CSE 427 Computational Biology

Lecture 4
Protein function prediction

Today

- **Protein function and structure**
- Gene Ontology: vocabulary of protein functions
- Protein function prediction

Today: what can we do by finding similar sequence?

- Protein function prediction
  - Find similar sequence
  - Supervised learning (e.g., k-NN) for function prediction

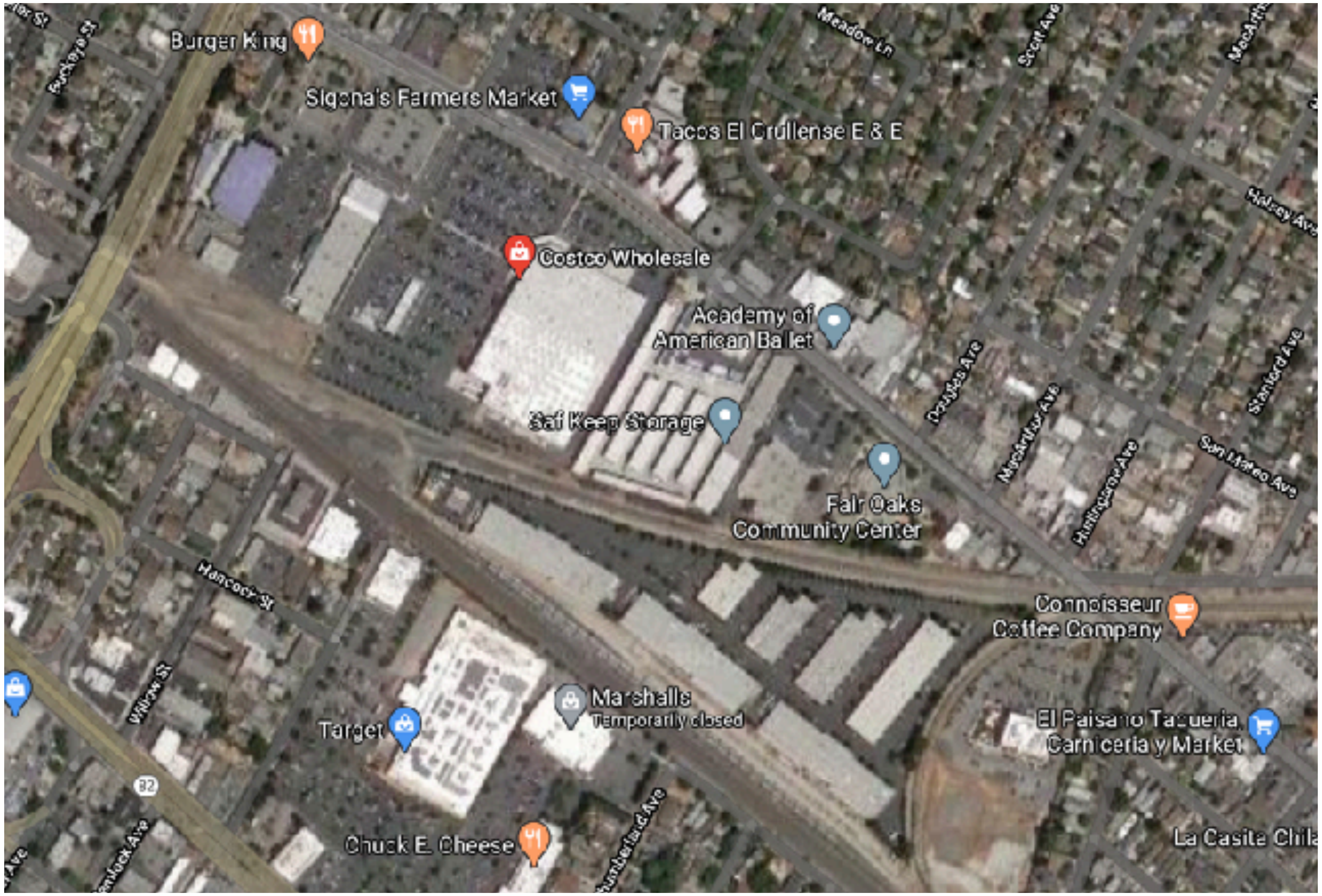
What is protein function prediction?

Human body = country

Single cell = town

Protein = brick, window, carpet, etc.

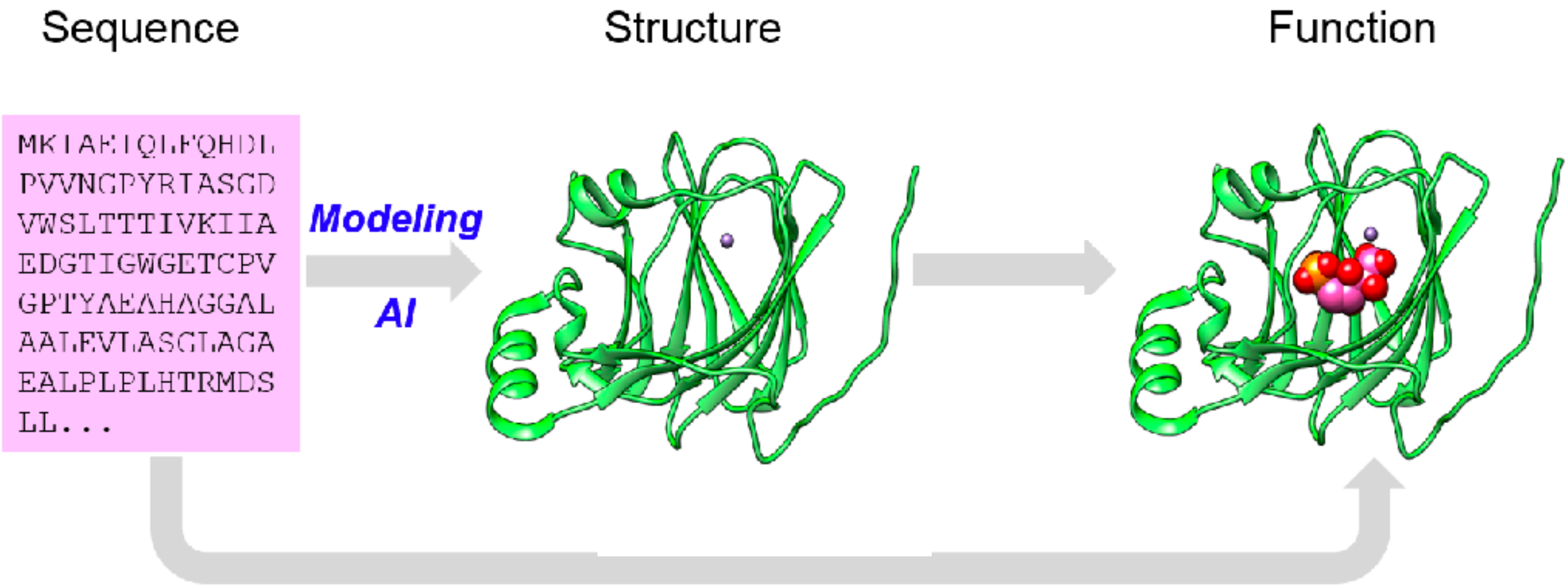
Protein function = fireproof, soundproof, etc.



Goal: classify each protein into its protein functions (multi-label)

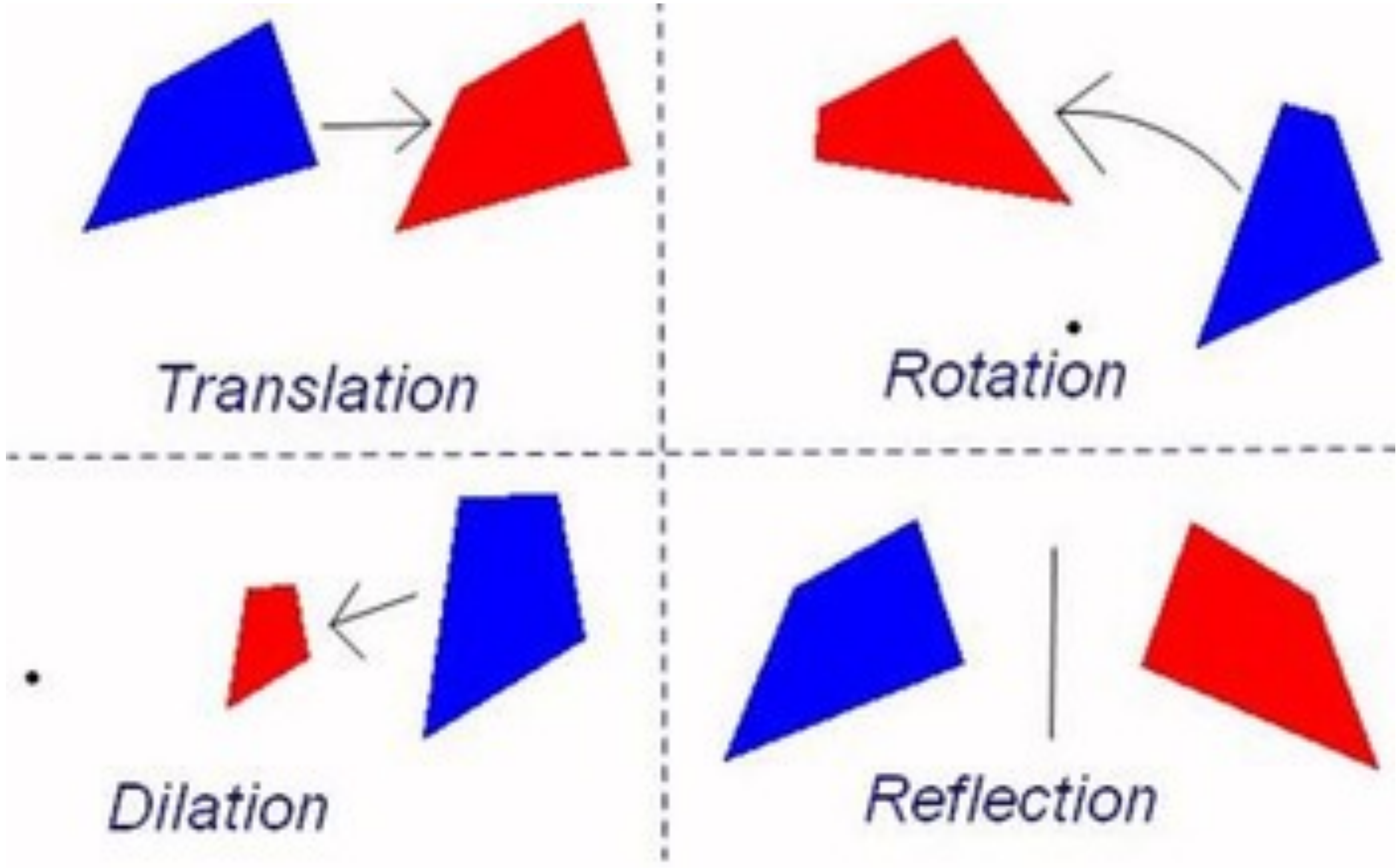
Solution: find proteins with similar sequences

Protein: sequence to structure to function



source: https://zhaolab.shanghaitech.edu.cn/research.html

What we want to maintain?

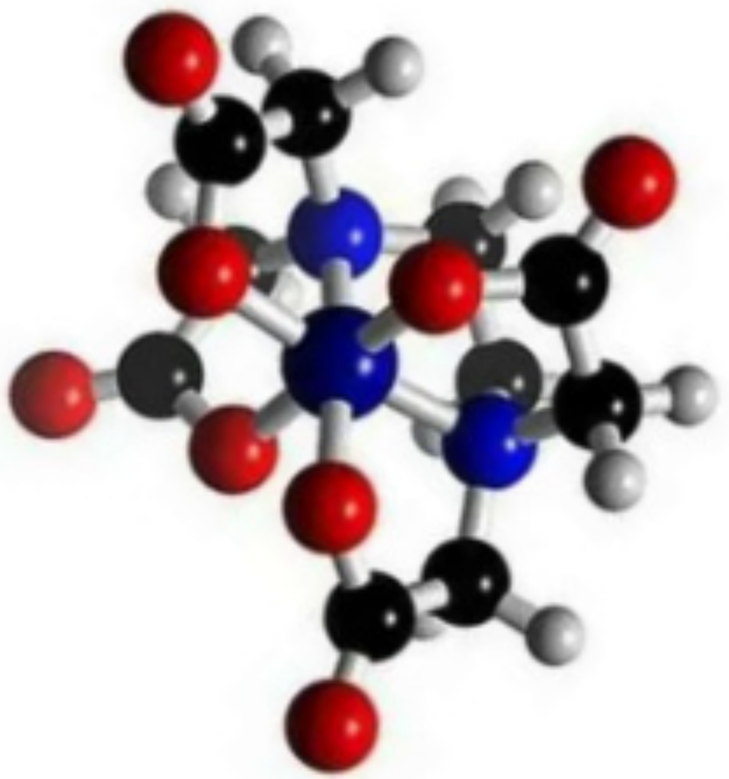


Translation equivariance: X + g
Rotation equivariance: QX
The embedding of the protein remains the same for any g and any Q.

translation, rotation and reflection invariance

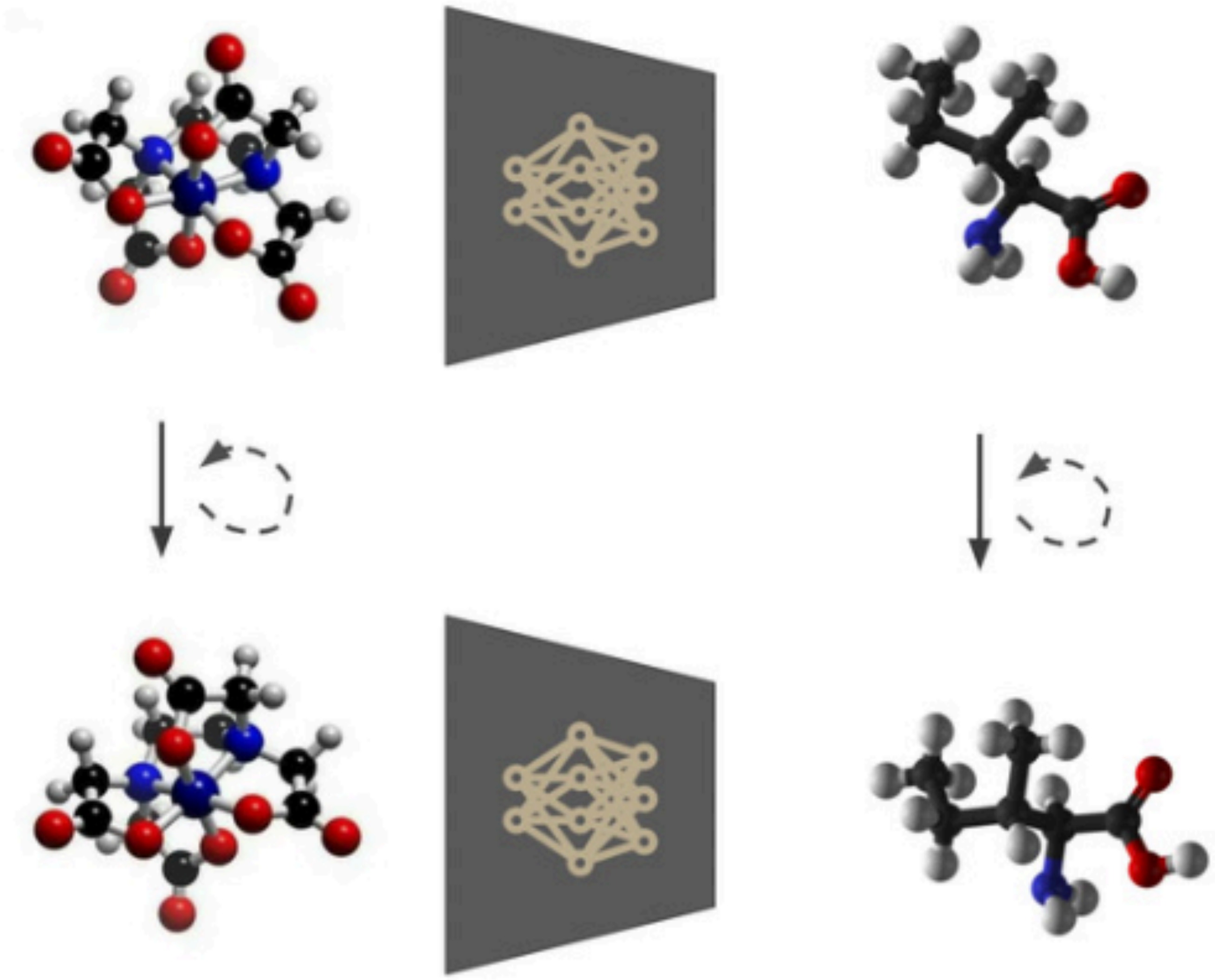
Satorras, Víctor Garcia, Emiel Hoogeboom, and Max Welling. "E (n) equivariant graph neural networks." International Conference on Machine Learning. PMLR, 2021.

translation, rotation and reflection invariance



Satorras, Víctor Garcia, Emiel Hoogeboom, and Max Welling. "E (n) equivariant graph neural networks." International Conference on Machine Learning. PMLR, 2021.

translation, rotation and reflection equivariance



Satorras, Víctor Garcia, Emiel Hoogeboom, and Max Welling. "E (n) equivariant graph neural networks." International Conference on Machine Learning. PMLR, 2021.

