

Applications of mOWL for Predicting Gene-Disease Associations

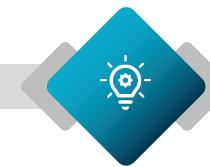
Azza Thamer Althagafi

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



Roadmap

Problem
motivation for predicting gene-disease associations



Dataset
Data sources: ontologies, annotations and associations



mOWL dataset
mOWL input files: convert data to axioms, OWL format



Prediction Methods
Graph-based & Syntactic embeddings in mOWL



Evaluation
Evaluating the prediction of the association

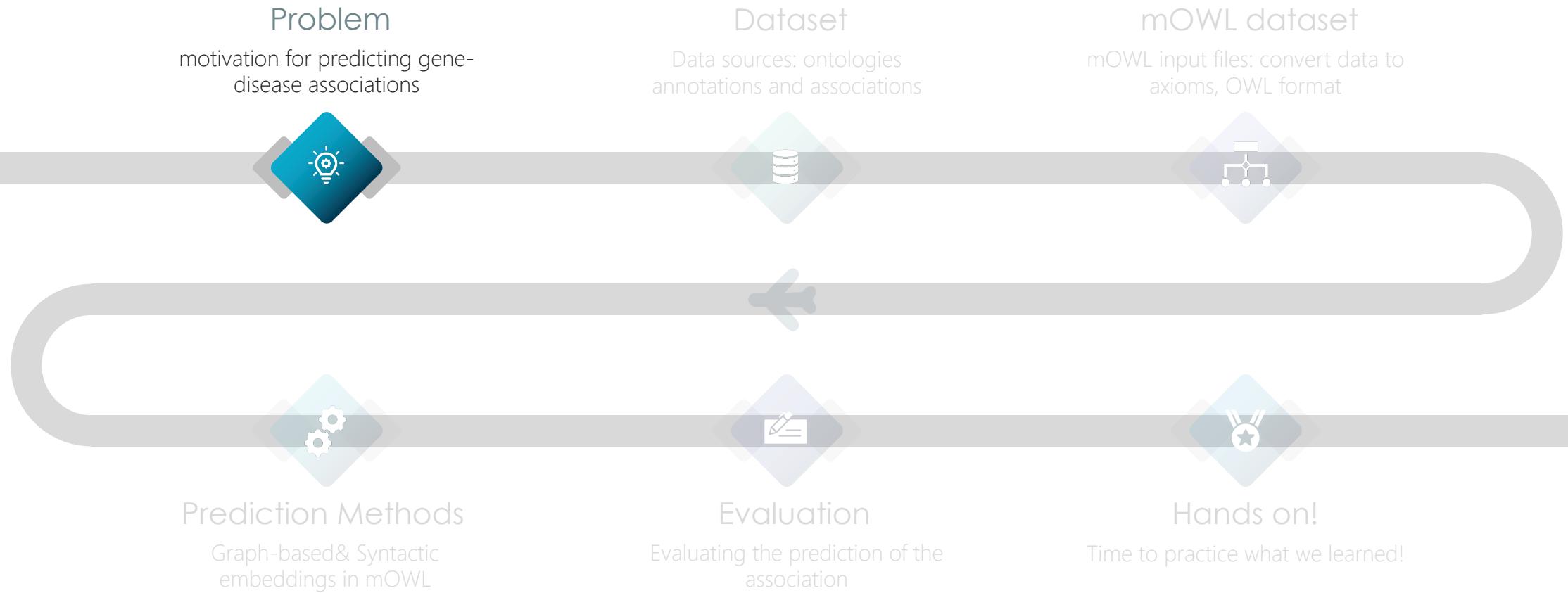


Hands on!
Time to practice what we learned!





Roadmap





Introduction: Problem



Patients Clinical Phenotypes &
Genomics Sequence Data

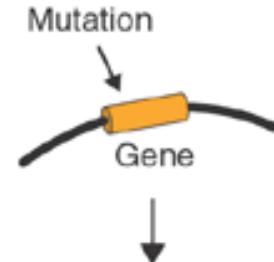


Find the causative genes



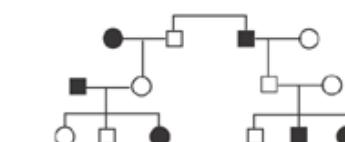
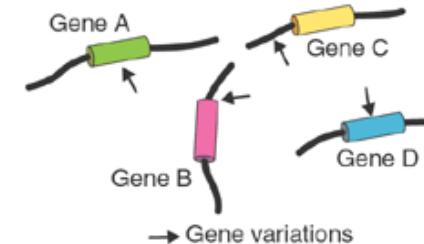
Prediction/Diagnosis

Rare/Monogenic Diseases



Inheritance pattern
(dominant or recessive)

Complex Diseases



Inheritance pattern (complex)



Introduction: Problem

- **STXBP1** Gene -> **Encephalopathy**

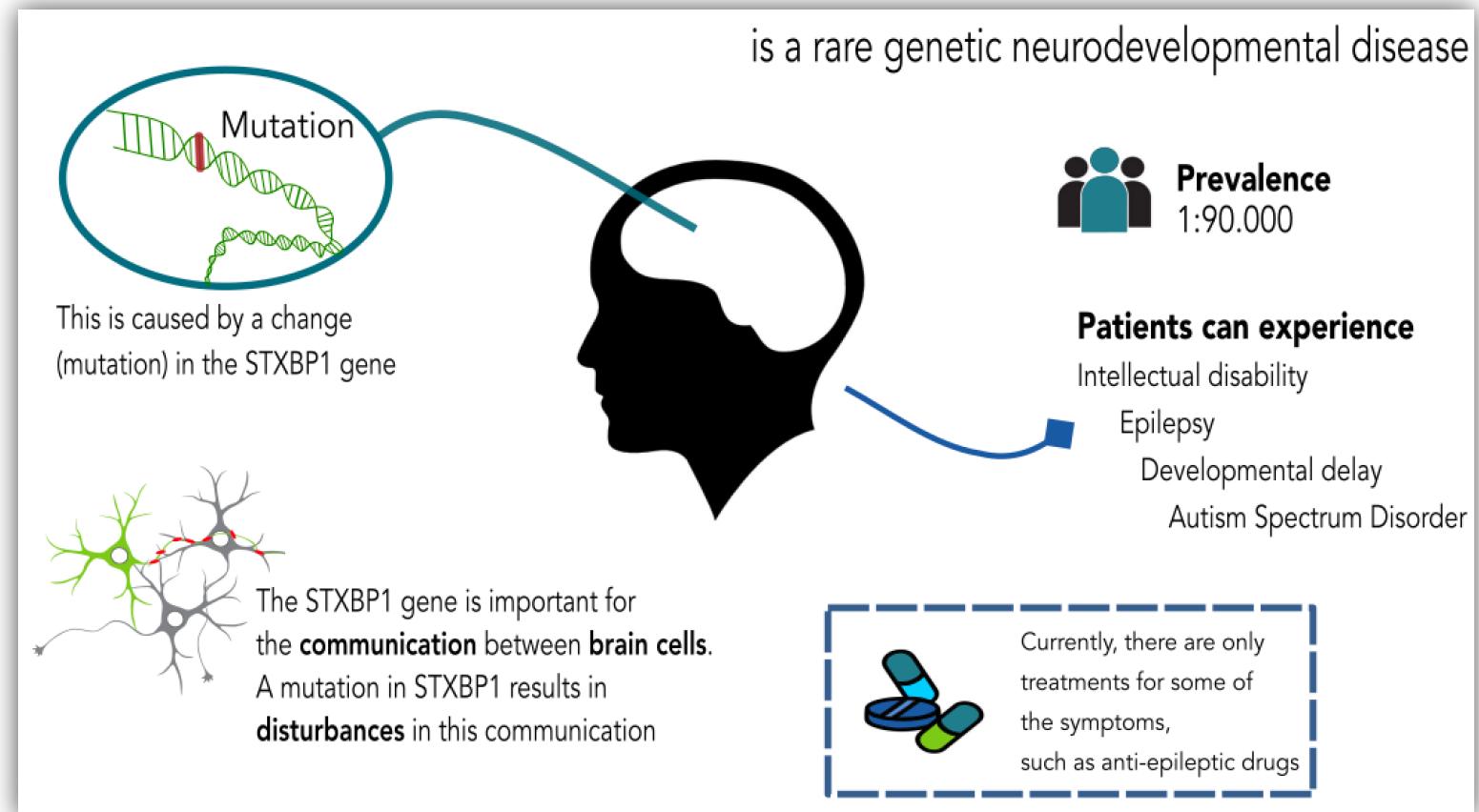


Fig: What is Encephalopathy?

¹ Information for patients, What is STXBP1-Encephalopathy, accessed by https://stxbp1.cnrc.nl/stxbp1_disorders



Introduction: Problem

- **STXBP1** Gene -> **Encephalopathy**
- **Disease gene prediction:** The process of determining which genes identified in the course of Whole-genome sequencing (WGS) are most likely to **damage** gene function and underlie the **disease phenotype**.
- Use disease phenotype to aid the prioritization.

