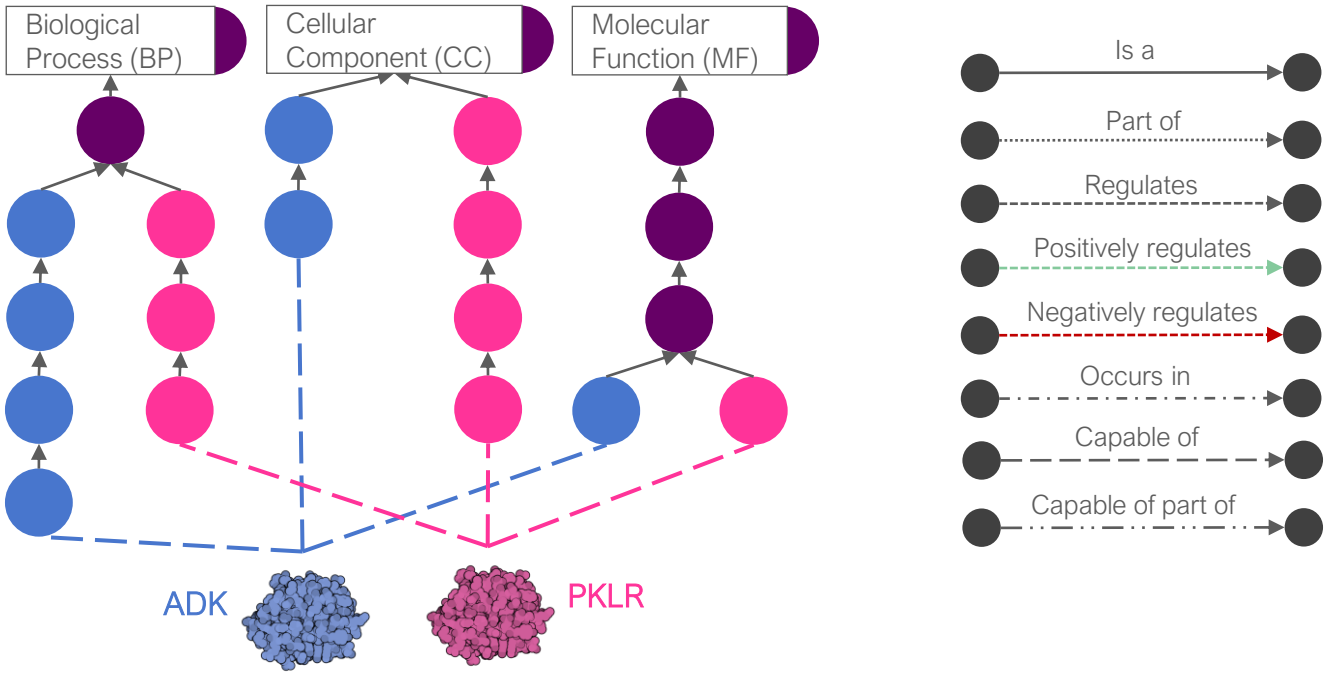
Adenosine kinase
P55263
ADK



Pyruvate kinase
P30613
PKLR

HOW CAN WE TAILOR SSMs TO FIT A SPECIFIC APPLICATION AND BIOLOGICAL PERSPECTIVE ON SIMILARITY?

