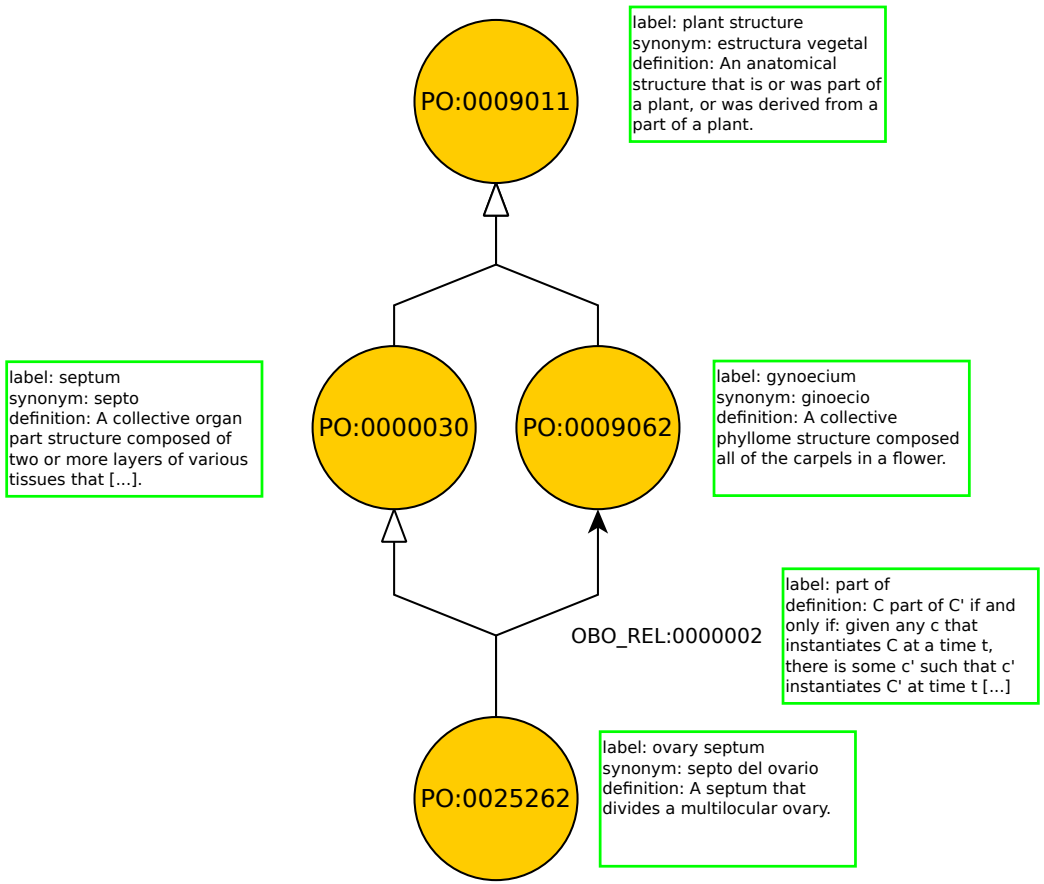
Machine learning with biomedical ontologies: applications in precision health

Robert Hoehndorf

Computational Bioscience Research Center
King Abdullah University of Science and Technology

Making sense of data... with ontologies

- ontology (philosophy) studies the nature of existence and categories of being
- an ontology (computer science) is the "explicit specification of a conceptualization of a domain" [Gruber, 1993]



Making sense of data... with ontologies

Ontologies provide

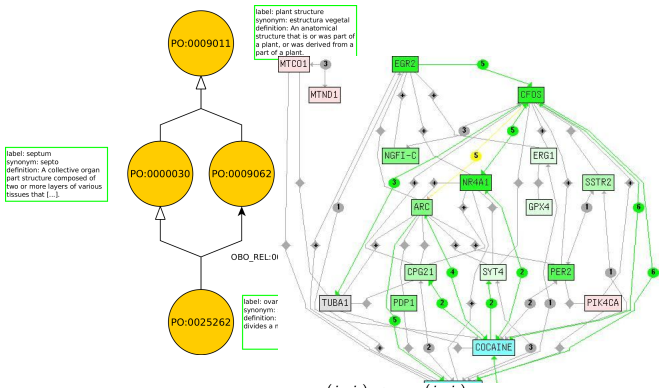
- standard identifiers — for integration of data
- terms — domain vocabulary
- definitions — for human understanding
- axioms — for machine understanding

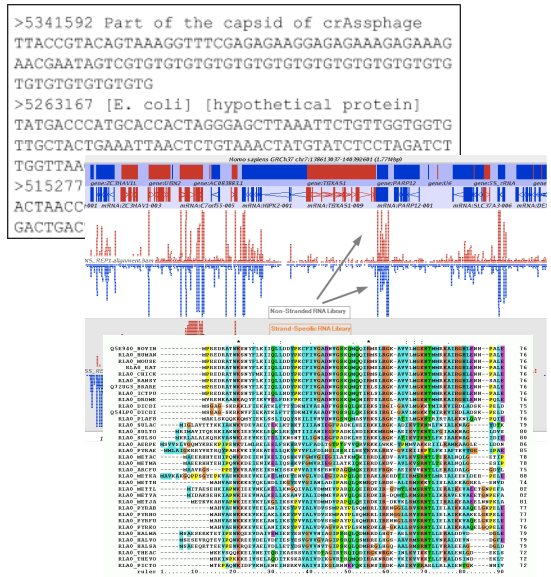
Making sense of data... with ontologies

Overview | Browse | DLQuery | Download

Annotation	Value
label	B cell apoptotic process
definition	Any apoptotic process in a B cell, a lymphocyte of B lineage with the phenotype CD19-positive and capable of B cell mediated immunity.
class	http://purl.obolibrary.org/obo/GO_0001783
ontology	GO-PLUS
Equivalent	apoptotic process and (occurs in some B cell)
SubClassOf	occurs in some B cell, lymphocyte apoptotic process
id	GO:0001783
has_obo_namespace	biological_process

Big Data + Big Knowledge





$conn(j_1, j_2) \rightarrow conn(j_2, j_1)$

$conn(j_1, j_2) \rightarrow j_1 \neq j_2$

$in(j_1, s_1) \wedge in(j_2, s_2) \wedge \neg soverlap(s_1, s_2) \rightarrow \neg conn(j_1, j_2)$

$conn(j_1, j_2) \wedge in(j_1, s) \rightarrow in(j_2, s)$

$\forall s \forall P(\forall x(P(x) \leftrightarrow in(x, s)) \wedge \forall Q(\exists a Q(a) \wedge \forall x(Q(x) \rightarrow P(x)) \wedge \forall u, v(Q(u) \wedge conn(u, v) \rightarrow Q(v)) \rightarrow \forall x(P(x) \rightarrow Q(x))))$

write

reasoning

read

Symbol system

embedding

extraction

Data

learning

- phenotype
- function
- disease

- genotype
- protein sequence
- expression

Embedding formal knowledge

Embedding

An embedding is a map (morphism) from one mathematical structure X into another structure Y:
$f : X \hookrightarrow Y$
such that X is preserved in Y.

- Y may be more suitable than X for some operations/algorithms.
  - similarity
  - gradients, optimization

Embedding formal knowledge

Embedding

An embedding is a map (morphism) from one mathematical structure X into another structure Y:
$f : X \hookrightarrow Y$
such that X is preserved in Y.

- Y may be more suitable than X for some operations/algorithms.
  - similarity
  - gradients, optimization

We want to embed formalized *knowledge bases* in $\mathbb{R}^n$. Approaches:

- graph-based
- syntactic
- model-theoretic

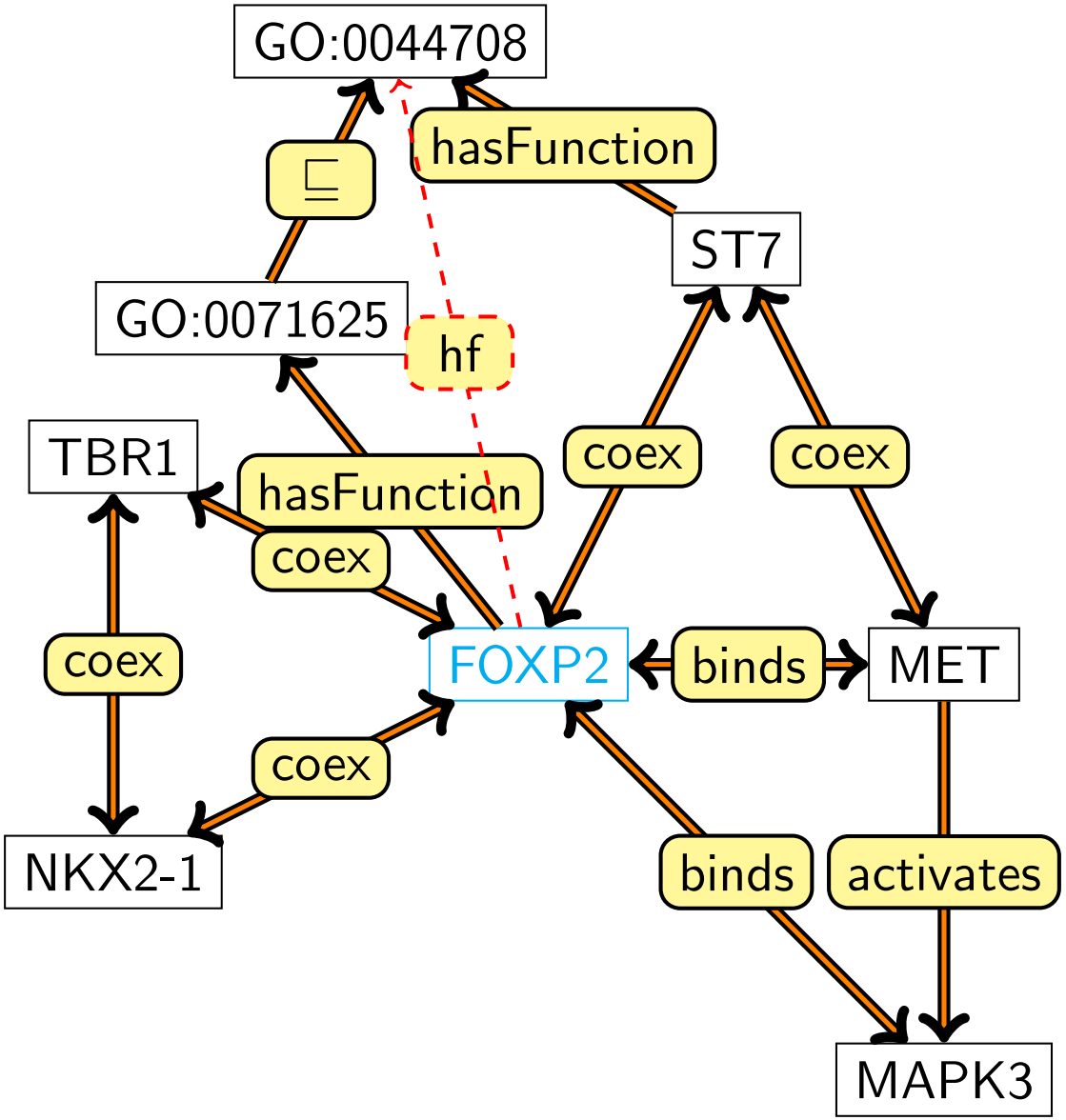
Knowledge to (knowledge) graphs

- $X \sqsubseteq Y$: $X \xrightarrow{\text{is-a}} Y$
- $X \sqsubseteq \exists \text{part-of}.Y$: $X \xrightarrow{\text{part-of}} Y$
- $X \sqsubseteq \exists \text{regulates}.Y$: $X \xrightarrow{\text{regulates}} Y$
- $X \sqcap Y \sqsubseteq \bot$: $X \xleftrightarrow{\text{disjoint}} Y$
- $X \equiv Y$: $X \overset{\equiv}{\leftrightarrow} Y$, $\{X, Y\}$

Asserted and inferred:

- $X \sqsubseteq \exists \text{part-of}.Y$
- $Y \sqsubseteq \exists \text{part-of}.Z$
- $\text{part-of} \circ \text{part-of} \subseteq \text{part-of}$
- entails: $X \sqsubseteq \exists \text{part-of}.Z$

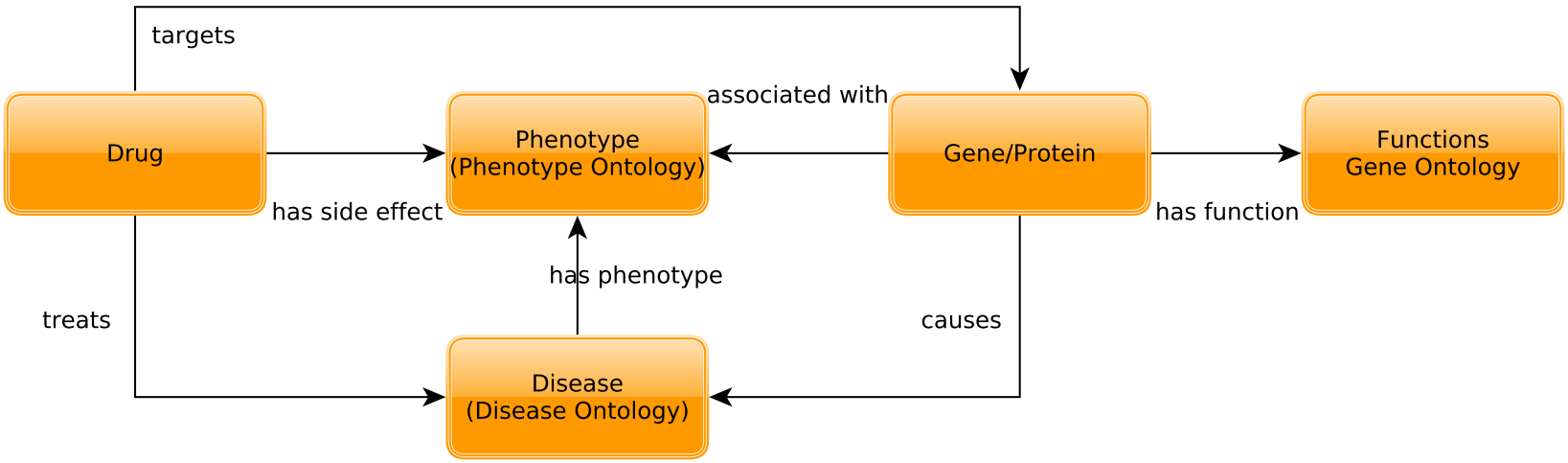
Knowledge graph embeddings



graph embedding, e.g.:

- iterated random walks generate "sentences"
- embedding through language model
- use embedding for clustering, edge prediction, etc.