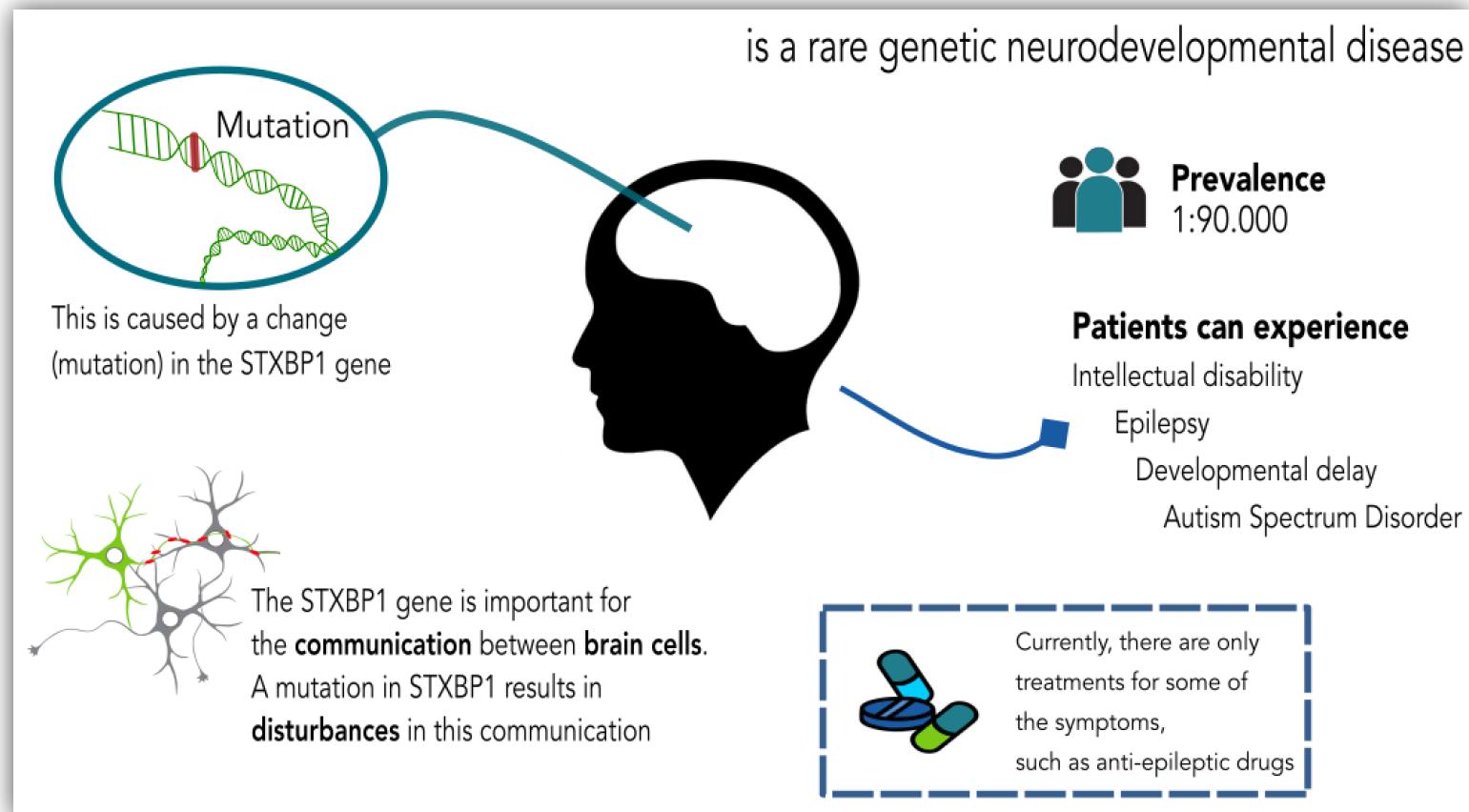


Fig: What is Encephalopathy?

¹ Information for patients, What is STXBP1-Encephalopathy, accessed by https://stxbp1.cnrc.nl/stxbp1_disorders



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.

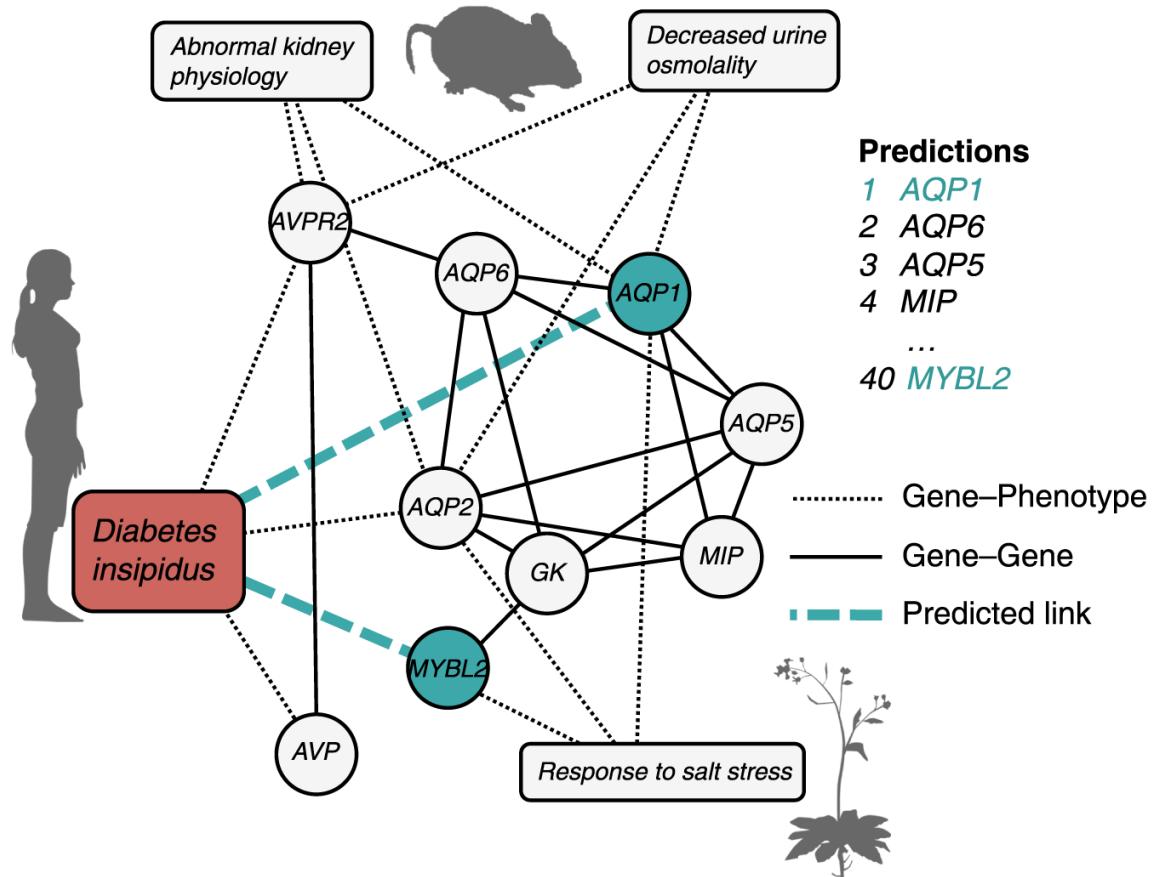
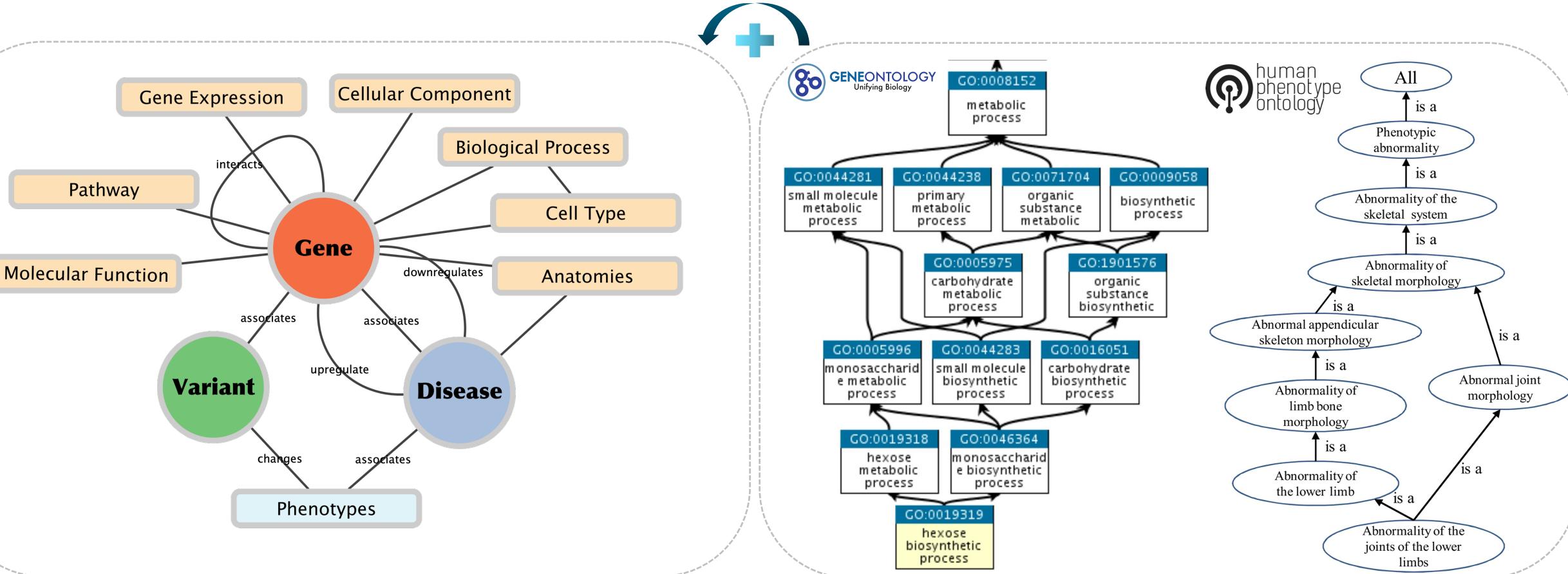


Fig: Prediction and Validation of Gene-Disease Associations Using Methods Inspired by Social Network Analyses¹



