Enhancing Region-Based Geometric Embedding for Gene-Disease Associations

Aniket Mitra, Vinu E. Venugopal
Scalable Data Science Lab
International Institute of Information Technology, Bangalore

Contact: aniket.mitra@iiitb.ac.in

1.INTRODUCTION

Motivation

- Any **abnormality** observed in **genes** leads to **diseases** in our bodies.
- Proper diagnosis needs the identification of the correct **gene-disease association**.
- **Biomedical ontologies** like *GO*, *HPO* etc. contain information about these biomedical entities as logical axioms (atomic concepts with logical operators).
- Particularly, **EL++ DL** is capable of fast reasoning over these large biomedical knowledge bases.

N-Ball Embedding

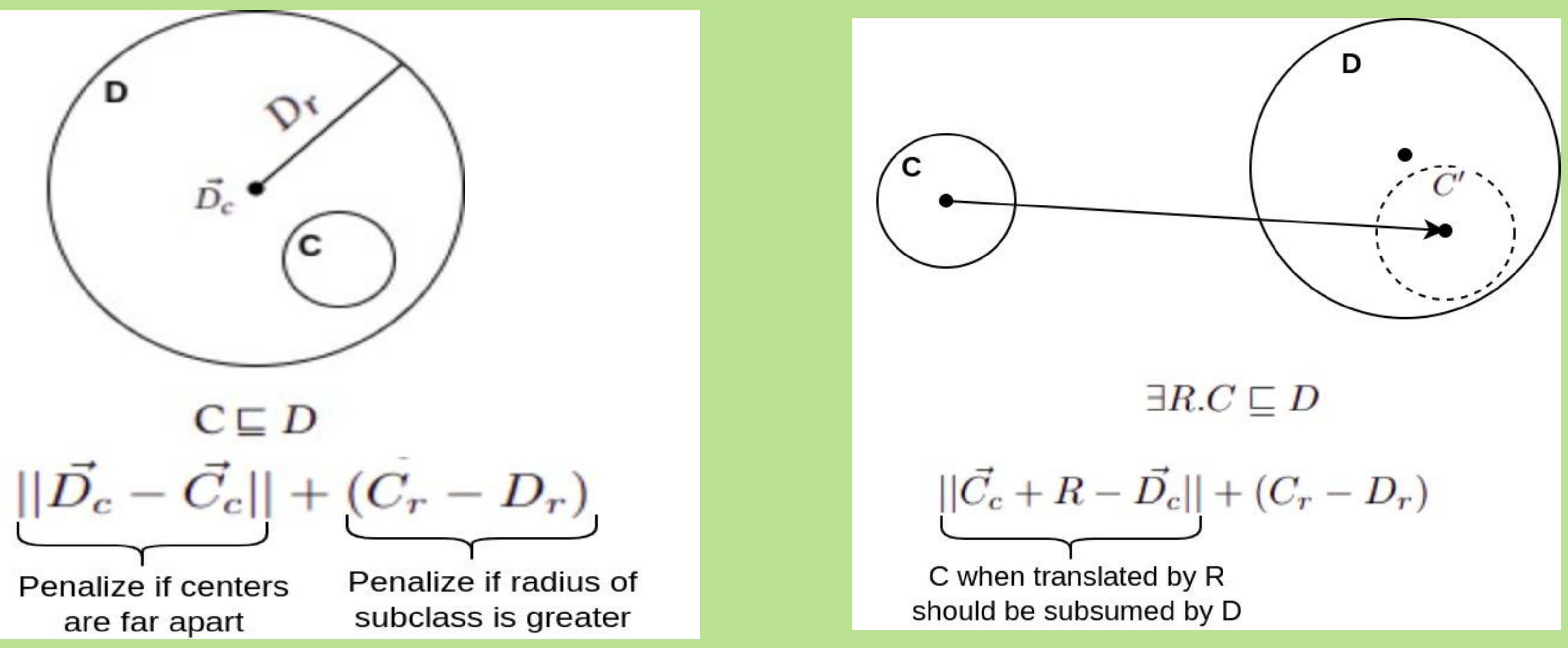
- [1] represented ontological concepts as **n-D balls** (*n-D center + scalar radius*). [2] shows the conversion of EL++ axioms to specific **normal forms(NFs)**. *Eg: C ⊑ D*
- The objective is to **formulate loss functions** for each NF preserving the semantics of EL++ geometry within R^n.
- The trained model will have minimum **Euclidean distance** between the parent and its child concept balls.