Relation prediction



Object property	Source type	Target type	Without reasoning		With reasoning	
			F-measure	AUC	F-measure	AUC
has target	Drug	Gene/Protein	0.94	0.97	0.94	0.98
has disease annotation	Gene/Protein	Disease	0.89	0.95	0.89	0.95
has side-effect*	Drug	Phenotype	0.86	0.93	0.87	0.94
has interaction	Gene/Protein	Gene/Protein	0.82	0.88	0.82	0.88
has function*	Gene/Protein	Function	0.85	0.95	0.83	0.91
has gene phenotype*	Gene/Protein	Phenotype	0.84	0.91	0.82	0.90
has indication	Drug	Disease	0.72	0.79	0.76	0.83
has disease phenotype*	Disease	Phenotype	0.72	0.78	0.70	0.77

More semantics

- walk- (and word-)based methods are not truly "semantic"
- semantics relies on interpretation (Σ-algebras)
- an interpretation is a *model* of T if all $C \sqsubseteq D$ are satisfied

Name	Syntax	Semantics
top	$\top$	$\Delta^{\mathcal{I}}$
bottom	$\bot$	$\emptyset$
nominal	$\{a\}$	$\{a^{\mathcal{I}}\}$
conjunction	$C \sqcap D$	$C^{\mathcal{I}} \cap D^{\mathcal{I}}$
existential restriction	$\exists r.C$	$\{x \in \Delta^{\mathcal{I}} \mid \exists y \in \Delta^{\mathcal{I}} : (x,y) \in r^{\mathcal{I}} \wedge y \in C^{\mathcal{I}}\}$
generalized concept inclusion	$C \sqsubseteq D$	$C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$
role inclusion	$r_1 \circ \ldots \circ r_n \sqsubseteq r$	$r_1^{\mathcal{I}} \circ \ldots \circ r_n^{\mathcal{I}} \subseteq r^{\mathcal{I}}$

EL Embeddings

- given a knowledge base T with signature $\Sigma(T)$

EL Embeddings

- given a knowledge base T with signature $\Sigma(T)$
- aim: find $f_e : \Sigma(T) \mapsto \mathbb{R}^n$ s.t. $f_e(\Sigma(T))$ is a model of T ($f_e(\Sigma(T)) \models T$)
  - maps symbols into $\mathbb{R}^n$ while preserving their model-theoretic semantics

EL Embeddings

- given a knowledge base T with signature $\Sigma(T)$
- aim: find $f_e : \Sigma(T) \mapsto \mathbb{R}^n$ s.t. $f_e(\Sigma(T))$ is a model of T ($f_e(\Sigma(T)) \models T$)
  - maps symbols into $\mathbb{R}^n$ while preserving their model-theoretic semantics
- preliminaries:
  - any consistent $\mathcal{EL}^{++}$theory has infinite models

EL Embeddings

- given a knowledge base T with signature $\Sigma(T)$
- aim: find $f_e : \Sigma(T) \mapsto \mathbb{R}^n$ s.t. $f_e(\Sigma(T))$ is a model of T ($f_e(\Sigma(T)) \models T$)
  - maps symbols into $\mathbb{R}^n$ while preserving their model-theoretic semantics
- preliminaries:
  - any consistent $\mathcal{EL}^{++}$theory has infinite models
  - any consistent $\mathcal{EL}^{++}$theory has models in $\mathbb{R}^n$ (Löwenheim-Skolem, upwards)

EL Embeddings

- for all $r \in \Sigma(T)$ and $C \in \Sigma(T)$, define $f_e(r)$ and $f_e(C)$

EL Embeddings

- for all $r \in \Sigma(T)$ and $C \in \Sigma(T)$, define $f_e(r)$ and $f_e(C)$
- $f_e(C)$ maps to points in an open n-ball such that $f_e(C) = C^{\mathcal{I}}$: $C^{\mathcal{I}} = \{x \in \mathbb{R}^n | \, \|f_e(C) - x\| < r_e(C)\}$
  - these are the *extension* of a unary predicate in $\mathbb{R}^n$

EL Embeddings

- for all $r \in \Sigma(T)$ and $C \in \Sigma(T)$, define $f_e(r)$ and $f_e(C)$
- $f_e(C)$ maps to points in an open n-ball such that $f_e(C) = C^{\mathcal{I}}$: $C^{\mathcal{I}} = \{x \in \mathbb{R}^n | \, \|f_e(C) - x\| < r_e(C)\}$
  - these are the *extension* of a unary predicate in $\mathbb{R}^n$
- $f_e(r)$ maps a binary predicate r to a vector such that $r^{\mathcal{I}} = \{(x, y) | x + f_e(r) = y\}$
- use axioms in T as constraints
- sufficient to focus on normal form of T

EL Embeddings

- eliminate the ABox:
  - rewrite role assertions $r(a, b)$ as $\{a\} \sqsubseteq \exists r.\{b\}$
  - rewrite class assertions $C(a)$ as $\{a\} \sqsubseteq C$
  - replace $\{a\}$ with N_a

EL Embeddings

- eliminate the ABox:
  - rewrite role assertions $r(a, b)$ as $\{a\} \sqsubseteq \exists r.\{b\}$
  - rewrite class assertions $C(a)$ as $\{a\} \sqsubseteq C$
  - replace $\{a\}$ with N_a
- normalize the TBox (Baader et al., 2005):
  - $C \sqsubseteq D$
  - $C \sqcap D \sqsubseteq E$
  - $C \sqsubseteq \exists R.D$
  - $\exists R.C \sqsubseteq D$

EL Embeddings

- eliminate the ABox:
  - rewrite role assertions $r(a, b)$ as $\{a\} \sqsubseteq \exists r.\{b\}$
  - rewrite class assertions $C(a)$ as $\{a\} \sqsubseteq C$
  - replace $\{a\}$ with N_a
- normalize the TBox (Baader et al., 2005):
  - $C \sqsubseteq D$
  - $C \sqcap D \sqsubseteq E$
  - $C \sqsubseteq \exists R.D$
  - $\exists R.C \sqsubseteq D$
- minimize loss
  - randomly initialize f_e and r_e
  - one loss function for each normal form

Algorithm: loss functions

$$
\begin{aligned}
&loss_{C \sqsubseteq \exists R.D}(c, d, r) = \\
&\max(0, \|f_e(c) + f_e(r) - f_e(d)\| + r_e(c) - r_e(d) - \gamma) \\
&+ |\,\|f_e(c)\| - 1| + |\,\|f_e(d)\| - 1|
\end{aligned} \tag{1}
$$

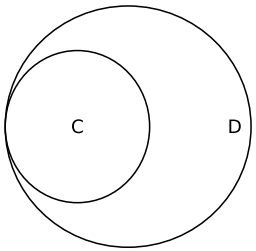
- for all $i \in C^{\mathcal{I}}$, $i + r^{\mathcal{I}} \in D^{\mathcal{I}}$
- $C^{\mathcal{I}} + r^{\mathcal{I}} \subseteq D^{\mathcal{I}}$

Algorithm: loss functions

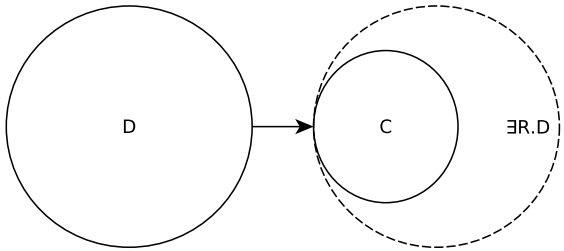
$$
\begin{aligned}
&loss_{C \sqsubseteq \exists R.D}(c, d, r) = \\
&\max(0, \|f_e(c) + f_e(r) - f_e(d)\| + r_e(c) - r_e(d) - \gamma) \\
&+ |\,\|f_e(c)\| - 1| + |\,\|f_e(d)\| - 1|
\end{aligned} \tag{1}
$$

- for all $i \in C^{\mathcal{I}}$, $i + r^{\mathcal{I}} \in D^{\mathcal{I}}$
- $C^{\mathcal{I}} + r^{\mathcal{I}} \subseteq D^{\mathcal{I}}$
- margin γ
- regularization

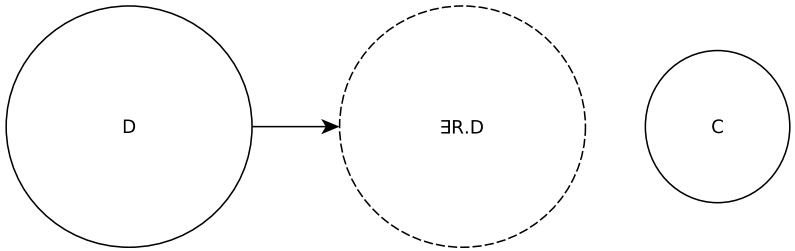
Algorithm: loss functions



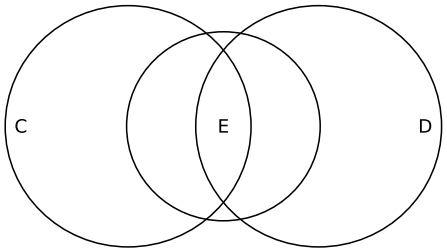
$C \sqsubseteq D$



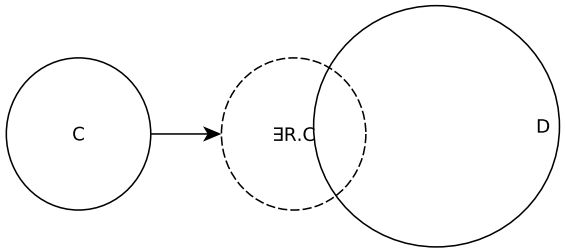
$C \sqsubseteq \exists R.D$



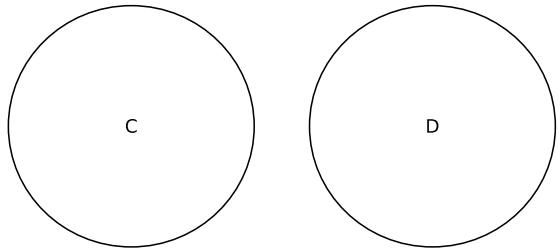
$C \not\sqsubseteq \exists R.D$



$C \sqcap D \sqsubseteq E$

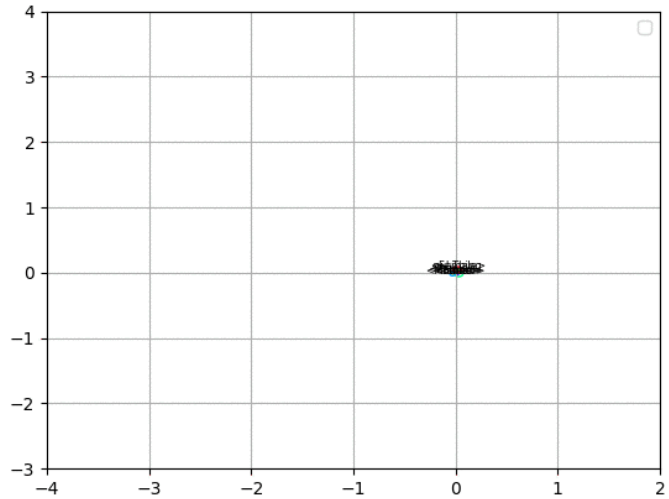


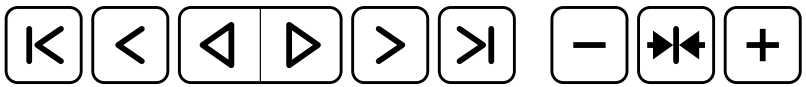
$\exists R.C \sqsubseteq D$



$C \sqcap D \sqsubseteq \bot$

EL Embeddings





Theorem (Soundness)

Let $\mathcal{T}$ be a theory in $\mathcal{EL}^{++}$. If $\gamma \leq 0$ and $loss_n(\eta(\mathcal{T})) = 0$ then $\mathcal{T}$ has a model.

$$
\begin{aligned}
\mathit{Male} &\sqsubseteq \mathit{Person} \\
\mathit{Female} &\sqsubseteq \mathit{Person} \\
\mathit{Father} &\sqsubseteq \mathit{Male} \\
\mathit{Mother} &\sqsubseteq \mathit{Female} \\
\mathit{Father} &\sqsubseteq \mathit{Parent} \\
\mathit{Mother} &\sqsubseteq \mathit{Parent} \\
\mathit{Female} \sqcap \mathit{Male} &\sqsubseteq \bot \\
\mathit{Female} \sqcap \mathit{Parent} &\sqsubseteq \mathit{Mother} \\
\mathit{Male} \sqcap \mathit{Parent} &\sqsubseteq \mathit{Father} \\
\exists \mathit{hasChild}.\mathit{Person} &\sqsubseteq \mathit{Parent} \\
\mathit{Parent} &\sqsubseteq \mathit{Person} \\
\mathit{Parent} &\sqsubseteq \exists \mathit{hasChild}.\top
\end{aligned}
$$

EL Box Embeddings

Limitations of EL Embeddings:

- intersection of two n-balls is no n-ball
  - solution: axis-parallel rectangles (boxes) instead of n-balls
- relation model (TransE) allows only 1:1 relations
  - solution: more complex relation models
- expressivity limited to EL++
  - how about (unrestricted) negation, union, universal quantifiers?
- no completeness result

mOWL

- software library for machine learning with Semantic Web (OWL) ontologies
- embedding methods
- constrained optimization
- semantic similarity