Introduction: Ontologies to Knowledge Graph

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

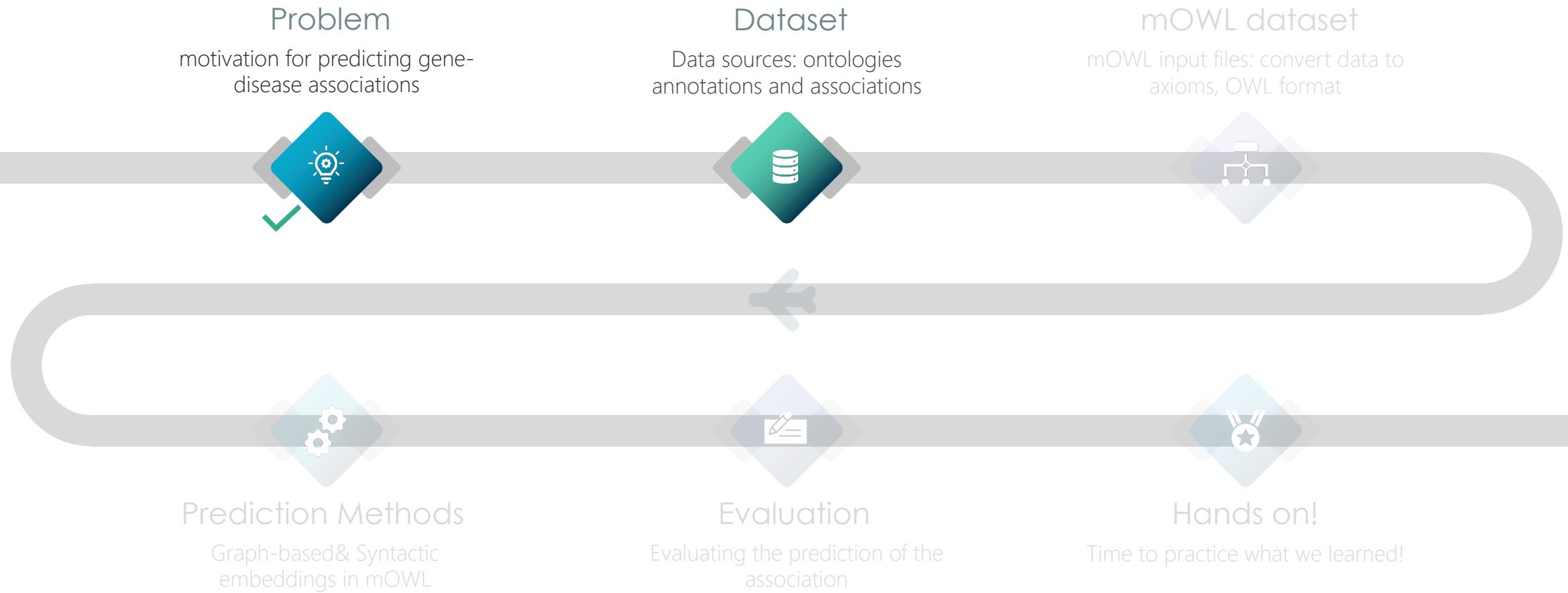


Knowledge Graph for all the phenotype annotations and information from different databases

Examples of ontologies: Gene Ontology (GO) "left" and Human Phenotype Ontology (HPO) "right"



Roadmap





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

Human phenotype	Mouse phenotype	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



Dataset: 1. Disease phenotype annotations

Diseases phenotypes:

- Obtained from the HPO annotations for rare diseases document ([phenotype.hpoa](#)).

```
#description: HPO annotations for rare diseases [8324: OMIM; 47: DECIPHER; 4264 ORPHANET]
#date: 2022-12-20
#tracker: https://github.com/obophenotype/human-phenotype-ontology
#HPO-version: http://purl.obolibrary.org/obo/hp/releases/2022-12-20/hp.json
```

#DatabaseID	DiseaseName	Qualifier	HPO_ID	Reference	Evidence	Onset	Frequency	Sex
OMIM:619340	Developmental and epileptic encephalopathy	96	HP:0011097		PMID:31675180	PCS		
OMIM:619340	Developmental and epileptic encephalopathy	96	HP:0002187		PMID:31675180	PCS		
OMIM:619340	Developmental and epileptic encephalopathy	96	HP:0001518		PMID:31675180	PCS		
OMIM:619340	Developmental and epileptic encephalopathy	96	HP:0032792		PMID:31675180	PCS		
OMIM:619340	Developmental and epileptic encephalopathy	96	HP:0011451		PMID:31675180	PCS		
OMIM:619340	Developmental and epileptic encephalopathy	96	HP:0010851		PMID:31675180	PCS		
OMIM:619340	Developmental and epileptic encephalopathy	96	HP:0001789		PMID:31675180	PCS		
OMIM:619340	Developmental and epileptic encephalopathy	96	HP:0200134		PMID:31675180	PCS		
OMIM:619340	Developmental and epileptic encephalopathy	96	HP:0001522		PMID:31675180	PCS		
OMIM:619340	Developmental and epileptic encephalopathy	96	HP:0000006		PMID:31675180	PCS		
OMIM:619340	Developmental and epileptic encephalopathy	96	HP:0002643		PMID:31675180	PCS		

Diseases ID

Human Phenotype Ontology (HPO)



Dataset: 2. Gene phenotypes annotations

- **Genes phenotypes:**

- Download Mouse/Human Ortholog with Phenotype Annotations from HPO database ([HMD_HumanPhenotype.rpt](#))

A1BG	1	A1bg	MGI:2152878
A1CF	29974	A1cf	MGI:1917115
A2M	2	A2m	MGI:2449119
A3GALT2	127550	A3galt2	MGI:2685279
A4GALT	53947	A4galt	MGI:3512453
A4GNT	51146	A4gnt	MGI:2143261
AAAS	8086	Aaas	MGI:2443767
AACS	65985	Aacs	MGI:1926144
AADAC	13	Aadac	MGI:1915008

EntrezGene ID

MP:0005367, MP:0005369, MP:0005370, MP:0005376, MP:0005378,
MP:0005376, MP:0005386, MP:0010768
MP:0002006, MP:0005381, MP:0005384, MP:0005385, MP:0005387
MP:0005378, MP:0005386, MP:0005389
MP:0005386
MP:0005367, MP:0005378, MP:0005379, MP:0005387, MP:0005389,

Mammalian Phenotype (MP) Ontology



Dataset: 3. Gene-Disease associations

Gene-Disease associations:

- Gene-disease associations were obtained from the Associations of Mouse Genes with DO Diseases file downloaded from ([MGI DO.rpt](#)).

DO Disease ID	DO Disease Name	OMIM IDs	Common Organism Name	NCBI Taxon ID	Symbol	EntrezGene ID	Mouse MGI ID
DOID:0112248	17-beta hydroxysteroid dehydrogenase 3 deficiency			OMIM:264300	human	9606	HSD17B3 3293
DOID:0112248	17-beta hydroxysteroid dehydrogenase 3 deficiency			OMIM:264300	mouse, laboratory	10090	Hsd17b3 15487 MGI:107177
DOID:0111453	2-amino adipic 2-oxoadipic aciduria	OMIM:204750		human	9606	DHTKD1	55526
DOID:0111453	2-amino adipic 2-oxoadipic aciduria	OMIM:204750		mouse, laboratory	10090	Dhtkd1	209692 MGI:2445096
DOID:0050573	2-hydroxyglutaric aciduria		human	9606	L2HGDH	79944	
DOID:0050573	2-hydroxyglutaric aciduria		human	9606	SLC25A1	6576	
DOID:0060575	3MC syndrome 1	OMIM:257920	human	9606	MASP1	5648	
DOID:0060576	3MC syndrome 2	OMIM:265050	human	9606	COLEC11	78989	
DOID:0060577	3MC syndrome 3	OMIM:248340	human	9606	COLEC10	10584	
DOID:0080579	3-Methylcrotonyl-CoA carboxylase 1 deficiency	OMIM:210200	human	9606	MCCC1	56922	
DOID:0080580	3-Methylcrotonyl-CoA carboxylase 2 deficiency	OMIM:210210	human	9606	MCCC2	64087	
DOID:0110002	3-methylglutaconic aciduria type 1	OMIM:250950	human	9606	AUH	549	
DOID:0110004	3-methylglutaconic aciduria type 3	OMIM:258501	human	9606	OPA3	80207	
DOID:0110004	3-methylglutaconic aciduria type 3	OMIM:258501	mouse, laboratory	10090	Opa3	403187	MGI:2686271
DOID:0110000	3-methylglutaconic aciduria type 5	OMIM:610198	human	9606	DNAJC19	131118	



Dataset: Example

❖ Developmental and epileptic encephalopathy (OMIM:615338)

- **Disease phenotype:**

Disease	Phenotypes
OMIM:615338	HP:0000572,HP:0000252,HP:0003676,HP:0002059,HP:0002133,HP:0001332,HP:0002376,HP:0006829, HP:0001263,HP:0005484,HP:0012448,HP:0001290,HP:0002071,HP:0000007,HP:0000648,HP:0200134, HP:0001269,HP:0001336

- **Gene phenotypes:**

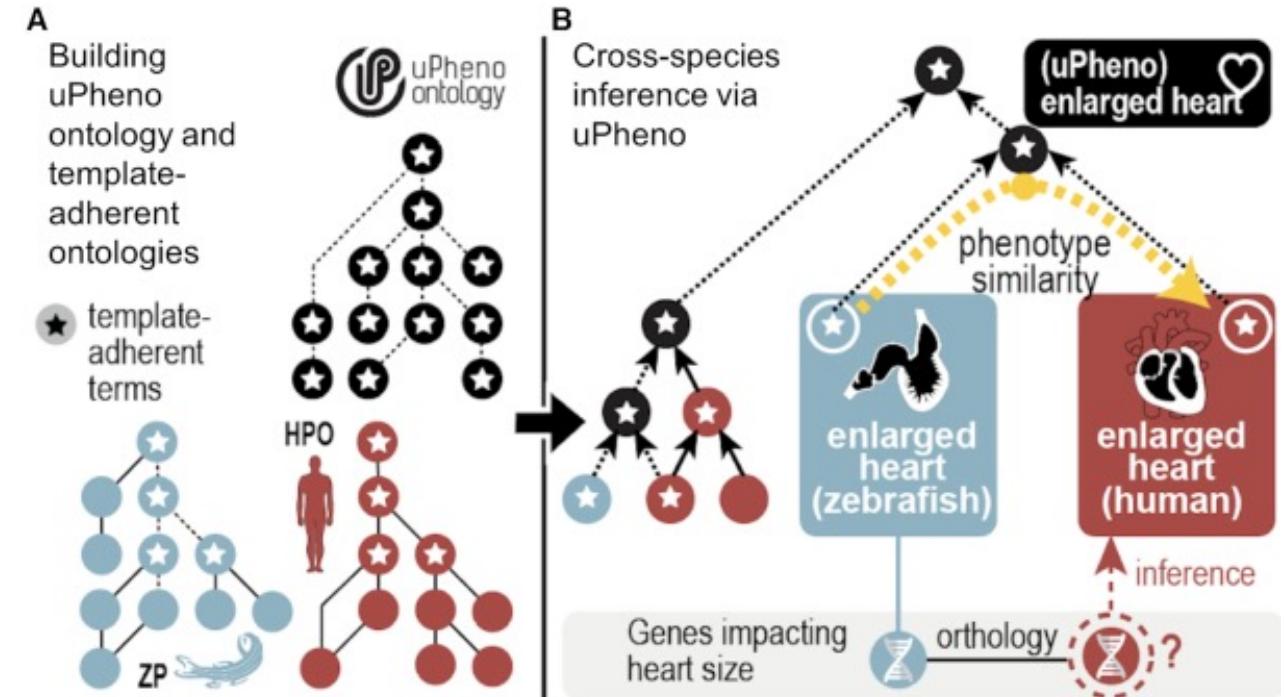
Gene	Phenotypes
Tbc1d24	MP:0003631, MP:0005384, MP:0005386, MP:0010768

- **Gene-Disease association:**

➤ OMIM:615338 linked with Mouse gene [Tbc1d24](#) (EntrezGene ID: 224617).