Predict the binding site between a protein and a compound

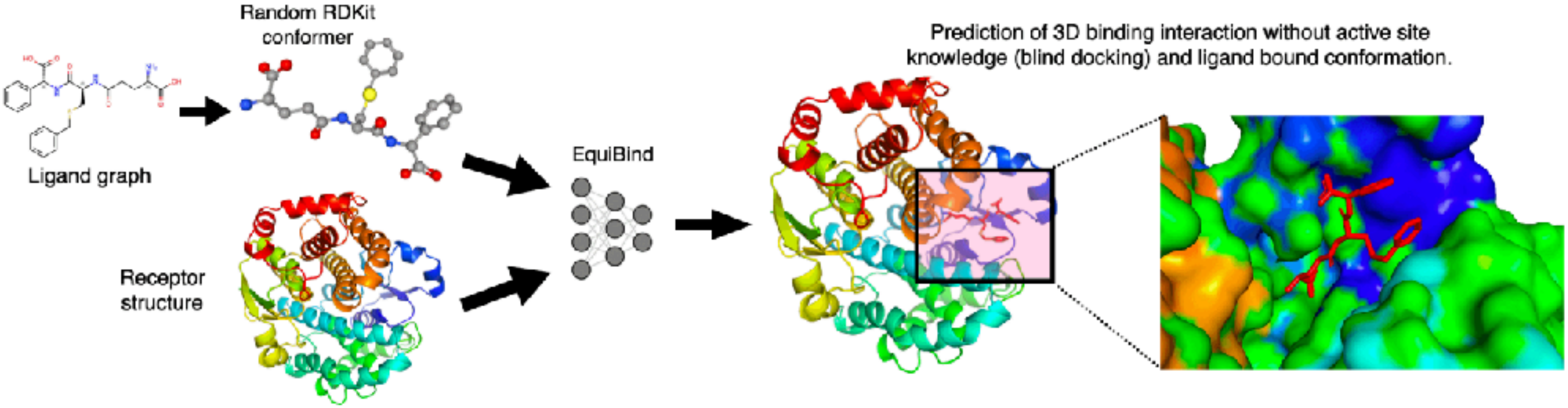
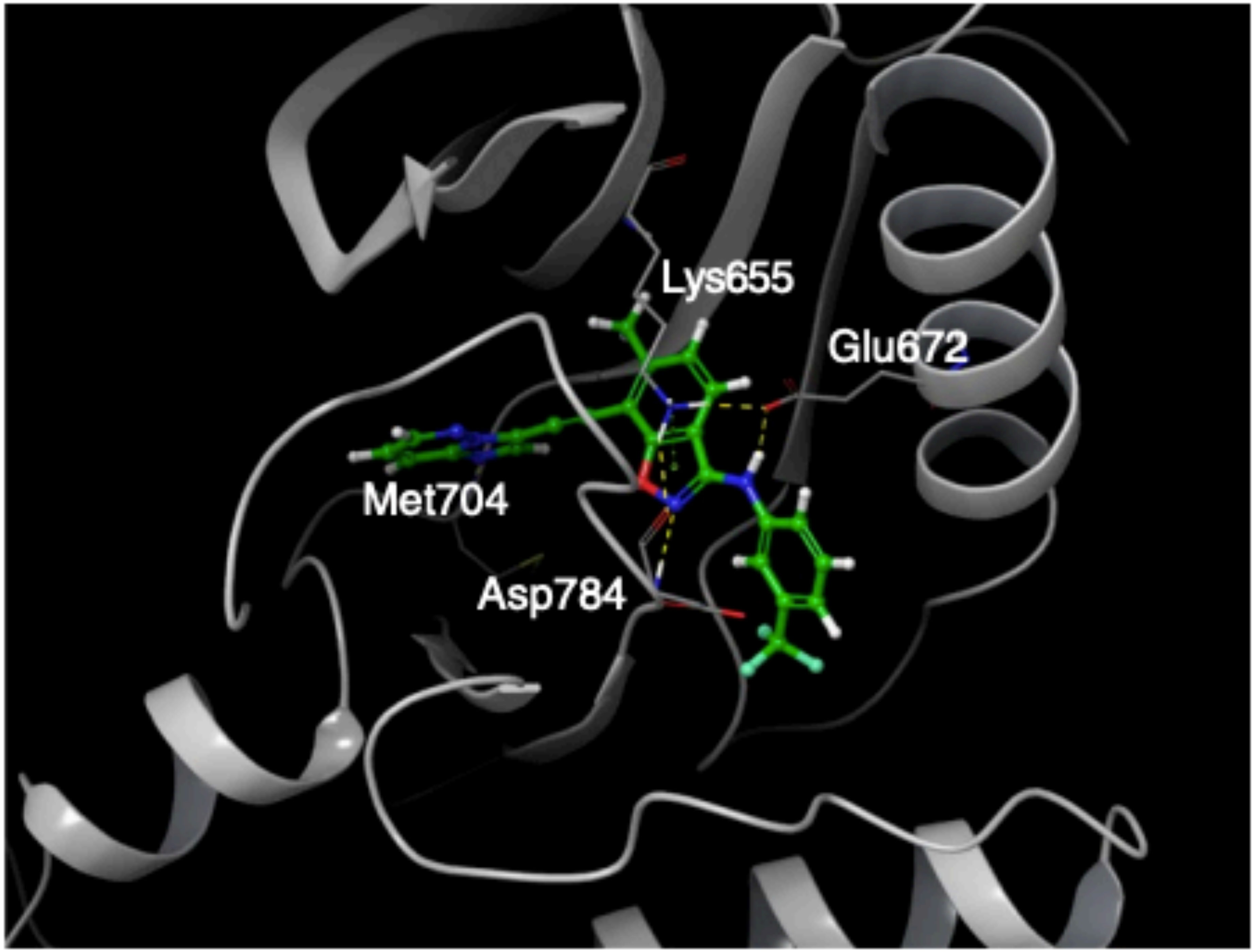

Random RDKit conformer
Ligand graph
Receptor structure
EquiBind
Prediction of 3D binding interaction without active site knowledge (blind docking) and ligand bound conformation.

Figure 1. High-level overview of the structural drug binding problem tackled by EQUIBIND.

Protein structure: drug binding



Today

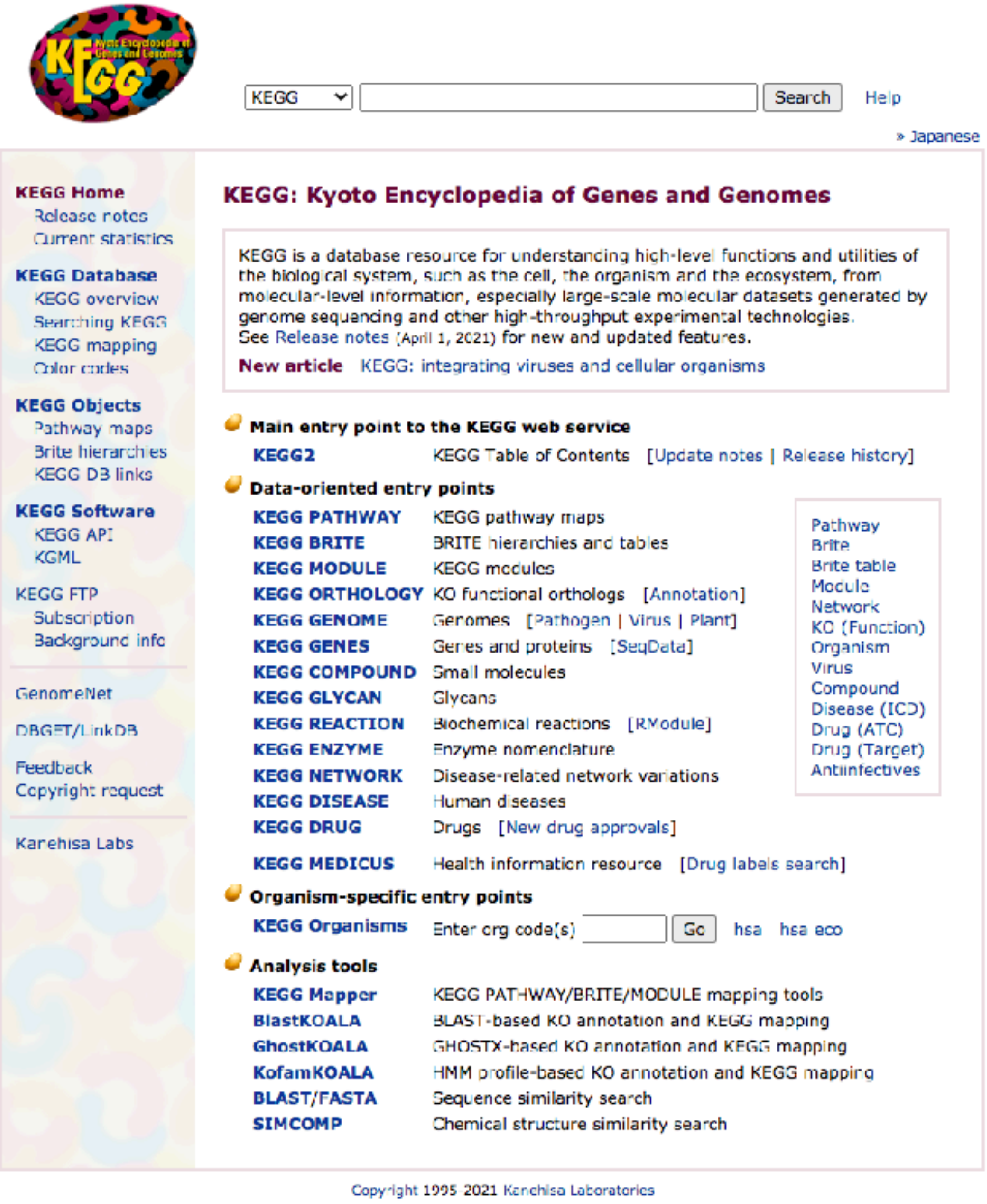
- Protein function and structure
- **Gene Ontology: vocabulary of protein functions**
- Protein function prediction

Three widely used databases for project functions

- Enzyme Commission (EC), Transporter Classification (TC)
- Kyoto Enclyclopedia of Genes and Genomes (KEGG)
- Gene Ontology (GO): molecular function, biological process, and cellular component.
  - More than 30K functions
  - many-to-many relationship between proteins and functions

Three widely used databases for project functions

- Enzyme Commission (EC), Transporter Classification (TC)
- **Kyoto Enclyclopedia of Genes and Genomes (KEGG)**
- Gene Ontology (GO): molecular function, biological process, and cellular component.
  - More than 30K functions
  - many-to-many relationship between proteins and functions



KEGG Home
Release notes
Current statistics

KEGG Database
KEGG overview
Searching KEGG
KEGG mapping
Color codes

KEGG Objects
Pathway maps
Brite hierarchies
KEGG DB links

KEGG Software
KEGG API
KGML

KEGG FTP
Subscription
Background info

GenomeNet

DBGET/LinkDB

Feedback
Copyright request

Kanehisa Labs

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.
See Release notes (April 1, 2021) for new and updated features.

New article KEGG: integrating viruses and cellular organisms

Main entry point to the KEGG web service

KEGG2	KEGG Table of Contents [Update notes | Release history]

Data-oriented entry points

KEGG PATHWAY	KEGG pathway maps
KEGG BRITE	BRITE hierarchies and tables
KEGG MODULE	KEGG modules
KEGG ORTHOLOGY	KO functional orthologs [Annotation]
KEGG GENOME	Genomes [Pathogen | Virus | Plant]
KEGG GENES	Genes and proteins [SeqData]
KEGG COMPOUND	Small molecules
KEGG GLYCAN	Glycans
KEGG REACTION	Biochemical reactions [RModule]
KEGG ENZYME	Enzyme nomenclature
KEGG NETWORK	Disease-related network variations
KEGG DISEASE	Human diseases
KEGG DRUG	Drugs [New drug approvals]
KEGG MEDICUS	Health information resource [Drug labels search]

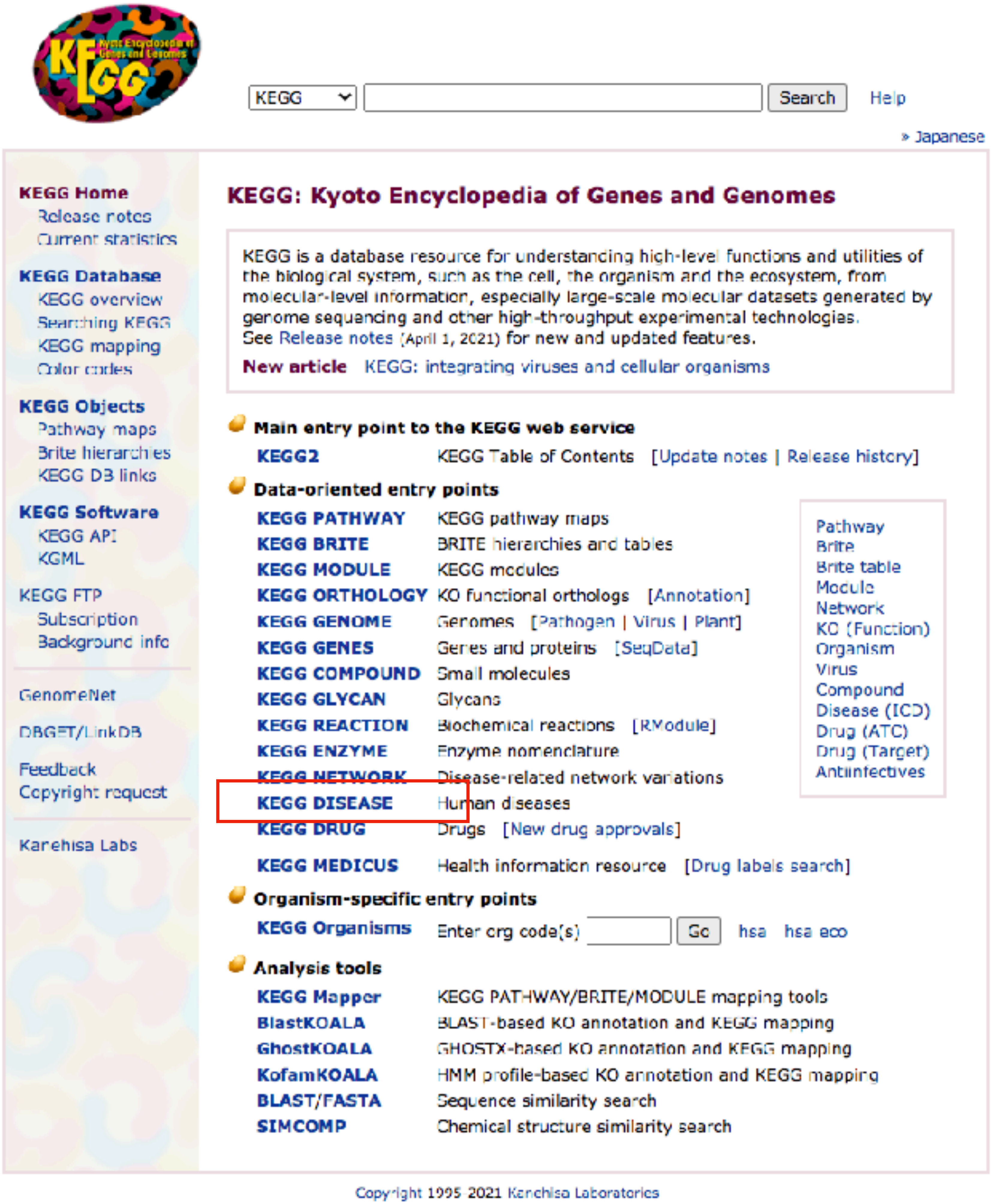
Pathway
Brite
Brite table
Module
Network
KO (Function)
Organism
Virus
Compound
Disease (ICD)
Drug (ATC)
Drug (Target)
Antiinfectives

Organism-specific entry points

KEGG Organisms	Enter org code(s) Go hsa hsa eco

Analysis tools

KEGG Mapper	KEGG PATHWAY/BRITE/MODULE mapping tools
BlastKOALA	BLAST-based KO annotation and KEGG mapping
GhostKOALA	GHOSTX-based KO annotation and KEGG mapping
KofamKOALA	HMM profile-based KO annotation and KEGG mapping
BLAST/FASTA	Sequence similarity search
SIMCOMP	Chemical structure similarity search



KEGG Home
Release notes
Current statistics

KEGG Database
KEGG overview
Searching KEGG
KEGG mapping
Color codes

KEGG Objects
Pathway maps
Brite hierarchies
KEGG DB links

KEGG Software
KEGG API
KGML

KEGG FTP
Subscription
Background info

GenomeNet

DBGET/LinkDB

Feedback
Copyright request

Kanehisa Labs

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies.
See Release notes (April 1, 2021) for new and updated features.