`https://github.com/bio-ontology-research-group/mowl`

Summary: embedding ontology axioms

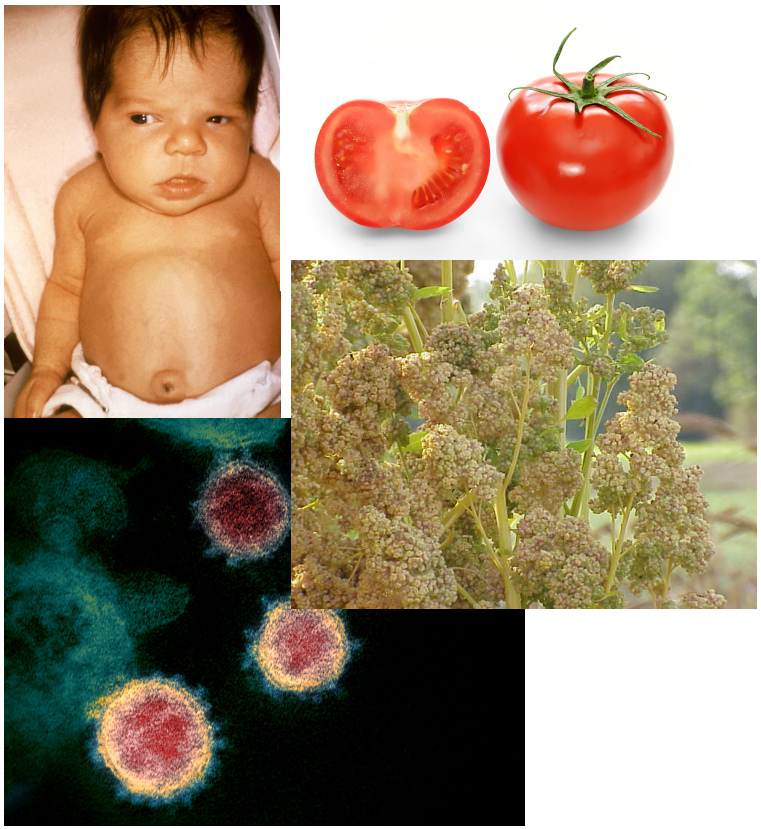
- set of neural embedding methods for formal knowledge bases
  - graph-based, syntactic, model-theoretic
- focus on Semantic Web ontologies (Description Logic)
  - major importance in life science
  - subsumes knowledge graphs
- embedding axioms and ontologies
  - link to model theory
  - proof of soundness — re-establishing theoretical results
- software library

Summary: embedding ontology axioms

- set of neural embedding methods for formal knowledge bases
  - graph-based, syntactic, model-theoretic
- focus on Semantic Web ontologies (Description Logic)
  - major importance in life science
  - subsumes knowledge graphs
- embedding axioms and ontologies
  - link to model theory
  - proof of soundness — re-establishing theoretical results
- software library

How can we use these methods in computational biology?

Variant effect prediction

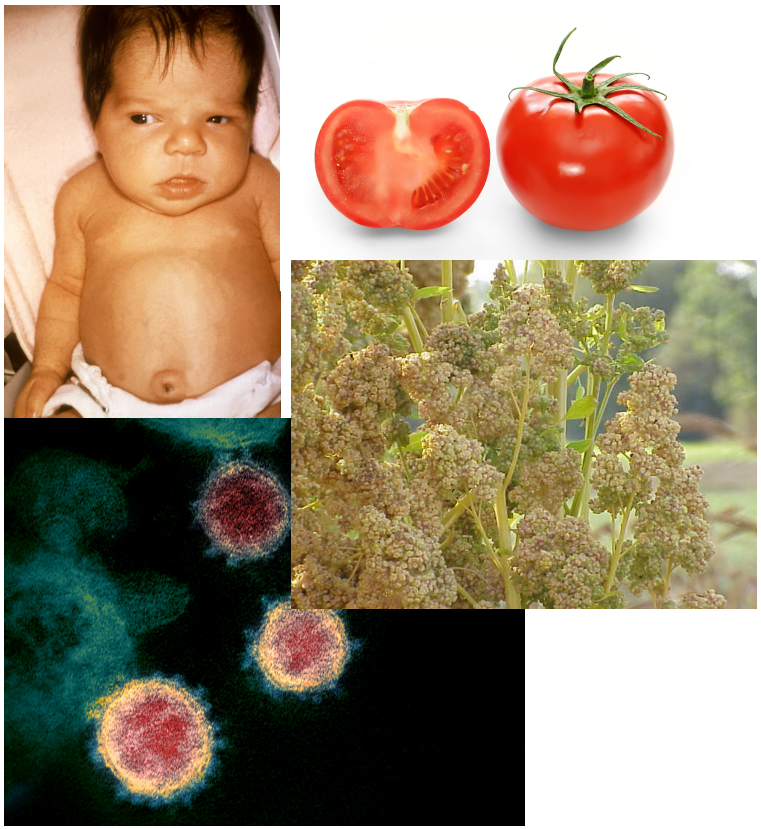


Source: Wikipedia

"Simple" case: Mendelian traits, Single Nucleotide Variants

- does a genomic variant affect gene function?
- which gene(s)?
- what pathways/networks/functions would be affected?
- how is physiology affected?
- any interactions with the environment?

Variant effect prediction



Source: Wikipedia

"Simple" case: Mendelian traits, Single Nucleotide Variants

- does a genomic variant affect gene function?
- which gene(s)?
- what pathways/networks/functions would be affected?
- how is physiology affected?
- any interactions with the environment?

None of these questions can be answered from the sequencing data alone!

- rely on medical knowledge about genetics, cell biology, anatomy, and physiology

Phenotypes

Phenotypes can encode for many kinds of interactions

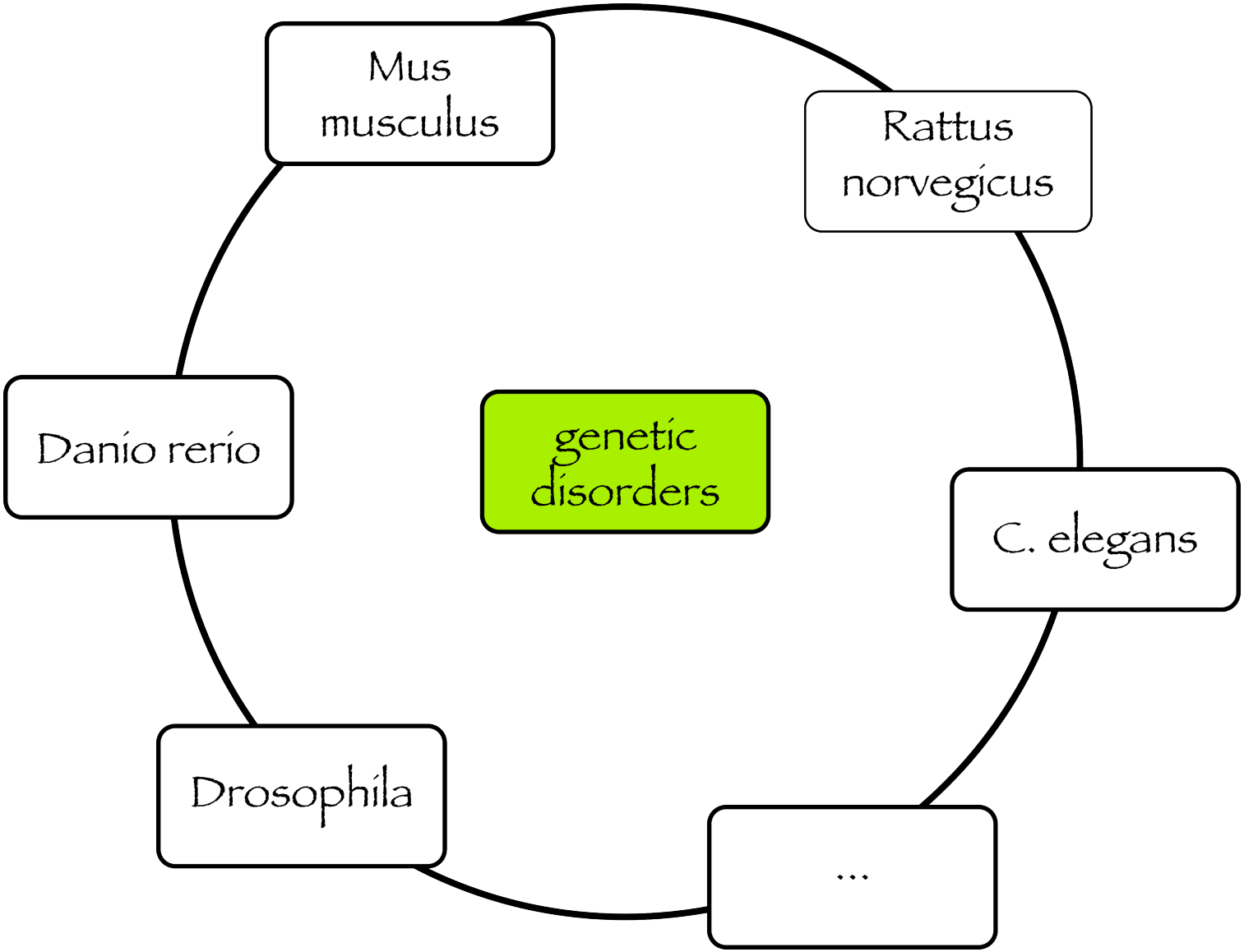
- forward genetics:
  - clinical observations
- reverse genetics:
  - model organisms
  - cell models
- similar phenotype – similar mechanism ?
  - patient–patient similarity
  - comparing patient to canonical disease phenotypes
  - only 4,209 single gene disorders known (OMIM, 2 Feb 2022)

Analyzing phenotypes

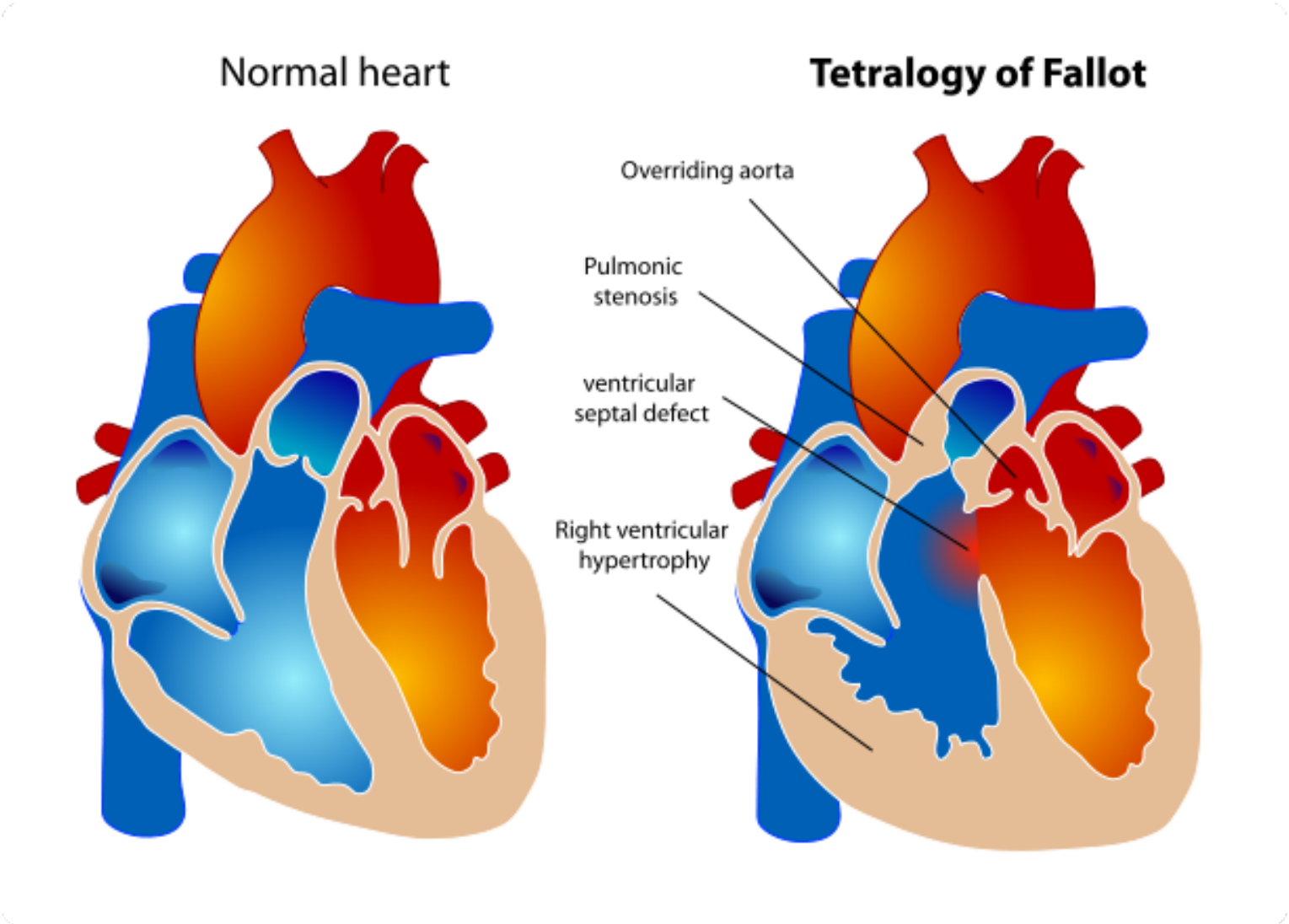
GENOTYPE		
ALSM1(NM_015120.4) [c.10775delC] + [-]	B6.Cg-Alms1$^{foz/fox}$/J	kcnj11$^{c14/c14}$; insr$^{t143/+}$(AB)
		
obesity, diabetes mellitus, insulin resistance	increased food intake, hyperglycemia, insulin resistance	increased weight, adipose tissue volume, glucose homeostasis altered
PHENOTYPE		

