Applications of mOWL for Predicting Gene-Disease Associations

Machine Learning with Biomedical Ontologies
SWAT4HCLS 2023 Conference, Basel, Feb 13-16

SWAT4LS

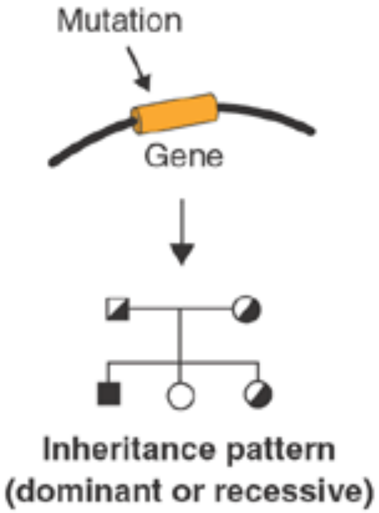
13 Feb 2023

Introduction: **Problem**

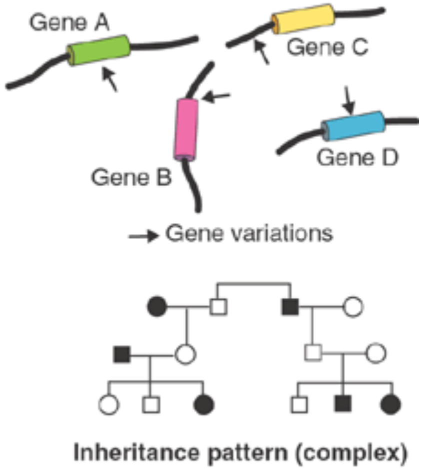
Patients Clinical Phenotypes **&** Genomics Sequence Data

Find the causative genes

Rare/Monogenic Diseases



Complex Diseases



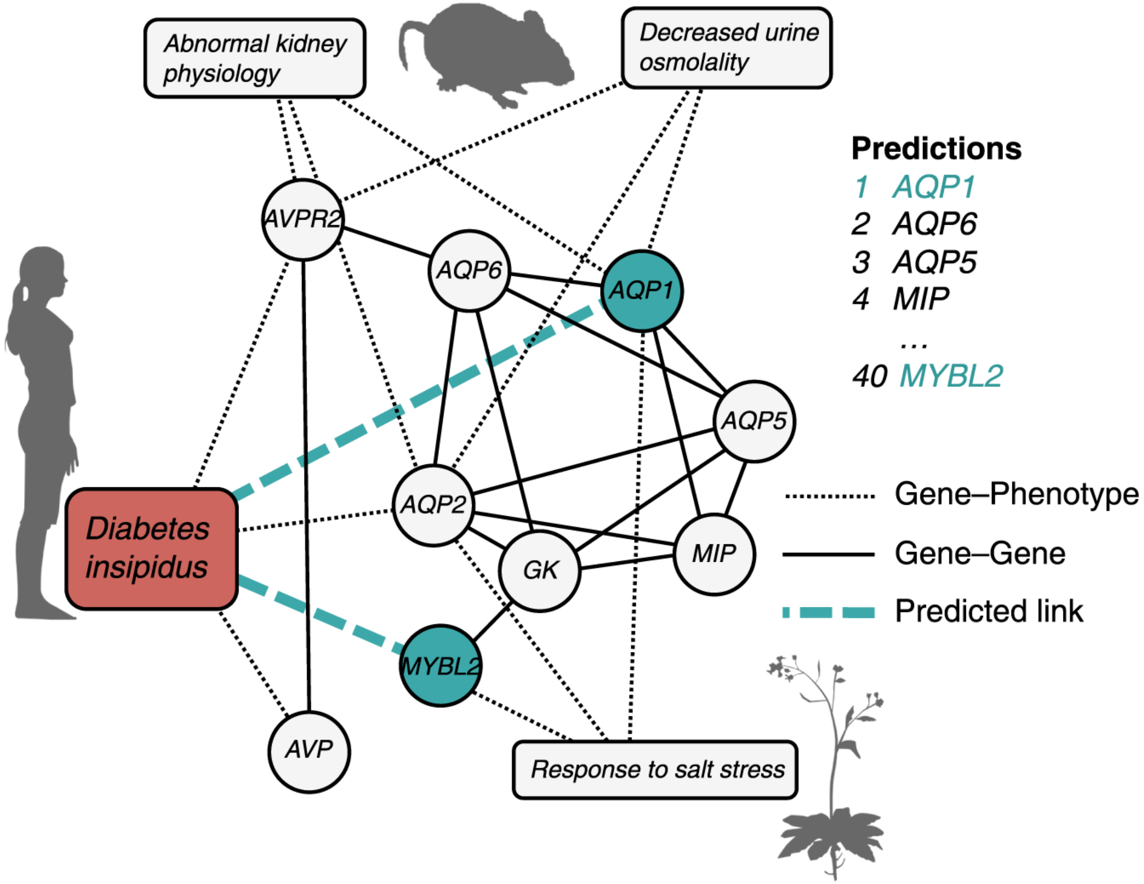
Prediction/Diagnosis

Introduction: **Problem**

- Predicting gene-disease associations based on phenotypic similarity
- Diagnosis of disease based on phenotypic similarity