New article KEGG: integrating viruses and cellular organisms

Main entry point to the KEGG web service

KEGG2	KEGG Table of Contents [Update notes | Release history]

Data-oriented entry points

KEGG PATHWAY	KEGG pathway maps
KEGG BRITE	BRITE hierarchies and tables
KEGG MODULE	KEGG modules
KEGG ORTHOLOGY	KO functional orthologs [Annotation]
KEGG GENOME	Genomes [Pathogen | Virus | Plant]
KEGG GENES	Genes and proteins [SeqData]
KEGG COMPOUND	Small molecules
KEGG GLYCAN	Glycans
KEGG REACTION	Biochemical reactions [RModule]
KEGG ENZYME	Enzyme nomenclature
KEGG NETWORK	Disease-related network variations
KEGG DISEASE	Human diseases
KEGG DRUG	Drugs [New drug approvals]
KEGG MEDICUS	Health information resource [Drug labels search]

Pathway
Brite
Brite table
Module
Network
KO (Function)
Organism
Virus
Compound
Disease (ICD)
Drug (ATC)
Drug (Target)
Antiinfectives

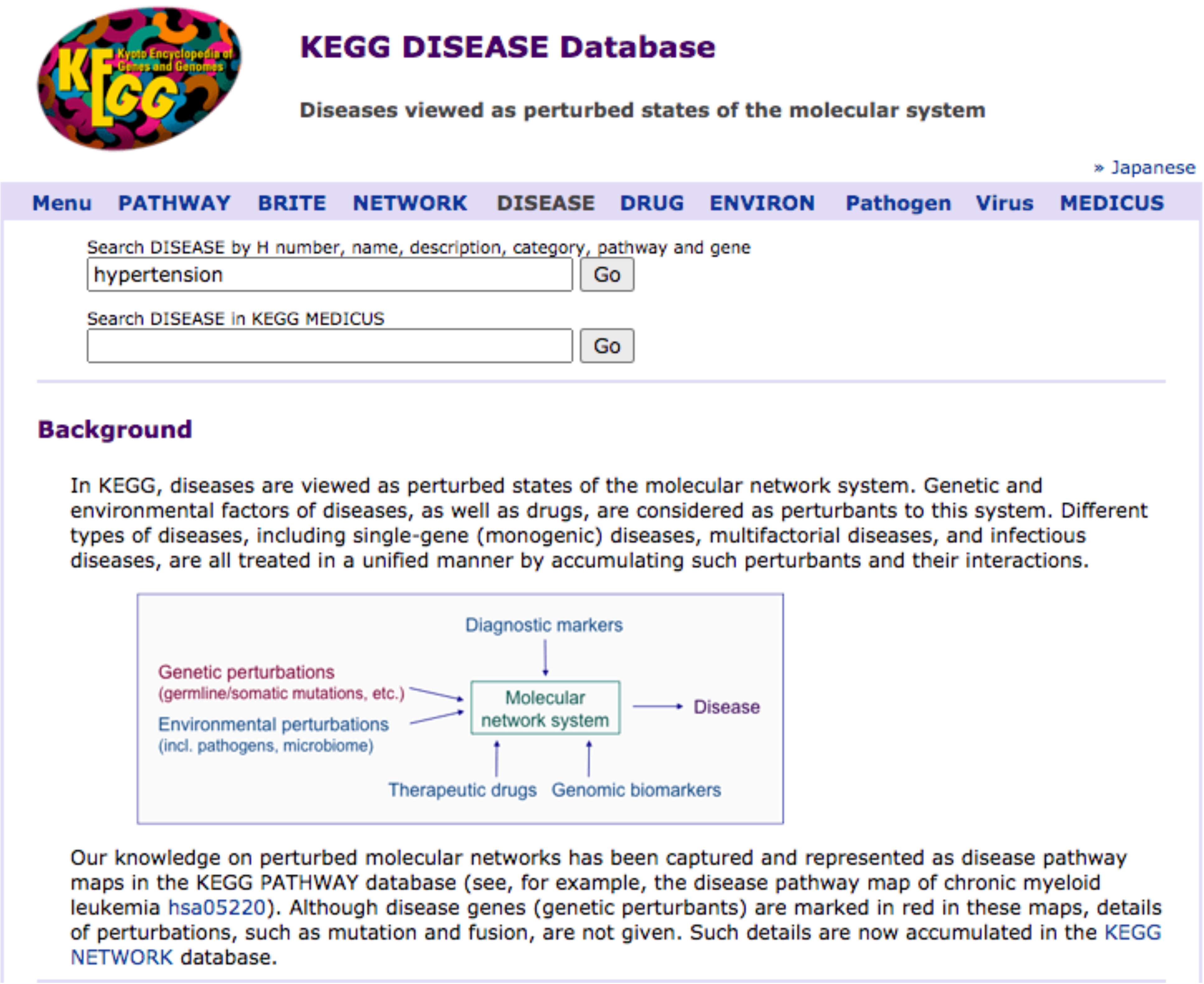
Organism-specific entry points

KEGG Organisms	Enter org code(s) Go hsa hsa eco

Analysis tools

KEGG Mapper	KEGG PATHWAY/BRITE/MODULE mapping tools
BlastKOALA	BLAST-based KO annotation and KEGG mapping
GhostKOALA	GHOSTX-based KO annotation and KEGG mapping
KofamKOALA	HMM profile-based KO annotation and KEGG mapping
BLAST/FASTA	Sequence similarity search
SIMCOMP	Chemical structure similarity search

Copyright 1995-2021 Kanehisa Laboratories



KEGG DISEASE Database

Diseases viewed as perturbed states of the molecular system

Search DISEASE by H number, name, description, category, pathway and gene

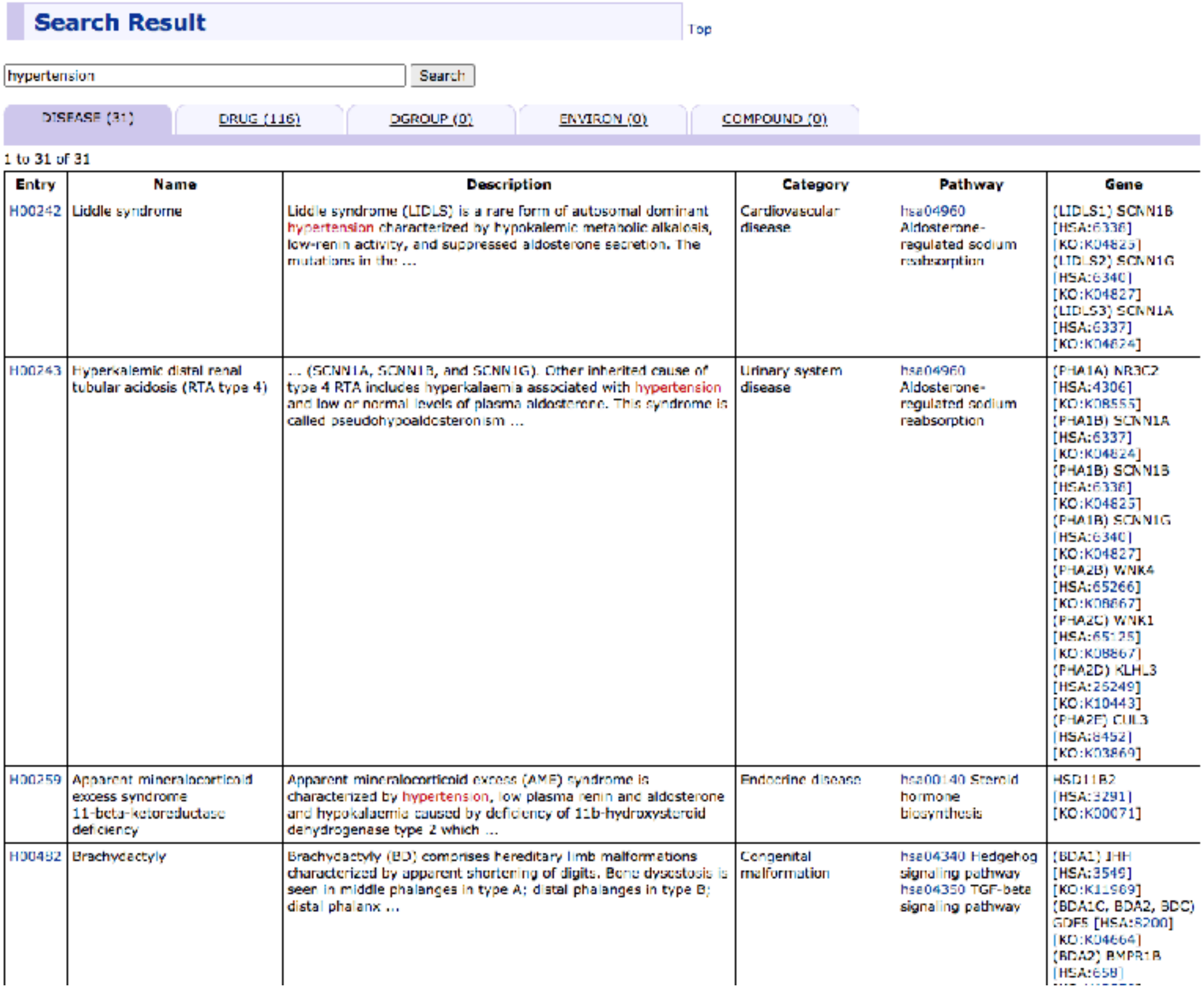
hypertension

Search DISEASE in KEGG MEDICUS

Background

In KEGG, diseases are viewed as perturbed states of the molecular network system. Genetic and environmental factors of diseases, as well as drugs, are considered as perturbants to this system. Different types of diseases, including single-gene (monogenic) diseases, multifactorial diseases, and infectious diseases, are all treated in a unified manner by accumulating such perturbants and their interactions.

Our knowledge on perturbed molecular networks has been captured and represented as disease pathway maps in the KEGG PATHWAY database (see, for example, the disease pathway map of chronic myeloid leukemia hsa05220). Although disease genes (genetic perturbants) are marked in red in these maps, details of perturbations, such as mutation and fusion, are not given. Such details are now accumulated in the KEGG NETWORK database.



Search Result

hypertension

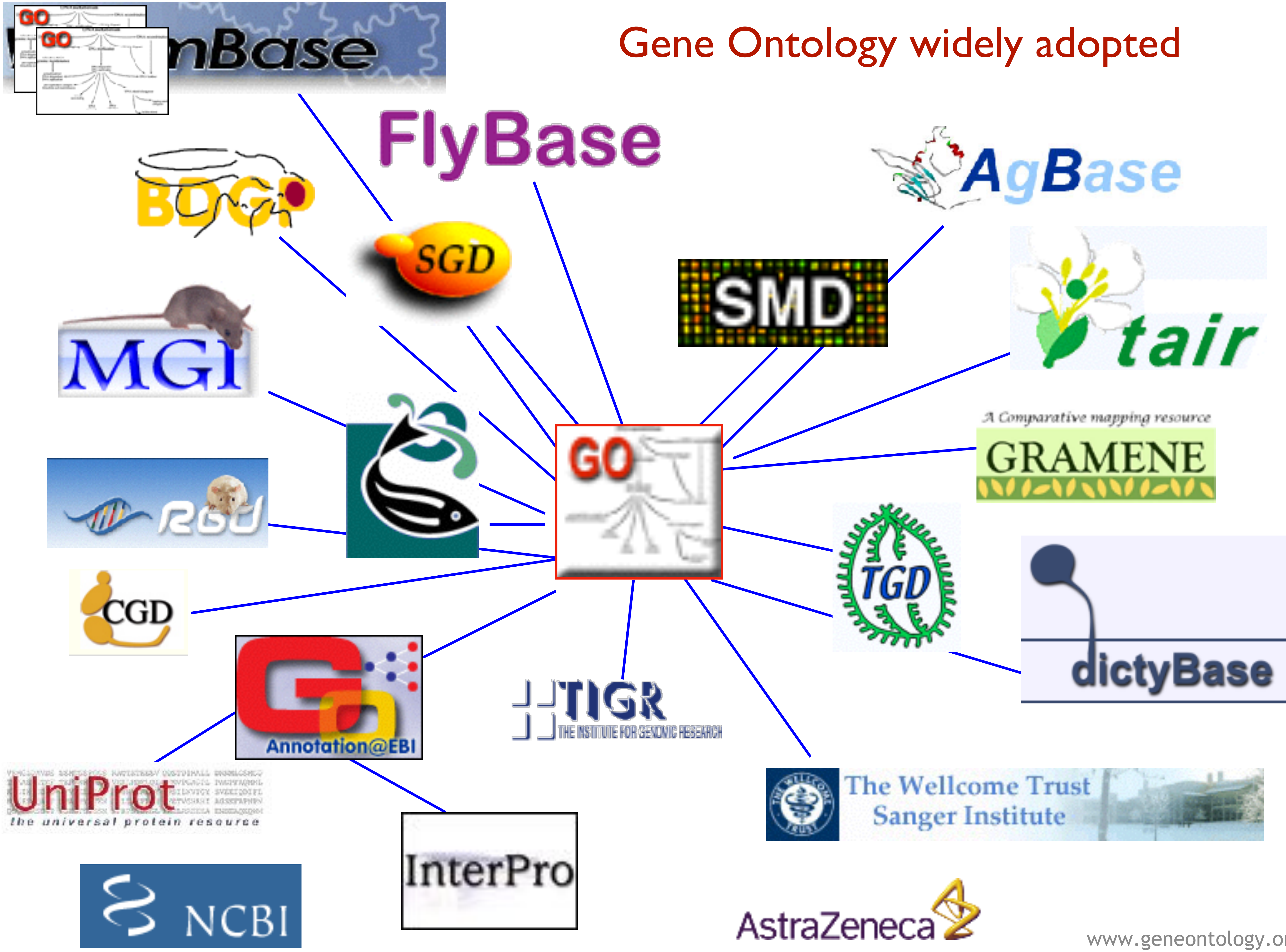
1 to 31 of 31

Entry	Name	Description	Category	Pathway	Gene
H00242	Liddle syndrome	Liddle syndrome (LIDLS) is a rare form of autosomal dominant hypertension characterized by hypokalemic metabolic alkalosis, low-renin activity, and suppressed aldosterone secretion. The mutations in the ...	Cardiovascular disease	hsa04960 Aldosterone-regulated sodium reabsorption	(LIDLS1) SCNN1B [HSA:6338] [KO:K04825] (LIDLS2) SCNN1G [HSA:6340] [KO:K04827] (LIDLS3) SCNN1A [HSA:6337] [KO:K04824]
H00243	Hyperkalemic distal renal tubular acidosis (RTA type 4)	... (SCNN1A, SCNN1B, and SCNN1G). Other inherited cause of type 4 RTA includes hyperkalaemia associated with hypertension and low or normal levels of plasma aldosterone. This syndrome is called pseudohypoaldosteronism ...	Urinary system disease	hsa04960 Aldosterone-regulated sodium reabsorption	(PHA1A) NR3C2 [HSA:4306] [KO:K08555] (PHA1B) SCNN1A [HSA:6337] [KO:K04824] (PHA1B) SCNN1B [HSA:6338] [KO:K04825] (PHA1B) SCNN1G [HSA:6340] [KO:K04827] (PHA2B) WNK4 [HSA:65266] [KO:K08867] (PHA2C) WNK1 [HSA:65125] [KO:K08867] (PHA2D) KLHL3 [HSA:26249] [KO:K10443] (PHA2E) CUL3 [HSA:8452] [KO:K03869]
H00259	Apparent mineralocorticoid excess syndrome 11-beta-ketoreductase deficiency	Apparent mineralocorticoid excess (AME) syndrome is characterized by hypertension, low plasma renin and aldosterone and hypokalaemia caused by deficiency of 11b-hydroxysteroid dehydrogenase type 2 which ...	Endocrine disease	hsa00140 Steroid hormone biosynthesis	HSD11B2 [HSA:3291] [KO:K00071]
H00482	Brachydactyly	Brachydactyly (BD) comprises hereditary limb malformations characterized by apparent shortening of digits. Bone dysostosis is seen in middle phalanges in type A; distal phalanges in type B; distal phalanx ...	Congenital malformation	hsa04340 Hedgehog signaling pathway hsa04350 TGF-beta signaling pathway	(BDA1) IHH [HSA:3549] [KO:K11989] (BDA1C, BDA2, BDC) GDF5 [HSA:8200] [KO:K04664] (BDA2) BMPR1B [HSA:658]

Three widely used databases for project functions

- Enzyme Commission (EC), Transporter Classification (TC)
- Kyoto Enclyclopedia of Genes and Genomes (KEGG)
- **Gene Ontology (GO): molecular function, biological process, and cellular component.**
  - **More than 30K functions**
  - **many-to-many relationship between proteins and functions**