1

Cross-species integration: PhenomeNET ontology



Analyzing phenotypes



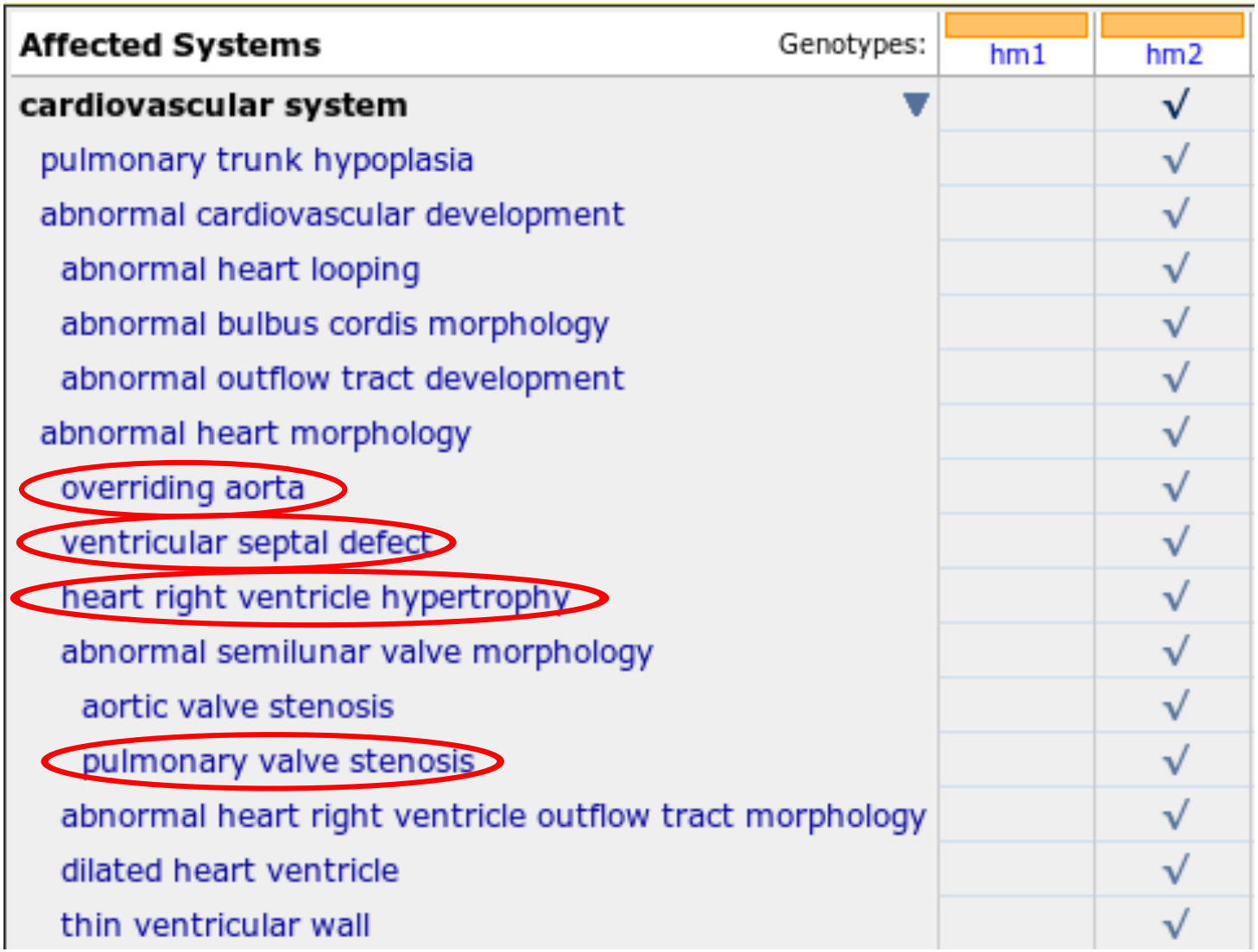
Analyzing phenotypes

Phc1 knockout mice

Affected Systems	Genotypes: hm1	hm2
cardiovascular system		√
pulmonary trunk hypoplasia		√
abnormal cardiovascular development		√
abnormal heart looping		√
abnormal bulbus cordis morphology		√
abnormal outflow tract development		√
abnormal heart morphology		√
overriding aorta		√
ventricular septal defect		√
heart right ventricle hypertrophy		√
abnormal semilunar valve morphology		√
aortic valve stenosis		√
pulmonary valve stenosis		√
abnormal heart right ventricle outflow tract morphology		√
dilated heart ventricle		√
thin ventricular wall		√

Analyzing phenotypes

Integration of phenotype ontologies enables identification of disease phenotypes in mice.



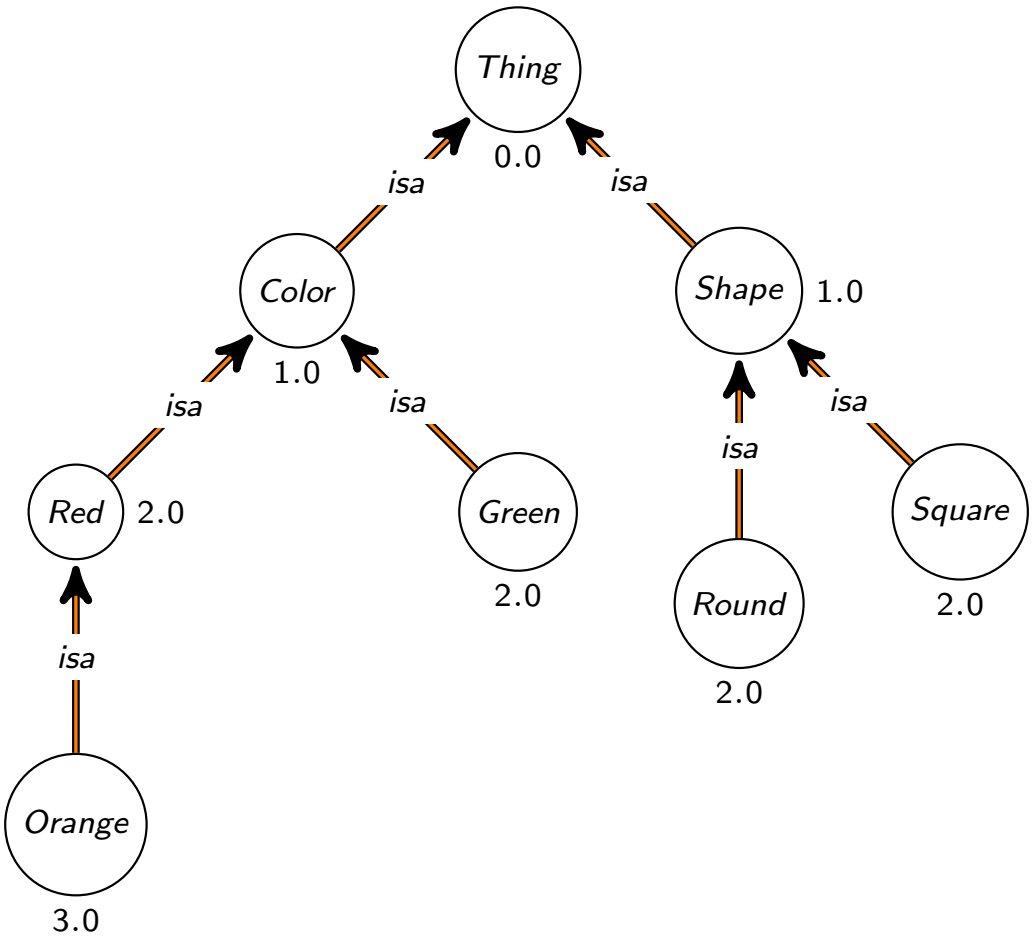
Affected Systems	Genotypes: hm1	hm2
cardiovascular system		√
pulmonary trunk hypoplasia		√
abnormal cardiovascular development		√
abnormal heart looping		√
abnormal bulbus cordis morphology		√
abnormal outflow tract development		√
abnormal heart morphology		√
overriding aorta		√
ventricular septal defect		√
heart right ventricle hypertrophy		√
abnormal semilunar valve morphology		√
aortic valve stenosis		√
pulmonary valve stenosis		√
abnormal heart right ventricle outflow tract morphology		√
dilated heart ventricle		√
thin ventricular wall		√

Analyzing phenotypes

Semantic similarity over phenotype ontologies measures phenotypic similarity

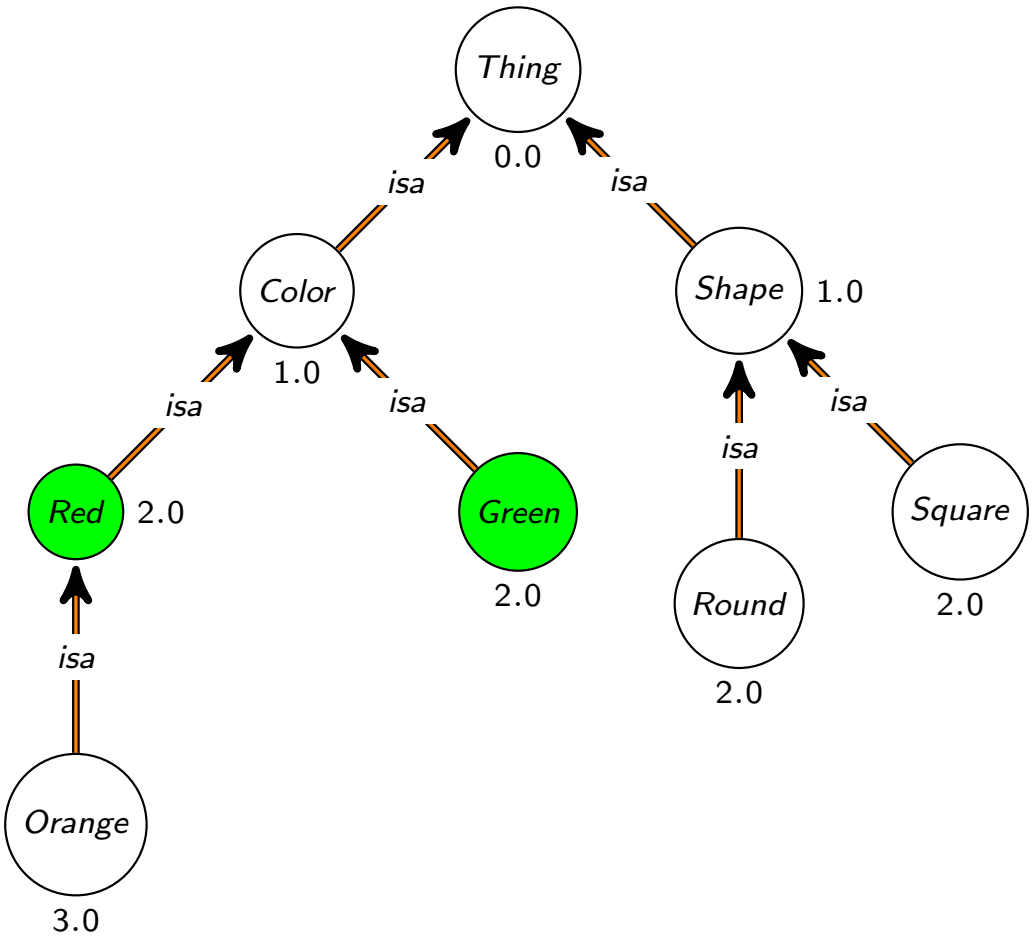
- semantic similarity: similarity measure based on information contained in the axioms/structure of an ontology
  - anatomy: front limb – hind limb vs. front limb – eye
  - function: detection of salty taste – detection of sweet taste vs. detection of salty taste – apoptosis
  - quality: red – orange vs. red – green vs. red – round
- $\Rightarrow$ semantic similarity over phenotype ontologies combines similarity between anatomy, function, and quality

How to measure similarity?



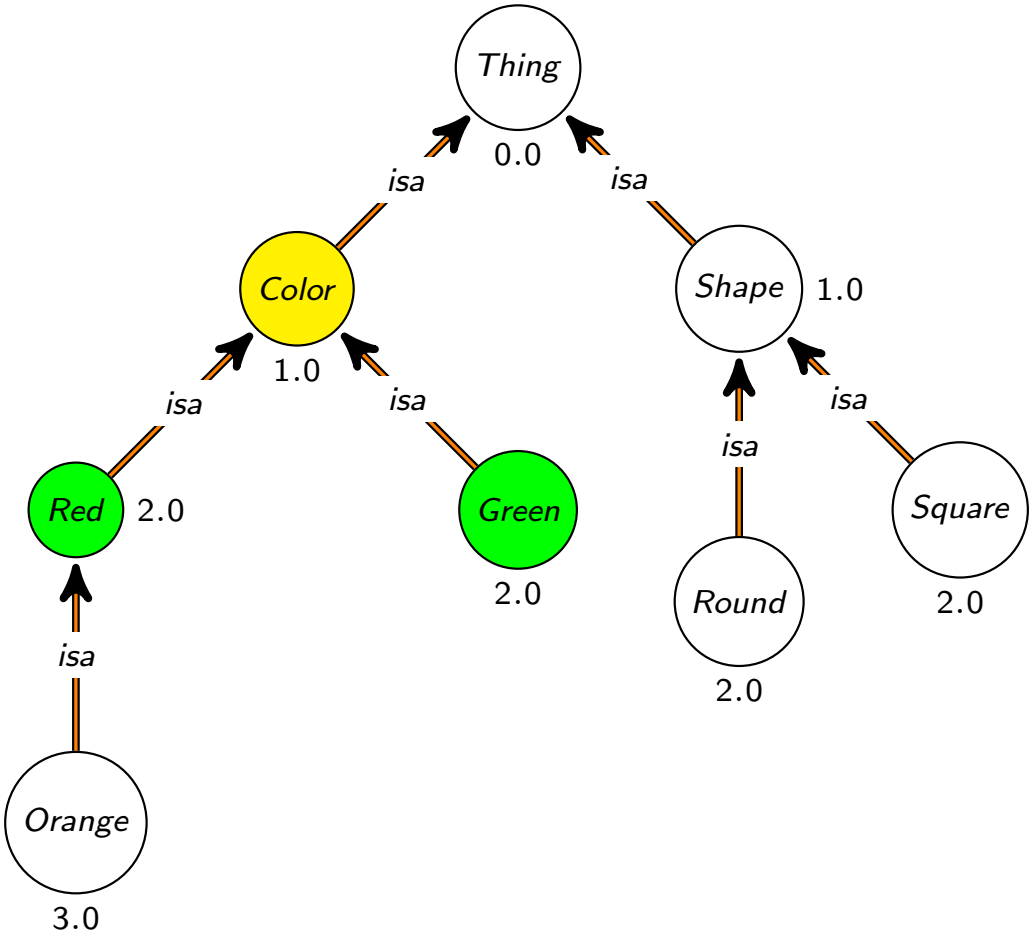
- Resnik 1995: similarity between x and y is the information content of the *most informative common ancestor*

How to measure similarity?