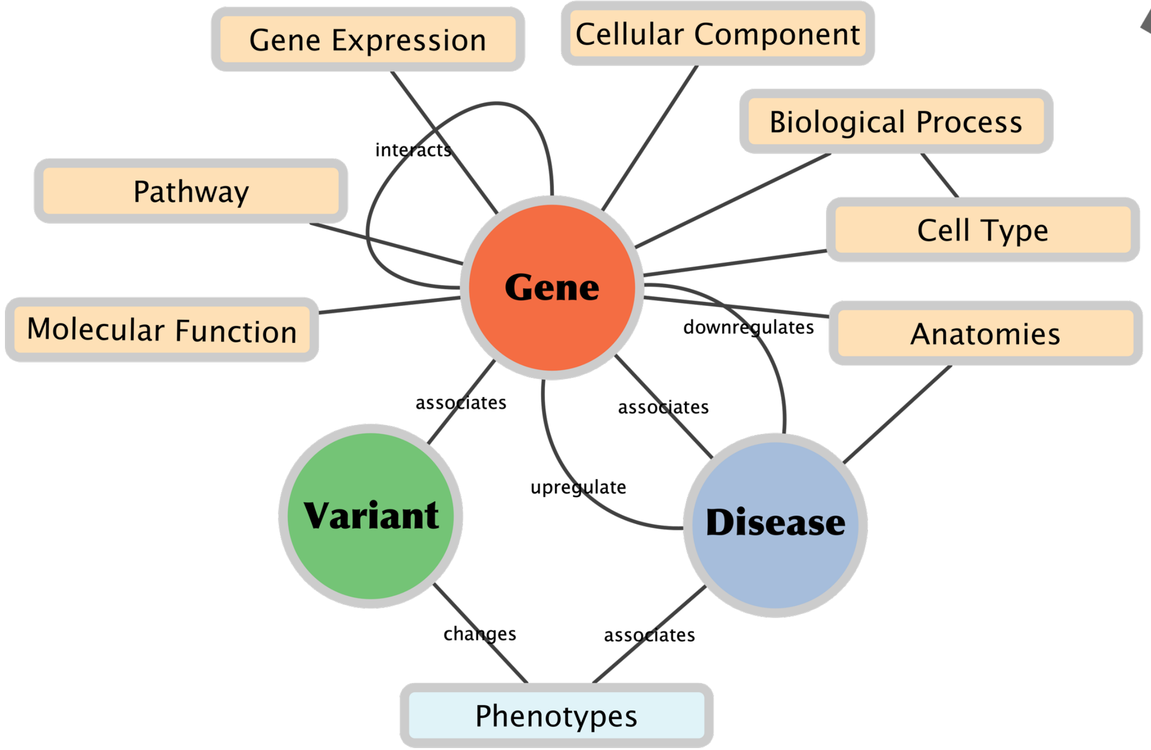
Introduction: **Phenotypic similarity**

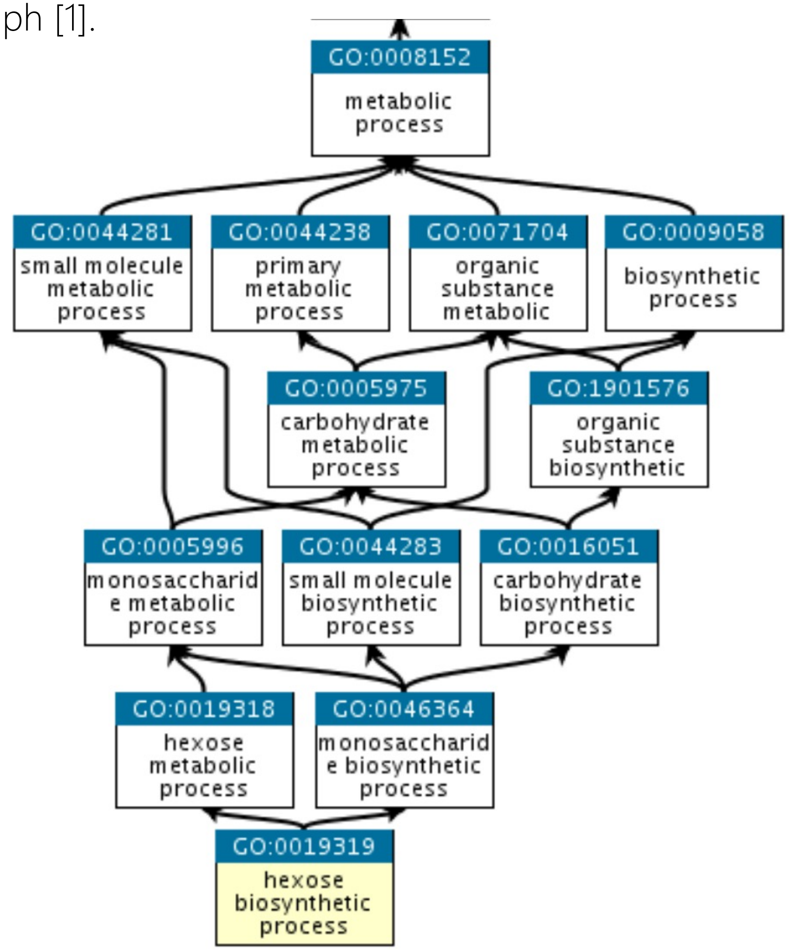
- Using the phenotypes of model organism genes and the diseases' phenotypes.



