Bi-weekly Colloquium

Özge Erten
Institute of Data Science - Maastricht University
17 February 2023

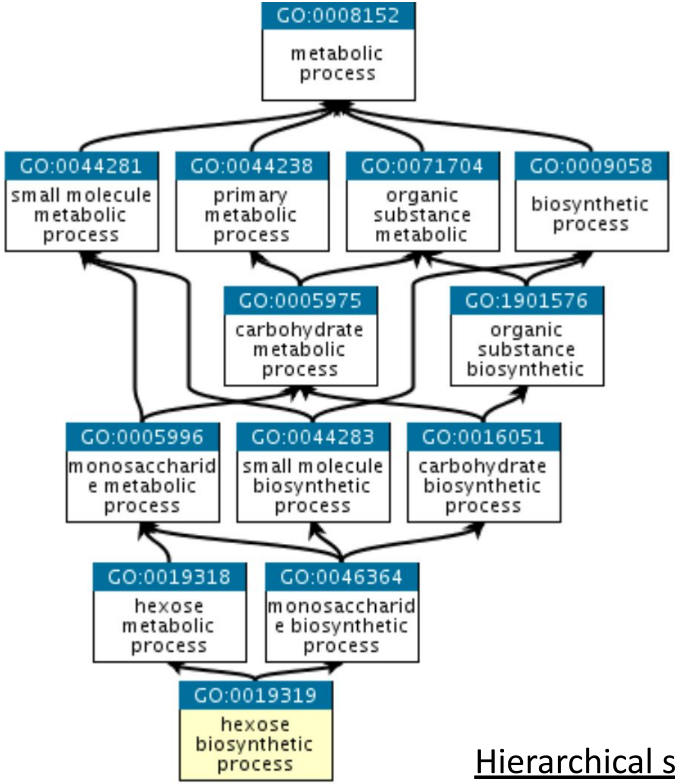
Introduction

- Problem
  - For many genes, information such as their functions and locations are lacking to understand of the mechanisms of disease.
- Significance
  - GO and GOA information were used in analyzes such as gene enrichment and gene pathway, to understand the functional roles and relationships of genes within a biological system.
- Hypothesis
  - We think that incomplete gene annotations can better predicted if the GO hierarchy were incorporated in a machine learning model using True Path Rule.

Gene Ontology - GO

Gene Ontology nodes are called ‘GO Terms’

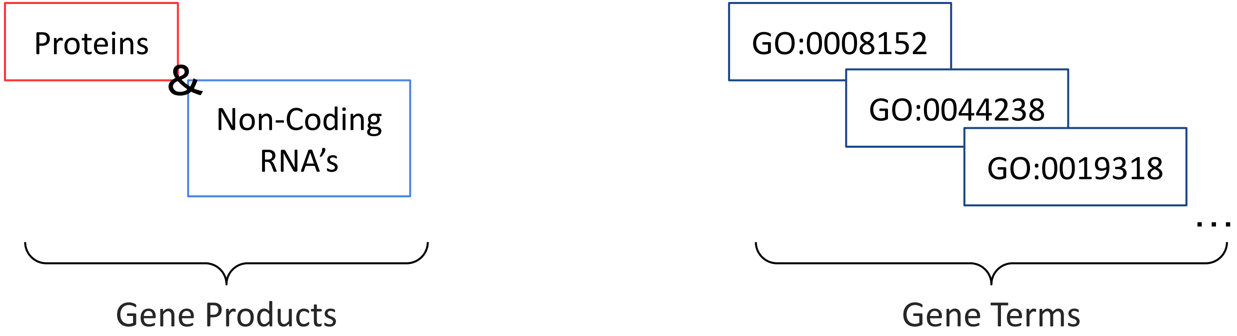
3 ASPECTS:
Molecular functions
Cellular components
Biological processes


GO:0008152
metabolic process
GO:0044281
small molecule metabolic process
GO:0044238
primary metabolic process
GO:0071704
organic substance metabolic
GO:0009058
biosynthetic process
GO:0005975
carbohydrate metabolic process
GO:1901576
organic substance biosynthetic
GO:0005996
monosaccharide metabolic process
GO:0044283
small molecule biosynthetic process
GO:0016051
carbohydrate biosynthetic process
GO:0019318
hexose metabolic process
GO:0046364
monosaccharide biosynthetic process
GO:0019319
hexose biosynthetic process