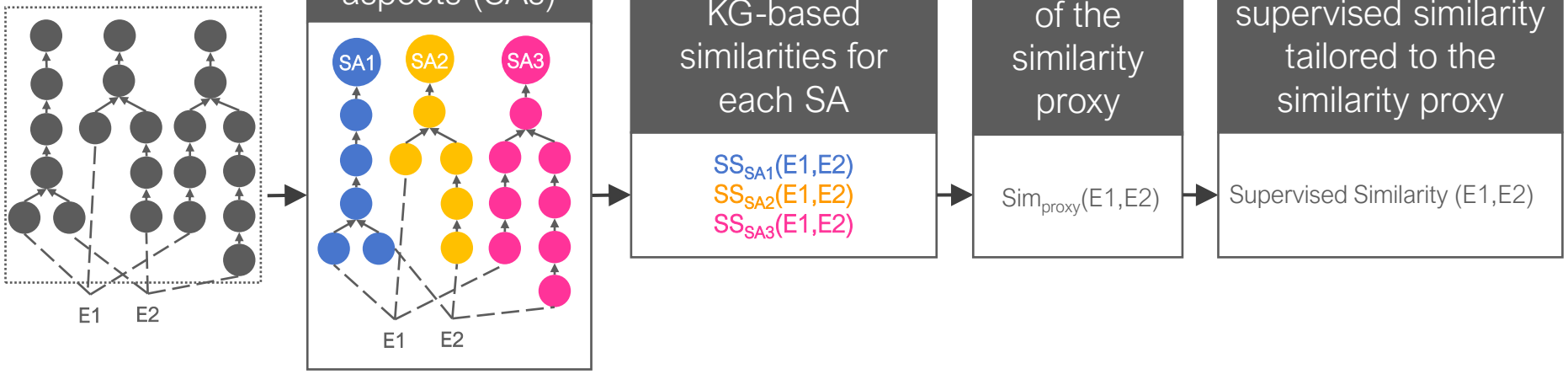
KGs DESCRIBE ENTITIES USING DIFFERENT SEMANTIC ASPECTS (SAs)



A SA is a perspective of the KG entities, and it can correspond to a given set of portions of the graph (e.g., describing a protein only through the BP subgraph) or property types (e.g., describing a protein only through *regulate* relation).

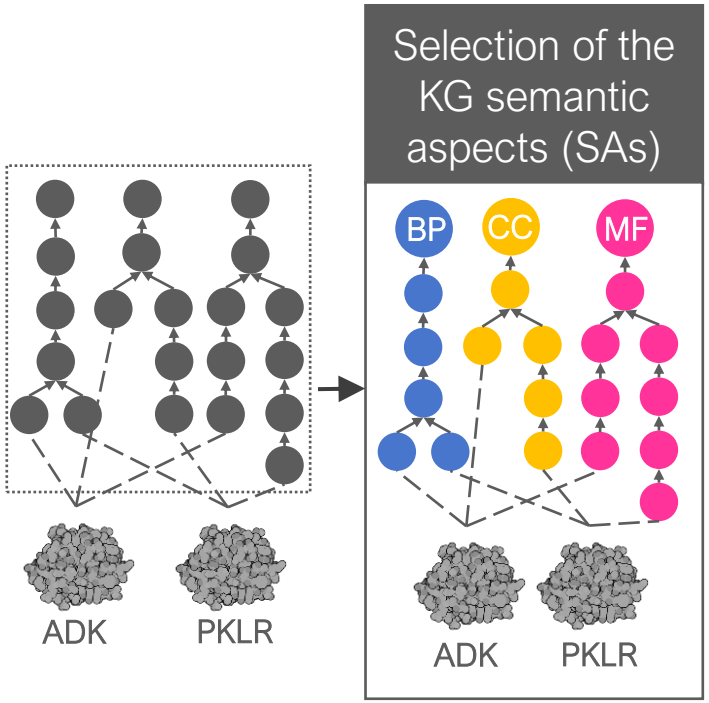
THE SUPERVISED SEMANTIC SIMILARITY TOOLKIT

Using supervised machine learning (ML) to tailor aspect-oriented semantic similarity measures to fit a particular view on biological similarity or relatedness.



THE SUPERVISED SEMANTIC SIMILARITY TOOLKIT

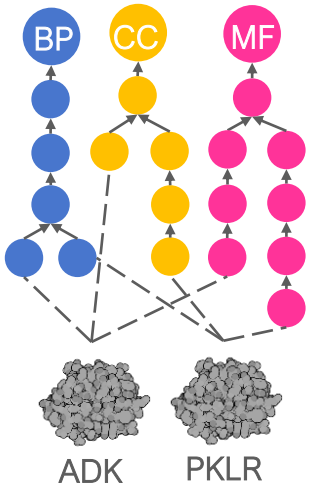
1



As pre-defined semantic aspects, the toolkit uses the subgraphs when the KGs have multiple roots or the subgraphs rooted in the classes at a distance of one from the KG root class.