Ontologies to Knowledge Graph

Converting knowledge axioms in phenotype ontologies to knowledge graph [1].




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

[1] Chen, Jun*, Azza Althagafi*, and Robert Hoehndorf. "Predicting candidate genes from phenotypes, functions and anatomical site of expression." Bioinformatics 37.6 (2021): 853-860.

Phenotypic characteristics for Genes

- **Using functional and phenotypic characteristics for genes in:**

Human phenome	Mouse phenome	Functions of the gene products	Gene expression in individual cell types	Anatomical site of expression from the GTEx tissue expression
Human Phenotype Ontology (HPO)	Mammalian Phenotype Ontology (MP)	Gene Ontology (GO)	Celltype Ontology (CL)	Uber-anatomy ontology (UBERON)
4,315 genes & 169,281 associations	13,529 genes & 168,550 associations	17,786 genes & 208,630 associations	6,559 genes & 17,149 associations	20,538 genes & 585,765 associations

Knowledge Graph Embeddings

- Generate the representation from the ontology graph (using mOWL).
- Collect features for the gene and disease using different method.

Syntactic embeddings	Onto2vec OPA2Vec
Graph-based embeddings	DL2vec OWL2Vec*

Knowledge Graph Embeddings

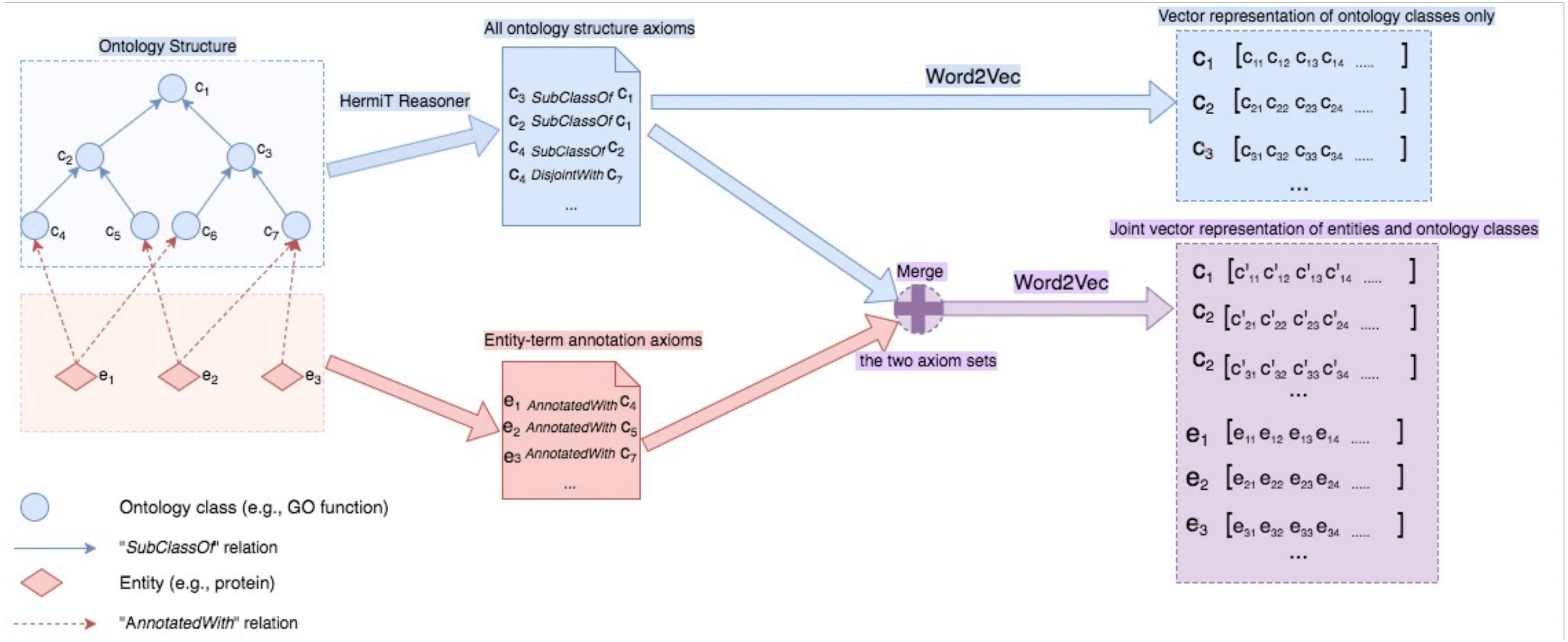

Ontology Structure
c_1
c_2
c_3
c_4
c_5
c_6
c_7
e_1
e_2
e_3
HermiT Reasoner
All ontology structure axioms
c_3 *SubClassOf* c_1
c_2 *SubClassOf* c_1
c_4 *SubClassOf* c_2
c_4 *DisjointWith* c_7
...
Entity-term annotation axioms
e_1 *AnnotatedWith* c_4
e_2 *AnnotatedWith* c_5
e_3 *AnnotatedWith* c_7
...
Word2Vec
Merge
the two axiom sets
Word2Vec
Vector representation of ontology classes only
c_1 $[c_{11}\ c_{12}\ c_{13}\ c_{14}\ .....\ ]$
c_2 $[c_{21}\ c_{22}\ c_{23}\ c_{24}\ .....\ ]$
c_3 $[c_{31}\ c_{32}\ c_{33}\ c_{34}\ .....\ ]$
...
Joint vector representation of entities and ontology classes
c_1 $[c'_{11}\ c'_{12}\ c'_{13}\ c'_{14}\ .....\ ]$
c_2 $[c'_{21}\ c'_{22}\ c'_{23}\ c'_{24}\ .....\ ]$
c_2 $[c'_{31}\ c'_{32}\ c'_{33}\ c'_{34}\ .....\ ]$
...
e_1 $[e_{11}\ e_{12}\ e_{13}\ e_{14}\ .....\ ]$
e_2 $[e_{21}\ e_{22}\ e_{23}\ e_{24}\ .....\ ]$
e_3 $[e_{31}\ e_{32}\ e_{33}\ e_{34}\ .....\ ]$
...
Ontology class (e.g., GO function)
"*SubClassOf*" relation
Entity (e.g., protein)
"*AnnotatedWith*" relation

Fig1: Onto2Vec Workflow.

Knowledge Graph Embeddings

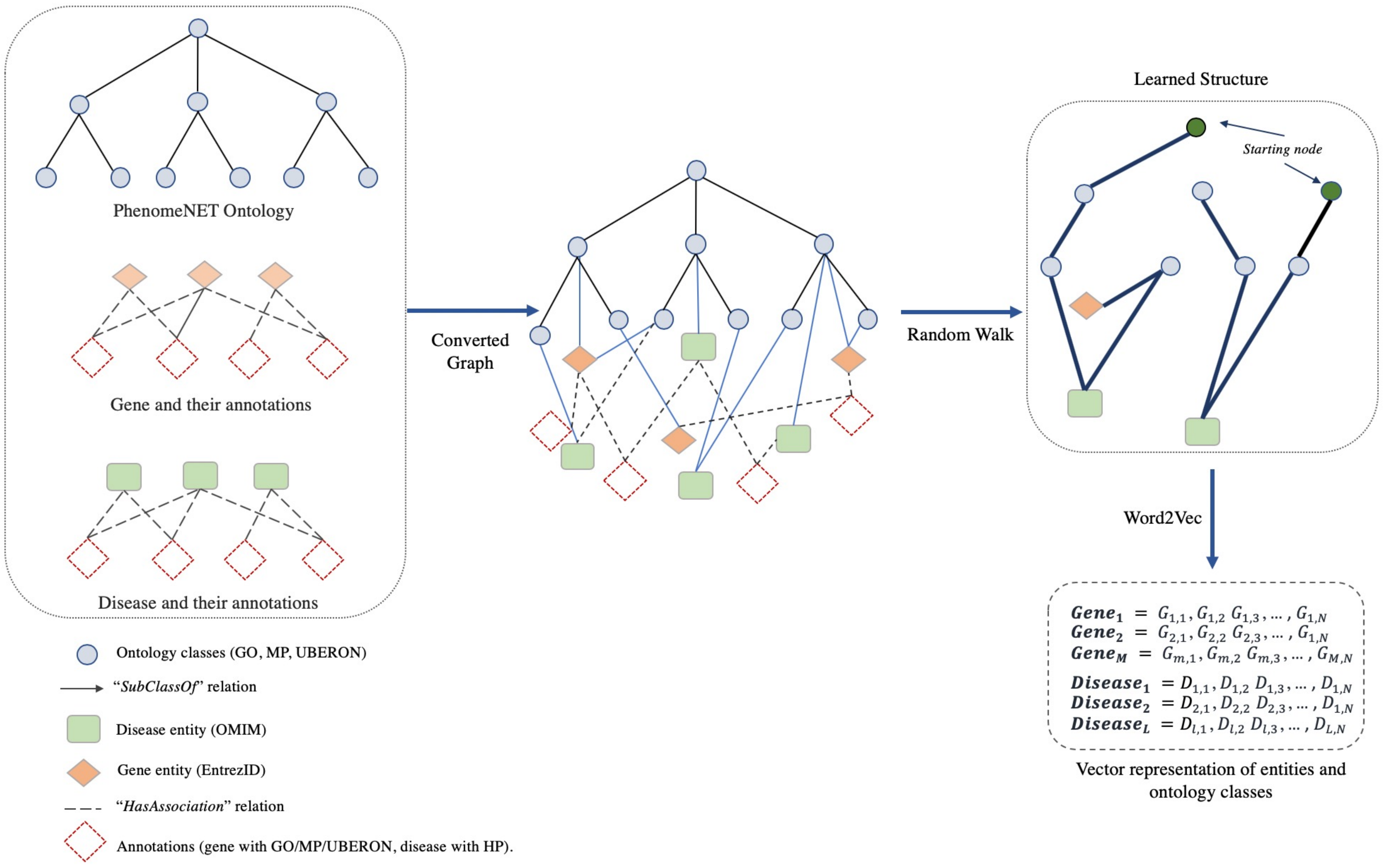
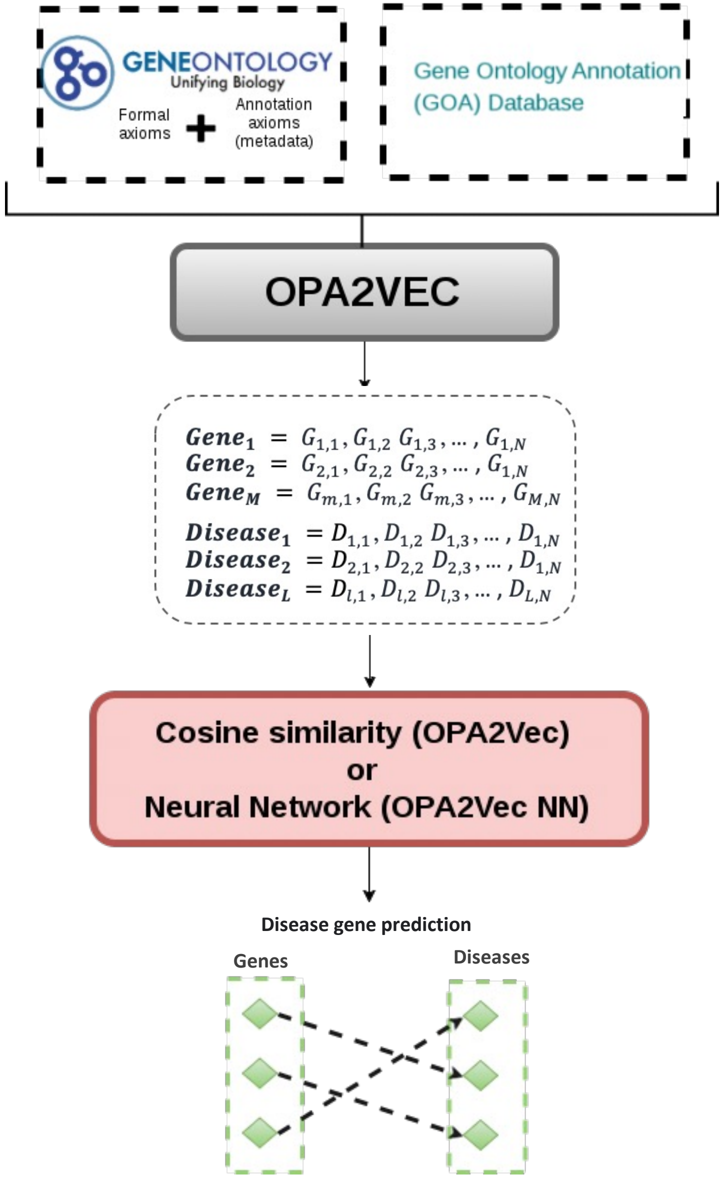