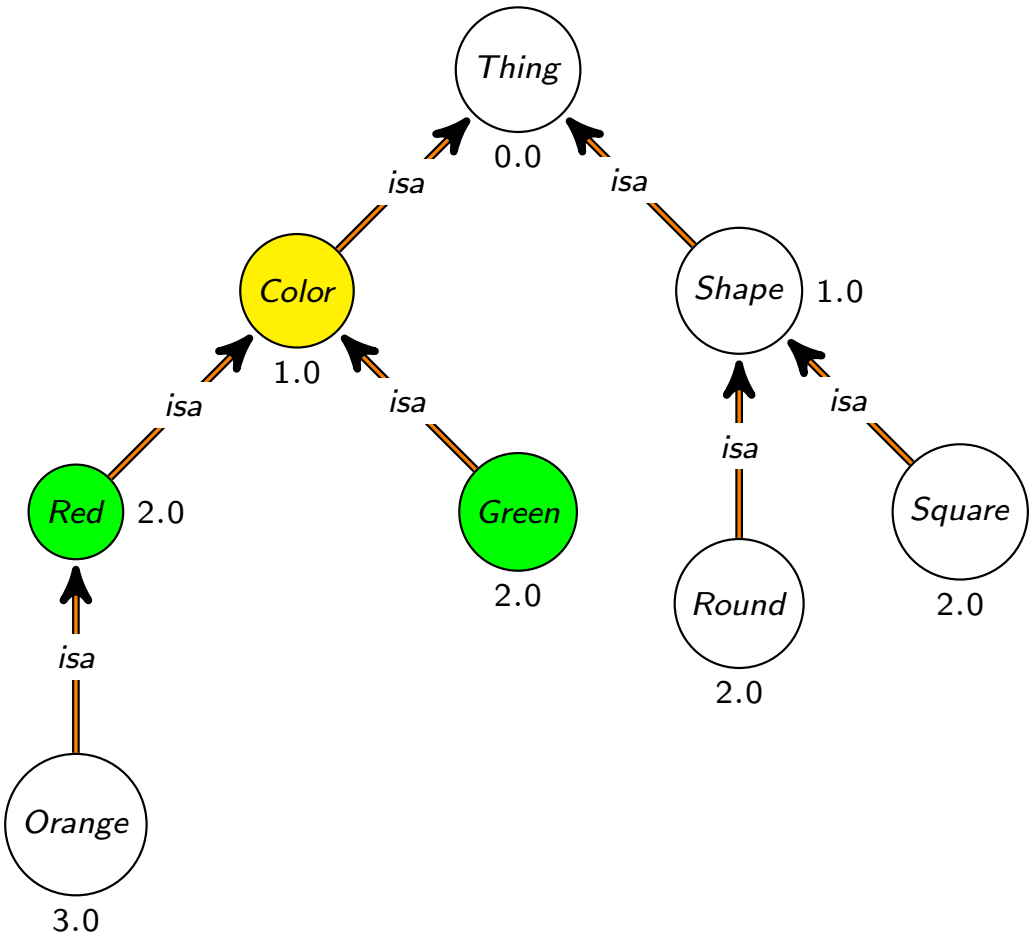
- Resnik 1995: similarity between x and y is the information content of the *most informative common ancestor*

How to measure similarity?



- Resnik 1995: similarity between x and y is the information content of the *most informative common ancestor*

How to measure similarity?



- Resnik 1995: similarity between x and y is the information content of the *most informative common ancestor*
- $sim_{Resnik}(Green, Red) = 1.0$

Analyzing phenotypes

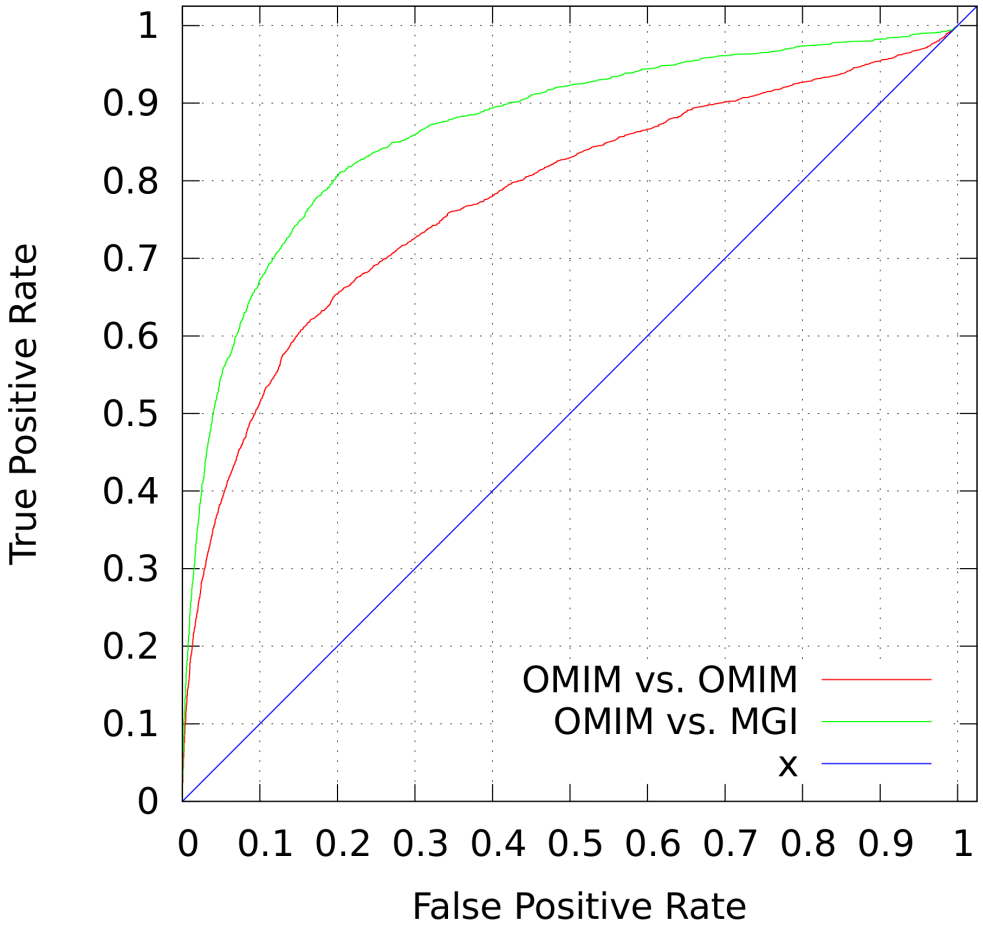
Information content of phenotype:

$$IC(x) = -\log(p(x))$$

Phenotype similarity:
Max average: $sim_{MA}(X, Y) = \frac{1}{|X|} \sum_{x \in X} \max_{y \in Y} sim_{Resnik}(x, y)$
Best match average: $sim_{BMA}(X, Y) = \frac{sim_{MA}(X,Y)+sim_{MA}(Y,X)}{2}$
$\Rightarrow$ systematic, pairwise comparison of disease and model organism phenotypes

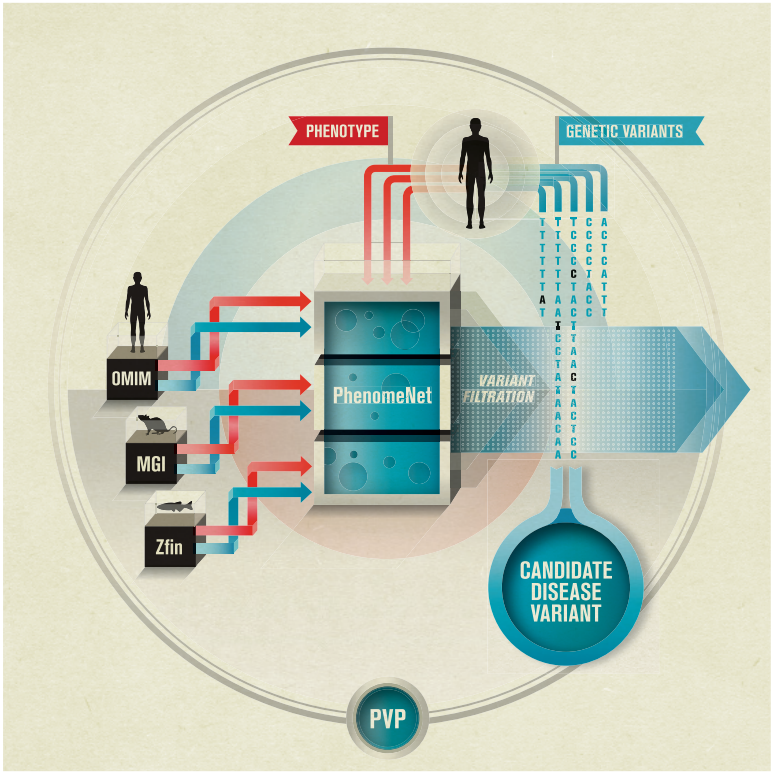
Evaluating the effect of annotation size on measures of semantic similarity. JBMS, 2017.

Analyzing phenotypes



- AUC (OMIM): 0.82
- AUC (MGI): 0.90

Variant effect prediction



DDIEM Home

	Top hit	Top 10 hits	Total
DeepPVP	4,096 (72.04%)	4,768 (83.86%)	5,686
PVP	3,619 (63.65%)	4,076 (71.68%)	5,686
Exomiser	2,910 (51.18%)	3,608 (63.45%)	5,686
CADD	1,060 (18.64%)	2,429 (42.72%)	5,686
DANN	170 (2.99%))	1,322 (23.25%))	5,686

The phenotype gap

- phenotype-based variant prioritization works quite well
  - but only if we have phenotypes associated with a gene

The phenotype gap

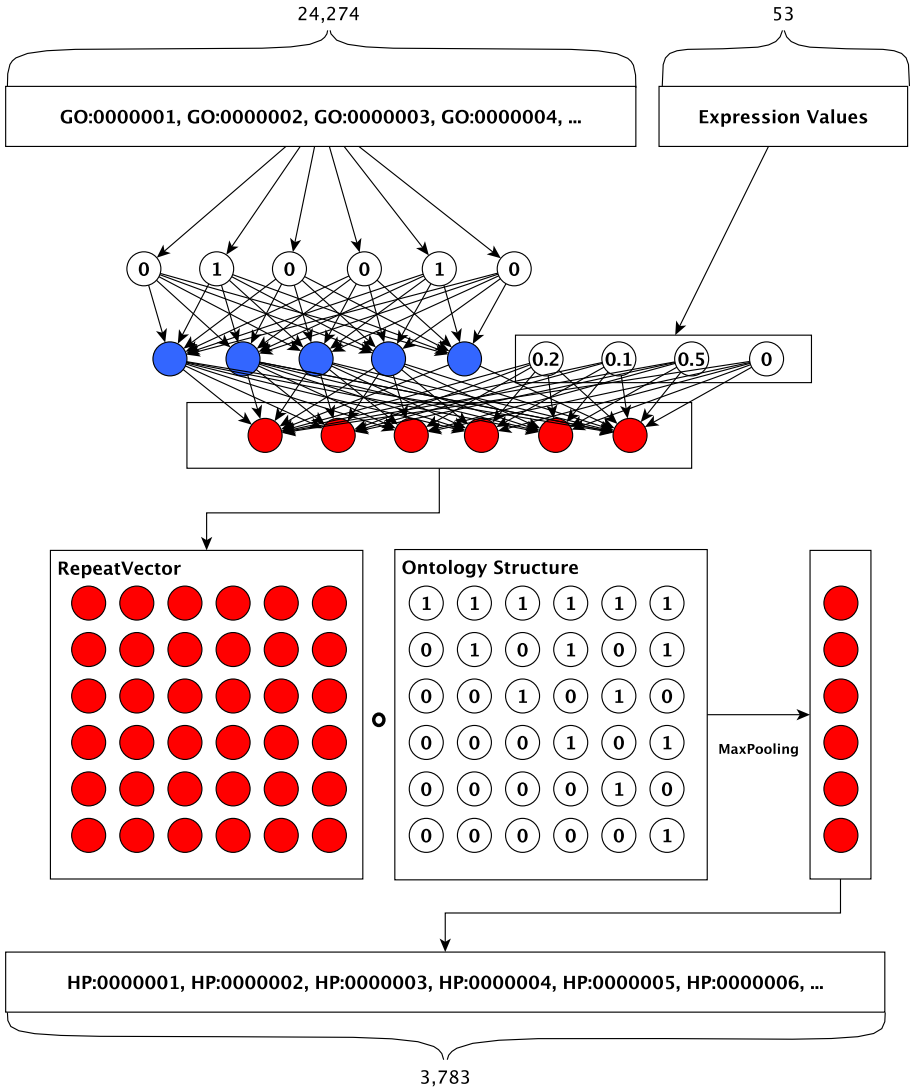
- phenotype-based variant prioritization works quite well
  - but only if we have phenotypes associated with a gene
- human + mouse genes with phenotypes: 60% of the human genome
  - 4,209 genes with human phenotypes, 14,899 in mouse
- what about the remaining $\approx$ 40% of the human genome?
  - genes without mouse orthologs
  - genes with different function in human

The phenotype gap

- phenotype-based variant prioritization works quite well
  - but only if we have phenotypes associated with a gene
- human + mouse genes with phenotypes: 60% of the human genome
  - 4,209 genes with human phenotypes, 14,899 in mouse
- what about the remaining $\approx$ 40% of the human genome?
  - genes without mouse orthologs
  - genes with different function in human

Can we predict phenotypes?