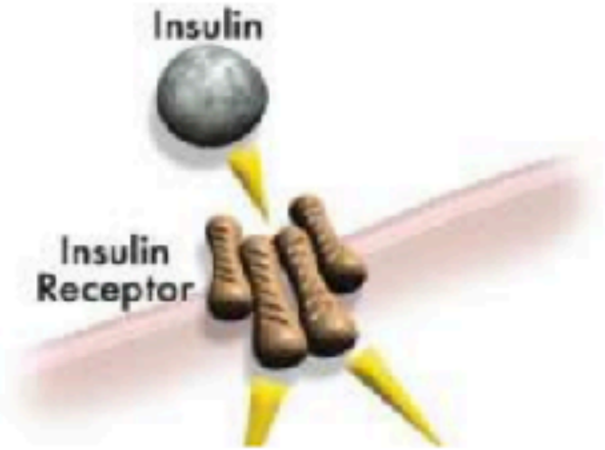
Gene Ontology widely adopted



www.geneontology.org

1. Molecular Function

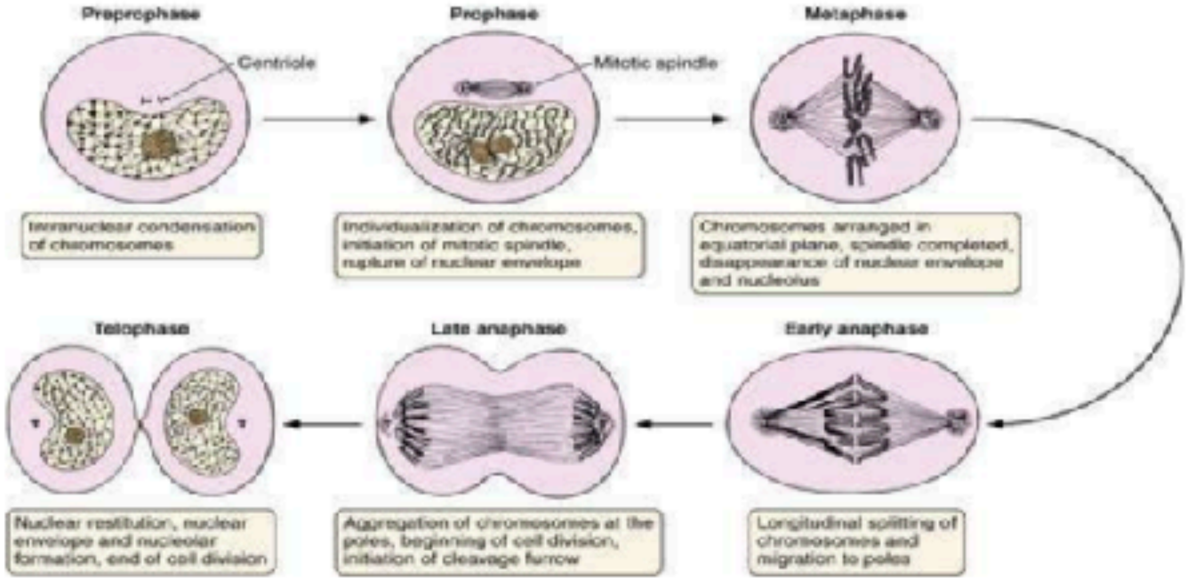
An elemental activity or task or job



- protein kinase activity
- insulin receptor activity

2. Biological Process

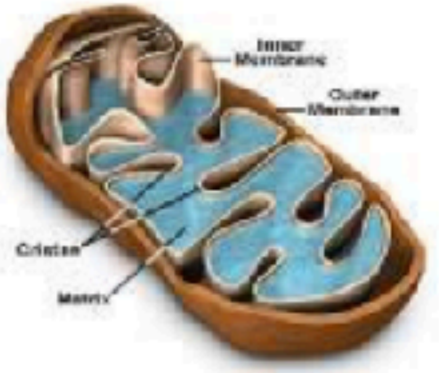
A commonly recognized series of events


Preprophase
Centriole
Prophase
Mitotic spindle
Metaphase
Intranuclear condensation of chromosomes
Individualization of chromosomes, initiation of mitotic spindle, rupture of nuclear envelope
Chromosomes arranged in equatorial plane, spindle completed, disappearance of nuclear envelope and nucleolus
Telophase
Late anaphase
Early anaphase
Nuclear restitution, nuclear envelope and nucleolar formation, end of cell division
Aggregation of chromosomes at the poles, beginning of cell division, initiation of cleavage furrow
Longitudinal splitting of chromosomes and migration to poles

- cell division

3. Cellular Component

Where a gene product is located

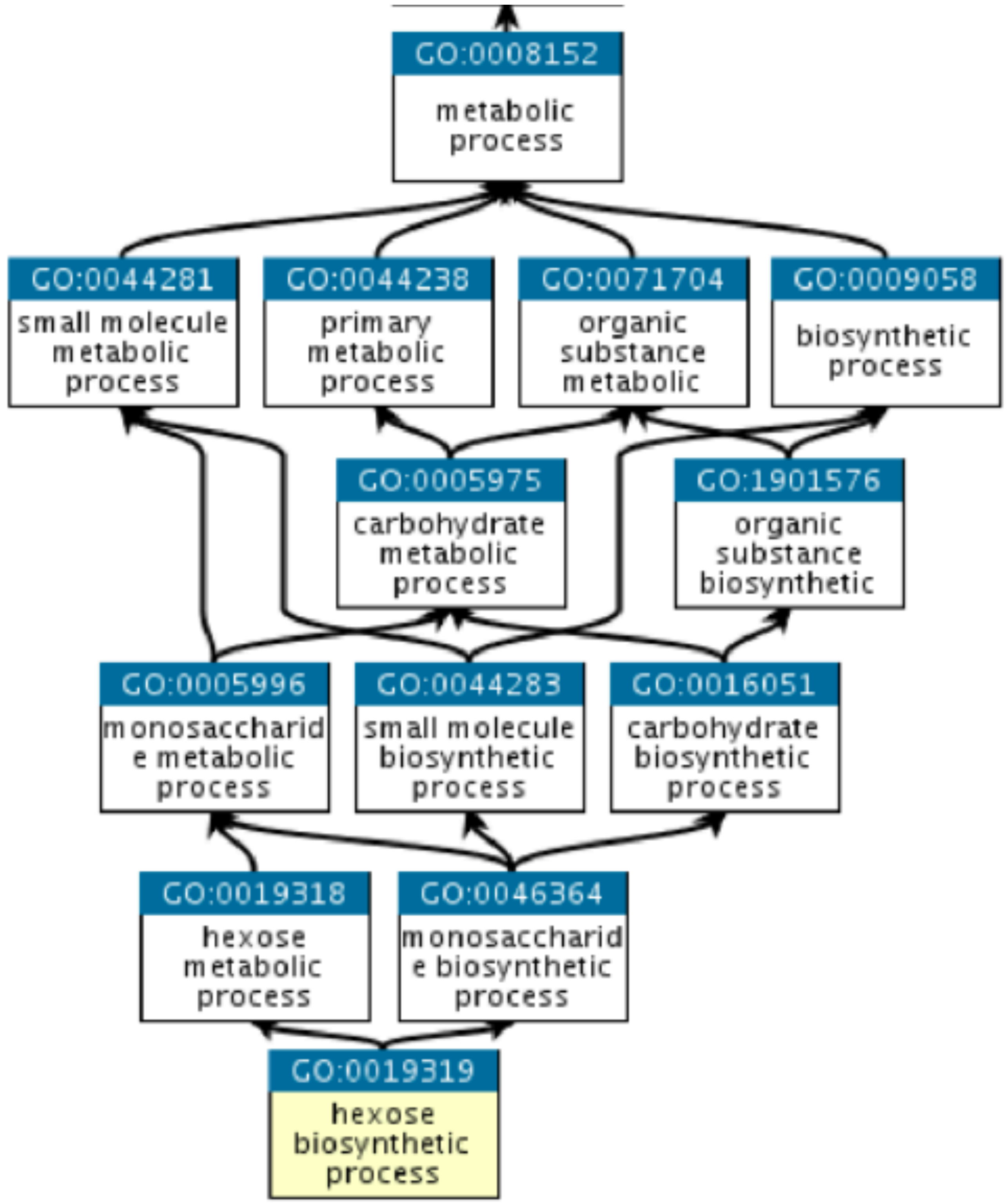


- mitochondrion
- mitochondrial matrix
- mitochondrial inner membrane

Gene Ontology: A directed acyclic graph

Less specific

More specific



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

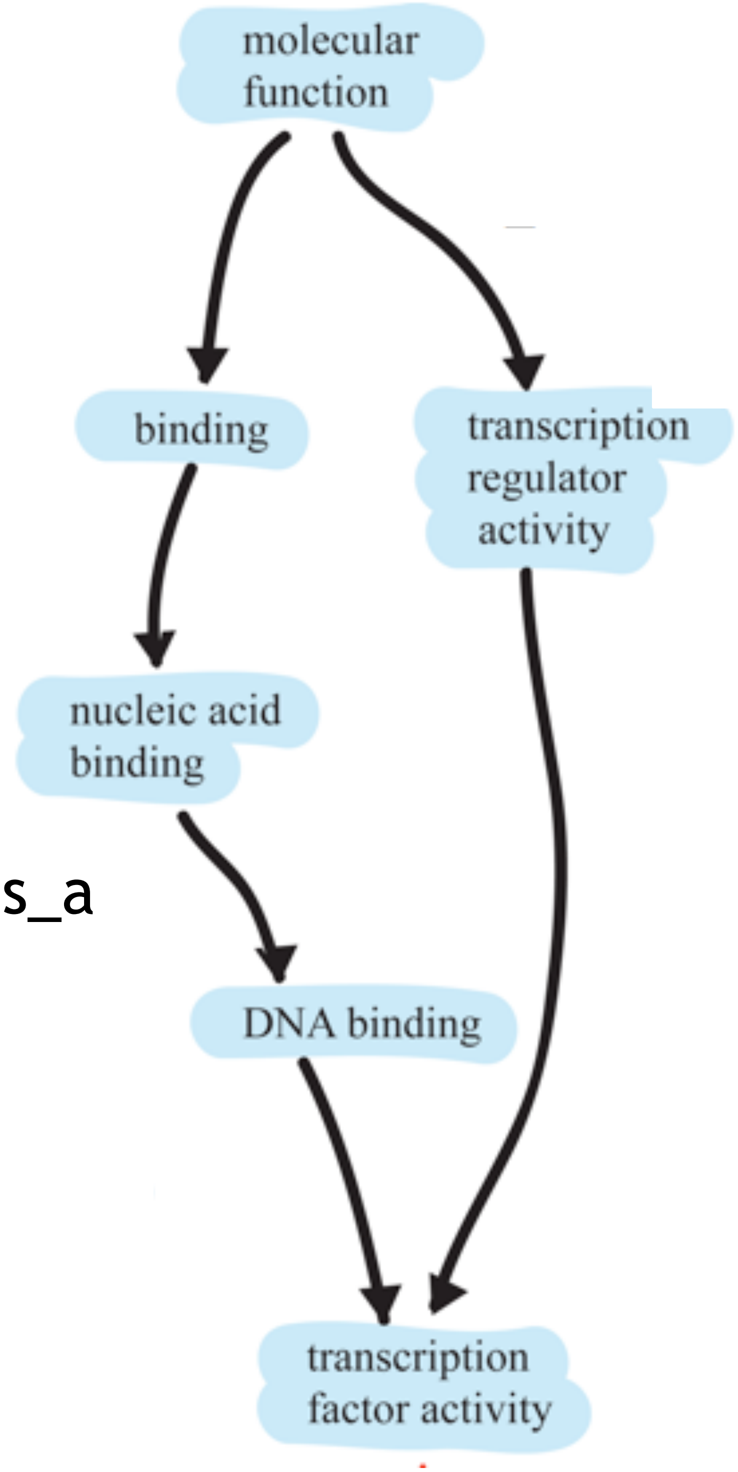
Molecular function ontology

Nucleic acid binding is a type of binding.

is_a

is_a

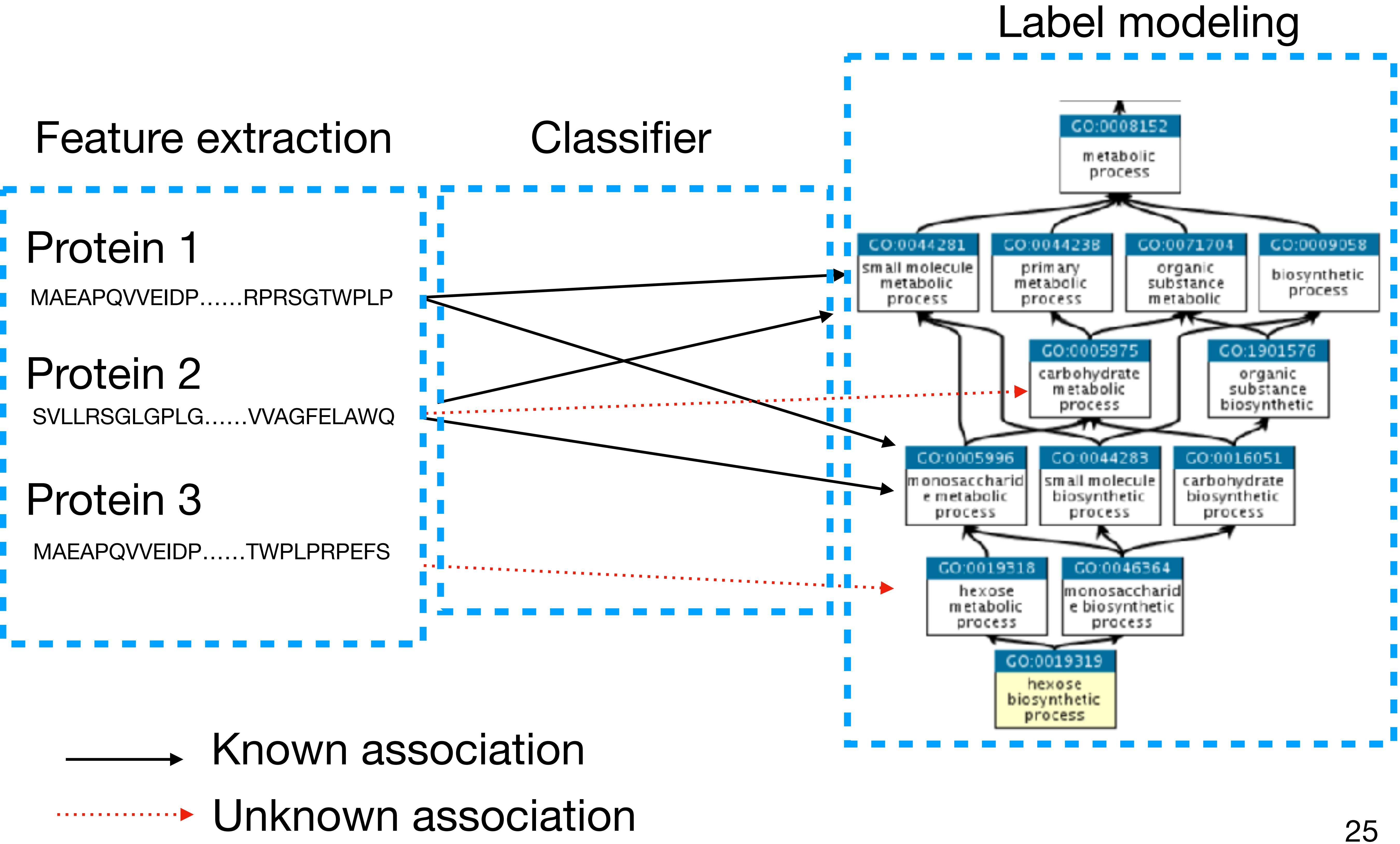
DNA binding is a type of nucleic acid binding.


molecular function
binding
transcription regulator activity
nucleic acid binding
DNA binding
transcription factor activity

Today

- Protein function and structure
- Gene Ontology: vocabulary of protein functions
- **Protein function prediction**