PhenomeNET Ontology
Gene and their annotations
Disease and their annotations
Converted Graph
Random Walk
Learned Structure
Starting node
Word2Vec
$Gene_1 = G_{1,1}, G_{1,2}\ G_{1,3}, \ldots, G_{1,N}$
$Gene_2 = G_{2,1}, G_{2,2}\ G_{2,3}, \ldots, G_{1,N}$
$Gene_M = G_{m,1}, G_{m,2}\ G_{m,3}, \ldots, G_{M,N}$
$Disease_1 = D_{1,1}, D_{1,2}\ D_{1,3}, \ldots, D_{1,N}$
$Disease_2 = D_{2,1}, D_{2,2}\ D_{2,3}, \ldots, D_{1,N}$
$Disease_L = D_{l,1}, D_{l,2}\ D_{l,3}, \ldots, D_{L,N}$
Vector representation of entities and ontology classes
Ontology classes (GO, MP, UBERON)
"SubClassOf" relation
Disease entity (OMIM)
Gene entity (EntrezID)
"HasAssociation" relation
Annotations (gene with GO/MP/UBERON, disease with HP).

Fig2: DL2vec Workflow

Knowledge Graph Embeddings



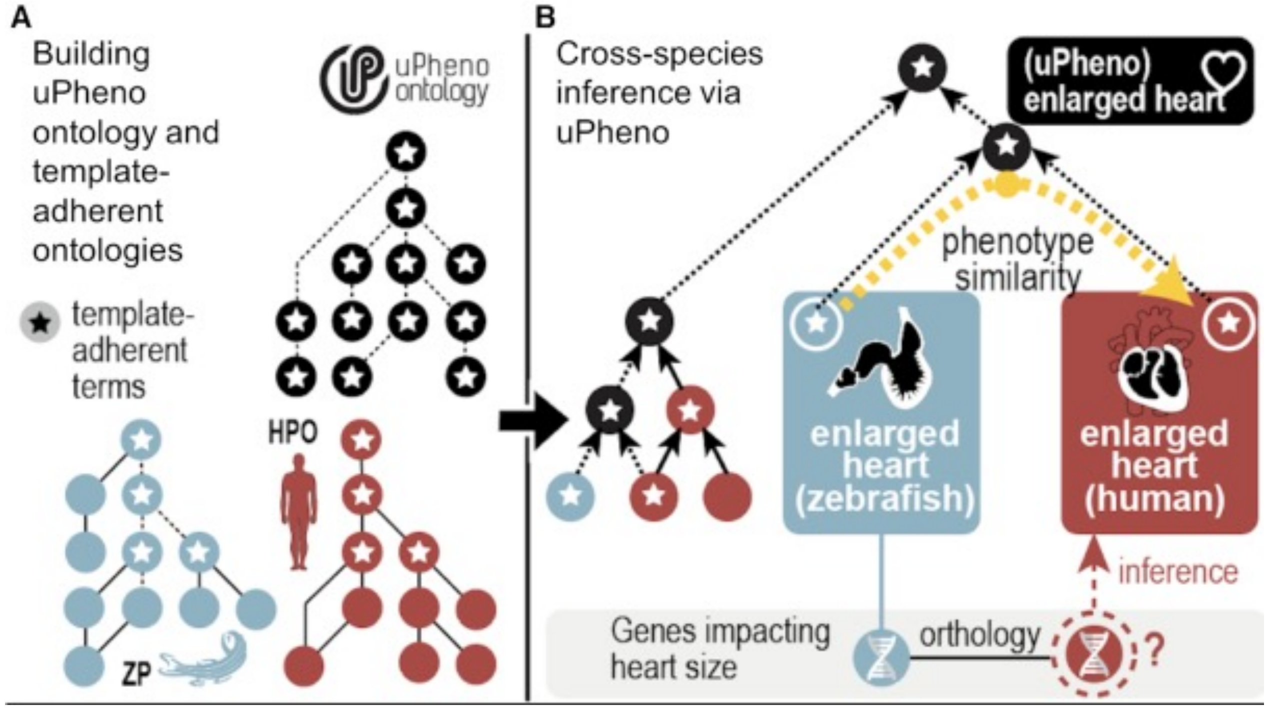
GENEONTOLOGY
Unifying Biology
Formal axioms
+
Annotation axioms (metadata)
Gene Ontology Annotation (GOA) Database
OPA2VEC