Dataset: Unified Phenotype Ontology (uPheno)

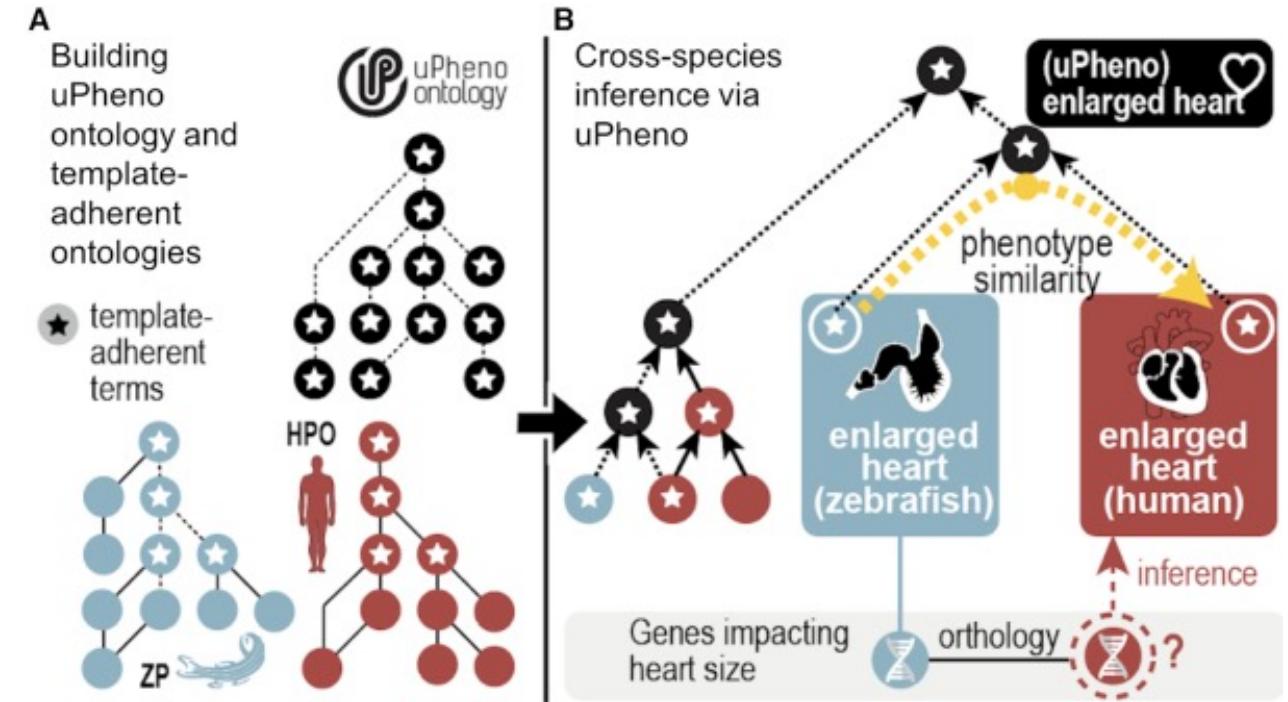
- The phenotypes are described using different organism-specific phenotype ontologies.
- Unified Phenotype Ontology (uPheno):
 - Include human phenotypes from the Human Phenotype Ontology (HPO) → relate mutant model organism phenotypes to human disease-associated phenotypes.



uPheno: Framework for standardized representation of phenotypes across species¹

Dataset: Unified Phenotype Ontology (uPheno)

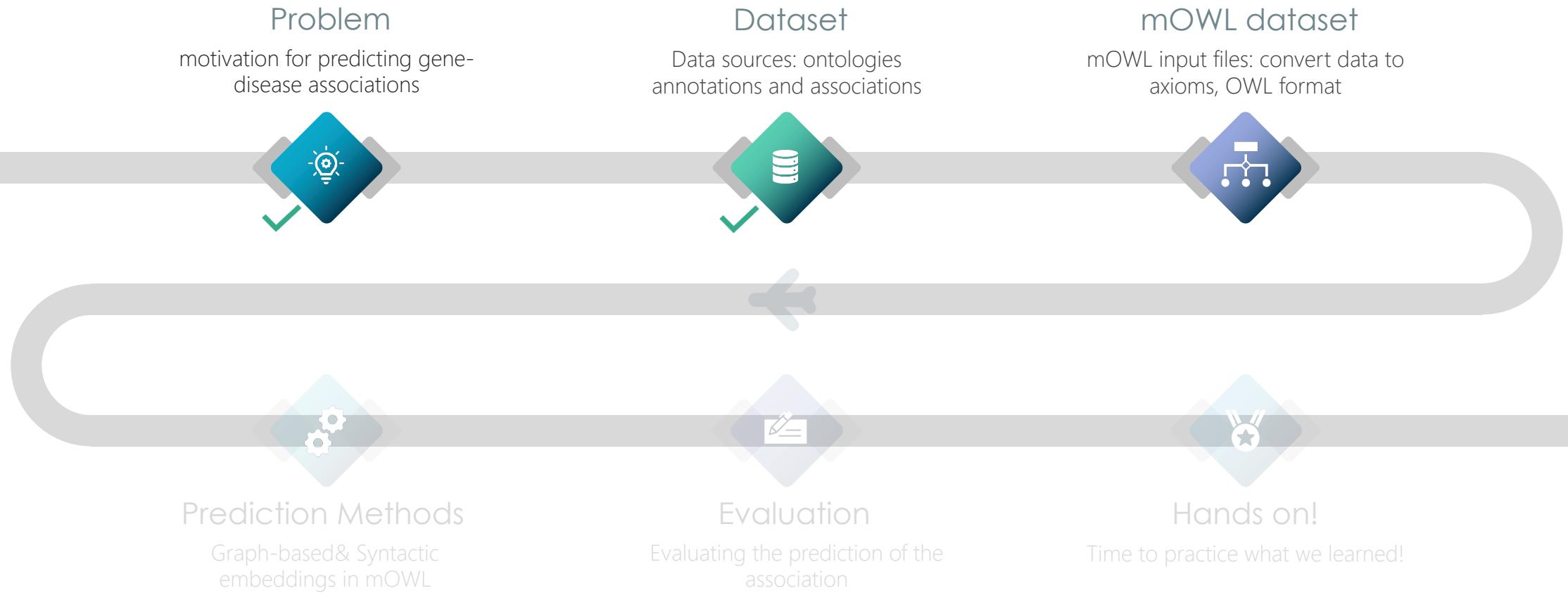
- The phenotypes are described using different organism-specific phenotype ontologies.
 - Unified Phenotype Ontology (uPheno):
 - Include human phenotypes from the Human Phenotype Ontology (HPO) → relate mutant model organism phenotypes to human disease-associated phenotypes.
- ❖ Gene and diseases annotations added to the uPheno ontology to build the training ontology. (Explain next...)



uPheno: Framework for standardized representation of phenotypes across species¹



Roadmap





- ❖ How to add the annotations to mOWL to generate the Gene and disease embedding?

- ❖ mOWL is designed to handle input in OWL format
 - you can input OWL ontologies.
 - A mOWL dataset contains 3 ontologies: training, validation and testing.





❖ Adding annotations to ontologies:

- All the annotations will be inserted into the ontology in the form

$$C \sqsubseteq \exists \mathcal{R}. D$$

where:

- C is the annotating entity (it can be a new ontology class),
- D is the annotated entity (usually is a class already existing in the ontology)
- and \mathcal{R} is the label of the relation, in our case, R relation can be "associated-with" or "has-annotation"



mOWL dataset: Convert annotations to axioms

❖ Adding annotations to ontologies: $C \sqsubseteq \exists \mathcal{R}. D$

Example:

<http://224617> $\sqsubseteq \exists \text{has-annotation. MP_0003631}$
<http://224617> $\sqsubseteq \exists \text{has-annotation. MP_0010771}$
<http://224617> $\sqsubseteq \exists \text{has-annotation. MP_0005386}$

Genes annotations axioms

http://OMIM_615338 $\sqsubseteq \exists \text{has-annotation. HP_0000572}$
http://OMIM_615338 $\sqsubseteq \exists \text{has-annotation. HP_0000252}$
http://OMIM_615338 $\sqsubseteq \exists \text{has-annotation. HP_0002059}$

Diseases annotations axioms



mOWL dataset: Convert annotations to axioms

❖ Adding annotations to ontologies: $C \sqsubseteq \exists \mathcal{R}. D$

Example:

<http://224617> $\sqsubseteq \exists \text{has-annotation. MP_0003631}$
<http://224617> $\sqsubseteq \exists \text{has-annotation. MP_0010771}$
<http://224617> $\sqsubseteq \exists \text{has-annotation. MP_0005386}$