Hierarchical structure in GO[1]

Gene Ontology Annotation - GOA

Information in GOA



Proteins

&

Non-Coding RNA’s

Gene Products

GO:0008152

GO:0044238

GO:0019318

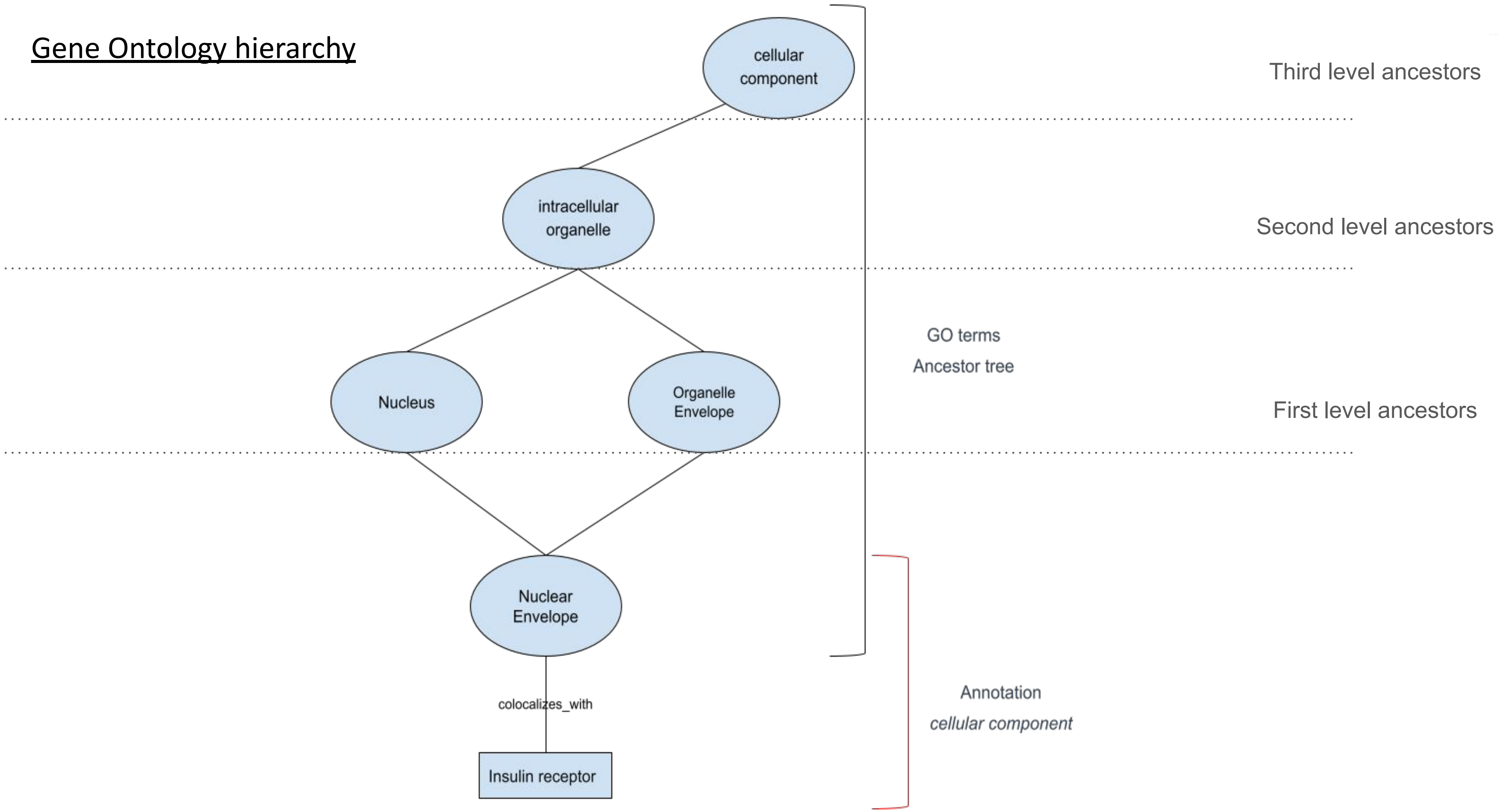
…

Gene Terms

Gene Ontology Annotations relate GENE PRODUCTs to GENE TERMs

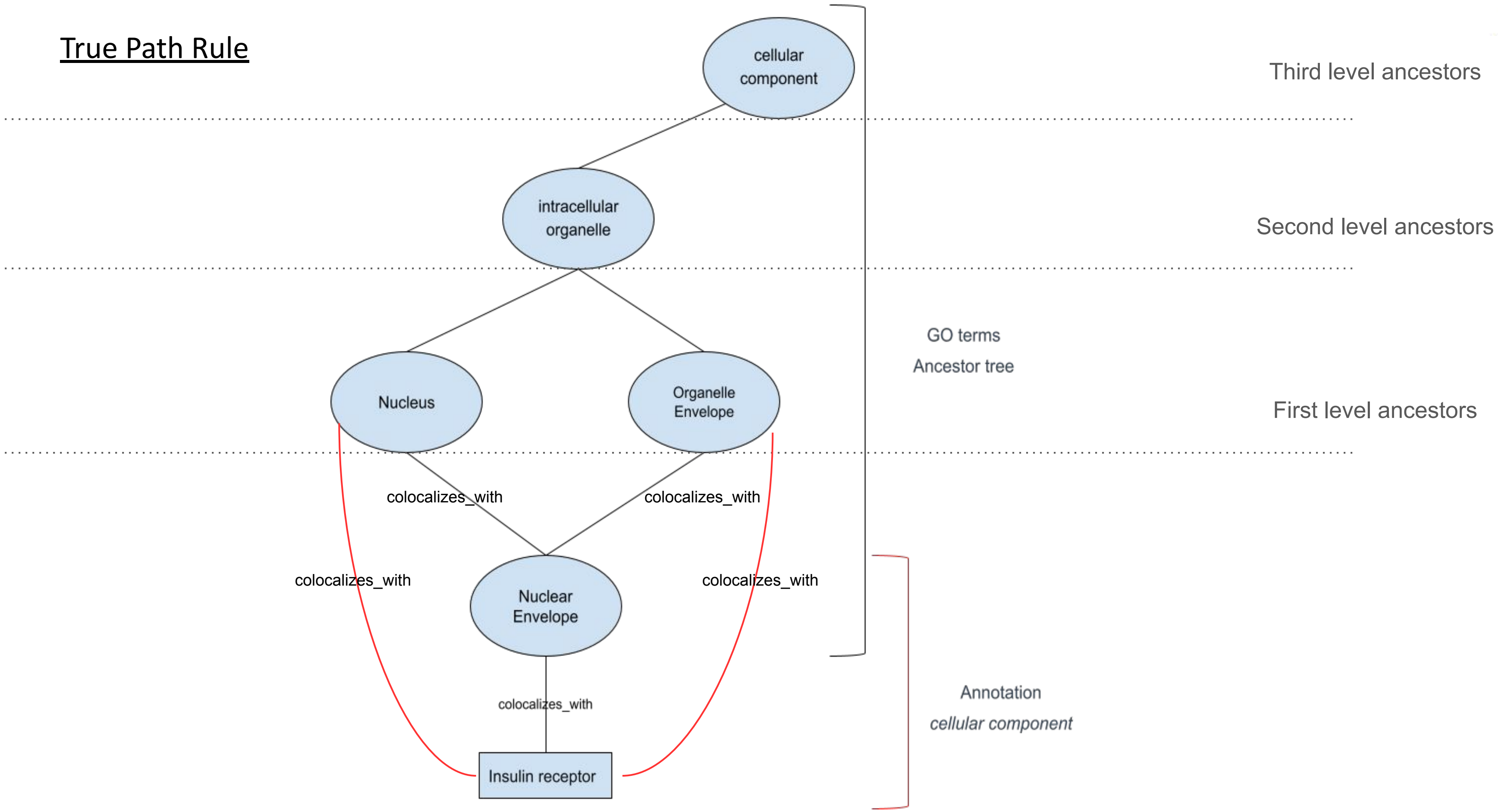