$$
\begin{aligned}
\boldsymbol{Gene_1} &= G_{1,1}, G_{1,2}\ G_{1,3}, \dots, G_{1,N} \\
\boldsymbol{Gene_2} &= G_{2,1}, G_{2,2}\ G_{2,3}, \dots, G_{1,N} \\
\boldsymbol{Gene_M} &= G_{m,1}, G_{m,2}\ G_{m,3}, \dots, G_{M,N} \\
\boldsymbol{Disease_1} &= D_{1,1}, D_{1,2}\ D_{1,3}, \dots, D_{1,N} \\
\boldsymbol{Disease_2} &= D_{2,1}, D_{2,2}\ D_{2,3}, \dots, D_{1,N} \\
\boldsymbol{Disease_L} &= D_{l,1}, D_{l,2}\ D_{l,3}, \dots, D_{L,N}
\end{aligned}
$$

Cosine similarity (OPA2Vec)
or
Neural Network (OPA2Vec NN)
Disease gene prediction
Genes
Diseases

Datasets: Ontology

- The phenotypes are described using different organism-specific phenotype ontologies.
- Unified Phenotype Ontology (uPheno) include human phenotypes from the Human Phenotype Ontology (HP), → relate mutant model organism phenotypes to human disease-associated phenotypes.


A
Building uPheno ontology and template-adherent ontologies
uPheno ontology
template-adherent terms
HPO
ZP
B
Cross-species inference via uPheno
(uPheno) enlarged heart
phenotype similarity
enlarged heart (zebrafish)
enlarged heart (human)
inference
Genes impacting heart size
orthology
?

Datasets: Genes and Diseases Phenotypes

- **Gene annotations:**
  - Mouse/Human Orthology with Phenotype Annotations from HPO database (HMD_HumanPhenotype.rpt).

- **Disease annotations:**
  - obtained from the HPO annotations for rare diseases document (phenotype.hpoa).

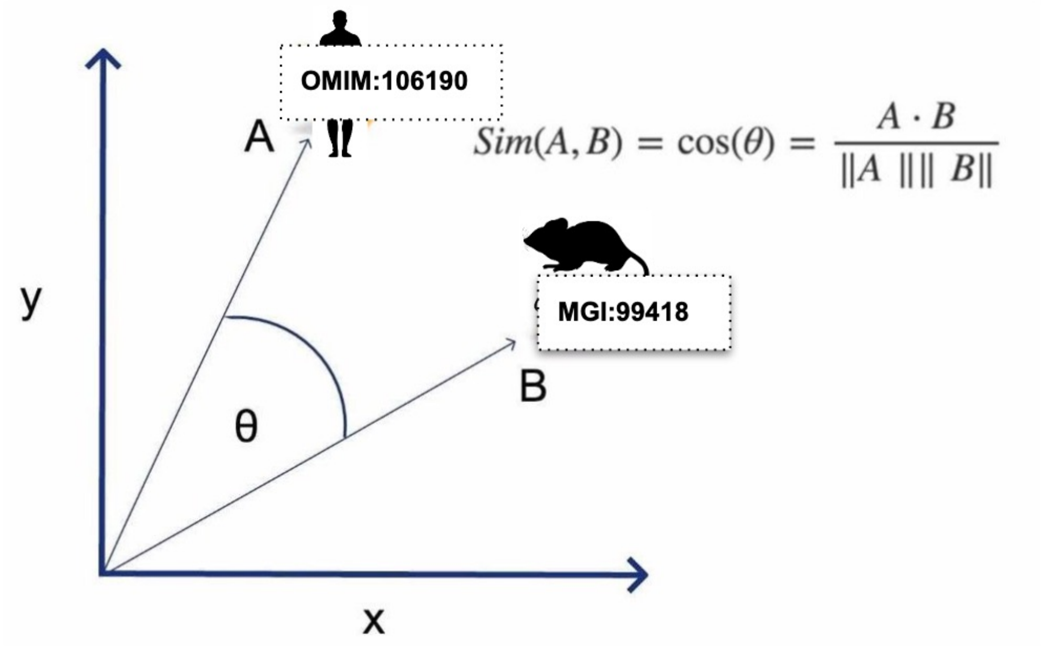
- These annotations added to the Unified Phenotype Ontology (uPheno) to build the training ontology.