THE SUPERVISED SEMANTIC SIMILARITY TOOLKIT

2



Computation of KG-based similarities for each SA

$$SS(ADK, PKLR) = \frac{GO(ADK) \cap GO(PKLR)}{GO(ADK) \cup GO(PKLR)}$$

$$SS_{BP}(ADK, PKLR) = {}^{2}/_{9}$$

$$SS_{CC}(ADK, PKLR) = {}^{1}/_{7}$$

$$SS_{MF}(ADK, PKLR) = {}^{4}/_{6}$$

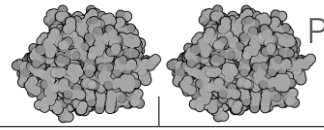
The toolkit employs 10 KG-based SSMs:

- SSMs
  - Taxonomic
    - IC approach
      - IC_{Seco}
      - IC_{Resnik}
    - Set approach
      - SimGIC
      - ResnikBMA
      - ResnikMAx
  - KG Embedding-based
    - TransE
    - RDF2Vec
    - distMult
    - OWL2Vec*

THE SUPERVISED SEMANTIC SIMILARITY TOOLKIT

3

Selection of the similarity proxy



Molecular Function Similarity

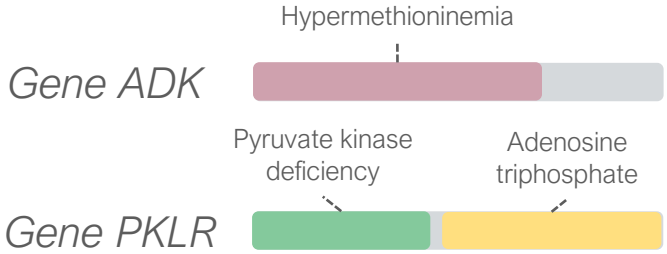
$$\text{Sim}_{\text{PFAM}}(ADK, PKLR) = {}^{0}/_{3} = 0$$

Sequence Similarity

```
P55263 ADK_HUMAN   1   MAAAEEEP--KPKKLKVEAPQAL
P30613 KPYR_HUMAN  1   MSIQENISSLQLRSWVSKSQRDL
                       *:  *:    : :.   :: : *
```

$$\text{Sim}_{\text{SEQ}}(ADK, PKLR) = 0.11$$

Phenotypic Series Similarity



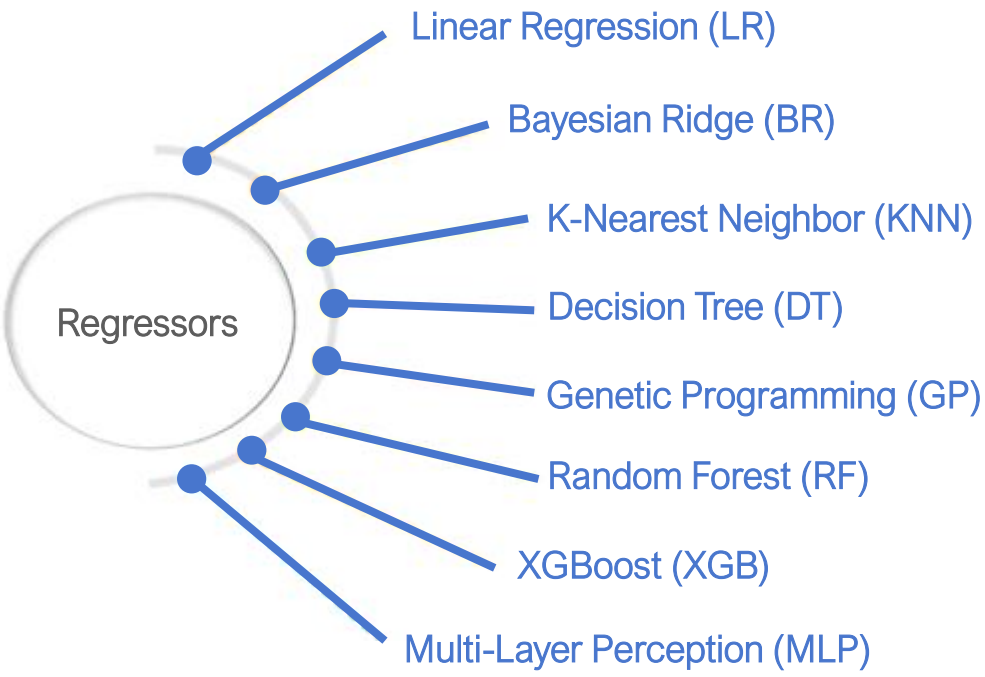
$$\text{Sim}_{\text{PS}}(ADK, PKLR) = {}^{0}/_{3} = 0$$