Methodology

Gene Ontology hierarchy



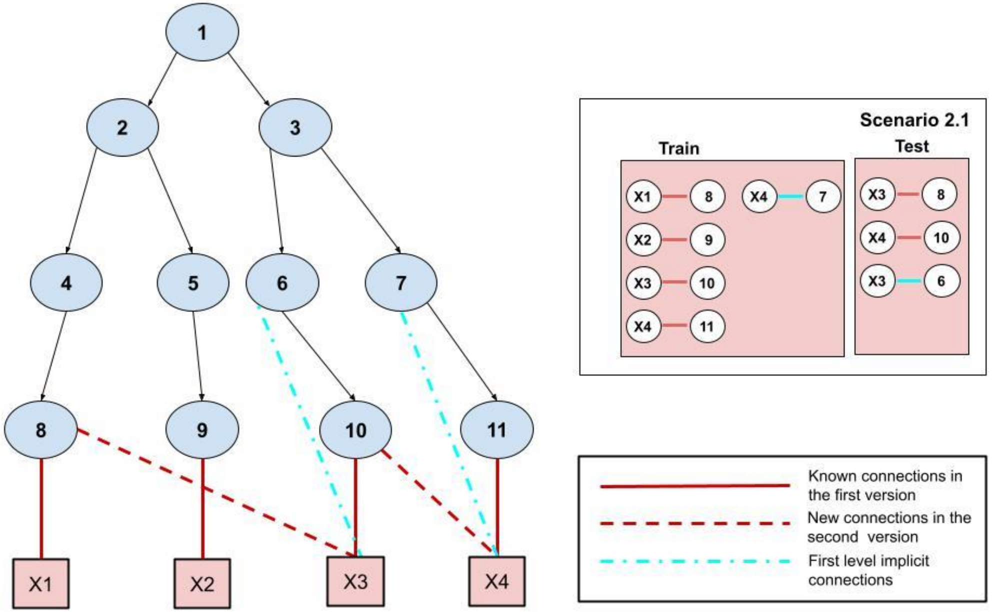
Methodology

True Path Rule



Train, test and validation sets structure

- Scenario 2.1

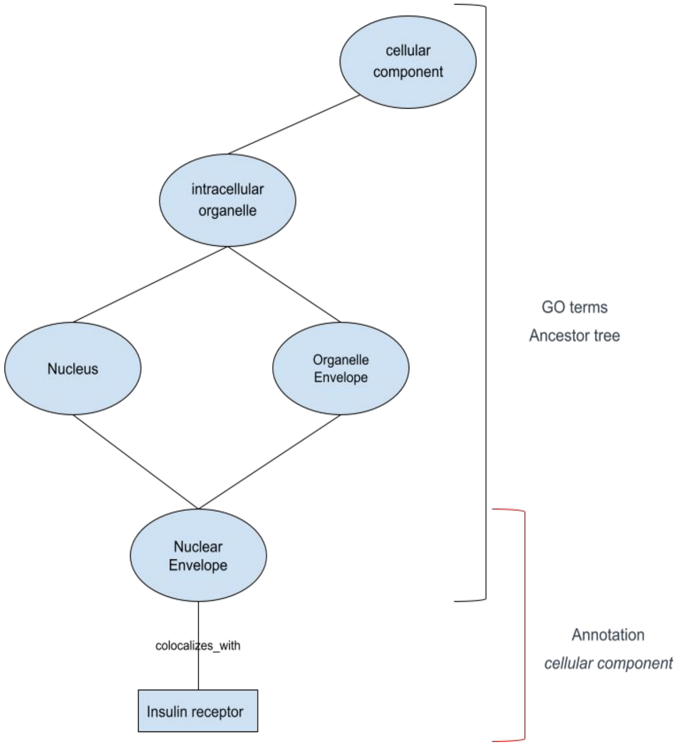


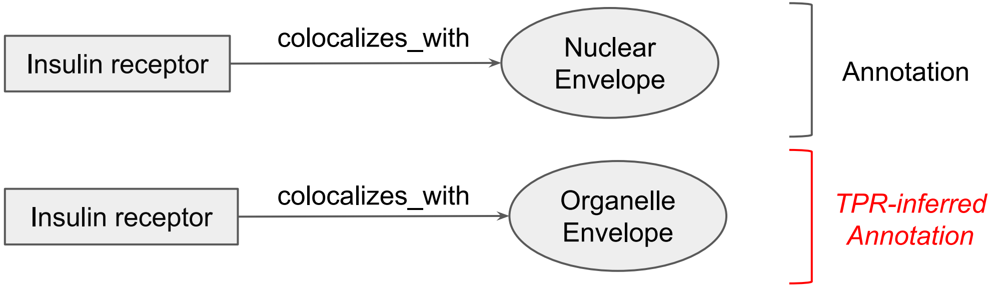
Known connections in the first version