Methods: Generate Embeddings

- mOWL is designed to handle input in OWL format.
  - A mOWL dataset contains 3 ontologies: training, validation and testing.
- Preparee the annoations file (`Genes -> Phenotypes , Diseasese -> Phenotypes`)
- Use mOWL methods to build dataset given an ontology file and the annotations to be inserted to the ontology.
  - Per each row, the first element is the annotated entity and the rest of the elements are the annotating entities (which are the entities in the ontology).
- Use different methods to generate the representation given the annotated ontology file.

Evaluation: Calculating Phenotypic Similarity Approaches

- **Unsupervised Approach**
  - Cosine similarity



$$Sim(A,B) = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|}$$

Hands-on

Hands On Tutorial

1. Prepare the dataset (Built-in / your own dataset)
2. Generate the embeddings using different methods
3. Prediction
4. Validation

Other application: **Prediction the causative variants**

Patients Clinical Phenotypes & Genomics Data

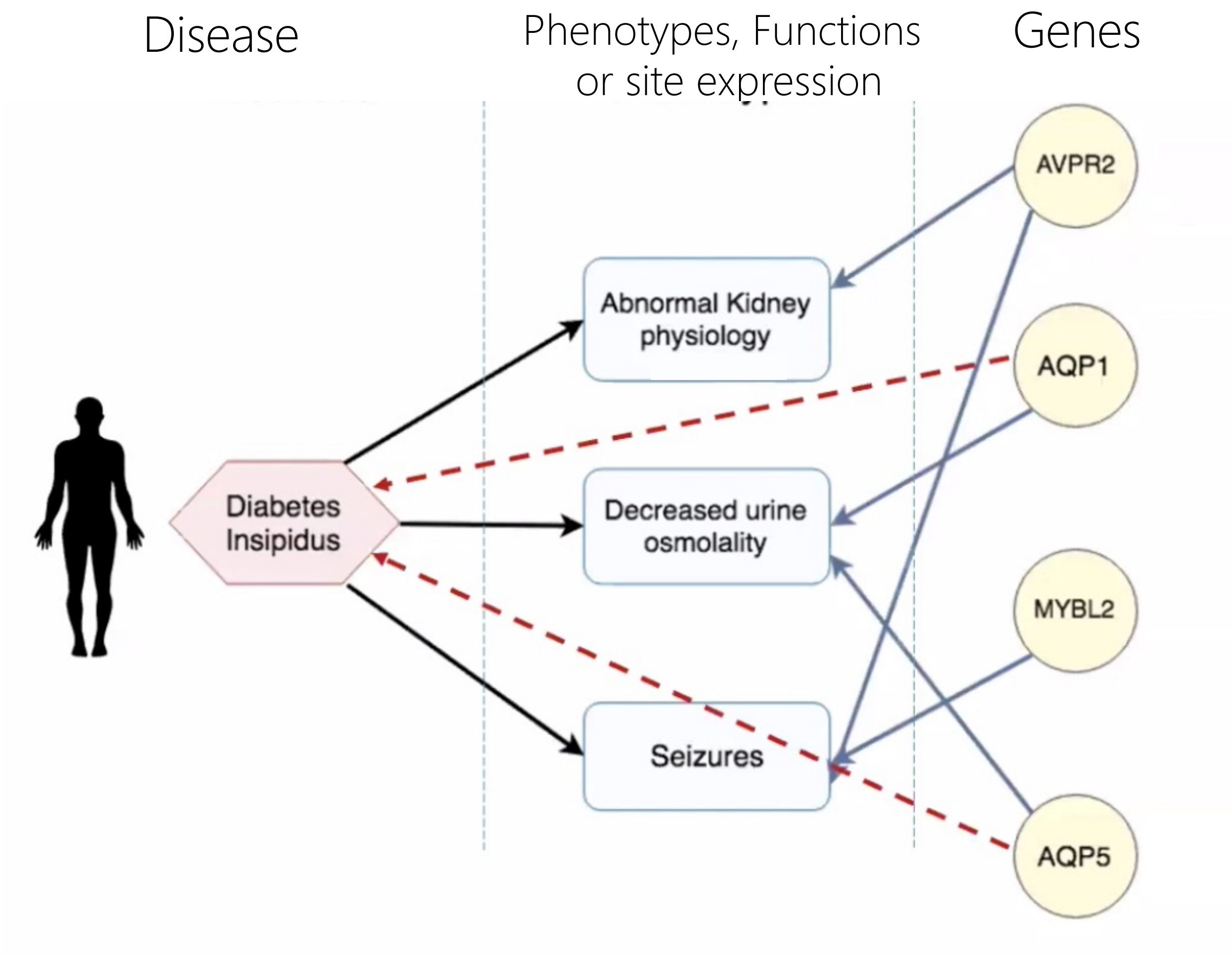
Can we find the causative variants associated with the phenotypes?