Genes annotations axioms

http://OMIM_615338 $\sqsubseteq \exists \text{has-annotation. HP_0000572}$
http://OMIM_615338 $\sqsubseteq \exists \text{has-annotation. HP_0000252}$
http://OMIM_615338 $\sqsubseteq \exists \text{has-annotation. HP_0002059}$

Diseases annotations axioms

<http://22931> $\sqsubseteq \exists \text{associated-with. http://OMIM_614222}$
<http://9440> $\sqsubseteq \exists \text{associated-with. http://OMIM_613668}$
<http://18023> $\sqsubseteq \exists \text{associated-with. http://OMIM_603694}$

Gene-diseases associations axioms



mOWL dataset: OWL format ontology

```
<?xml version="1.0"?>
<rdf:RDF xmlns="http://ontology.com/someuri.owl#"
  xmlns:base="http://ontology.com/someuri.owl"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:go="http://purl.obolibrary.org/obo/go#"
  xmlns:hp="http://purl.obolibrary.org/obo/hp#"
  xmlns:ns="http://creativecommons.org/ns#"
  xmlns:pr="http://purl.obolibrary.org/obo/pr#"
  xmlns:so="http://purl.obolibrary.org/obo/so#"
  xmlns:apo="http://purl.obolibrary.org/obo/apo#"
  xmlns:obo="http://purl.obolibrary.org/obo/"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:xml="http://www.w3.org/XML/1998/namespace"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:FBcv="http://purl.obolibrary.org/obo/FBcv#"
  xmlns:bspo="http://purl.obolibrary.org/obo/bspo#"
  xmlns:cito="http://purl.org/spar/cito/"
  <!-- http://purl.obolibrary.org/obo/BFO_0000179 -->

  <owl:AnnotationProperty rdf:about="http://purl.obolibrary.org/obo/BFO_0000179">
    <obo:IAO_0000115 xml:lang="en">Relates an entity in the ontology to the
    name of the variable that is used to represent it in the code that generates the
    BFO OWL file from the lisp specification.</obo:IAO_0000115>
    <obo:IAO_0000232 xml:lang="en">Really of interest to developers
    only</obo:IAO_0000232>
    <rdfs:label xml:lang="en">BFO OWL specification label</rdfs:label>
  </owl:AnnotationProperty>

  <!-- http://purl.obolibrary.org/obo/BFO_0000180 -->

  <owl:AnnotationProperty rdf:about="http://purl.obolibrary.org/obo/BFO_0000180">
    <obo:IAO_0000115 xml:lang="en">Relates an entity in the ontology to the
    term that is used to represent it in the CLIF specification of
    BFO</obo:IAO_0000115>
    <obo:IAO_0000119>Person:Alan Ruttenberg</obo:IAO_0000119>
    <obo:IAO_0000232 xml:lang="en">Really of interest to developers
    only</obo:IAO_0000232>
    <rdfs:label xml:lang="en">BFO CLIF specification label</rdfs:label>
  </owl:AnnotationProperty>
  <!-- http://purl.obolibrary.org/obo/HP_0010818 -->
  <owl:Class rdf:about="http://purl.obolibrary.org/obo/HP_0010818">
    <owl:annotatedSource
      rdf:resource="http://purl.obolibrary.org/obo/HP_0010818"/>
    <owl:annotatedSource
      rdf:resource="http://purl.obolibrary.org/obo/HP_0010818"/>
    <owl:annotatedSource
      rdf:resource="http://purl.obolibrary.org/obo/HP_0010818"/>
```

OMIM_615338	http://purl.obolibrary.org/obo/HP_0000572 http://purl.obolibrary.org/obo/HP_0003676 etc...	http://purl.obolibrary.org/obo/HP_0000252 http://purl.obolibrary.org/obo/HP_0002059
OMIM_303110	http://purl.obolibrary.org/obo/HP_0001133 http://purl.obolibrary.org/obo/HP_0000662 http://purl.obolibrary.org/obo/HP_0001256	http://purl.obolibrary.org/obo/HP_0000365 http://purl.obolibrary.org/obo/HP_0001256 http://purl.obolibrary.org/obo/HP_0001263

Diseases annotation in diseases_annotation.tsv file



mOWL dataset: OWL format ontology

```

<?xml version="1.0"?>
<rdf:RDF xmlns="http://ontology.com/someuri.owl#"
  xmlns:base="http://ontology.com/someuri.owl"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:go="http://purl.obolibrary.org/obo/go#"
  xmlns:hp="http://purl.obolibrary.org/obo/hp#"
  xmlns:ns="http://creativecommons.org/ns#"
  xmlns:pr="http://purl.obolibrary.org/obo/pr#"
  xmlns:so="http://purl.obolibrary.org/obo/so#"
  xmlns:apo="http://purl.obolibrary.org/obo/apo#"
  xmlns:obo="http://purl.obolibrary.org/obo/"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:xml="http://www.w3.org/XML/1998/namespace"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:FBcv="http://purl.obolibrary.org/obo/FBcv#"
  xmlns:bspo="http://purl.obolibrary.org/obo/bspo#"
  xmlns:cito="http://purl.org/spar/cito/"
  <!-- http://purl.obolibrary.org/obo/BFO_0000179 -->

  <owl:AnnotationProperty rdf:about="http://purl.obolibrary.org/obo/BFO_0000179">
    <obo:IAO_0000115 xml:lang="en">Relates an entity in the ontology to the
    name of the variable that is used to represent it in the code that generates the
    BFO OWL file from the lispy specification.</obo:IAO_0000115>
    <obo:IAO_0000232 xml:lang="en">Really of interest to developers
    only</obo:IAO_0000232>
    <rdfs:label xml:lang="en">BFO OWL specification label</rdfs:label>
  </owl:AnnotationProperty>
  <!-- http://purl.obolibrary.org/obo/BFO_0000180 -->

  <owl:AnnotationProperty rdf:about="http://purl.obolibrary.org/obo/BFO_0000180">
    <obo:IAO_0000115 xml:lang="en">Relates an entity in the ontology to the
    term that is used to represent it in the CLIF specification of
    BFO</obo:IAO_0000115>
    <obo:IAO_0000119>Person:Alan Ruttenberg</obo:IAO_0000119>
    <obo:IAO_0000232 xml:lang="en">Really of interest to developers
    only</obo:IAO_0000232>
    <rdfs:label xml:lang="en">BFO CLIF specification label</rdfs:label>
  </owl:AnnotationProperty>
  <!-- http://purl.obolibrary.org/obo/HP_0010818 -->
  <owl:Class rdf:about="http://purl.obolibrary.org/obo/HP_0010818">
    <owl:annotatedSource
      rdf:resource="http://purl.obolibrary.org/obo/HP_0010818"/>
    <owl:annotatedSource
      rdf:resource="http://purl.obolibrary.org/obo/HP_0010818"/>
    <owl:annotatedSource
      rdf:resource="http://purl.obolibrary.org/obo/HP_0010818"/>

```

OMIM_615338	http://purl.obolibrary.org/obo/HP_0000572 http://purl.obolibrary.org/obo/HP_0003676 etc...	http://purl.obolibrary.org/obo/HP_0000252 http://purl.obolibrary.org/obo/HP_0002059
OMIM_303110	http://purl.obolibrary.org/obo/HP_0001133 http://purl.obolibrary.org/obo/HP_0000662 http://purl.obolibrary.org/obo/HP_0001256	http://purl.obolibrary.org/obo/HP_0000365 http://purl.obolibrary.org/obo/HP_0001256 http://purl.obolibrary.org/obo/HP_0001263
Diseases annotation in diseases_annot.tsv file		
http://224617	http://purl.obolibrary.org/obo/MP_0003631 http://purl.obolibrary.org/obo/MP_0010771	http://purl.obolibrary.org/obo/MP_0005384 http://purl.obolibrary.org/obo/MP_0005386
http://60496	http://purl.obolibrary.org/obo/MP_0001186 http://purl.obolibrary.org/obo/MP_0005386 http://purl.obolibrary.org/obo/MP_0005391 http://purl.obolibrary.org/obo/MP_0010768	http://purl.obolibrary.org/obo/MP_0005376

Genes annotation in genes_annot.tsv file



mOWL dataset: OWL format ontology

```

<?xml version="1.0"?>
<rdf:RDF xmlns="http://ontology.com/someuri.owl#"
  xmlns:base="http://ontology.com/someuri.owl"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:go="http://purl.obolibrary.org/obo/go#"
  xmlns:hp="http://purl.obolibrary.org/obo/hp#"
  xmlns:ns="http://creativecommons.org/ns#"
  xmlns:pr="http://purl.obolibrary.org/obo/pr#"
  xmlns:so="http://purl.obolibrary.org/obo/so#"
  xmlns:apo="http://purl.obolibrary.org/obo/apo#"
  xmlns:obo="http://purl.obolibrary.org/obo/"
  xmlns:owl="http://www.w3.org/2002/07/owl#"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:xml="http://www.w3.org/XML/1998/namespace"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:FBcv="http://purl.obolibrary.org/obo/FBcv#"
  xmlns:bspo="http://purl.obolibrary.org/obo/bspo#"
  xmlns:cito="http://purl.org/spar/cito/"
  <!-- http://purl.obolibrary.org/obo/BFO_0000179 -->

  <owl:AnnotationProperty rdf:about="http://purl.obolibrary.org/obo/BFO_0000179">
    <obo:IAO_0000115 xml:lang="en">Relates an entity in the ontology to the
    name of the variable that is used to represent it in the code that generates the
    BFO OWL file from the lispy specification.</obo:IAO_0000115>
    <obo:IAO_0000232 xml:lang="en">Really of interest to developers
    only</obo:IAO_0000232>
    <rdfs:label xml:lang="en">BFO OWL specification label</rdfs:label>
  </owl:AnnotationProperty>
  <!-- http://purl.obolibrary.org/obo/BFO_0000180 -->

  <owl:AnnotationProperty rdf:about="http://purl.obolibrary.org/obo/BFO_0000180">
    <obo:IAO_0000115 xml:lang="en">Relates an entity in the ontology to the
    term that is used to represent it in the CLIF specification of
    BFO</obo:IAO_0000115>
    <obo:IAO_0000119>Person:Alan Ruttenberg</obo:IAO_0000119>
    <obo:IAO_0000232 xml:lang="en">Really of interest to developers
    only</obo:IAO_0000232>
    <rdfs:label xml:lang="en">BFO CLIF specification label</rdfs:label>
  </owl:AnnotationProperty>
  <!-- http://purl.obolibrary.org/obo/HP_0010818 -->
  <owl:Class rdf:about="http://purl.obolibrary.org/obo/HP_0010818">
    <owl:annotatedSource
      rdf:resource="http://purl.obolibrary.org/obo/HP_0010818"/>
    <owl:annotatedSource
      rdf:resource="http://purl.obolibrary.org/obo/HP_0010818"/>
    <owl:annotatedSource
      rdf:resource="http://purl.obolibrary.org/obo/HP_0010818"/>