- **Link Prediction– O(n)**

2. RESEARCH QUESTIONS

1. How does n-Ball Algorithm perform in Gene-Disease (g-d) Association Prediction involving highly complicated and expressive **Human Phenotype Ontology (HPO)?**
2. Is n-Ball Algorithm capable enough to distinguish between positive and negative link associations distinctively?
3. How to integrate external knowledge as formal axioms to enhance the knowledge content of ontology for better performance?

N-Ball Loss Functions Explained


D
D_r
$\vec{D}_c$
C
C ⊑ D
$\underbrace{||\vec{D}_c - \vec{C}_c||}$ + $\underbrace{(C_r - D_r)}$
Penalize if centers are far apart
Penalize if radius of subclass is greater
C
D
C'
∃R.C ⊑ D
$\underbrace{||\vec{C}_c + R - \vec{D}_c||}$ + $(C_r - D_r)$
C when translated by R should be subsumed by D

4. EXPERIMENTAL RESULTS

BENCHMARKS

- TransE
- OWL2Vec
- OPA2Vec
- DistMult
- RDF2Vec

EVALUATION METRICS

- **Hits@10,@100-** Proportion of positive test cases with rank within top 10 & 100 respectively.
- **Median, 90th Percentile Rank-** The rank below which 50th and 90th percentage of positive test cases lie.

Metric	n-Ball	TransE	OWL2Vec	OPA2Vec	DistMult	RDF2Vec
Test Split 1						
hits@10	**0.267**	0.013	0.005	0.011	0.258	**0.267**
hits@100	**0.483**	0.076	0.049	0.083	0.415	0.456
Median Rank	**113**	915	936	769	219	132
90th-P Rank	1003(2nd)	1629	1640	1609	1351	**975**
MWU p-val	2.13×10^{-226}(2nd)	0.001	0.729	0.082	1.33×10^{-177}	$\mathbf{4.14\times10^{-228}}$
Test Split 2						
hits@10	**0.292**	0.010	0.005	0.007	0.272	0.291
hits@100	**0.513**	0.076	0.046	0.072	0.438	0.460
Median Rank	**93**	868	920	728	197	136
90th-P Rank	**911**	1620	1622	1538	1294	1001
MWU p-val	$\mathbf{1.30\times10^{-250}}$	0.015	0.423	0.124	2.50×10^{-198}	7.02×10^{-221}
Test Split 3						
hits@10	0.263 (2nd)	0.011	0.006	0.013	0.258	**0.283**
hits@100	0.467 (2nd)	0.078	0.052	0.078	0.420	**0.473**
Median Rank	130 (2nd)	895	929	793	227	**122**
90th-P Rank	**958**	1624	1629	1518	1267	997
MWU p-val	9.34×10^{-228}(2nd)	0.092	0.848	0.499	7.93×10^{-194}	$\mathbf{3.48\times10^{-234}}$
Test Split 4						
hits@10	**0.284**	0.011	0.002	0.011	0.268	0.279
hits@100	**0.494**	0.070	0.052	0.082	0.436	0.454
Median Rank	**106**	879	927	866	187	134
90th-P Rank	1049 (2nd)	1618	1626	1630	1303	**1040**
MWU p-val	$\mathbf{6.59\times10^{-239}}$	2.08×10^{-6}	0.464	0.883	1.33×10^{-193}	1.24×10^{-212}

Our method performed extremely well retaining top two spots across all evaluation metrics in all test sets !!!