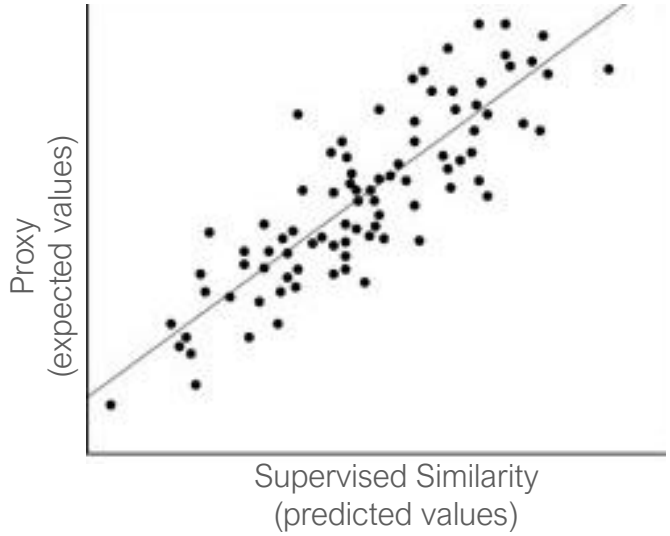
THE SUPERVISED SEMANTIC SIMILARITY TOOLKIT

4

Learning of supervised similarity tailored to the similarity proxy

Supervised $Similarity$ $(ADK, PKLR)$





EVALUATION

Benchmark datasets[1] exploit 3 similarity proxies for biomedical entity similarity and include data from Gene Ontology (GO) and Human Phenotype Ontology (HP).

Dataset	Number datasets	Proxies	KGs
Protein	10	Sim_{SEQ} and Sim_{PFAM}	GO KG
Gene	1	Sim_{PS}	GO KG and HP KG

These datasets cover multiple species and present two levels of annotation completion.

E. coli

D. melanogaster

H. sapiens

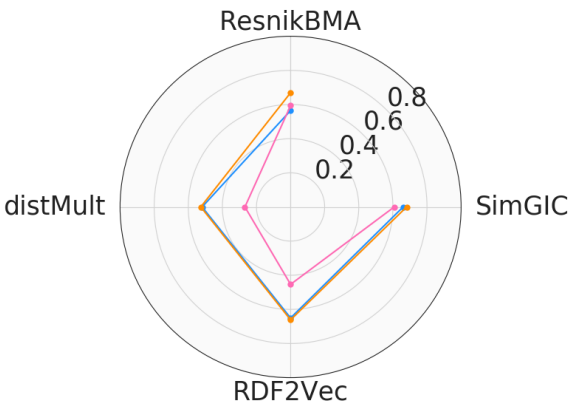
S. cerevisiae

All species

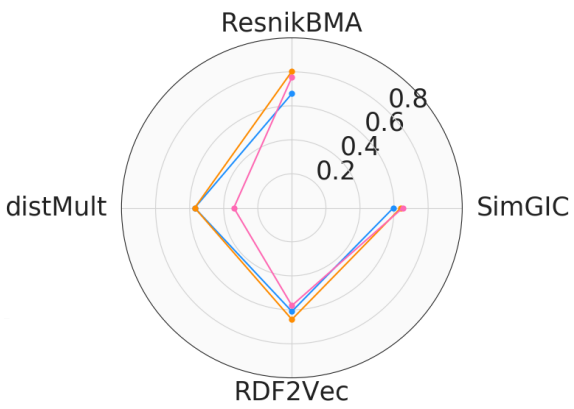
[1] Cardoso, C., Sousa, R. T., Köhler, S., & Pesquita, C. (2020). A Collection of Benchmark Data Sets for Knowledge Graph-Based Similarity in the Biomedical Domain. In European Semantic Web Conference (pp. 50-55). Springer, Cham.

Radar charts with the median Pearson's correlation between similarity proxy and supervised similarity.