```

OMIM_615338	http://purl.obolibrary.org/obo/HP_0000572 http://purl.obolibrary.org/obo/HP_0003676 etc...	http://purl.obolibrary.org/obo/HP_0000252 http://purl.obolibrary.org/obo/HP_0002059
OMIM_303110	http://purl.obolibrary.org/obo/HP_0001133 http://purl.obolibrary.org/obo/HP_0000662 http://purl.obolibrary.org/obo/HP_0001256	http://purl.obolibrary.org/obo/HP_0000365 http://purl.obolibrary.org/obo/HP_0001256 http://purl.obolibrary.org/obo/HP_0001263
Diseases annotation in diseases_annot.tsv file		
http://224617	http://purl.obolibrary.org/obo/MP_0003631 http://purl.obolibrary.org/obo/MP_0010771	http://purl.obolibrary.org/obo/MP_0005384 http://purl.obolibrary.org/obo/MP_0005386
http://60496	http://purl.obolibrary.org/obo/MP_0001186 http://purl.obolibrary.org/obo/MP_0005386 http://purl.obolibrary.org/obo/MP_0005391 http://purl.obolibrary.org/obo/MP_0010768	http://purl.obolibrary.org/obo/MP_0005376
Genes annotation in genes_annot.tsv file		
http://22931	OMIM_614222	
http://9440	OMIM_613668	

Associations in gene_disease_associations.tsv file



Our annotations classes (e.g. http://purl.obolibrary.org/obo/HP_0010818) that are in relation with the already classes in the **uPheno** ontology.
Add the classes with relation, which must be a proper URI (we will use http://has_annotation) and for the gene-diseases associations (we will use http://is_associated_with).



mOWL dataset: OWL format ontology

- How to add the annotations using mOWL?

- To add that information to the ontology use the following (`insert_annotations`) function:

```
from mowl.ontology.extend import insert_annotations

diseases_annotations = ("diseases_annot.tsv", "http://has_annotation", True)
genes_annotations = ("genes_annot.tsv", "http://has_annotation", True)
gene_disease_associations = ("gene_disease_associations.tsv", "http://is_associated_with/", True)

annotations = [diseases_annotations, genes_annotations, gene_disease_associations] # There
could be more than 1 annotations file.

insert_annotations("upheno.owl", annotations, out_file = "upheno_with_annotations.owl", )
```



- Gene and diseases annotations added to uPheno.
- Associations for human and mouse extracted and randomly split 80:10:10, added to the **training** ontology and created the **validation** and **testing** ontologies



- Gene and diseases annotations added to uPheno.
- Associations for human and mouse extracted and randomly split 80:10:10, added to the **training** ontology and created the **validation** and **testing** ontologies
- **GDAHumanDataset** and **GDAMouseDataset** are available as a built-in datasets in mOWL.

`class mowl.datasets.builtin.GDAHumanDataset` [\[source\]](#)

Bases: `GDADataset`

`class mowl.datasets.builtin.GDAMouseDataset` [\[source\]](#)

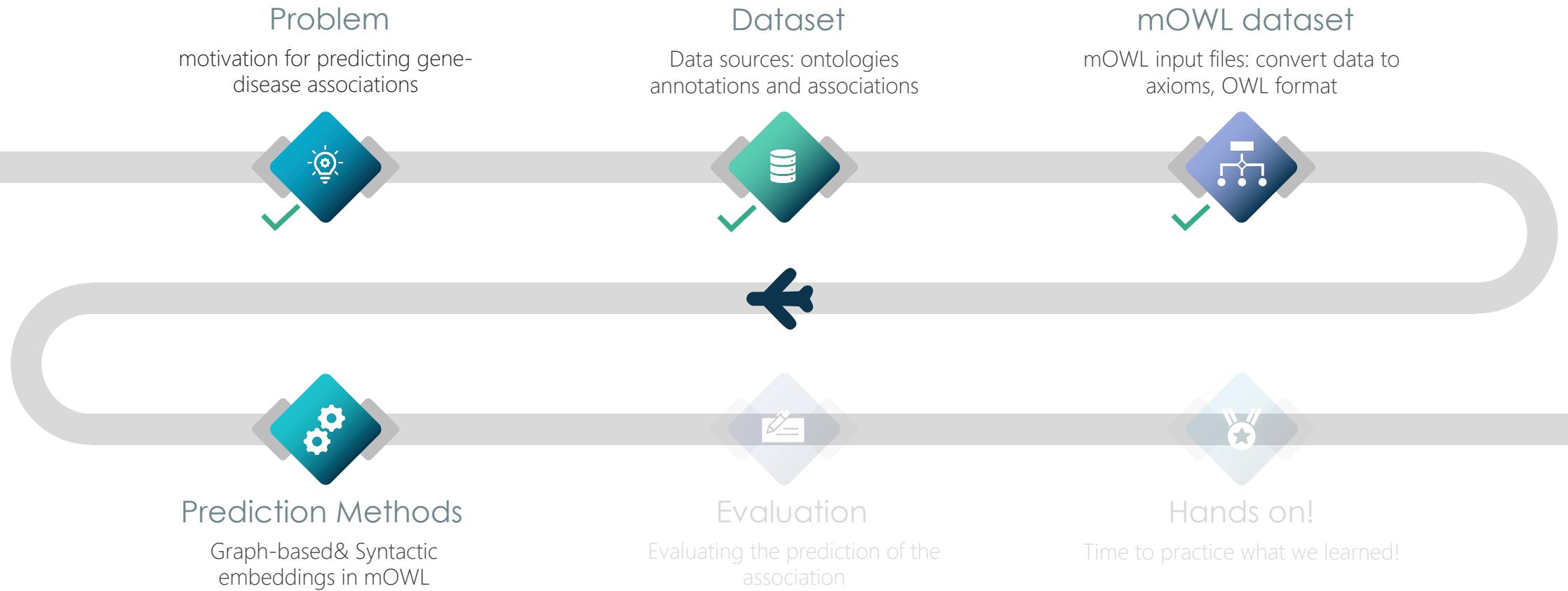
Bases: `GDADataset`



- Gene and diseases annotations added to uPheno.
 - Associations for human and mouse extracted and randomly split 80:10:10, added to the training ontology and created the validation and testing ontologies
 - **GDAHumanDataset** and **GDAMouseDataset** are available as a built-in datasets in mOWL.
- Now we can use the dataset as input to the prediction methods...

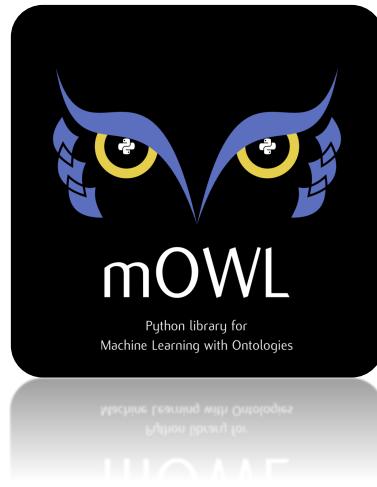


Roadmap





mOWL Prediction Methods



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

Syntactic embeddings

Onto2vec

OPA2Vec

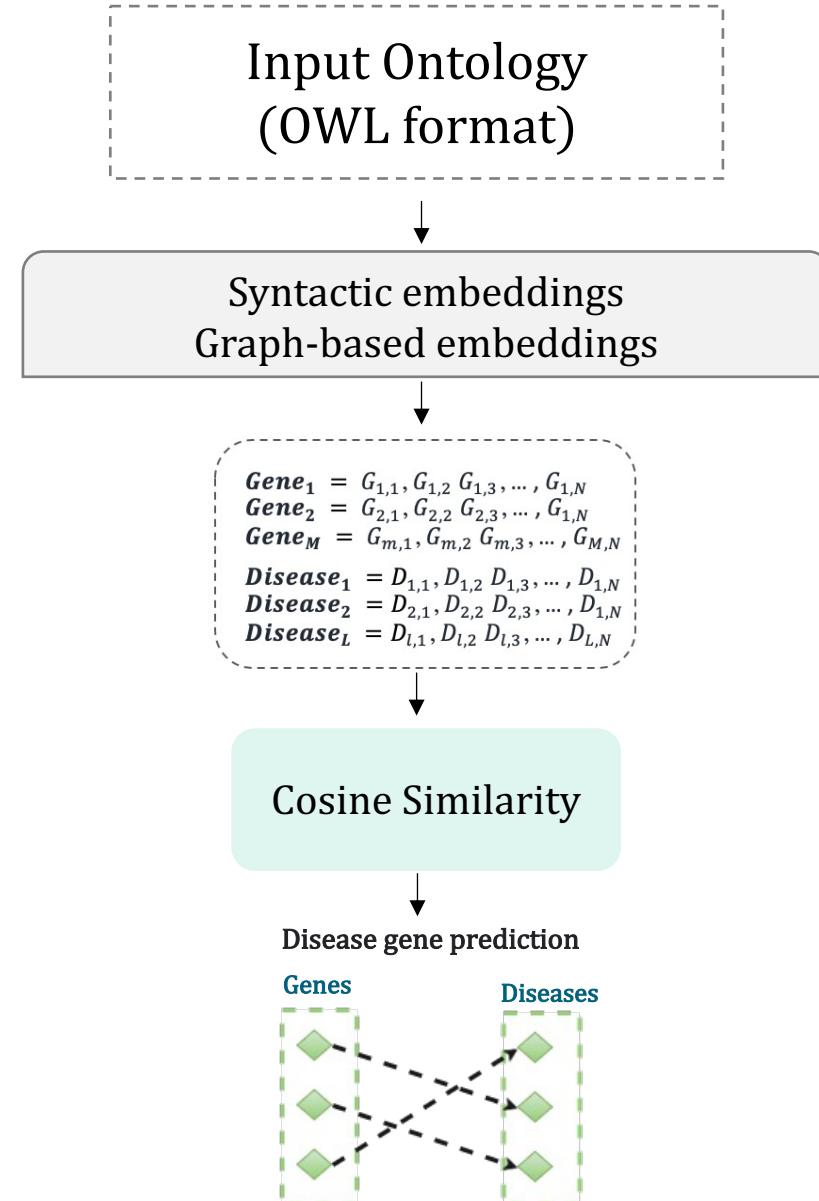
Graph-based embeddings

DL2vec

OWL2Vec*



mOWL Prediction Methods: workflow



- Use different methods to generate the representation given the annotated ontology file.