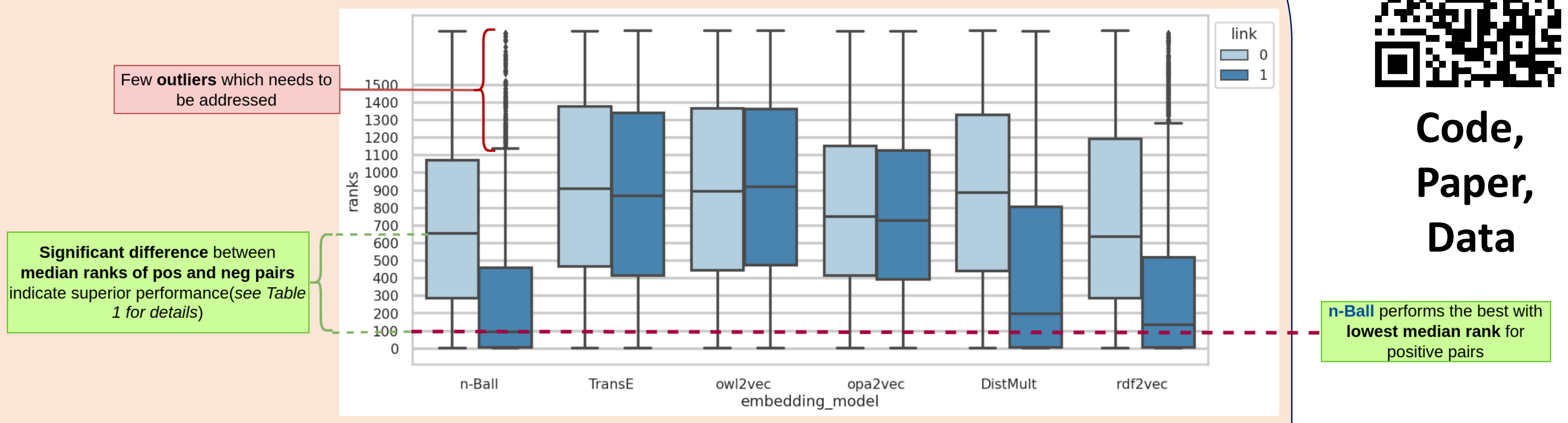
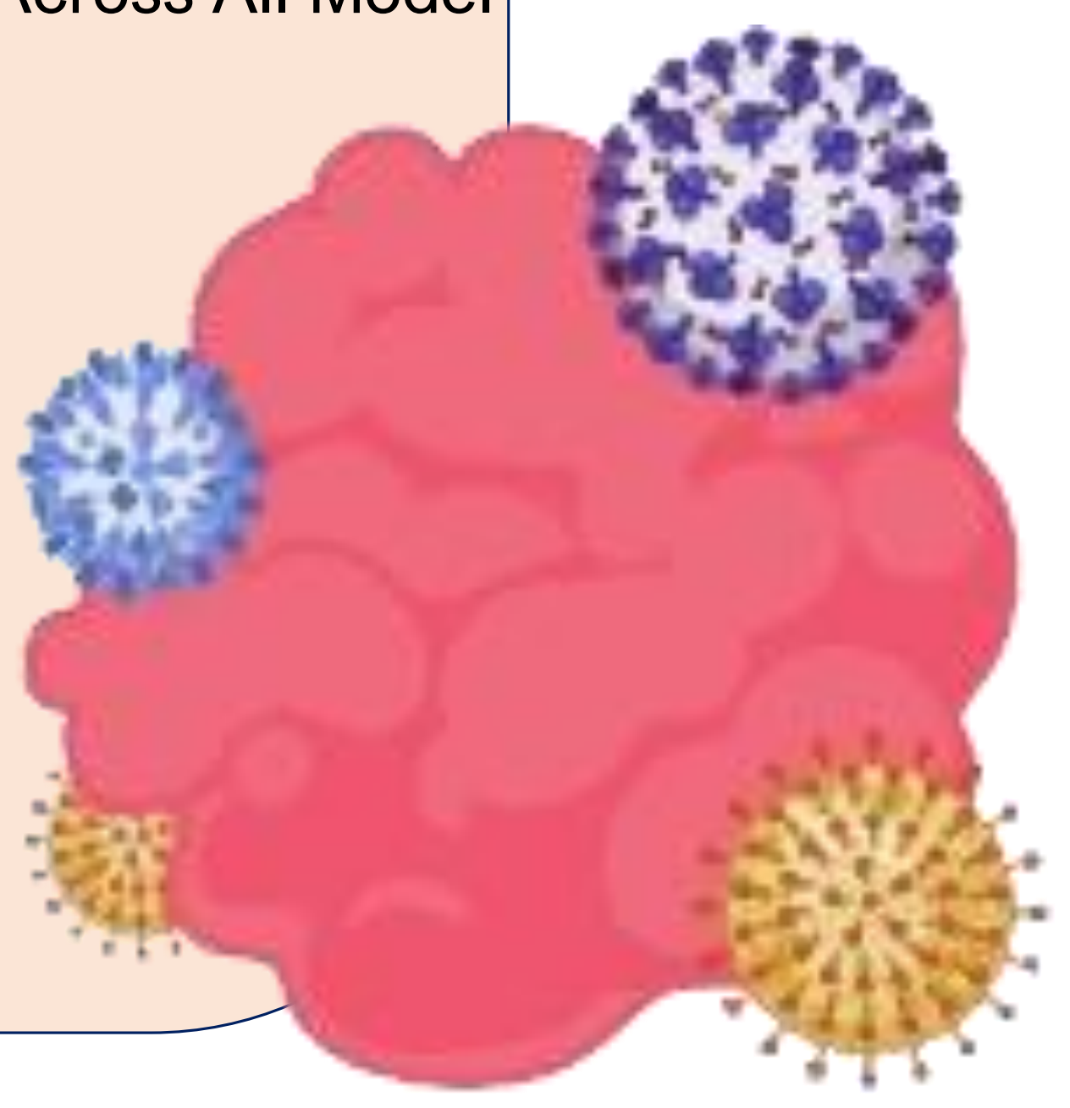