New connections in the second version

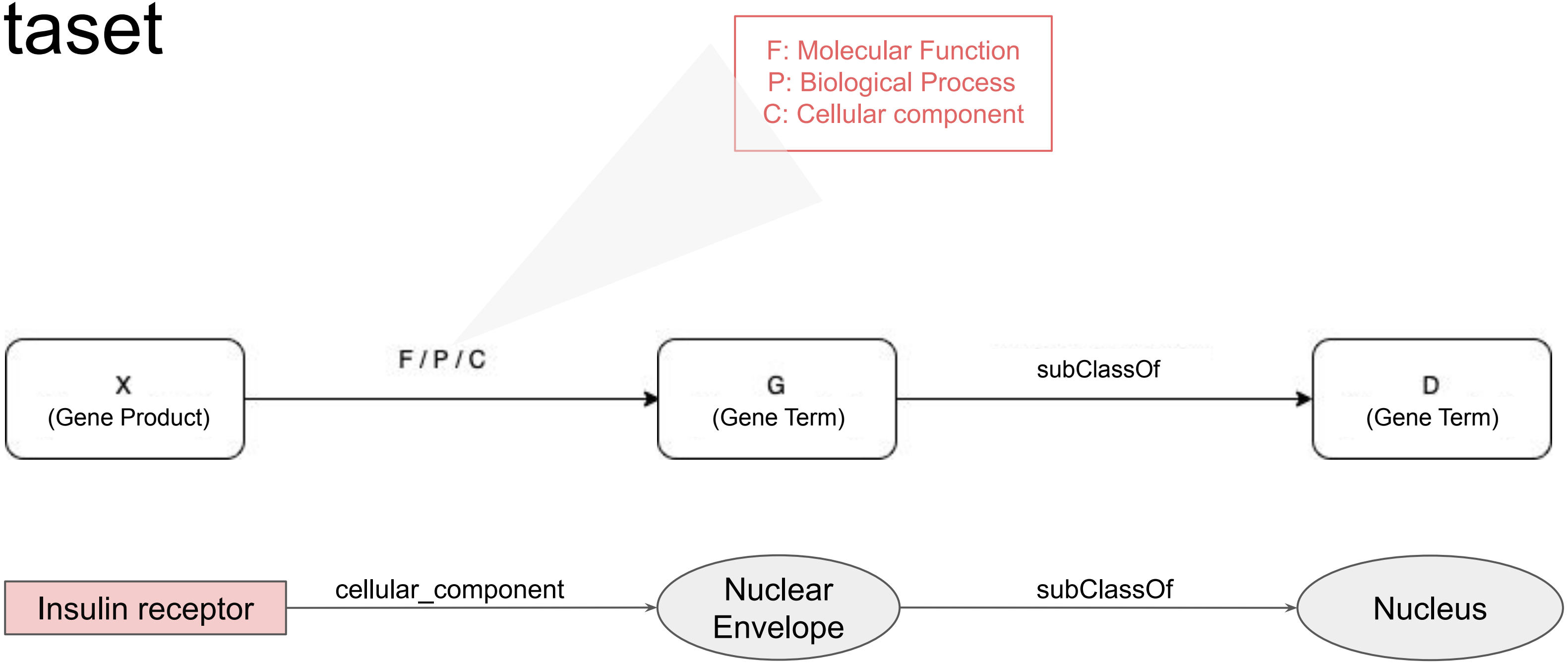
First level implicit connections

For instance,





Dataset



F: Molecular Function
P: Biological Process
C: Cellular component

X
(Gene Product)

F / P / C

G
(Gene Term)

subClassOf

D
(Gene Term)

Insulin receptor

cellular_component

Nuclear Envelope

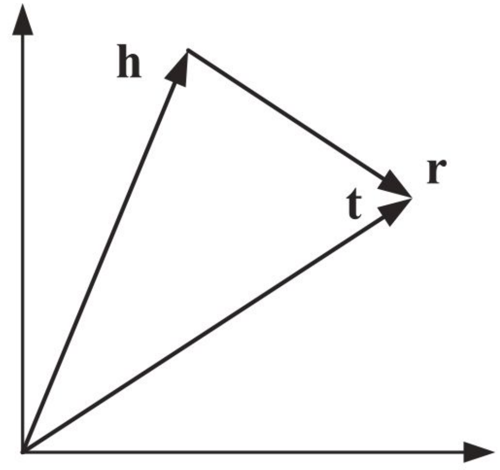
subClassOf

Nucleus

Knowledge Graph Embeddings (KGEs)

- Typically, KG embedding methods embeds entities and relations onto a vector space directly where each triple (head entity, relation, tail entity) in the KG is assigned a score based on its validity.