1) Projecting the ontology

- Project the ontology using the DL2Vec/OWL2vec Projector class, with the specific rules used to project the ontology.
- The outcome of the projection algorithm is an edgelist.

```
http://OMIM_615338, http://purl.obolibrary.org/obo/HP_0000007'  
http://224617, http://purl.obolibrary.org/obo/MP_0003631
```

...

...

- Use different methods to generate the representation given the annotated ontology file.

1) Projecting the ontology

- Project the ontology using the DL2Vec/OWL2vec Projector class, with the specific rules used to project the ontology.
- The outcome of the projection algorithm is an edgelist.

2) Generating random walks

- The random walks are generated using the DeepWalk.

```
http://OMIM_615338 http://has_annotation http://purl.obolibrary.org/obo/HP_0006829 http://subclassof  
http://purl.obolibrary.org/obo/HP_0001252 http://superclassof http://purl.obolibrary.org/obo/HP_0008947  
http://subclassof http://purl.obolibrary.org/obo/HP_0001252  
http://20682 http://is_associated_with OMIM:114290 http://has_annotation  
http://purl.obolibrary.org/obo/HP_0001263 http://superclassof http://purl.obolibrary.org/obo/HP_0011344  
http://subclassof http://purl.obolibrary.org/obo/HP_0001263  
...
```

- Use different methods to generate the representation given the annotated ontology file.

1) Projecting the ontology to generate the graph

- Project the ontology using the DL2Vec/OWL2vec Projector class, with the specific rules used to project the ontology.
- The outcome of the projection algorithm is an edgelist.

2) Generating random walks

- The random walks are generated using the DeepWalk.

3) Training the Word2Vec model

- To train the Word2Vec model, we rely on the Gensim library

```
http://OMIM_615338 array([-0.69632155, -0.05665049,  0.32989377, -0.05275345,  0.01977283, 0.08355428, -0.30948249,  0.33214468,  0.2840028 ,  0.15992194,  
http://20682 array([-0.3579754 ,  0.13698728, -0.0022221 ,  0.5450176 , -0.42158118,  
0.20895344,  1.2265201 ,  0.8927095 , -0.6502107 , -0.6157604 ,...]
```



- **We need three components:**

1. The reasoner
2. The corpus generator
3. The Word2Vec model.

1) Inferring new axioms

- Onto2Vec/OPA2vec uses an ontology reasoner to infer new axioms as a preprocessing step (e.g. ELK reasoner).

```
http://OMIM_615338 SubClassOf http://has_annotation some http://purl.obolibrary.org/obo/HP_0001336
http://OMIM_615338 SubClassOf http://has_annotation some http://purl.obolibrary.org/obo/HP_0002059
http://OMIM_615338 SubClassOf http://has_annotation some http://purl.obolibrary.org/obo/HP_0003676
http://OMIM_615338 SubClassOf http://has_annotation some http://purl.obolibrary.org/obo/HP_0000252
http://OMIM_615338 SubClassOf http://is_associated_with some http://224616
http://OMIM_114290 SubClassOf http://is_associated_with some http://20682
```

...



1) Inferring new axioms

- Onto2Vec/OPA2vec uses an ontology reasoner to infer new axioms as a preprocessing step (e.g. ELK reasoner).

2) Generating the corpus and training the model

- To train the Word2Vec model on the generated corpus.

```
http://OMIM_615338 http://has_annotation http://purl.obolibrary.org/obo/HP_0006829 http://subclassof  
http://purl.obolibrary.org/obo/HP_0001252 http://superclassof http://purl.obolibrary.org/obo/HP_0008947  
http://subclassof http://purl.obolibrary.org/obo/HP_0001252 http://20682 http://is_associated_with OMIM:114290  
http://has_annotation http://purl.obolibrary.org/obo/HP_0001263 http://superclassof  
http://purl.obolibrary.org/obo/HP_0011344 http://subclassof http://purl.obolibrary.org/obo/HP_0001263  
...
```



mOWL Syntactic models

1) Inferring new axioms

- Onto2Vec/OPA2vec uses an ontology reasoner to infer new axioms as a preprocessing step (e.g. ELK reasoner).

2) Generating the corpus and training the model

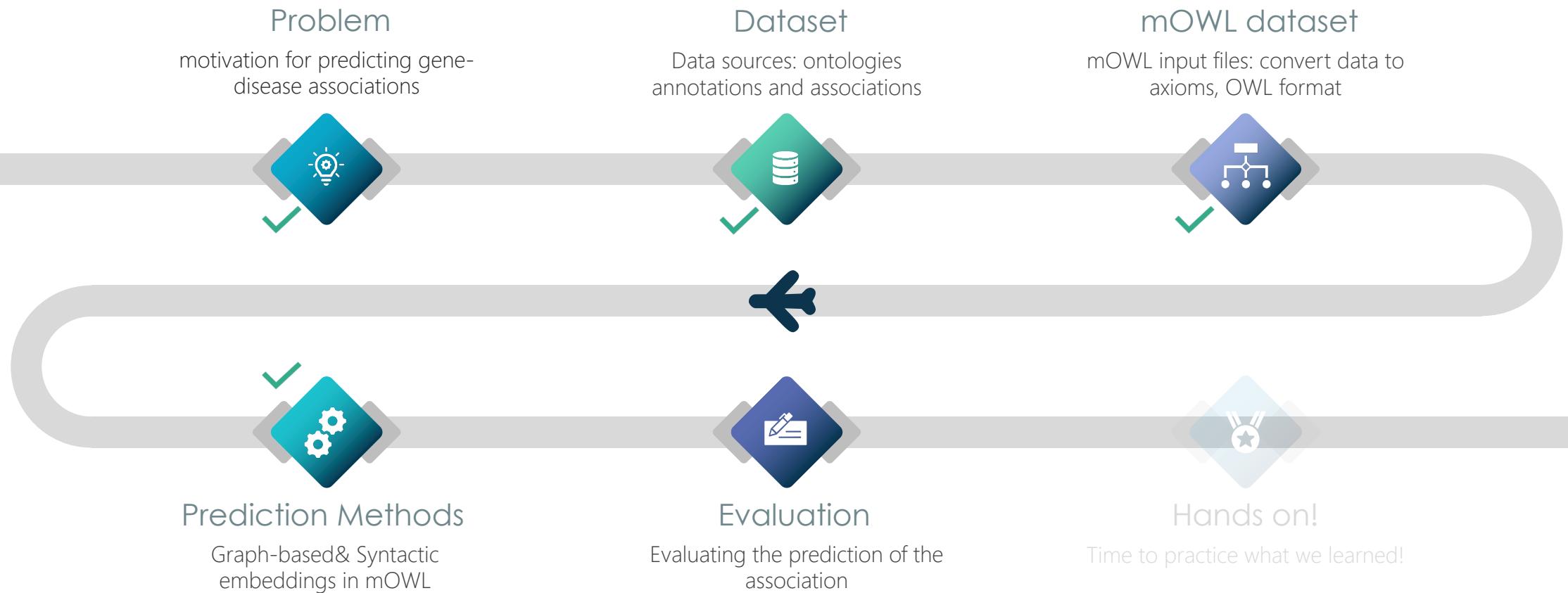
- To train the Word2Vec model on the generated corpus.

```
http://OMIM_615338 http://has_annotation http://purl.obolibrary.org/obo/HP_0006829 http://subclassof  
http://purl.obolibrary.org/obo/HP_0001252 http://superclassof http://purl.obolibrary.org/obo/HP_0008947  
http://subclassof http://purl.obolibrary.org/obo/HP_0001252 http://20682 http://is_associated_with OMIM:114290  
http://has_annotation http://purl.obolibrary.org/obo/HP_0001263 http://superclassof  
http://purl.obolibrary.org/obo/HP_0011344 http://subclassof http://purl.obolibrary.org/obo/HP_0001263  
...
```

```
http://OMIM_615338 array([-0.69632155, -0.05665049,  0.32989377, -0.05275345,  0.01977283, 0.08355428, -  
0.30948249,  0.33214468,  0.2840028 ,  0.15992194,  
http://20682 array([-0.3579754 ,  0.13698728, -0.0022221 ,  0.5450176 , -0.42158118,  
0.20895344,  1.2265201 ,  0.8927095 , -0.6502107 , -0.6157604 ,...]
```



Roadmap





Evaluation

- We are going to evaluate the plausibility of an association gene-disease with a gene against all possible diseases and check the rank of the true disease association.

```
genes, diseases = dataset.evaluation_classes
projector = TaxonomyWithRelationsProjector(taxonomy=False,
                                             relations=[http://is\_associated\_with])

evaluation_edges= projector.project(dataset.testing)
filterig_edges= projector.project(dataset.ontology)

assert len(evaluation_edges) > 0
```