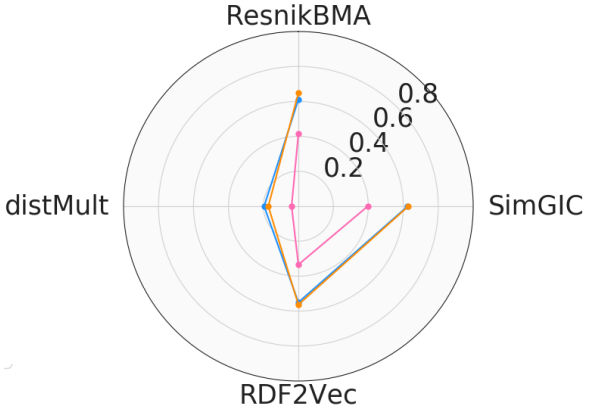
Dataset: Protein
Species: All
Proxy: Sim_{PFAM}



Dataset: Protein
Species: All
Proxy: Sim_{SEQ}



Dataset: Gene
Species: Human
Proxy: Sim_{PS}

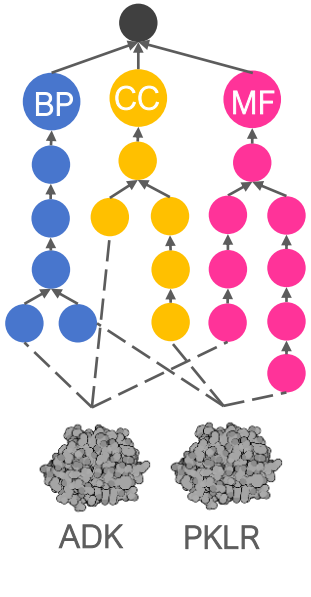


- Taxonomic similarity performs well across many evaluations and, in most of the datasets, has better performance than embedding similarity.

SUPERVISED SIMILARITY VS STATIC SIMILARITY

Baselines:

- whole KG similarity
- the single SA similarities
- 2 well-known strategies for combining the single aspect scores



$SS_{ALL}(ADK, PKLR)$

$SS_{BP}(ADK, PKLR)$

$SS_{CC}(ADK, PKLR)$

$SS_{MF}(ADK, PKLR)$

$SS_{MAX}(ADK, PKLR) = max(SS_{BP} + SS_{CC} + SS_{MF})$

$SS_{AVG}(ADK, PKLR) = {(SS_{BP}+SS_{CC}+SS_{MF})}/{3}$

SUPERVISED SIMILARITY VS STATIC SIMILARITY

Pearson's correlation coefficient between the similarity proxy using ResnikBMA for the baselines and the median Pearson's correlation coefficient between the similarity proxy and the supervised similarity.

Proxy	Dataset	Static Similarity							Supervised Similarity	
		All	HP	BP	CC	MF	Avg	Max	XGB	RF
Sim_{PFAM}	Protein (158 512)	0.534		0.448	0.370	0.456	0.525	0.500	**0.669**	0.638
Sim_{SEQ}	Protein (158 512)	0.510		0.528	0.373	0.291	0.481	0.399	**0.803**	0.746
Sim_{PS}	Gene (12 000)	0.524	0.601	0.210	0.142	0.055	0.413	0.552	**0.648**	**0.648**

Improvements over the whole graph similarity and the single aspect similarities are consistent for all datasets and also clear when considering the combination of single aspects.

SUPERVISED SIMILARITY

CLOSING REMARKS

- Our approach is able to learn a supervised semantic similarity that outperforms static semantic similarity in capturing biological similarity both using KG embeddings and standard taxonomic SSMs.

- Combining a taxonomic SSM with an ensemble method is a good choice.

Proxy	SSM	ML Algorithm
Sim_{PFAM}	ResnikBMA	RF
Sim_{SEQ}	SimGIC	RF
Sim_{PS}	ResnikBMA	XGB

- As future work, supervised similarity tailored to relevant biological similarities can be transferred to other predictive tasks.

THANK YOU FOR YOUR ATTENTION.

risousa@ciencias.ulisboa.pt

@RitaTorresSousa

GitHub

https://github.com/liseda-lab/Supervised-SS

This work was funded by FCT through LASIGE Research Unit (UIDB/00408/2020, UIDP/00408/2020); projects GADgET (DSAIPA/DS/0022/2018) and BINDER (PTDC/CCI-INF/29168/2017); PhD grant SFRH/BD/145377/2019.