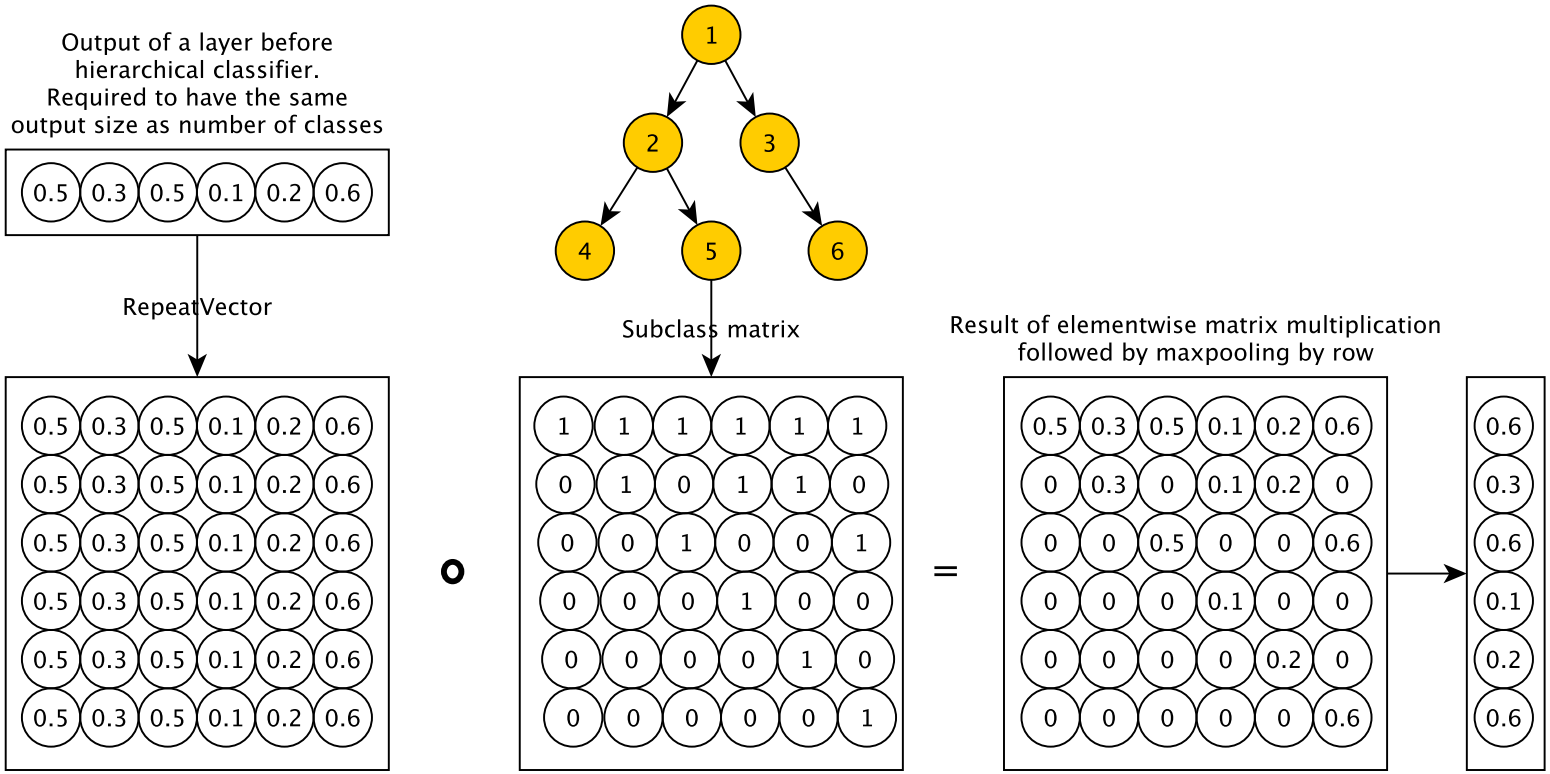
DeepPheno: Predicting single gene loss of function phenotypes



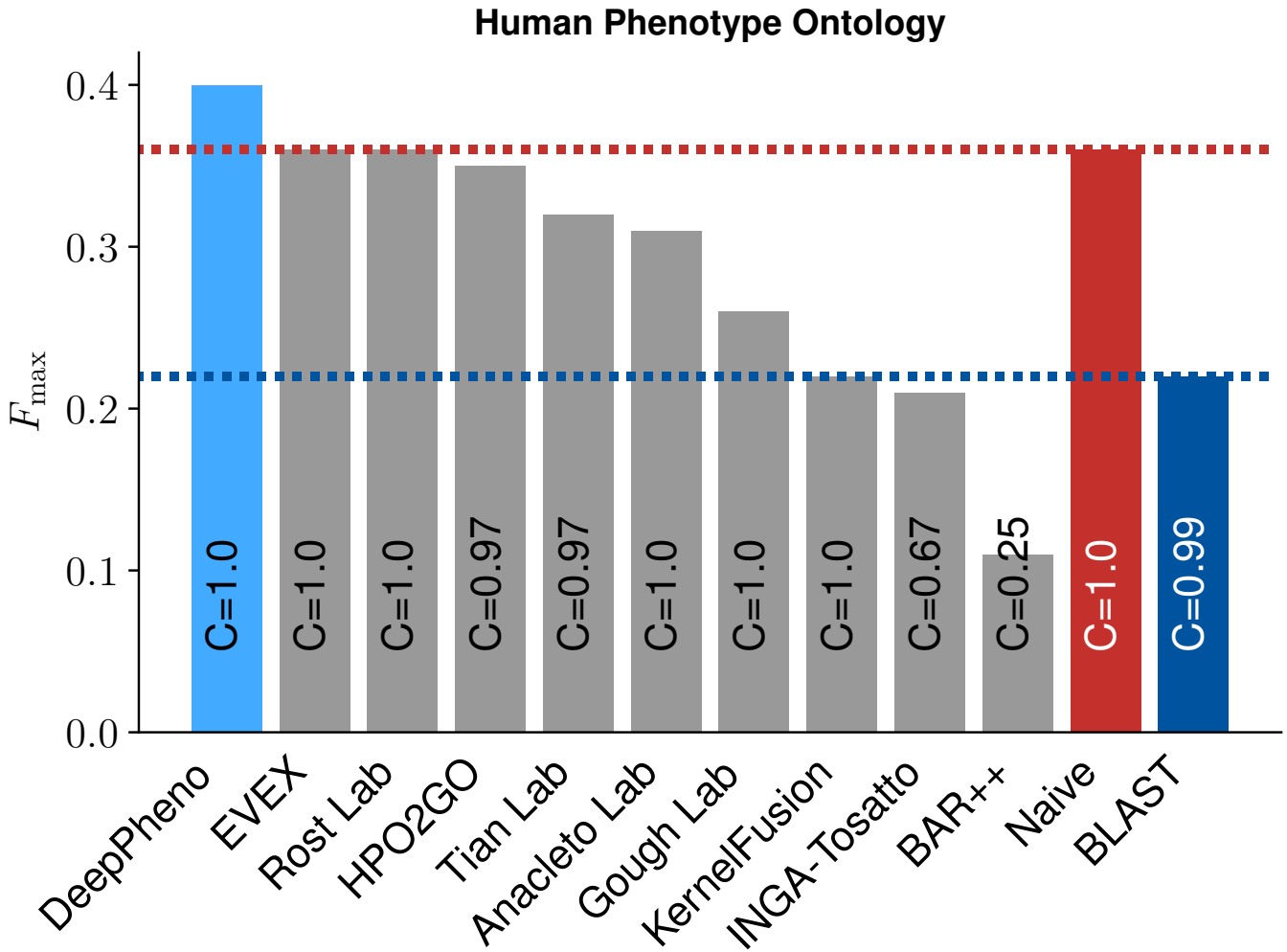
Based on Łukasiewic-Tarski logic:

- consistency constraints during training and prediction
- e.g.: for a gene G and phenotype (HPO) classes C and D: $C \sqsubseteq D$ implies $\sigma(P, C) \leq \sigma(P, D)$

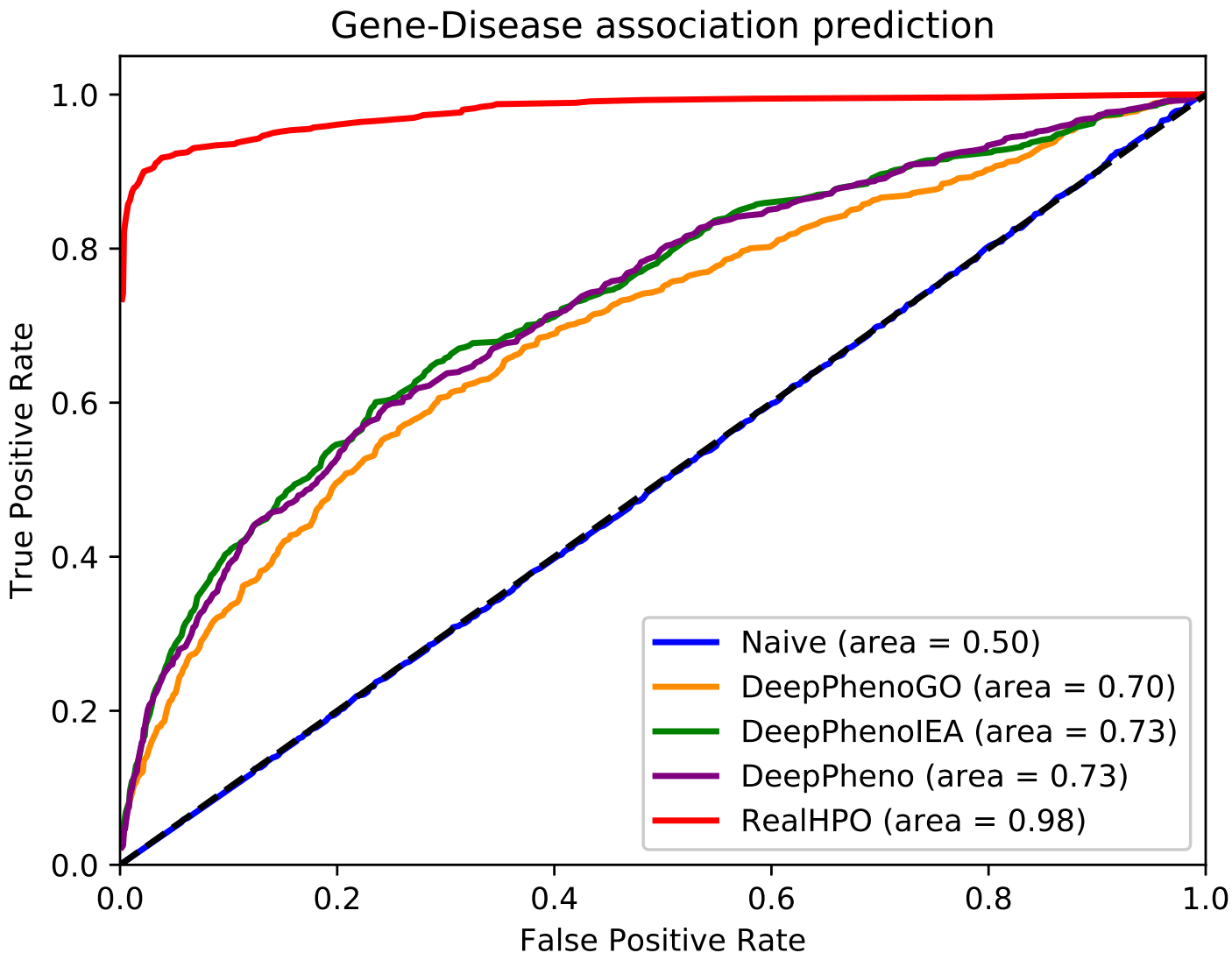
DeepPheno: Predicting single gene loss of function phenotypes



DeepPheno: Predicting single gene loss of function phenotypes



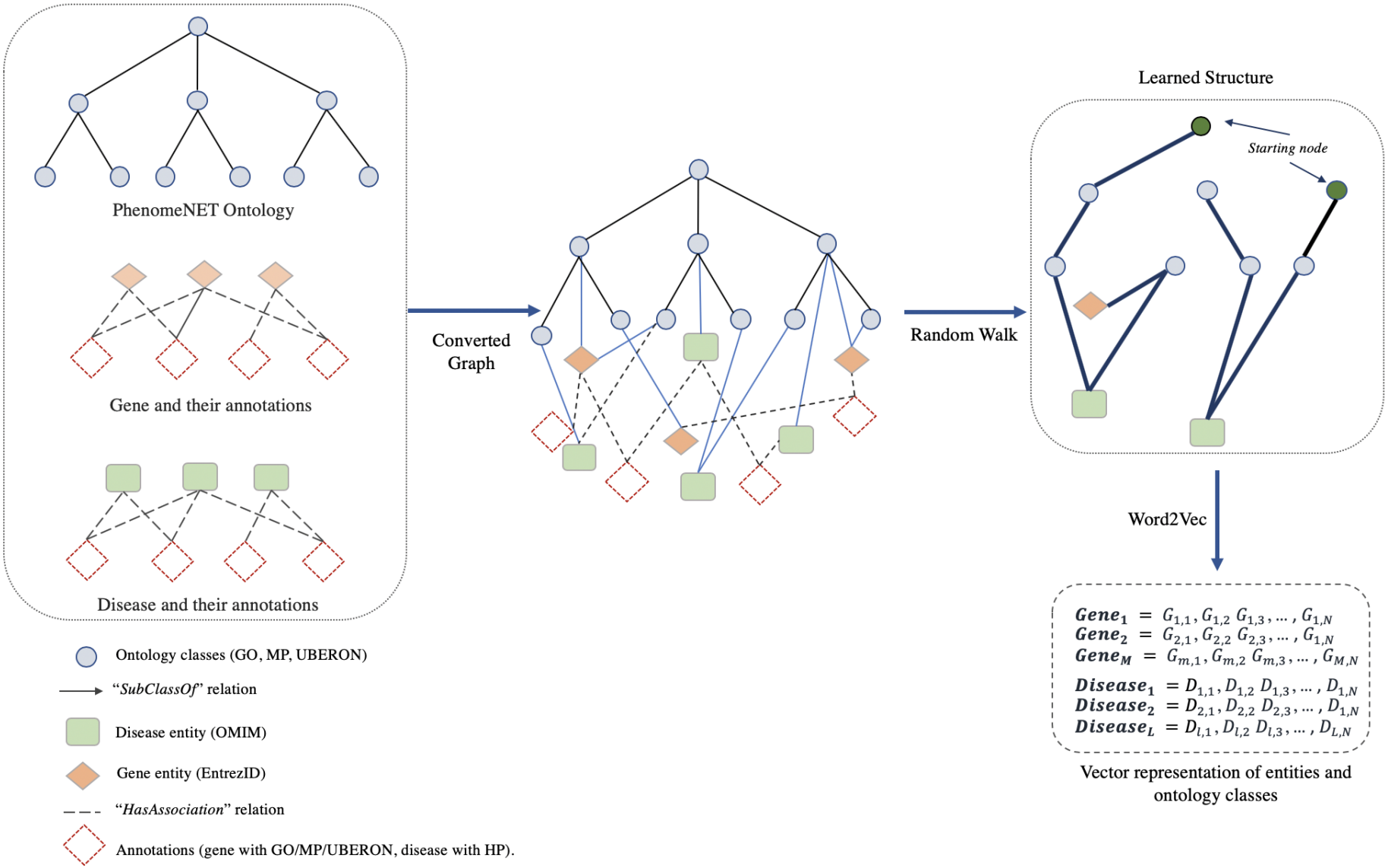
Predicted phenotypes predict gene–disease associations