In TransE embedding;
If a fact ***(h,r,t)*** holds, then ***h +r*** should be close to ***t***
If not, ***h+r*** should be distant to ***t***



Entity and relation space in TransE[2]

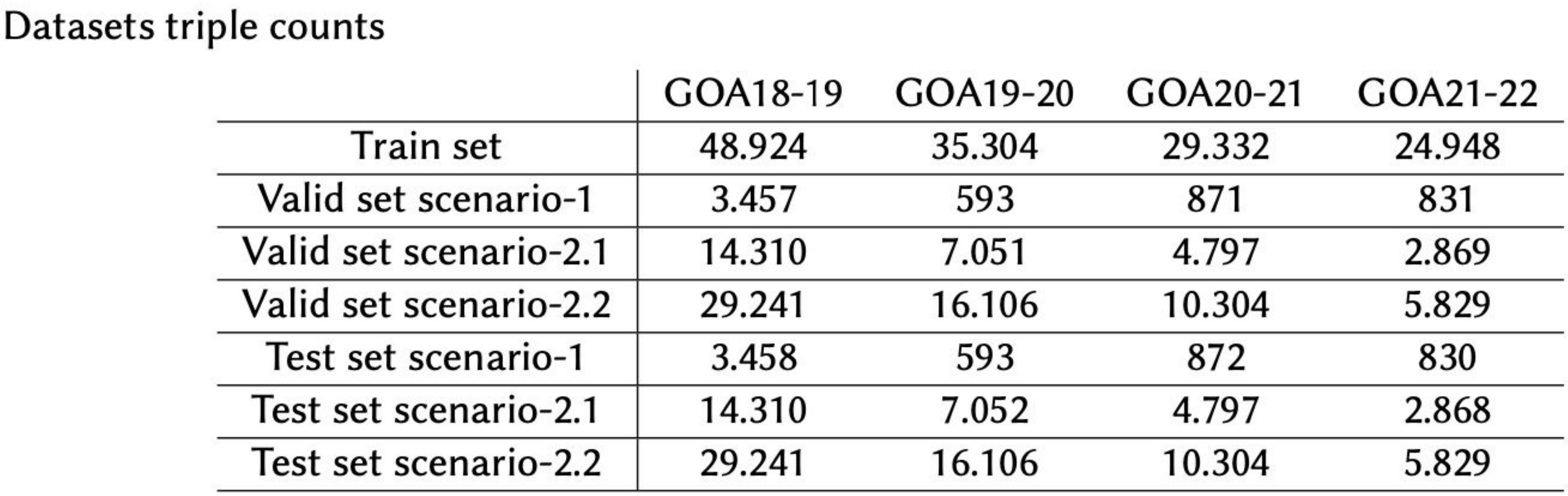
If a knowledge graph schema contains hierarchy; hereditary features can be incorporated in a machine learning model

- ◆ Hereditary features: The properties or characteristics that an entity inherits from its ancestor entities in the hierarchy.
- ◆ True Path Rule is one of the methods to identify implicit hereditary relations.
- ◆ These inferred relations can be used to increase samples in the training and test sets.
- ◆ With this way, it’s possible to improve the performance of the link prediction model by introducing more positive samples.

➔ To further capture and embed the TPR, we generate and incorporate samples using the TPR in the training data.

➔ We refer this method as TransE+TPR.