Fig: Rank Distribution of Positive & Negative Test Links Across All Model

Code, Paper, Data

EVALUATION METRICS

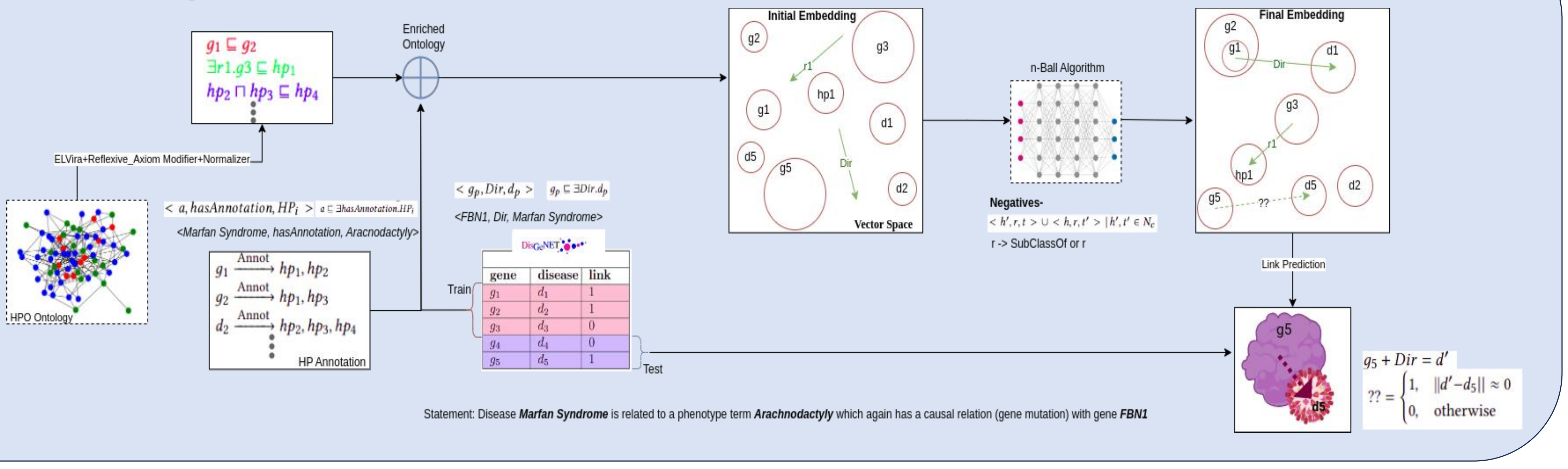
- **Mann-Whitney U (MWU) Test-** Statistical test to determine the capability of models to distinguish between positive and negative links.
  1. The best rank for a test data is 1.
  2. The overall test dataset is split into 4 disjoint sets with equal number of positive and negative examples.