Evaluation: Calculating Phenotypic Similarity Approaches

- The gene-disease associations will be scored using cosine similarity
- For that reason we use the **CosineSimilarity** class.

```
vectors = model.wv
evaluator = EmbeddingsRankBasedEvaluator(
    vectors,
    evaluation_edges,
    CosineSimilarity,
    training_set=filtering_edges,
    head_entities = genes.as_str,
    tail_entities = diseases.as_str,
    device = 'cpu'
)

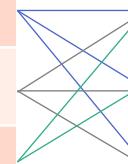
evaluator.evaluate(show=True)
```



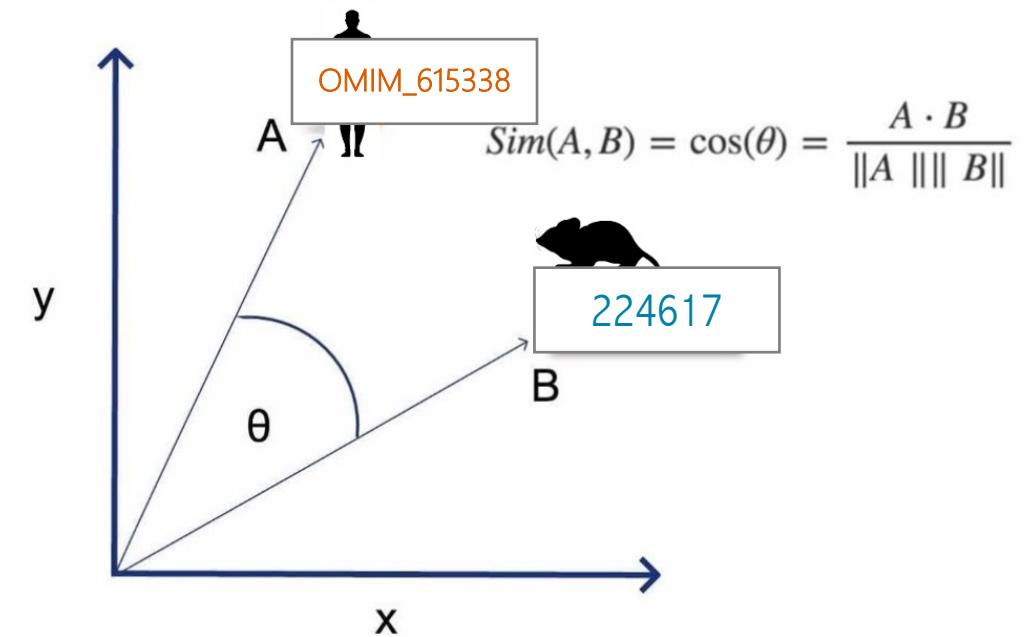
Evaluation: Calculating Phenotypic Similarity Approaches

- The gene-disease associations will be scored using cosine similarity

Disease	Vectors	Genes	Vectors
<i>Disease</i> ₁	$D_{1,1}, D_{1,2}, D_{1,3}, \dots, D_{1,N}$	<i>Gene</i> ₁	$G_{1,1}, G_{1,2}, G_{1,3}, \dots, G_{1,N}$
<i>Disease</i> ₂	$D_{2,1}, D_{2,2}, D_{2,3}, \dots, D_{2,N}$	<i>Gene</i> ₂	$G_{2,1}, G_{2,2}, G_{2,3}, \dots, G_{2,N}$
<i>Disease</i> _L	$D_{L,1}, D_{L,2}, D_{L,3}, \dots, D_{L,N}$	<i>Gene</i> _M	$G_{M,1}, G_{M,2}, G_{M,3}, \dots, G_{M,N}$

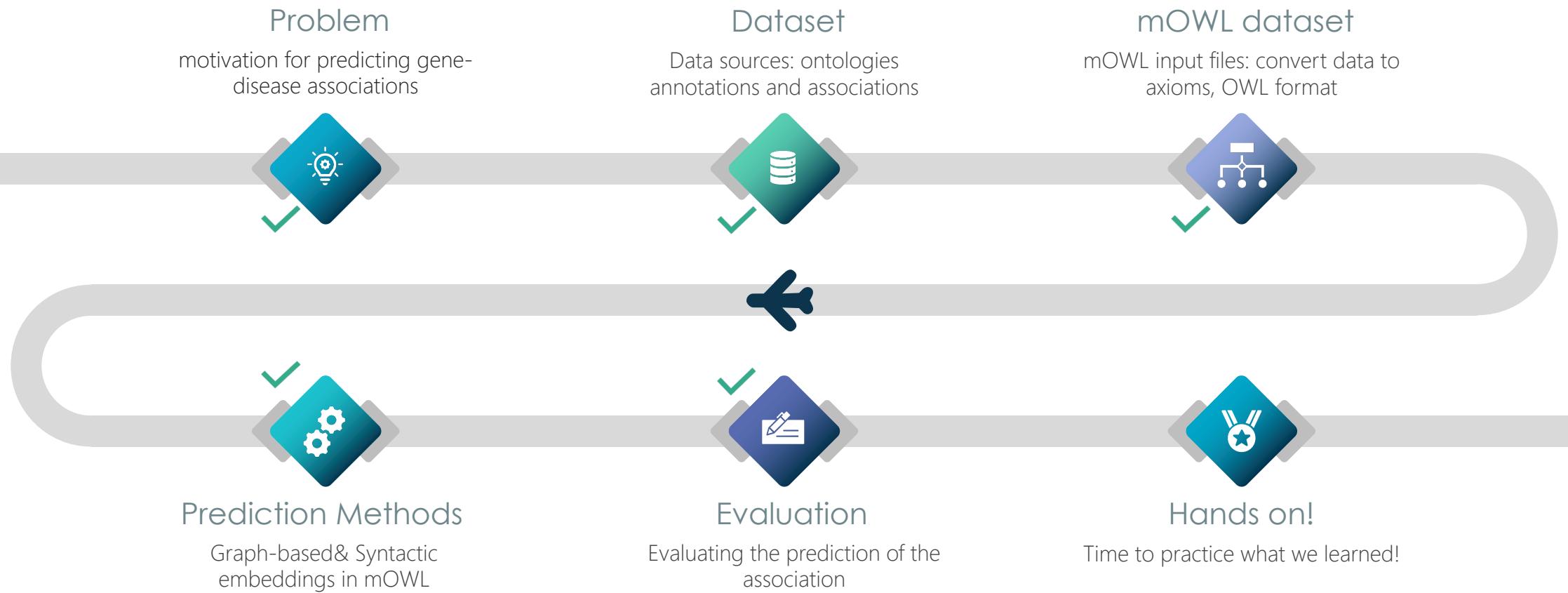


Predict associated genes to disease *Disease*_l :
 $\text{MAX(similarity}(\textit{Disease}_l, \textit{Gene}_M)\text{)}$





Roadmap





Hands On Tutorial

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

Another Application: Prediction the causative variants



**Patients Clinical Phenotypes &
Genomics Data**

**Can we find the causative variants
associated with the phenotypes?**

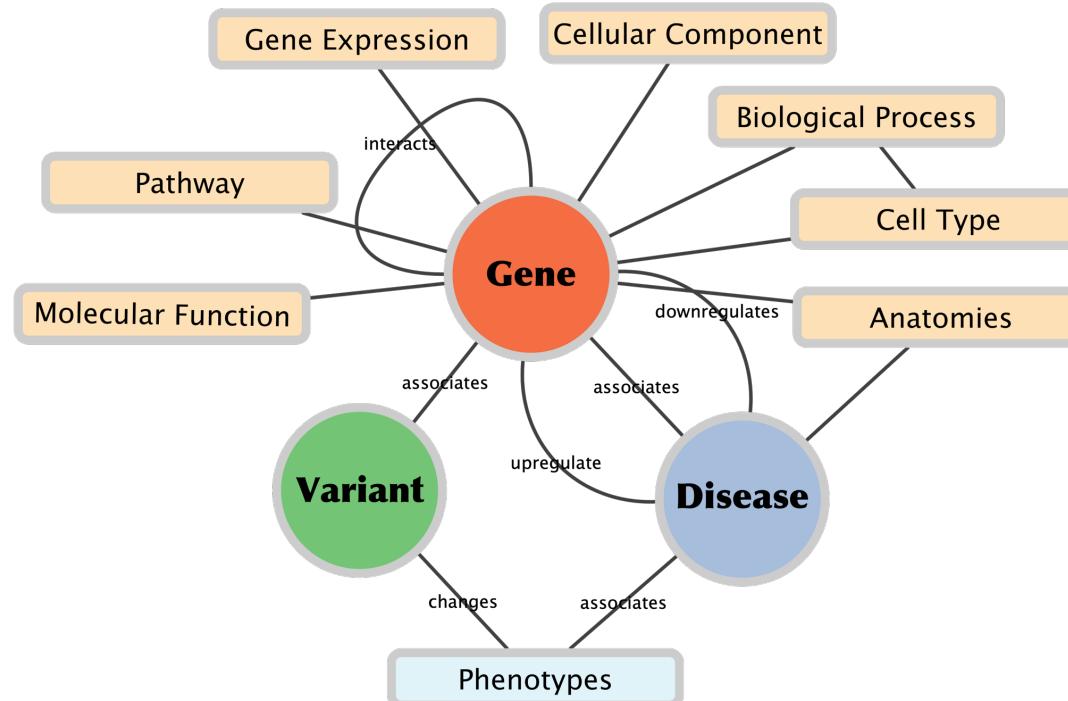


Other application: Prediction the causative variants

- 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



- Gene-Phenotype annotations
 - Gene-Function annotations
 - Gene-Celltype annotations
 - Gene-Anatomical site of gene expression annotations
 - Disease-Phenotypes annotations
- **Prioritize the candidate genes/variants**



Lessons to be learned

- mOWL is a Python library for Machine Learning with Ontologies.
 - Provide functionalities to manipulate ontologies
 - Use them as data for several methods that generate embeddings of ontology entities
- Predicting **gene-disease associations** is one practical examples using the library.
 - Use the available data sources to learn and infer associations
- Other applications:
 - predict protein-protein interactions based on their functional similarity, predict the causative variants, predict the drug-targets...etc.



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!