Problem setting for protein function prediction



Label modeling

Feature extraction

Classifier

Protein 1

MAEAPQVVEIDP……RPRSGTWPLP

Protein 2

SVLLRSGLGPLG……VVAGFELAWQ

Protein 3

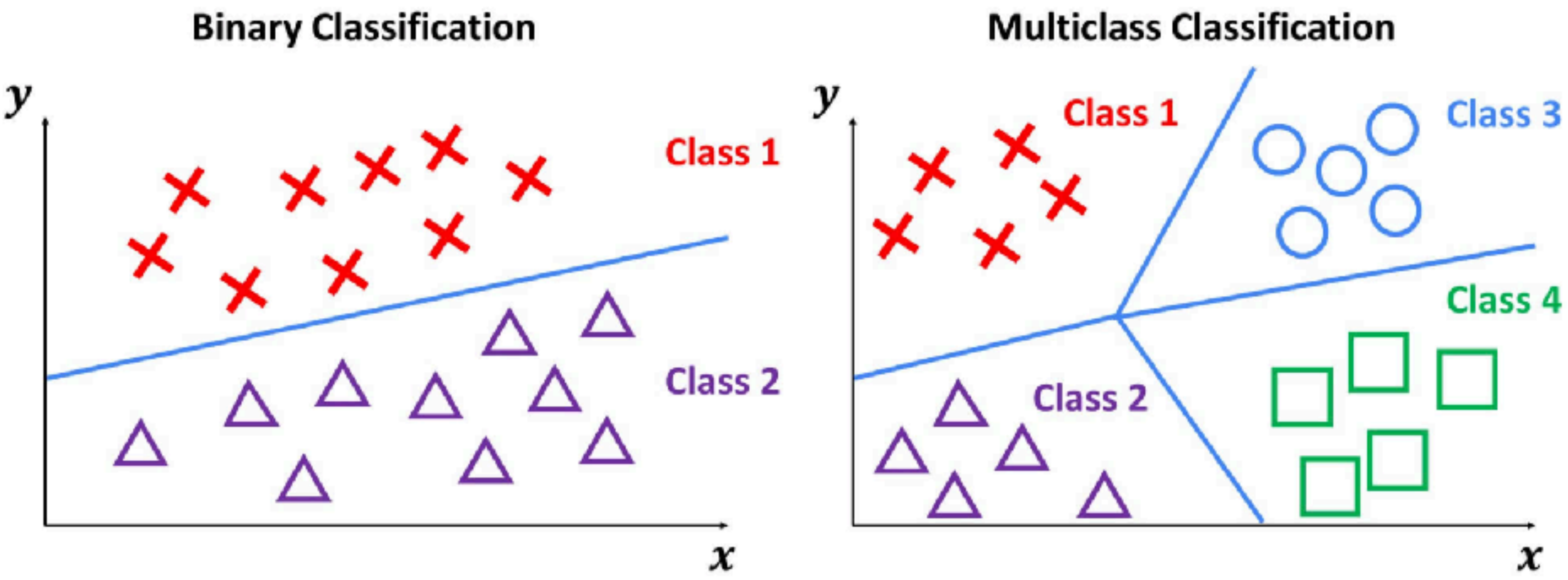
MAEAPQVVEIDP……TWPLPRPEFS

Known association

Unknown association

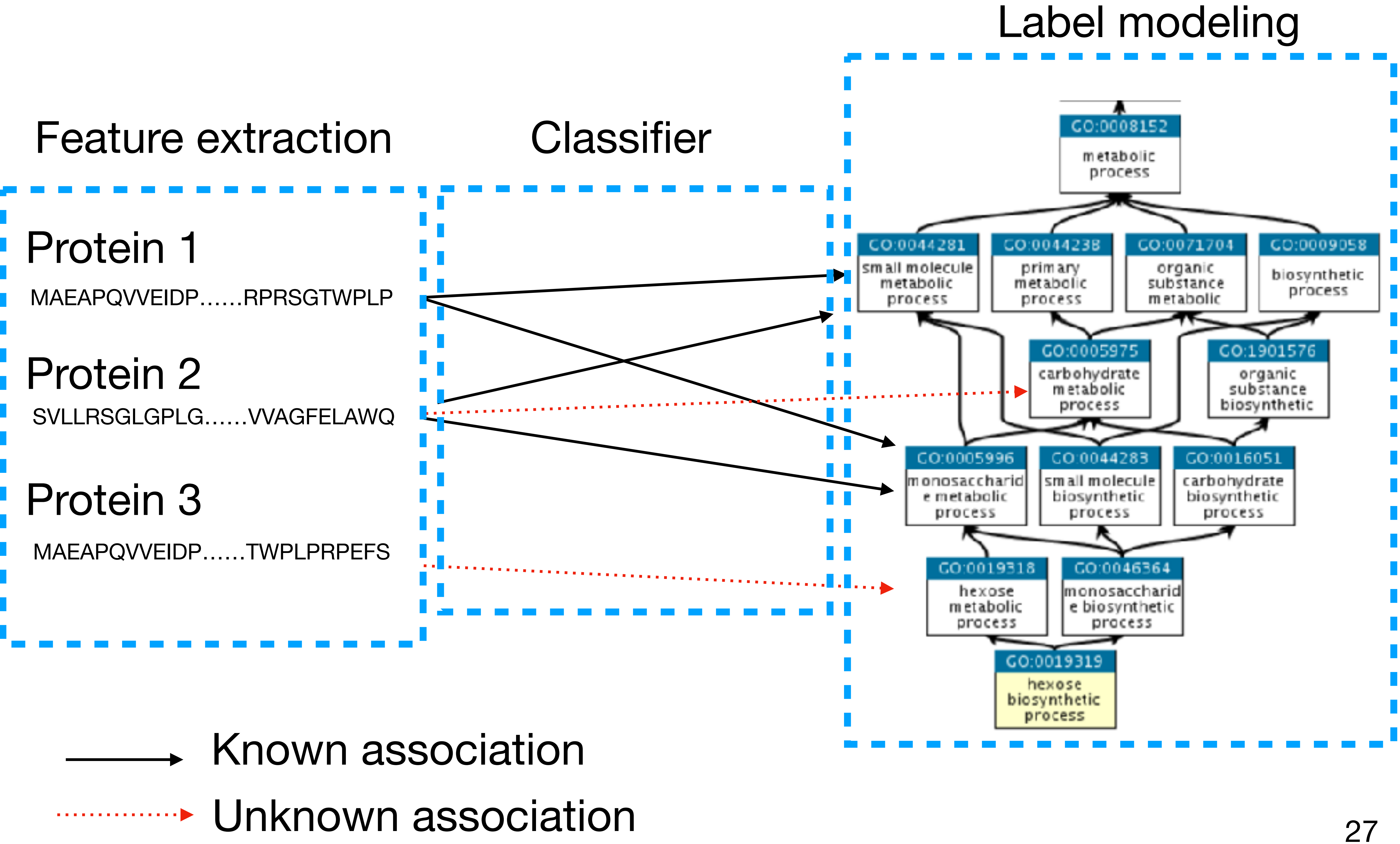
GO:0008152
metabolic process
GO:0044281
small molecule metabolic process
GO:0044238
primary metabolic process
GO:0071704
organic substance metabolic
GO:0009058
biosynthetic process
GO:0005975
carbohydrate metabolic process
GO:1901576
organic substance biosynthetic
GO:0005996
monosaccharide metabolic 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

Introduction to machine learning classification



source: https://datahacker.rs/008-machine-learning-multiclass-classification-and-softmax-function/

Problem setting for protein function prediction



Label modeling

Feature extraction

Classifier

Protein 1

MAEAPQVVEIDP……RPRSGTWPLP

Protein 2

SVLLRSGLGPLG……VVAGFELAWQ

Protein 3

MAEAPQVVEIDP……TWPLPRPEFS

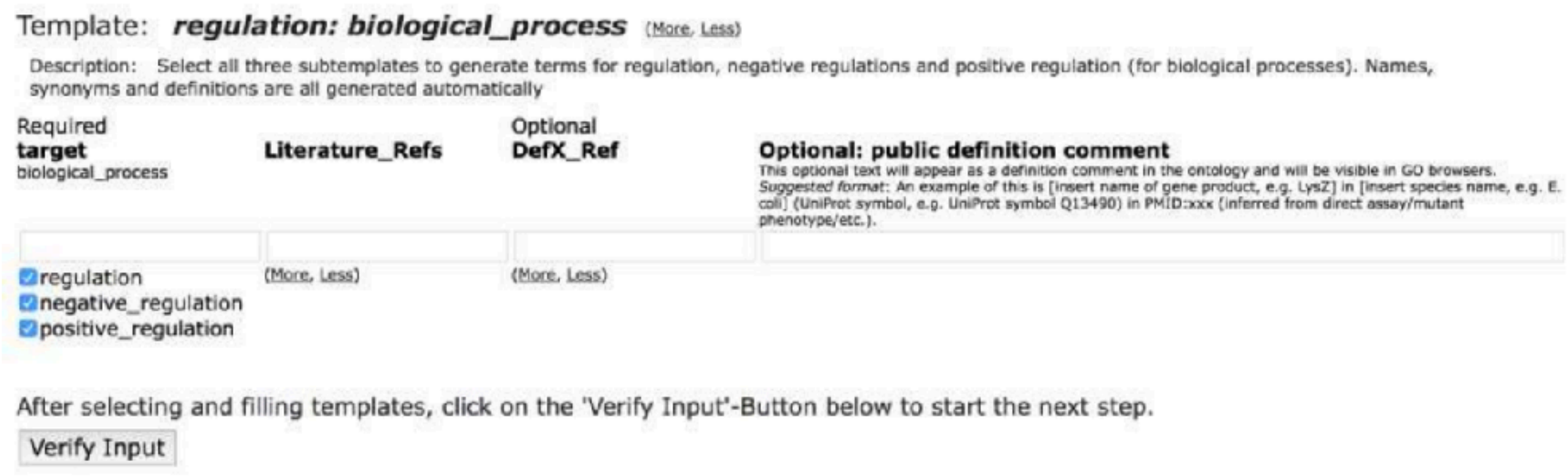
Known association

Unknown association

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

How did we get the known associations (training data)?

- The GO editorial team
- Submission via GitHub, https://github.com/geneontology/
- Submissions via TermGenie, http://go.termgenie.org
    - ~80% terms are now created this way



Template: ***regulation: biological_process*** (More, Less)

Description: Select all three subtemplates to generate terms for regulation, negative regulations and positive regulation (for biological processes). Names, synonyms and definitions are all generated automatically

Required **target** biological_process	**Literature_Refs**	Optional **DefX_Ref**	**Optional: public definition comment** This optional text will appear as a definition comment in the ontology and will be visible in GO browsers. *Suggested format*: An example of this is [insert name of gene product, e.g. LysZ] in [insert species name, e.g. E. coli] (UniProt symbol, e.g. UniProt symbol Q13490) in PMID:xxx (inferred from direct assay/mutant phenotype/etc.).
☑ regulation ☑ negative_regulation ☑ positive_regulation	(More, Less)	(More, Less)	

After selecting and filling templates, click on the 'Verify Input'-Button below to start the next step.

source: EMBL-EBI industry workshop 2016

The Gene Ontology is like a dictionary

Each concept has:

- Name
- Definition
- ID
- Parent nodes

Term: transcription initiation

ID: GO:0006352

Definition: Processes involved in the assembly of the RNA polymerase complex at the promoter region of a DNA template resulting in the subsequent synthesis of RNA from that promoter.

Parent nodes: GO:0002221, is-a

A GO annotation is …

…a statement that a gene product;

Accession	Name	GO ID	GO term name	Reference	Evidence code
P00505	GOT2	GO:0004069	aspartate transaminase activity	PMID:2731362	IDA

A GO annotation is …

…a statement that a gene product;

1. has a particular **molecular function**
or is involved in a particular **biological process**
or is located within a certain **cellular component**

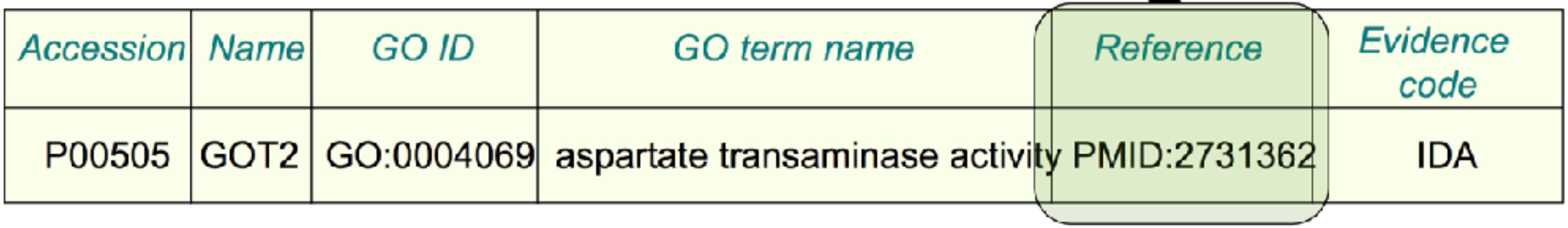
Accession	*Name*	*GO ID*	*GO term name*	*Reference*	*Evidence code*
P00505	GOT2	GO:0004069	aspartate transaminase activity	PMID:2731362	IDA

NLP could be very helpful here!

…a statement that a gene product;

1. has a particular **molecular function**
or is involved in a particular **biological process**
or is located within a certain **cellular component**

2. as described in a particular reference



Accession	*Name*	*GO ID*	*GO term name*	*Reference*	*Evidence code*
P00505	GOT2	GO:0004069	aspartate transaminase activity	PMID:2731362	IDA

Manual annotation: high-quality labelled data, key for ML

- Time-consuming process producing lower numbers of annotations (~2,800 taxons covered)
- More specific GO terms
- Manual annotation is essential for creating predictions

Aleksandra Shypitsyna

Elena Speretta

Alex Holmes

Tony Sawford

Electronic annotation

- Quick way of producing large numbers of annotations
- Annotations use less-specific GO terms
- Only source of annotation for ~438,000 non-model organism species