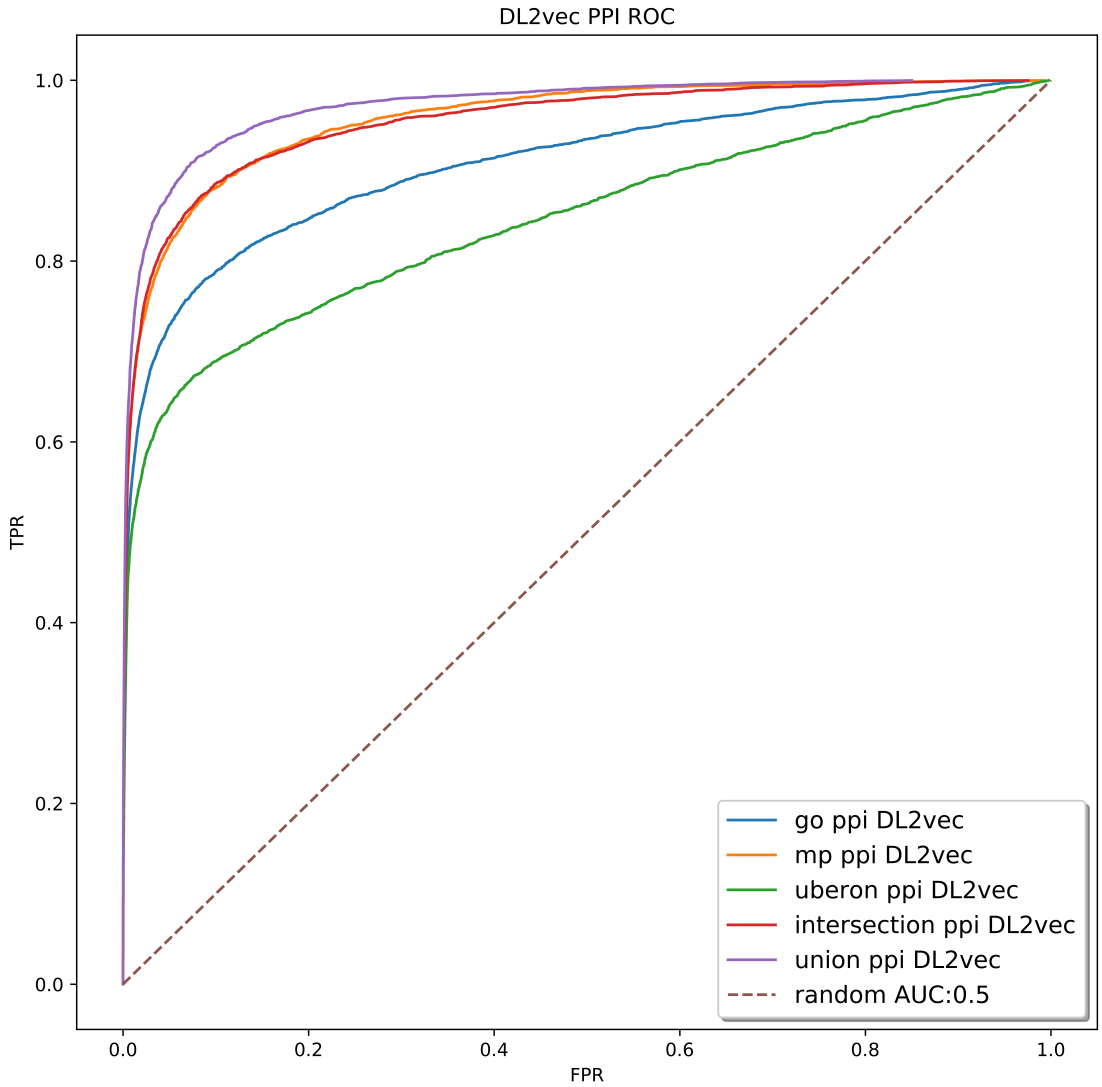
DL2Vec predictions

- DeepPheno uses (predicted) GO functions
  - not available for all gene products
  - can be inaccurate
- more information about phenotypes:
  - celltype of expression
  - anatomical site of expression
  - protein-protein interactions (phenotype modules)
  - and their interrelations (from ontologies)
- knowledge bases and ontologies!
- allows us to employ our embeddings directly

DL2Vec


PhenomeNET Ontology
Gene and their annotations
Disease and their annotations
Converted Graph
Random Walk
Learned Structure
Starting node
Word2Vec
$Gene_1 = G_{1,1}, G_{1,2}\ G_{1,3}, \dots, G_{1,N}$
$Gene_2 = G_{2,1}, G_{2,2}\ G_{2,3}, \dots, G_{1,N}$
$Gene_M = G_{m,1}, G_{m,2}\ G_{m,3}, \dots, G_{M,N}$
$Disease_1 = D_{1,1}, D_{1,2}\ D_{1,3}, \dots, D_{1,N}$
$Disease_2 = D_{2,1}, D_{2,2}\ D_{2,3}, \dots, D_{1,N}$
$Disease_L = D_{l,1}, D_{l,2}\ D_{l,3}, \dots, D_{L,N}$
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).

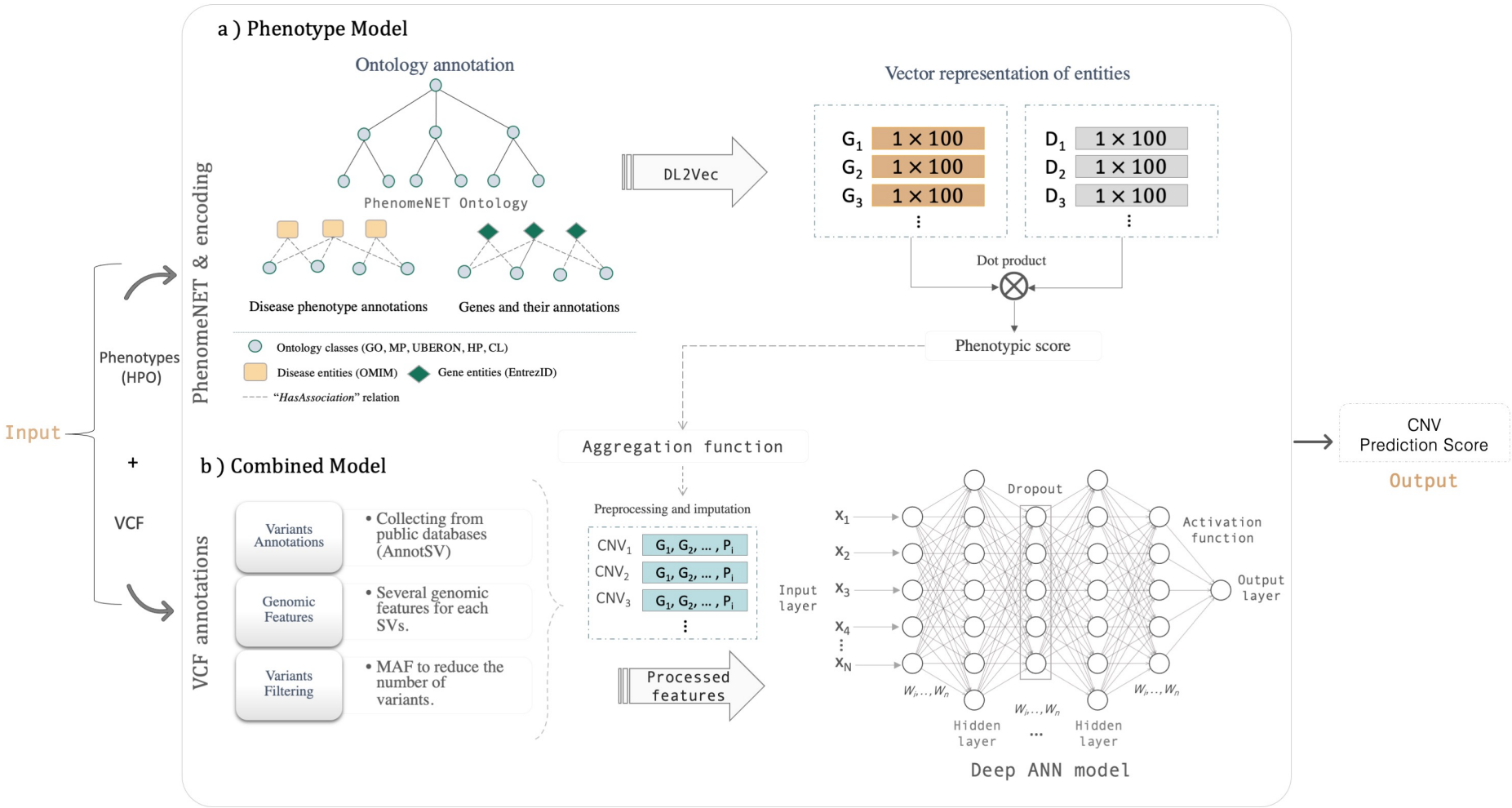
DL2Vec predictions for gene–disease associations



Comparison

- DeepPheno + DL2Vec predict phenotypes from background knowledge
  - significant improvement over state of the art in phenotype-based prediction of gene–disease associations
  - covers almost the entire human genome
- learn to account for some "biases" in data
- data-driven mapping between anatomy, molecular function, and endophenotypes
  - learns some (multi-scale) physiology/pathophysiology
- may allow us to look at more complex diseases

Structural Variants



Variant classification: structural variants

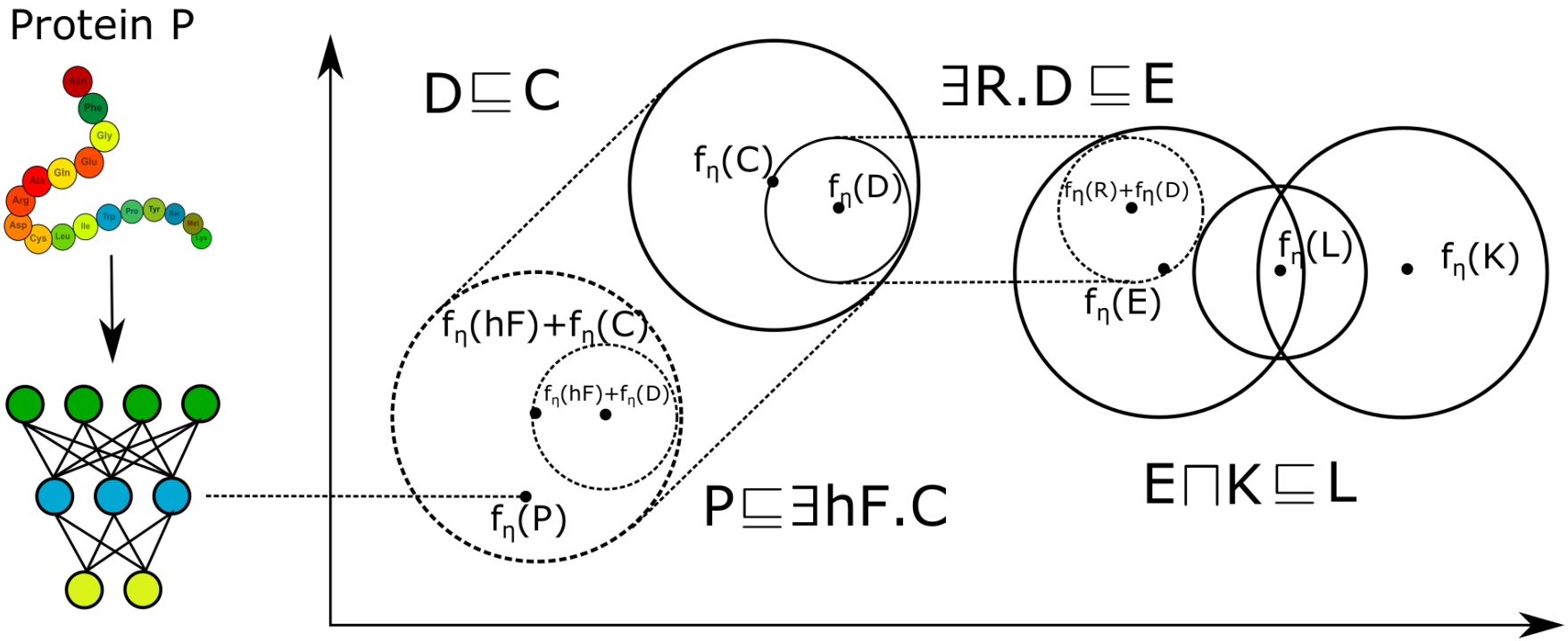
		Synthetic dataset					Synthetic dataset (novel diseases)				
		Recall@1	**Recall@10**	**Recall@30**	**ROCAUC**	**PRAUC**	**Recall@1**	**Recall@10**	**Recall@30**	**ROCAUC**	**PRAUC**
DeepSVP models using maximum score	GO	487 (0.3240)	969 (0.6447)	969 (0.6447)	0.9930	0.0285	98 (0.1531)	414 (0.6469)	590 (0.9219)	0.9944	**0.0504**
	MP	352 (0.2342)	894 (0.5948)	1222 (0.8130)	0.9913	0.0310	101 (0.1578)	355 (0.5547)	536 (0.8375)	0.9928	0.0423
	HP	337 (0.2242)	1117 (0.7432)	1363 (0.9069)	0.9932	0.0586	111 (0.1734)	519 (0.8109)	609 (0.9516)	0.9956	0.0806
	CL	227 (0.1510)	**1154 (0.7678)**	1371 (0.9122)	0.9934	0.0667	72 (0.1125)	**533 (0.8328)**	611 (0.9547)	0.9960	0.0905
	UBERON	444 (0.2954)	1052 (0.6999)	1417 (0.9428)	0.9935	0.0393	92 (0.1437)	477 (0.7453)	615 (0.9609)	0.9953	0.0669
	Union	431 (0.2868)	1148 (0.7638)	1383 (0.9202)	**0.9937**	0.0476	67 (0.1047)	335 (0.5234)	522 (0.8156)	0.9959	0.0708
DeepSVP models using average score	GO	416 (0.2768)	1071 (0.7126)	1391 (0.9255)	0.9932	**0.0724**	128 (0.2000)	475 (0.7422)	615 (0.9609)	**0.9960**	**0.1070**
	MP	360 (0.2395)	678 (0.4511)	1356 (0.9022)	0.9918	0.0253	182 (0.2844)	367 (0.5734)	608 (0.9500)	0.9941	0.0232
	HP	330 (0.2196)	893 (0.5941)	1409 (0.9375)	0.9924	0.0395	78 (0.1219)	370 (0.5781)	614 (0.9594)	0.9939	0.0436
	CL	222 (0.1477)	860 (0.5722)	1367 (0.9095)	0.9918	0.0430	69 (0.1078)	360 (0.5625)	611 (0.9547)	0.9940	0.0575
	UBERON	193 (0.1284)	716 (0.4764)	1302 (0.8663)	0.9882	0.0389	53 (0.0828)	384 (0.6000)	585 (0.9141)	0.9927	0.0542
	Union	**504 (0.3353)**	1018 (0.6773)	**1431 (0.9521)**	0.9932	0.0258	**149 (0.2328)**	460 (0.7188)	**618 (0.9656)**	0.9959	0.0464
SV pathogenicity prediction	StrVCTVRE	38 (0.0252)	620 (0.4125)	1022 (0.6799)	0.9820	0.0270	9 (0.0140)	162 (0.2531)	373 (0.5828)	0.9850	0.0209
	CADD-SV	72 (0.0471)	223 (0.1458)	403 (0.2634)	0.9191	0.0075	34 (0.0531)	120 (0.1875)	210 (0.3281)	0.9307	0.0144
SV ranking	AnnotSV (best)	930 (0.6188)	930 (0.6188)	1493 (0.9933)	0.9956	0.0035	248 (0.3875)	248 (0.3875)	636 (0.99375)	0.9937	0.0039
	AnnotSV (average)	-	930 (0.6188)	1493 (0.9933)	0.9750	0.0054	-	248 (0.3875)	636 (0.99375)	0.9658	0.0086
	AnnotSV (worst)	-	930 (0.6188)	930 (0.6188)	0.9543	0.0049	-	248 (0.3875)	248 (0.3875)	0.9379	0.0081