Other application: **Prediction the causative variants**

- Prioritizing the causative variants.
  - Determining which variants identified using Whole-exome Sequencing (WES) or Whole-genome sequencing (WGS) are most likely to damage gene function and underlie the disease phenotype.

Other application: **Prediction the causative variants**

Phenotype-based prioritization of candidate genes



Disease
Phenotypes, Functions or site expression
Genes
Diabetes Insipidus
Abnormal Kidney physiology
Decreased urine osmolality
Seizures
AVPR2
AQP1
MYBL2
AQP5

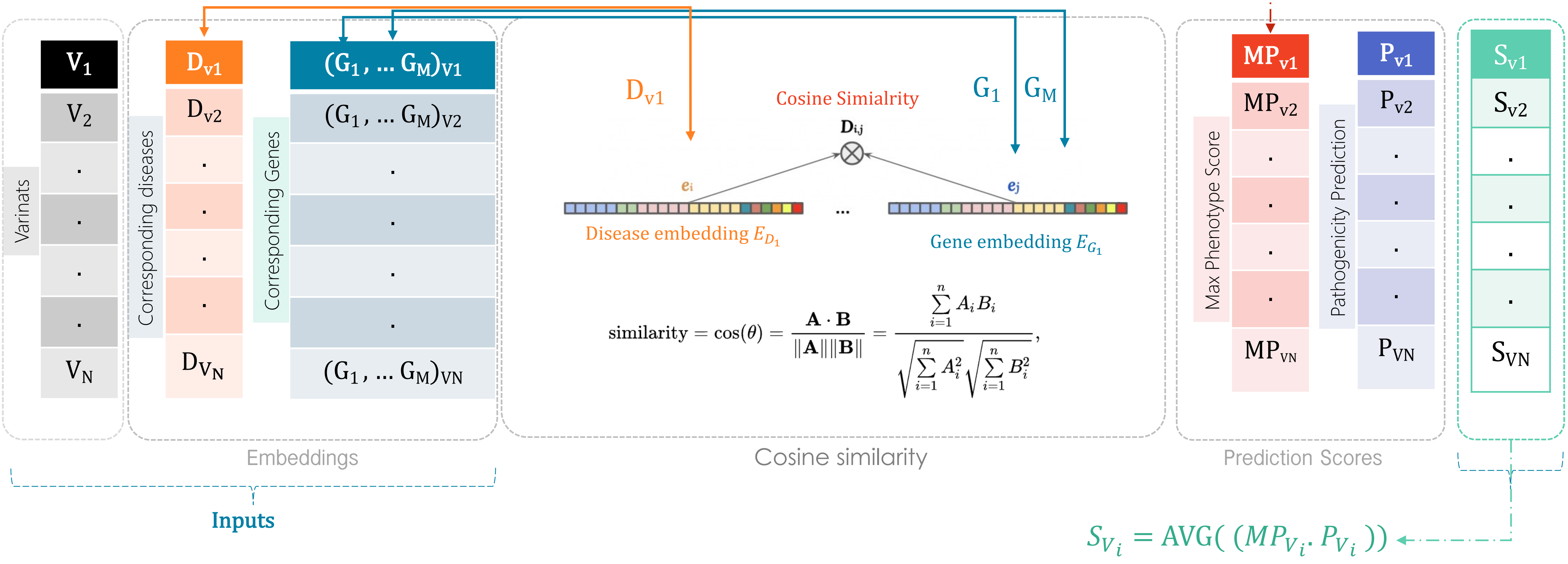
- Gene-Phenotype annotations
- Gene-Function annotations
- Gene-Celltype annotations
- Gene-Anatomical site of gene expression annotations

- Disease-Phenotypes annotations

➢ Prioritize the candidate genes

Other application: Prediction the causative variants

$$MP_{V_i} = \text{Max}(\, (D_{V_i}.\, G_{1_{V_i}}\,)\,, (D_{V_i}.\, G_{2_{V_i}}\,), \ldots, (D_{V_i}.\, G_{M_{V_i}})\,)$$



$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}},$$

$$S_{V_i} = \text{AVG}(\,(MP_{V_i}.\, P_{V_i}\,))$$

Resources

- Zhapa-Camacho, Fernando, Maxat Kulmanov, and Robert Hoehndorf. "mOWL: Python library for machine learning with biomedical ontologies." Bioinformatics 39.1 (2023): btac811.
- Chen, Jun, Azza Althagafi, and Robert Hoehndorf. "Predicting candidate genes from phenotypes, functions and anatomical site of expression." Bioinformatics 37.6 (2021): 853-860.
- Smaili, Fatima Zohra, Xin Gao, and Robert Hoehndorf. "OPA2Vec: combining formal and informal content of biomedical ontologies to improve similarity-based prediction." Bioinformatics 35.12 (2019): 2133-2140.
- Smaili, Fatima Zohra, Xin Gao, and Robert Hoehndorf. "Onto2vec: joint vector-based representation of biological entities and their ontology-based annotations." Bioinformatics 34.13 (2018): i52-i60.
- Chen, Jiaoyan, et al. "Owl2vec*: Embedding of owl ontologies." Machine Learning 110.7 (2021): 1813-1845.

THANK YOU!