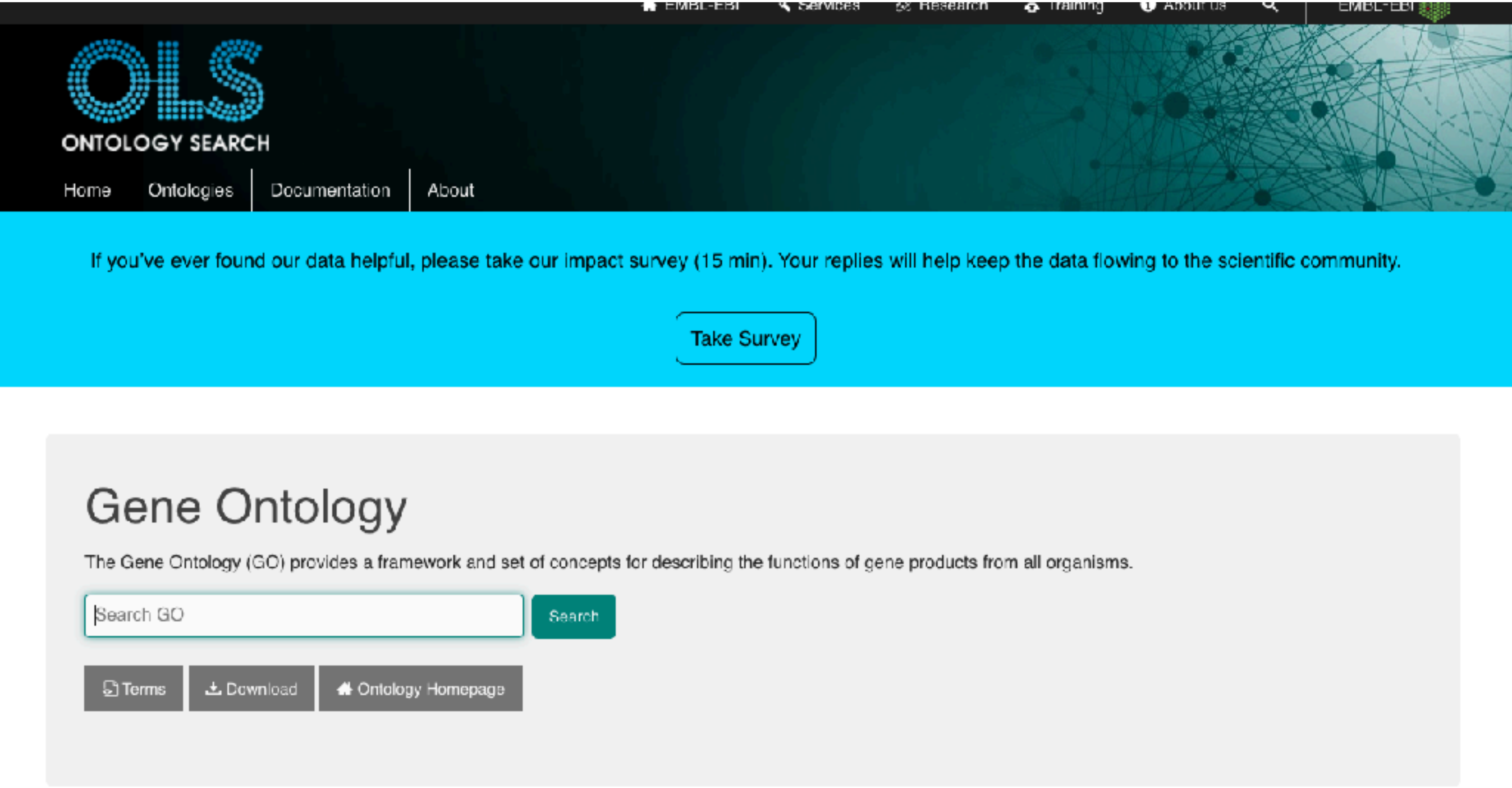
orthology

InterPro2GO

Protein sequence analysis & classification

taxon constraints

Let’s take a look at this database

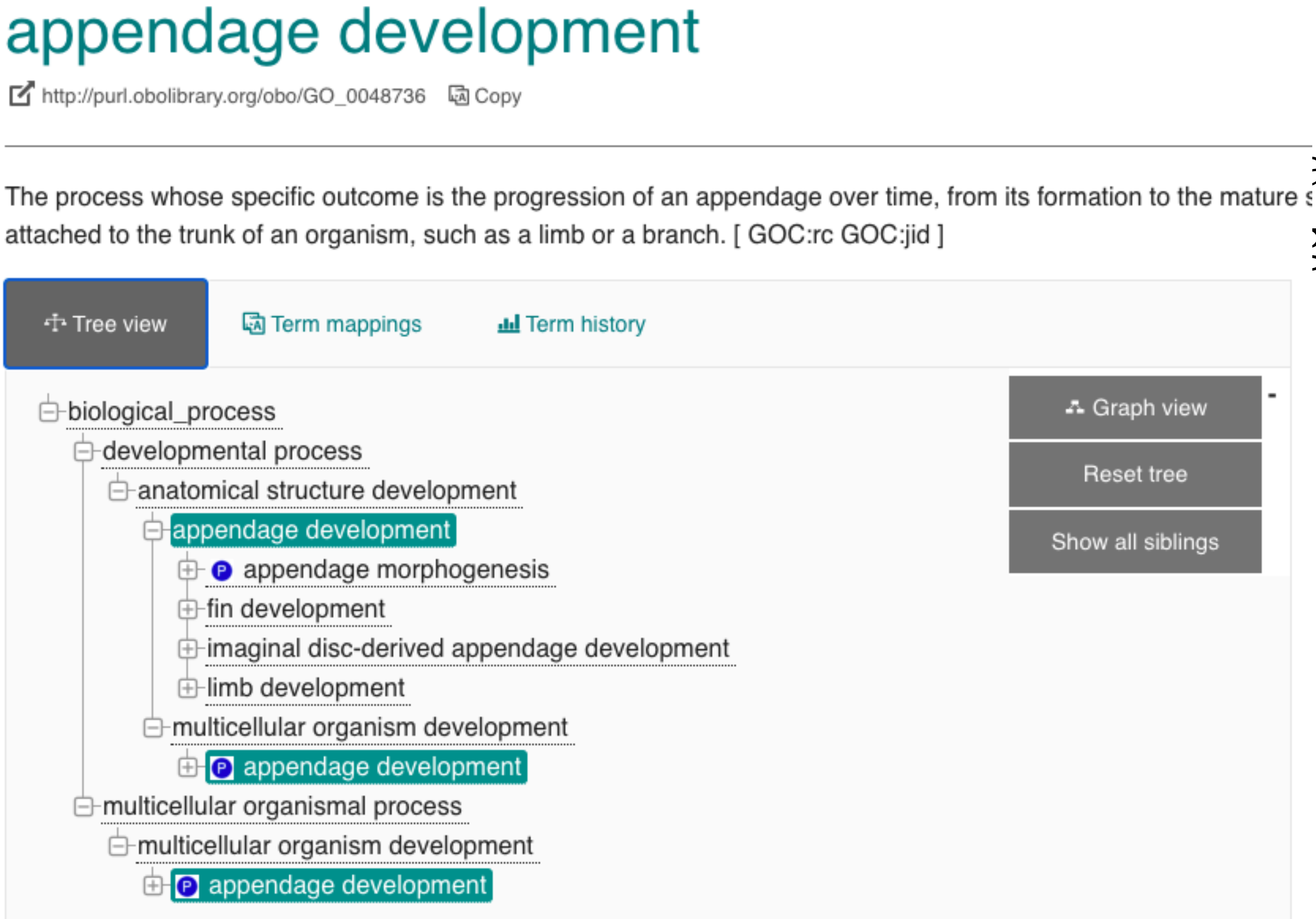


If you’ve ever found our data helpful, please take our impact survey (15 min). Your replies will help keep the data flowing to the scientific community.

Gene Ontology

The Gene Ontology (GO) provides a framework and set of concepts for describing the functions of gene products from all organisms.

https://www.ebi.ac.uk/ols/ontologies/go



appendage development

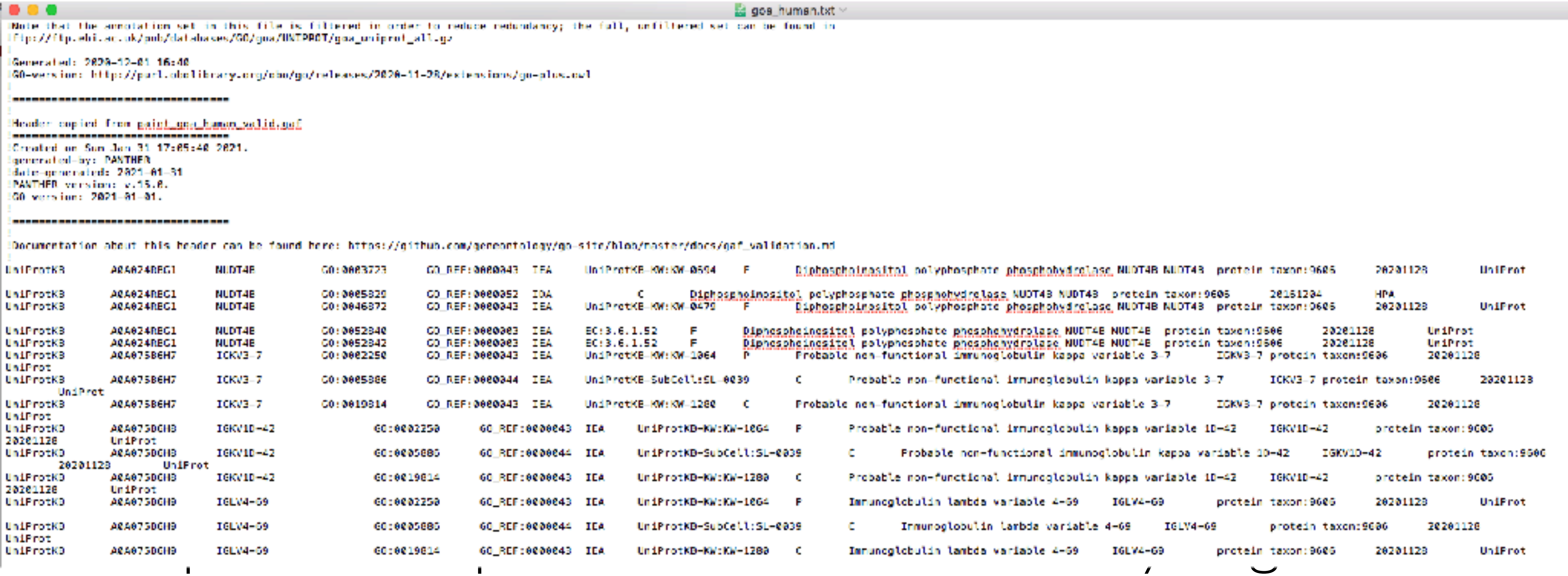
http://purl.obolibrary.org/obo/GO_0048736 Copy

The process whose specific outcome is the progression of an appendage over time, from its formation to the mature s
attached to the trunk of an organism, such as a limb or a branch. [GOC:rc GOC:jid]

- biological_process
  - developmental process
    - anatomical structure development
      - appendage development
        - P appendage morphogenesis
        - fin development
        - imaginal disc-derived appendage development
        - limb development
    - multicellular organism development
      - P appendage development
  - multicellular organismal process
    - multicellular organism development
      - P appendage development

at
ge

http://current.geneontology.org/products/pages/downloads.html



```
!Note that the annotation set in this file is filtered in order to reduce redundancy; the full, unfiltered set can be found in
!ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/goa_uniprot_all.gz
!
!Generated: 2020-12-01 16:40
!GO-version: http://purl.obolibrary.org/obo/go/releases/2020-11-28/extensions/go-plus.owl
!
!===================================
!
!Header copied from paint_goa_human_valid.gaf
!===================================
!Created on Sun Jan 31 17:05:40 2021.
!generated-by: PANTHER
!date-generated: 2021-01-31
!PANTHER version: v.15.0.
!GO version: 2021-01-01.
!
!===================================
!
!Documentation about this header can be found here: https://github.com/geneontology/go-site/blob/master/docs/gaf_validation.md
!
UniProtKB       A0A024RBG1      NUDT4B          GO:0003723      GO_REF:0000043  IEA     UniProtKB-KW:KW-0694    F       Diphosphoinositol polyphosphate phosphohydrolase NUDT4B NUDT4B  protein taxon:9606      20201128        UniProt

UniProtKB       A0A024RBG1      NUDT4B          GO:0005829      GO_REF:0000052  IDA             C       Diphosphoinositol polyphosphate phosphohydrolase NUDT4B NUDT4B  protein taxon:9606      20161204        HPA
UniProtKB       A0A024RBG1      NUDT4B          GO:0046872      GO_REF:0000043  IEA     UniProtKB-KW:KW-0479    F       Diphosphoinositol polyphosphate phosphohydrolase NUDT4B NUDT4B  protein taxon:9606      20201128        UniProt

UniProtKB       A0A024RBG1      NUDT4B          GO:0052840      GO_REF:0000003  IEA     EC:3.6.1.52     F       Diphosphoinositol polyphosphate phosphohydrolase NUDT4B NUDT4B  protein taxon:9606      20201128        UniProt
UniProtKB       A0A024RBG1      NUDT4B          GO:0052842      GO_REF:0000003  IEA     EC:3.6.1.52     F       Diphosphoinositol polyphosphate phosphohydrolase NUDT4B NUDT4B  protein taxon:9606      20201128        UniProt
UniProtKB       A0A075B6H7      IGKV3-7         GO:0002250      GO_REF:0000043  IEA     UniProtKB-KW:KW-1064    P       Probable non-functional immunoglobulin kappa variable 3-7       IGKV3-7 protein taxon:9606      20201128
UniProt
UniProtKB       A0A075B6H7      IGKV3-7         GO:0005886      GO_REF:0000044  IEA     UniProtKB-SubCell:SL-0039       C       Probable non-functional immunoglobulin kappa variable 3-7       IGKV3-7 protein taxon:9606      20201128
        UniProt
UniProtKB       A0A075B6H7      IGKV3-7         GO:0019814      GO_REF:0000043  IEA     UniProtKB-KW:KW-1280    C       Probable non-functional immunoglobulin kappa variable 3-7       IGKV3-7 protein taxon:9606      20201128
UniProt
UniProtKB       A0A075B6H8      IGKV1D-42               GO:0002250      GO_REF:0000043  IEA     UniProtKB-KW:KW-1064    F       Probable non-functional immunoglobulin kappa variable 1D-42     IGKV1D-42       protein taxon:9606
20201128        UniProt
UniProtKB       A0A075B6H8      IGKV1D-42               GO:0005886      GO_REF:0000044  IEA     UniProtKB-SubCell:SL-0039       C       Probable non-functional immunoglobulin kappa variable 1D-42     IGKV1D-42       protein taxon:9606
        20201128        UniProt
UniProtKB       A0A075B6H8      IGKV1D-42               GO:0019814      GO_REF:0000043  IEA     UniProtKB-KW:KW-1280    C       Probable non-functional immunoglobulin kappa variable 1D-42     IGKV1D-42       protein taxon:9606
20201128        UniProt
UniProtKB       A0A075B6H9      IGLV4-69                GO:0002250      GO_REF:0000043  IEA     UniProtKB-KW:KW-1064    F       Immunoglobulin lambda variable 4-69     IGLV4-69        protein taxon:9606      20201128        UniProt

UniProtKB       A0A075B6H9      IGLV4-69                GO:0005886      GO_REF:0000044  IEA     UniProtKB-SubCell:SL-0039       C       Immunoglobulin lambda variable 4-69     IGLV4-69        protein taxon:9606      20201128
UniProt
UniProtKB       A0A075B6H9      IGLV4-69                GO:0019814      GO_REF:0000043  IEA     UniProtKB-KW:KW-1280    C       Immunoglobulin lambda variable 4-69     IGLV4-69        protein taxon:9606      20201128        UniProt
```

- Enable transferring knowledge across species

A good dataset for ML

Number of annotations in UniProt-GOA database (March 2016)

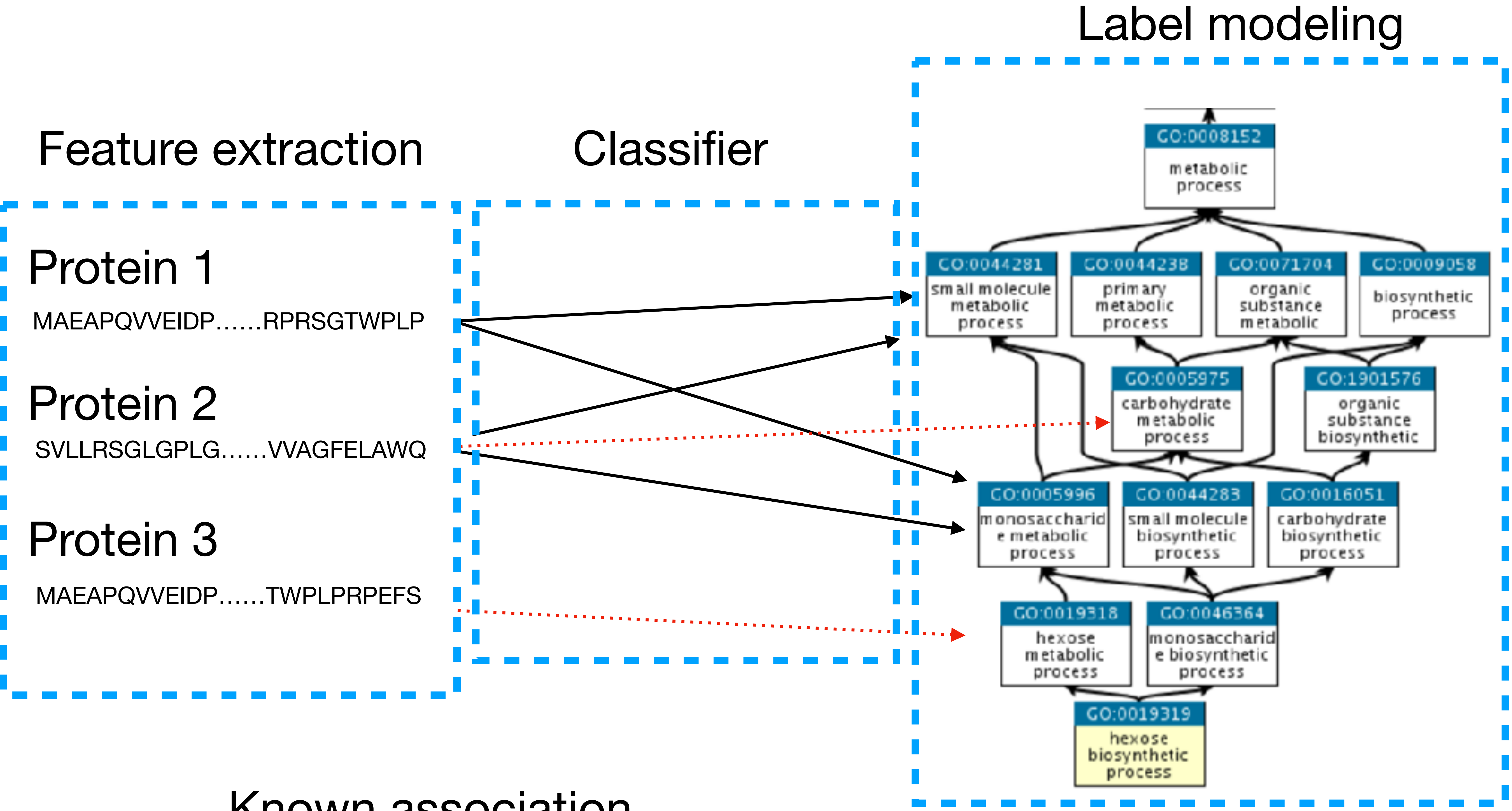
Electronic annotations	269,207,317
Manual annotations*	2,752,604