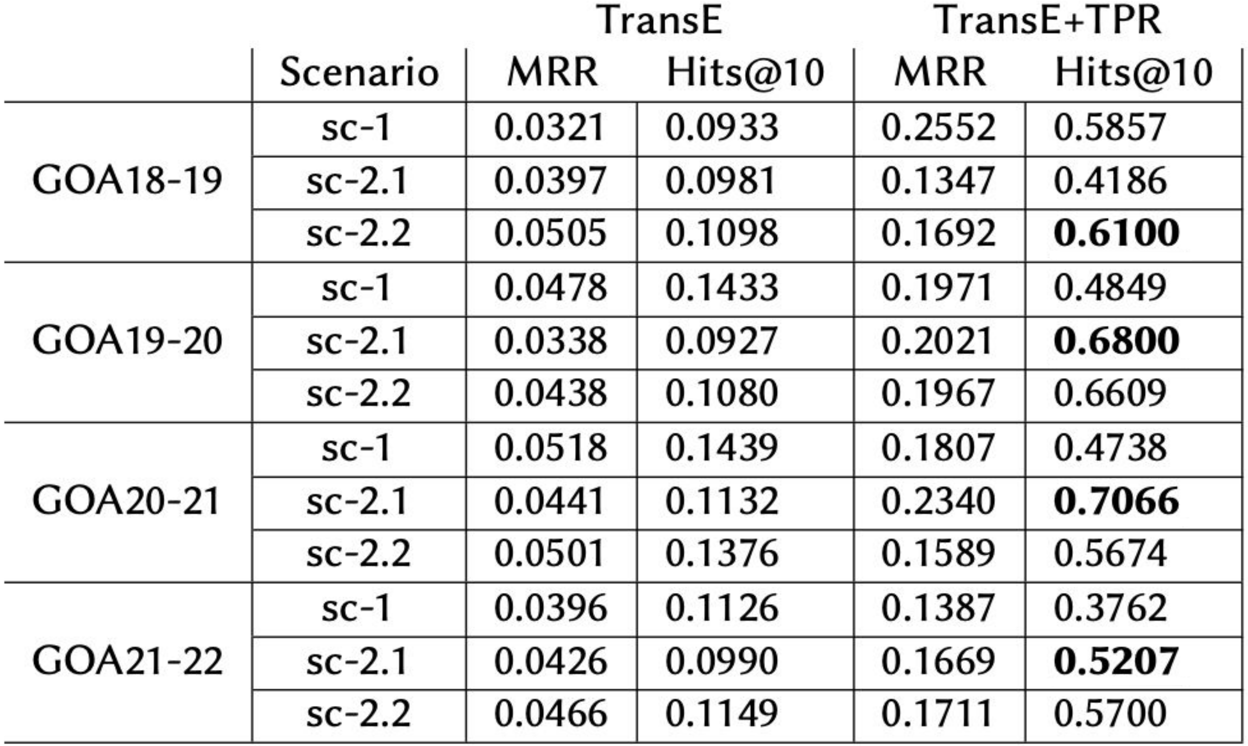
Experiments



Datasets triple counts

	GOA18-19	GOA19-20	GOA20-21	GOA21-22
Train set	48.924	35.304	29.332	24.948
Valid set scenario-1	3.457	593	871	831
Valid set scenario-2.1	14.310	7.051	4.797	2.869
Valid set scenario-2.2	29.241	16.106	10.304	5.829
Test set scenario-1	3.458	593	872	830
Test set scenario-2.1	14.310	7.052	4.797	2.868
Test set scenario-2.2	29.241	16.106	10.304	5.829

Prediction accuracy results for TransE and TransE+TPR:



		TransE		TransE+TPR	
	Scenario	MRR	Hits@10	MRR	Hits@10
GOA18-19	sc-1	0.0321	0.0933	0.2552	0.5857
	sc-2.1	0.0397	0.0981	0.1347	0.4186
	sc-2.2	0.0505	0.1098	0.1692	**0.6100**
GOA19-20	sc-1	0.0478	0.1433	0.1971	0.4849
	sc-2.1	0.0338	0.0927	0.2021	**0.6800**
	sc-2.2	0.0438	0.1080	0.1967	0.6609
GOA20-21	sc-1	0.0518	0.1439	0.1807	0.4738
	sc-2.1	0.0441	0.1132	0.2340	**0.7066**
	sc-2.2	0.0501	0.1376	0.1589	0.5674
GOA21-22	sc-1	0.0396	0.1126	0.1387	0.3762
	sc-2.1	0.0426	0.0990	0.1669	**0.5207**
	sc-2.2	0.0466	0.1149	0.1711	0.5700

Thank you!

References

1 - Kulmanov, M., Smaili, F. Z., Gao, X., & Hoehndorf, R. (2020). Machine learning with biomedical ontologies. biorxiv.

2- Gusmão, A. C., Correia, A. H. C., De Bona, G., & Cozman, F. G. (2018). Interpreting embedding models of knowledge bases: a pedagogical approach. arXiv preprint arXiv:1806.09504.