Imane Boudellioua et al., PLoS Comp Bio, 2017; Scientific Reports, 2019; BMC Bioinformatics, 2019; Marwa Abdelhakim et al., Orphanet Journal, 2020; Jun Chen et al., Bioinformatics, 2020; Azza Althagafi et al., bioRxiv, 2021.

Zero-shot prediction with EL Embeddings

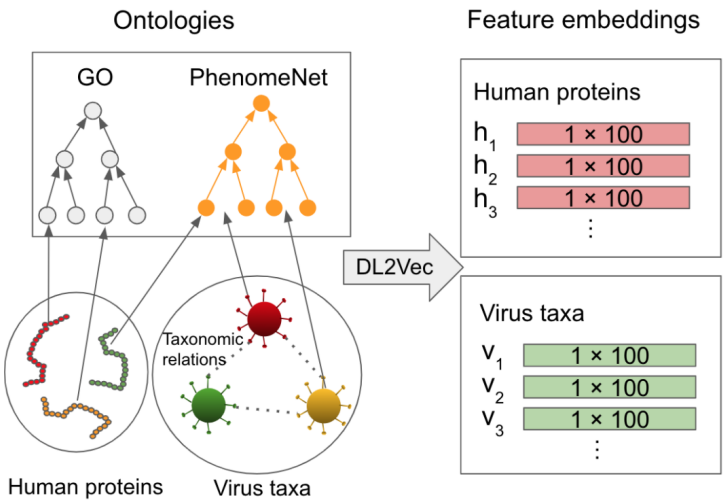
- zero-shot phenotype prediction:
  - no gene/protein/variant has *ever* been observed to produce phenotype P
  - can we predict that a gene/protein/variant results in P?
- no training data
  - but maybe we can exploit the axioms?

Zero-shot prediction with EL Embeddings

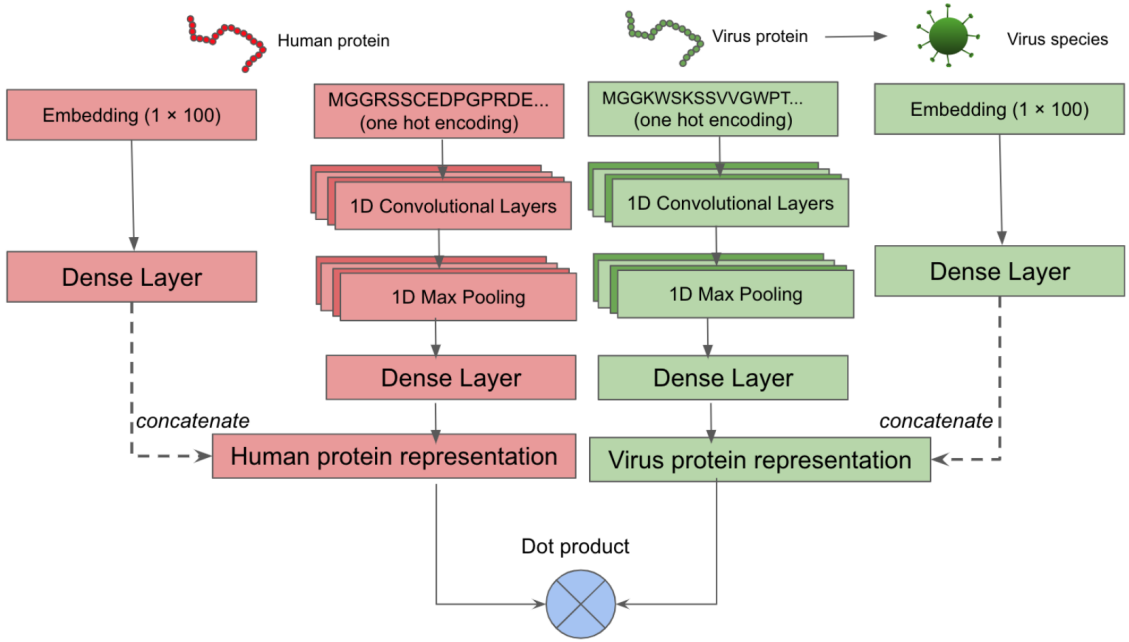


Kulmanov & Hoehndorf, DeepGOZero: Improving protein function prediction from sequence and zero-shot learning based on ontology axioms. BioRxiv, 2022.

DeepViral

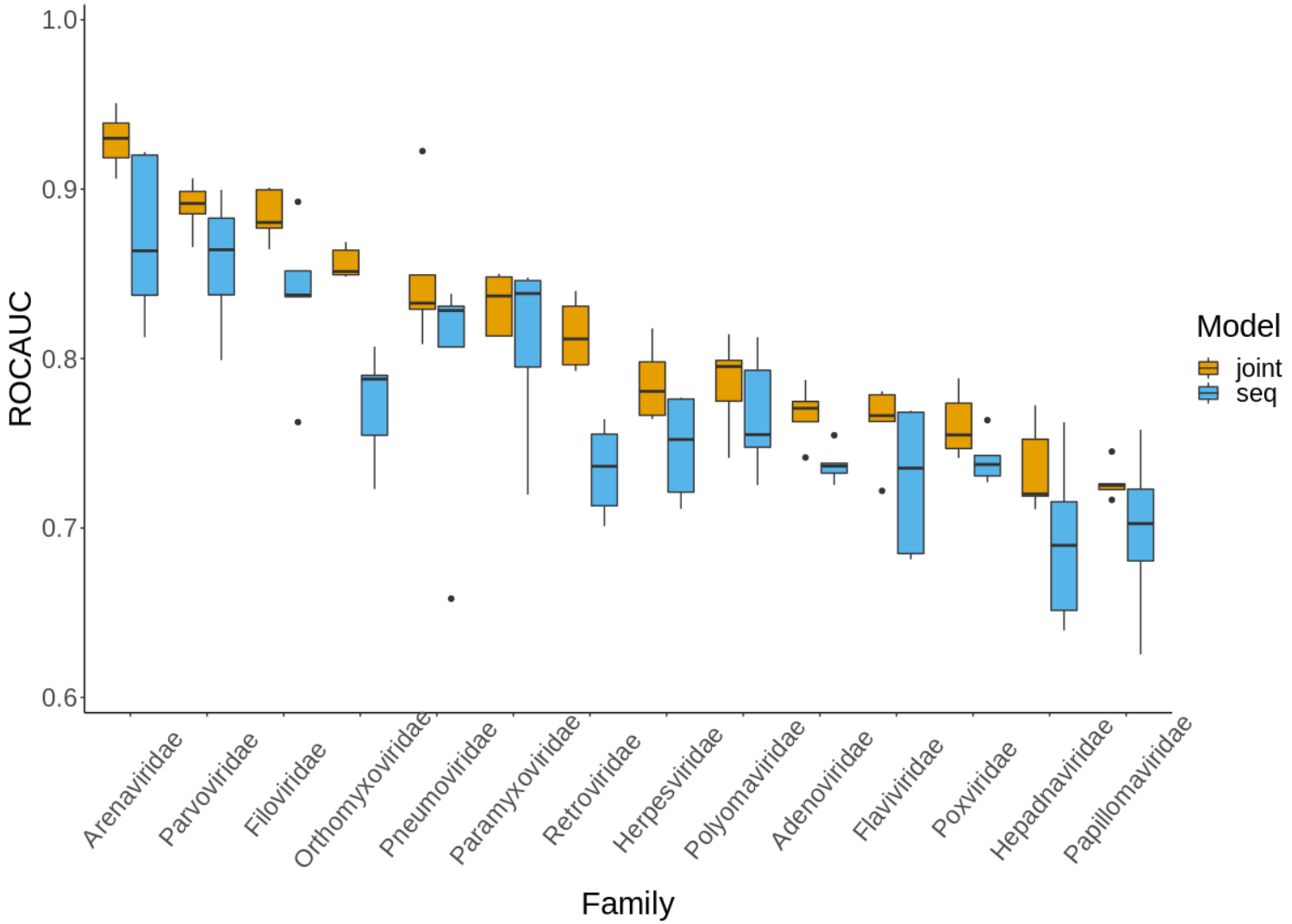


(a) Generation of feature embeddings

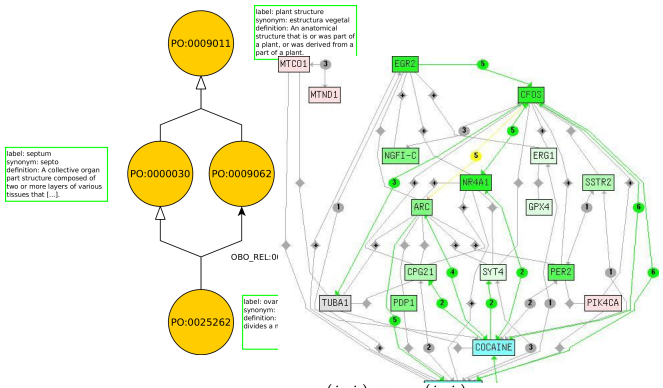


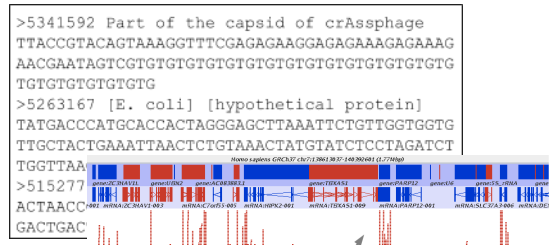
(b) Joint prediction model from embeddings and sequences

DeepViral predicts targets of novel viruses



Knowledge-based methods and machine learning





$conn(j_1, j_2) \rightarrow conn(j_2, j_1)$

$conn(j_1, j_2) \rightarrow j_1 \neq j_2$

$in(j_1, s_1) \wedge in(j_2, s_2) \wedge \neg soverlap(s_1, s_2) \rightarrow \neg conn(j_1, j_2)$

$conn(j_1, j_2) \wedge in(j_1, s) \rightarrow in(j_2, s)$

$\forall s \forall P(\forall x(P(x) \leftrightarrow in(x, s)) \wedge \forall Q(\exists a Q(a) \wedge \forall x(Q(x) \rightarrow P(x)) \wedge \forall u, v(Q(u) \wedge conn(u, v) \rightarrow Q(v)) \rightarrow \forall x(P(x) \rightarrow Q(x))))$

write

reasoning

read

Symbol system

embedding

extraction

Data

learning

- phenotype
- function
- disease

- genotype
- protein sequence
- expression

Summary

Feigenbaum, 1977

[The domain-specific knowledge] plays a critical role in organizing and constraining search. The theme is that in the knowledge is the power.

- AI models in biology rely on biological background knowledge
  - from $>$ 100 years of experiments and observation
  - needs appropriate "embedding" algorithms
- (symbolically) constrained optimization using background knowledge allows using smaller and smaller cohorts (up to individuals)
  - precision/personalized medicine

Acknowledgements

- Paul Schofield (Cambridge)
- George Gkoutos (Birmingham)

Thank you

Amyloid beta
Protein classified with blood
coagulation.

A Semantic Haiku

generated from the UniProt Knowledgebase

`http://borg.kaust.edu.sa`
robert.hoehndorf@kaust.edu.sa