* Includes manual annotations integrated from external model organism and specialist groups

https://www.ebi.ac.uk/QuickGO/

http://www.ebi.ac.uk/GOA

Problem setting for protein function prediction



Label modeling

Feature extraction

Classifier

Protein 1

MAEAPQVVEIDP……RPRSGTWPLP

Protein 2

SVLLRSGLGPLG……VVAGFELAWQ

Protein 3

MAEAPQVVEIDP……TWPLPRPEFS

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

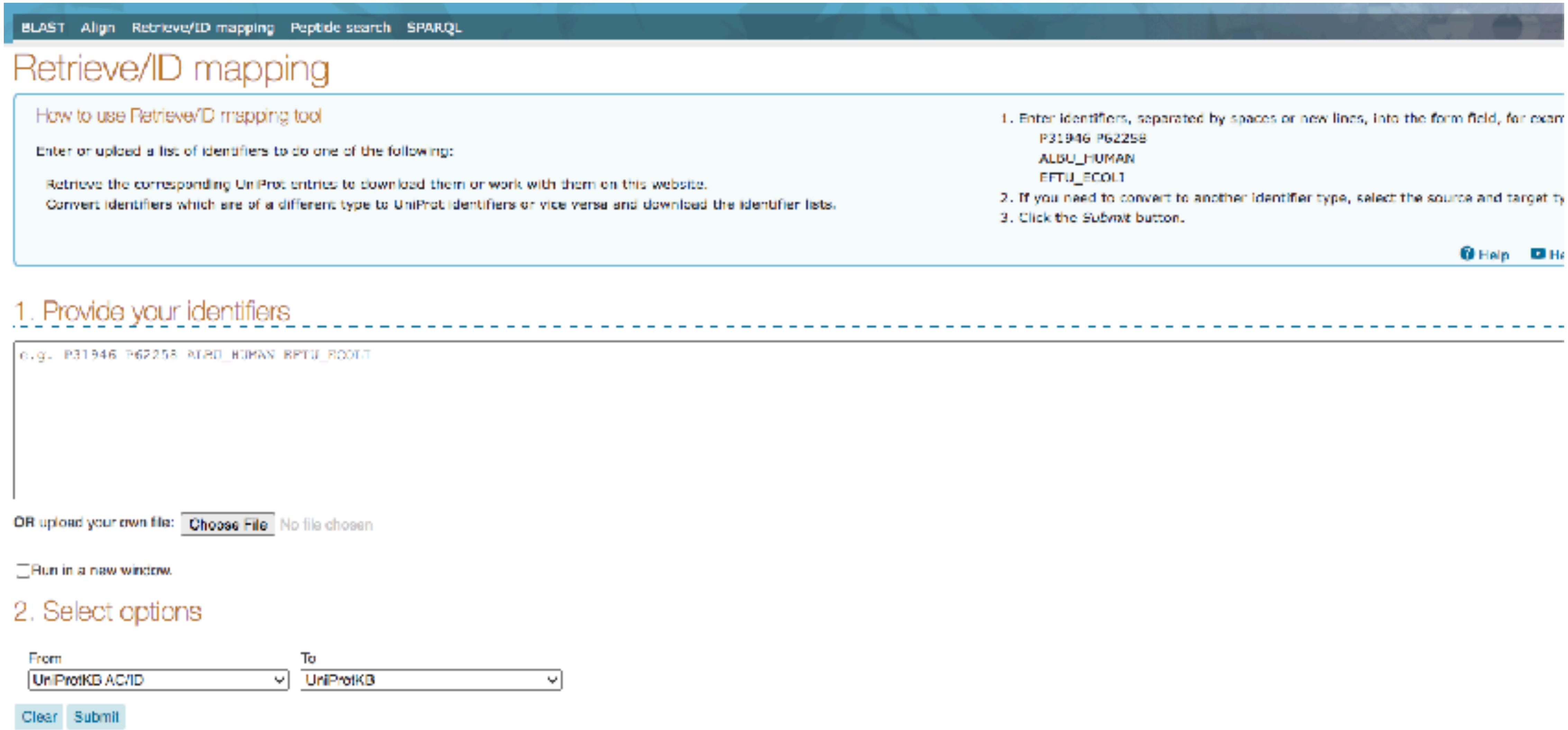
→ Known association

⇢ Unknown association

Feature extraction

- Step 1: what features are we going to use to represent a protein
  - Sequence
  - Structure
  - Network
- Step 2: How to convert these features into numeral vectors that computer can understand?
  - Feature embedding

Gene id name mapping tool
https://www.uniprot.org/uploadlists/



Retrieve/ID mapping

How to use Retrieve/ID mapping tool

Enter or upload a list of identifiers to do one of the following:

Retrieve the corresponding UniProt entries to download them or work with them on this website.
Convert identifiers which are of a different type to UniProt identifiers or vice versa and download the identifier lists.

1. Enter identifiers, separated by spaces or new lines, into the form field, for exam
   P31946 P62258
   ALBU_HUMAN
   EFTU_ECOLI
2. If you need to convert to another identifier type, select the source and target ty
3. Click the Submit button.

1. Provide your identifiers

e.g. P31946 P62258 ALBU_HUMAN EFTU_ECOLI

OR upload your own file: No file chosen

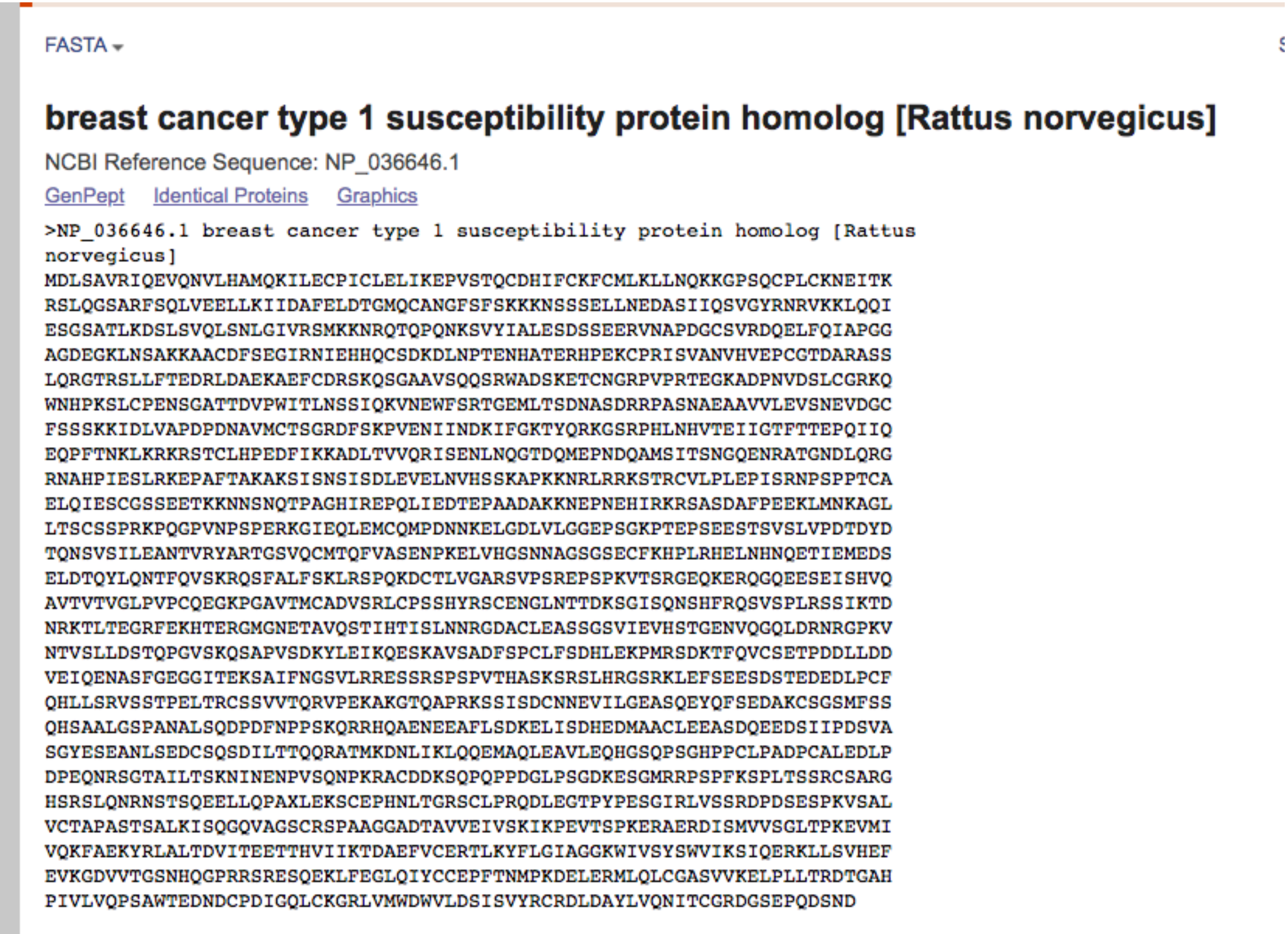
□ Run in a new window

2. Select options

From: UniProtKB AC/ID

To: UniProtKB

Sequence of BRCA1
https://www.ncbi.nlm.nih.gov/protein/NP_036646?report=fasta



breast cancer type 1 susceptibility protein homolog [Rattus norvegicus]

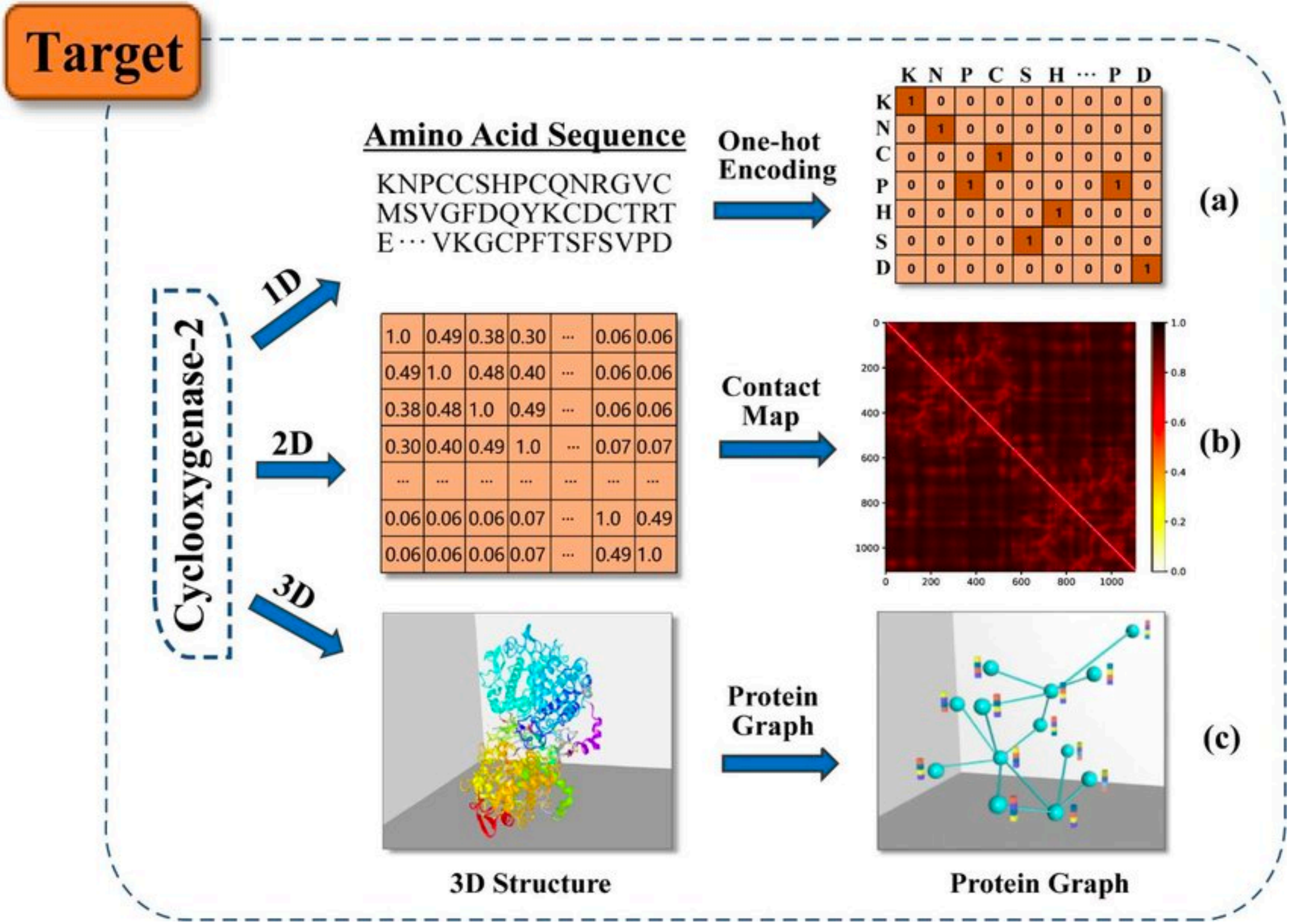
NCBI Reference Sequence: NP_036646.1

```
>NP_036646.1 breast cancer type 1 susceptibility protein homolog [Rattus
norvegicus]
MDLSAVRIQEVQNVLHAMQKILECPICLELIKEPVSTQCDHIFCKFCMLKLLNQKKGPSQCPLCKNEITK
RSLQGSARFSQLVEELLKIIDAFELDTGMQCANGFSFSKKKNSSSELLNEDASIIQSVGYRNRVKKLQQI
ESGSATLKDSLSVQLSNLGIVRSMKKNRQTQPQNKSVYIALESDSSEERVNAPDGCSVRDQELFQIAPGG
AGDEGKLNSAKKAACDFSEGIRNIEHHQCSDKDLNPTENHATERHPEKCPRISVANVHVEPCGTDARASS
LQRGTRSLLFTEDRLDAEKAEFCDRSKQSGAAVSQQSRWADSKETCNGRPVPRTEGKADPNVDSLCGRKQ
WNHPKSLCPENSGATTDVPWITLNSSIQKVNEWFSRTGEMLTSDNASDRRPASNAEAAVVLEVSNEVDGC
FSSSKKIDLVAPDPDNAVMCTSGRDFSKPVENIINDKIFGKTYQRKGSRPHLNHVTEIIGTFTTEPQIIQ
EQPFTNKLKRKRSTCLHPEDFIKKADLTVVQRISENLNQGTDQMEPNDQAMSITSNGQENRATGNDLQRG
RNAHPIESLRKEPAFTAKAKSISNSISDLEVELNVHSSKAPKKNRLRRKSTRCVLPLEPISRNPSPPTCA
ELQIESCGSSEETKKNNSNQTPAGHIREPQLIEDTEPAADAKKNEPNEHIRKRSASDAFPEEKLMNKAGL
LTSCSSPRKPQGPVNPSPERKGIEQLEMCQMPDNNKELGDLVLGGEPSGKPTEPSEESTSVSLVPDTDYD
TQNSVSILEANTVRYARTGSVQCMTQFVASENPKELVHGSNNAGSGSECFKHPLRHELNHNQETIEMEDS
ELDTQYLQNTFQVSKRQSFALFSKLRSPQKDCTLVGARSVPSREPSPKVTSRGEQKERQGQEESEISHVQ
AVTVTVGLPVPCQEGKPGAVTMCADVSRLCPSSHYRSCENGLNTTDKSGISQNSHFRQSVSPLRSSIKTD
NRKTLTEGRFEKHTERGMGNETAVQSTIHTISLNNRGDACLEASSGSVIEVHSTGENVQGQLDRNRGPKV
NTVSLLDSTQPGVSKQSAPVSDKYLEIKQESKAVSADFSPCLFSDHLEKPMRSDKTFQVCSETPDDLLDD
VEIQENASFGEGGITEKSAIFNGSVLRRESSRSPSPVTHASKSRSLHRGSRKLEFSEESDSTEDEDLPCF
QHLLSRVSSTPELTRCSSVVTQRVPEKAKGTQAPRKSSISDCNNEVILGEASQEYQFSEDAKCSGSMFSS
QHSAALGSPANALSQDPDFNPPSKQRRHQAENEEAFLSDKELISDHEDMAACLEEASDQEEDSIIPDSVA
SGYESEANLSEDCSQSDILTTQQRATMKDNLIKLQQEMAQLEAVLEQHGSQPSGHPPCLPADPCALEDLP
DPEQNRSGTAILTSKNINENPVSQNPKRACDDKSQPQPPDGLPSGDKESGMRRPSPFKSPLTSSRCSARG
HSRSLQNRNSTSQEELLQPAXLEKSCEPHNLTGRSCLPRQDLEGTPYPESGIRLVSSRDPDSESPKVSAL
VCTAPASTSALKISQGQVAGSCRSPAAGGADTAVVEIVSKIKPEVTSPKERAERDISMVVSGLTPKEVMI
VQKFAEKYRLALTDVITEETTHVIIKTDAEFVCERTLKYFLGIAGGKWIVSYSWVIKSIQERKLLSVHEF
EVKGDVVTGSNHQGPRRSRESQEKLFEGLQIYCCEPFTNMPKDELERMLQLCGASVVKELPLLTRDTGAH
PIVLVQPSAWTEDNDCPDIGQLCKGRLVMWDWVLDSISVYRCRDLDAYLVQNITCGRDGSEPQDSND
```

Feature extraction

- Step 1: what features are we going to use to represent a protein
    - Sequence
    - Structure
    - Network
- Step 2: How to convert these features into numeral vectors that computer can understand?
    - Feature embedding