3. METHOD

Step-By-Step Procedure

1. HPO ontology was reduced to EL++ equivalent using **ELVira** [3].
2. The **reflexive axioms** were modified as $< a, r, a >\sim (< a, r, b > |b \equiv a)$ and then the overall ontology was reduced to **normal forms**.
3. Enriched the ontology with **annotation data** introducing them as <*a, hasAnnotation, HP*>; where *a* denotes the disease/gene term and *HP* denotes their annotation(s).
4. We divide the gene-disease dataset into train and test(70-30 ratio) set and **introduced the positive associations** from train set as $< g_p, Dir, d_p >$ into the ontology.
5. We generated **negatives** for n-Ball algorithm by corrupting NF1 and NF3 axioms.
6. Applied the **n-Ball algorithm** and obtained the final embedding for each entity. Then link prediction for test set (g-d) is done by calculating the distance between *(g+Dir)* and *d*.


$g_1 \sqsubseteq g_2$
$\exists r1.g3 \sqsubseteq hp_1$
$hp_2 \sqcap hp_3 \sqsubseteq hp_4$
Enriched Ontology
ELVira+Reflexive_Axiom Modifier+Normalizer
HPO Ontology
$< a, hasAnnotation, HP_i >$
$a \sqsubseteq \exists hasAnnotation.HP_i$
<Marfan Syndrome, hasAnnotation, Aracnodactyly>
$g_1 \xrightarrow{Annot} hp_1, hp_2$
$g_2 \xrightarrow{Annot} hp_1, hp_3$
$d_2 \xrightarrow{Annot} hp_2, hp_3, hp_4$
HP Annotation
$< g_p, Dir, d_p >$
$g_p \sqsubseteq \exists Dir.d_p$
<FBN1, Dir, Marfan Syndrome>
DisGeNET
Train
Test
Initial Embedding
g2
g3
r1
hp1
g1
d1
d5
g5
Dir
d2
Vector Space
n-Ball Algorithm
Negatives-
$< h', r, t > \cup < h, r, t' > | h', t' \in N_c$
r -> SubClassOf or r
Final Embedding
g2
g1
Dir
d1
g3
r1
hp1
g5
??
d5
d2
Link Prediction
g5
d5
$g_5 + Dir = d'$
$?? = \begin{cases} 1, & ||d' - d_5|| \approx 0 \\ 0, & \text{otherwise} \end{cases}$

gene	disease	link
g_1	d_1	1
g_2	d_2	1
g_3	d_3	0
g_4	d_4	0
g_5	d_5	1

Statement: Disease ***Marfan Syndrome*** is related to a phenotype term ***Arachnodactyly*** which again has a causal relation (gene mutation) with gene ***FBN1***

DATASET

- **HPO** and **HPO annotations** were downloaded from HPO website+ *latest version as of May 2023*.
- Gene-disease associations were collected from **DisGeNET!** .

+ https://hpo.jax.org/app/ ! https://www.disgenet.org/downloads

5. KEY TAKEAWAYS

- Region-based KGE Algorithms can represent ontological data **accurately** in vector space.
- Addition of relevant **external knowledge** enhances the reasoning capability.
- Has the potential to make **disease diagnosis** and treatment much safer by **accurate** and **trustable** identification of genetic causes.
- Can be used in **personalized medicine** and **differential diagnostics.**

6. FUTURE WORK

- Reduction to EL++ DL removes lots of expressive axioms, so we look to extend Region-based Embedding to **more expressive languages** like ***SROIQ***.
- Ontology contains relevant information in the form of **lexical annotations** like *definition*, *synonyms*, etc which can be integrated as formal axioms to increase knowledge content.

REFERENCES:

[1] **Kulmanov, Maxat, et al. "El embeddings: Geometric construction of models for the description logic el++."** ***arXiv preprint arXiv:1902.10499*** **(2019).**

[2] **Baader, Franz, Sebastian Brandt, and Carsten Lutz. "Pushing the EL envelope." (2005): 364-369.**

[3] **Hoehndorf, Robert, et al. "A common layer of interoperability for biomedical ontologies based on OWL EL."** ***Bioinformatics*** **27.7 (2011): 1001-1008.**