Converting proteins to numeral features



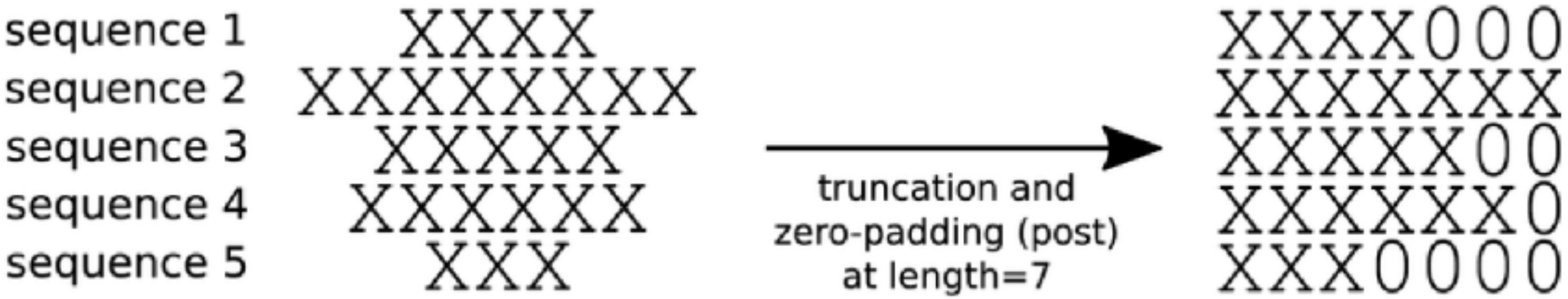
KNPCCSHPCQNRGVC
MSVGFDQYKCDCTRT
E··· VKGCPFTSFSVPD

	K	N	P	C	S	H	···	P	D
K	1	0	0	0	0	0	0	0	0
N	0	1	0	0	0	0	0	0	0
C	0	0	0	1	0	0	0	0	0
P	0	0	1	0	0	0	0	1	0
H	0	0	0	0	0	1	0	0	0
S	0	0	0	0	1	0	0	0	0
D	0	0	0	0	0	0	0	0	1

1.0	0.49	0.38	0.30	···	0.06	0.06
0.49	1.0	0.48	0.40	···	0.06	0.06
0.38	0.48	1.0	0.49	···	0.06	0.06
0.30	0.40	0.49	1.0	···	0.07	0.07
···	···	···	···	···	···	···
0.06	0.06	0.06	0.07	···	1.0	0.49
0.06	0.06	0.06	0.07	···	0.49	1.0

source: Deep learning for drug repurposing: methods, databases, and applications

Truncation and zero-padding to have a matched length



sequence 1 XXXX

sequence 2 XXXXXXXX

sequence 3 XXXXX

sequence 4 XXXXXX

sequence 5 XXX

truncation and
zero-padding (post)
at length=7

XXXX000

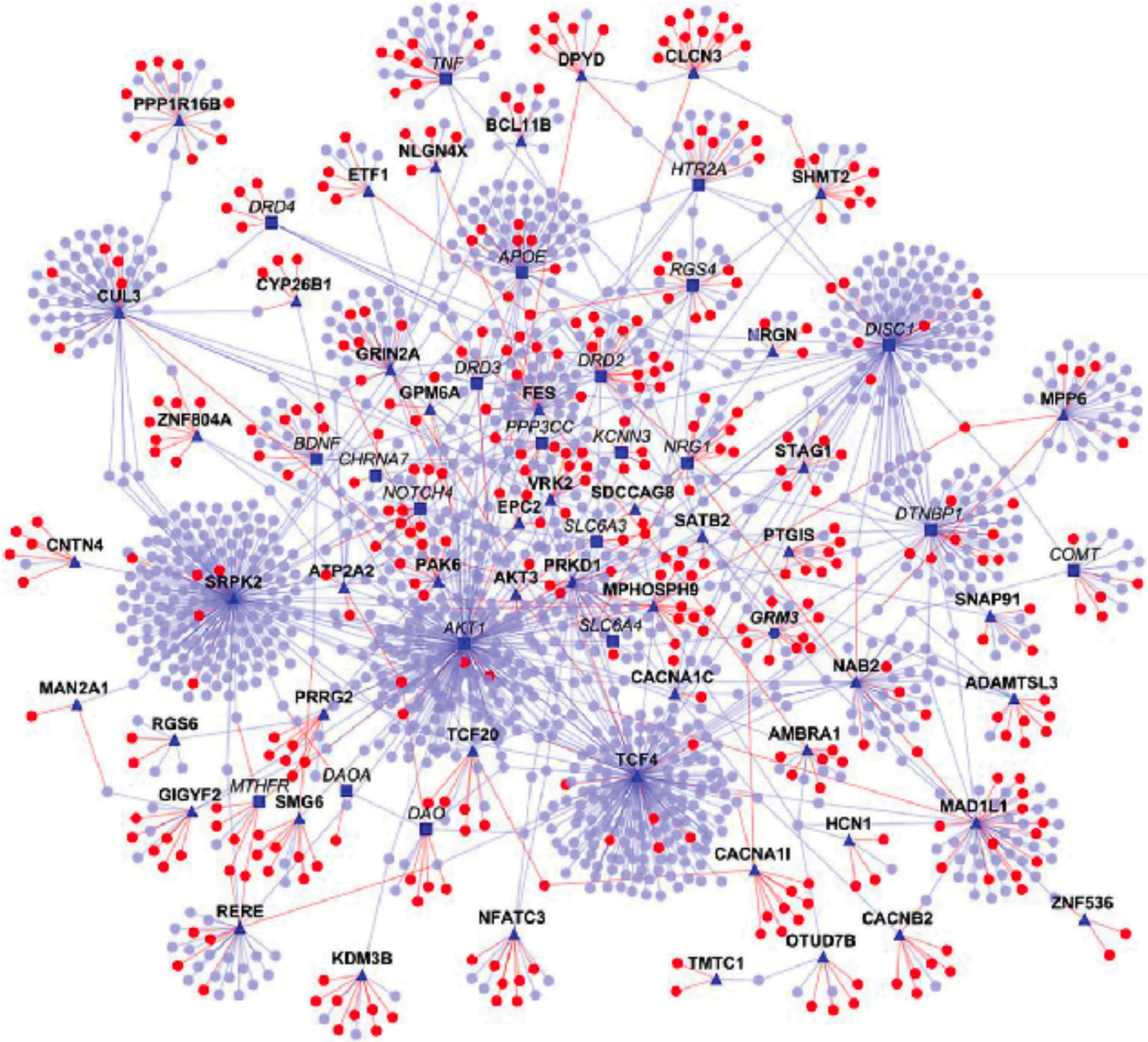
XXXXXXX

XXXXX00

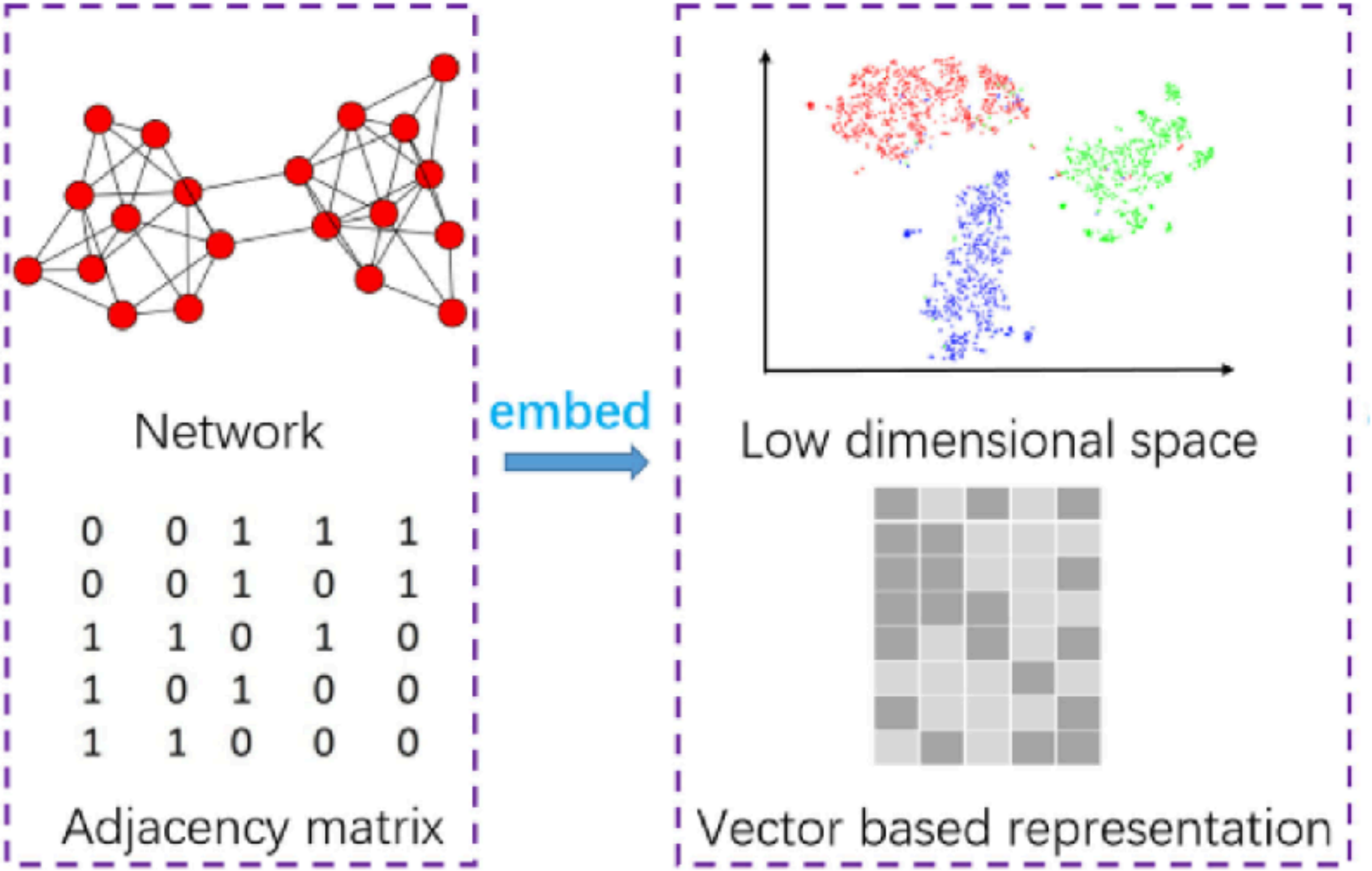
XXXXXX0

XXX0000

Protein protein network



Protein protein network

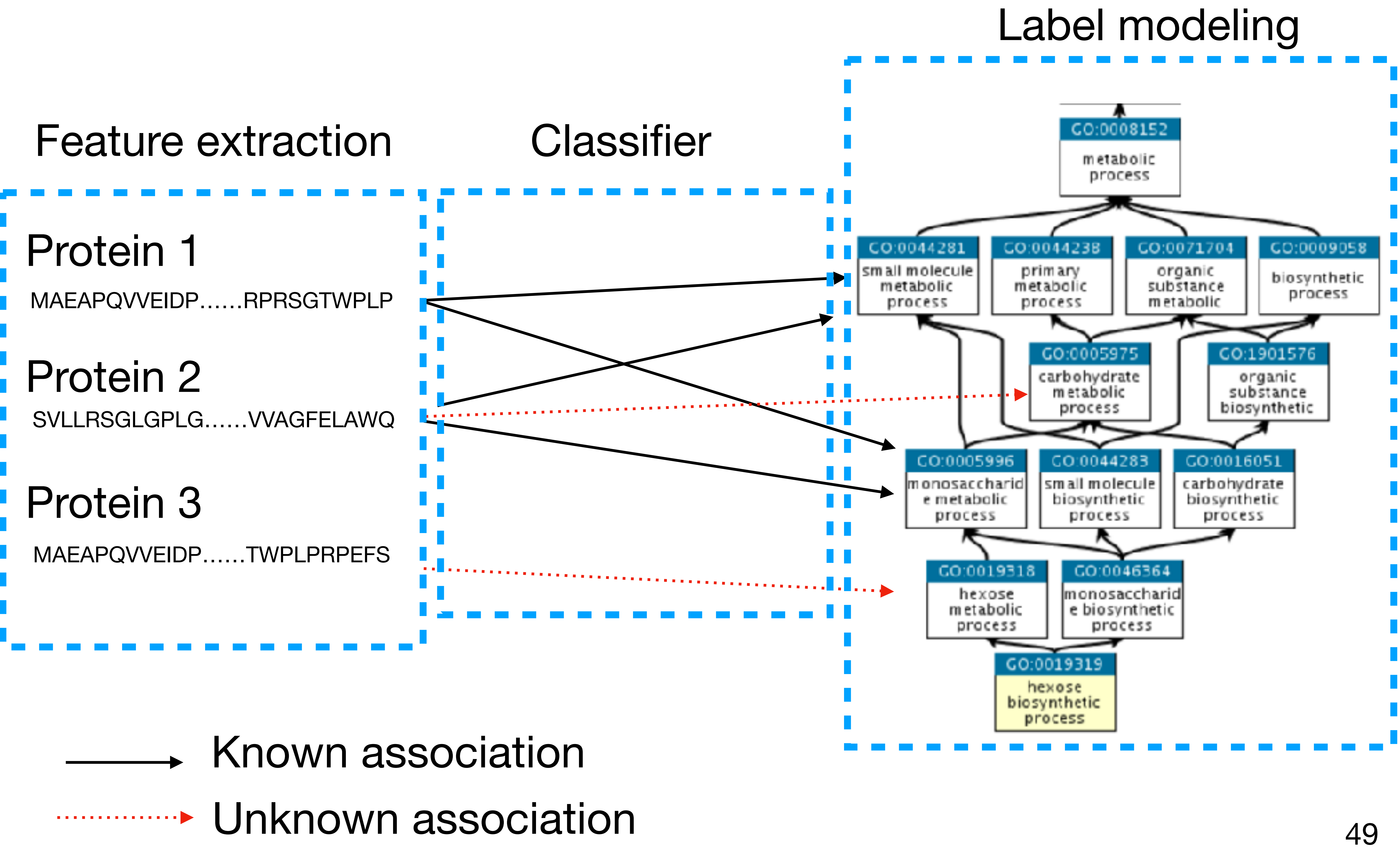


$$\begin{matrix} 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{matrix}$$

source: A Survey on Network Embedding

Classifier

Problem setting for protein function prediction



Label modeling

Feature extraction

Classifier

Protein 1

MAEAPQVVEIDP……RPRSGTWPLP

Protein 2

SVLLRSGLGPLG……VVAGFELAWQ

Protein 3

MAEAPQVVEIDP……TWPLPRPEFS

Known association

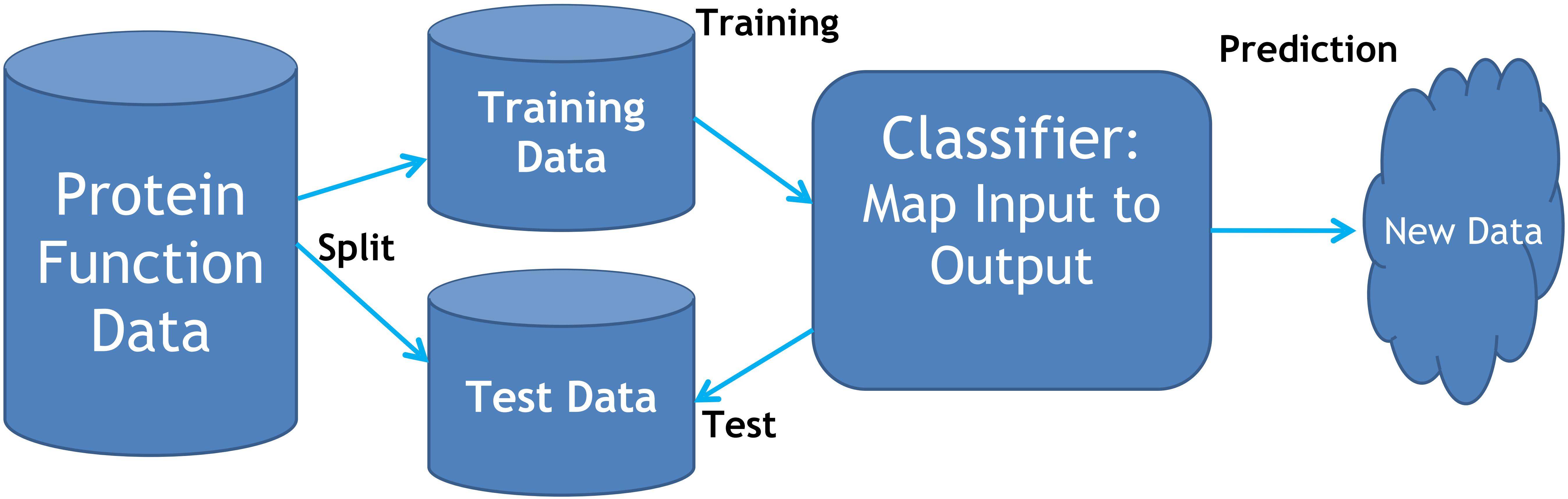
Unknown association

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

Problem setting

- Input:
  - Features: sequence of known proteins
  - Known annotations: <Gene Ontology i, protein j>
  - Label graph: gene ontology graph
- Output:
  - Unknown annotations: Should we annotate protein k to gene ontology q?

Data Driven Machine Learning Approach



Input: sequence features
Output: function category

Training: Build a classifier
Test: Test the model

Key idea: Learn from known data and Generalize to unseen data

kNN-based (fiind similar proteins)

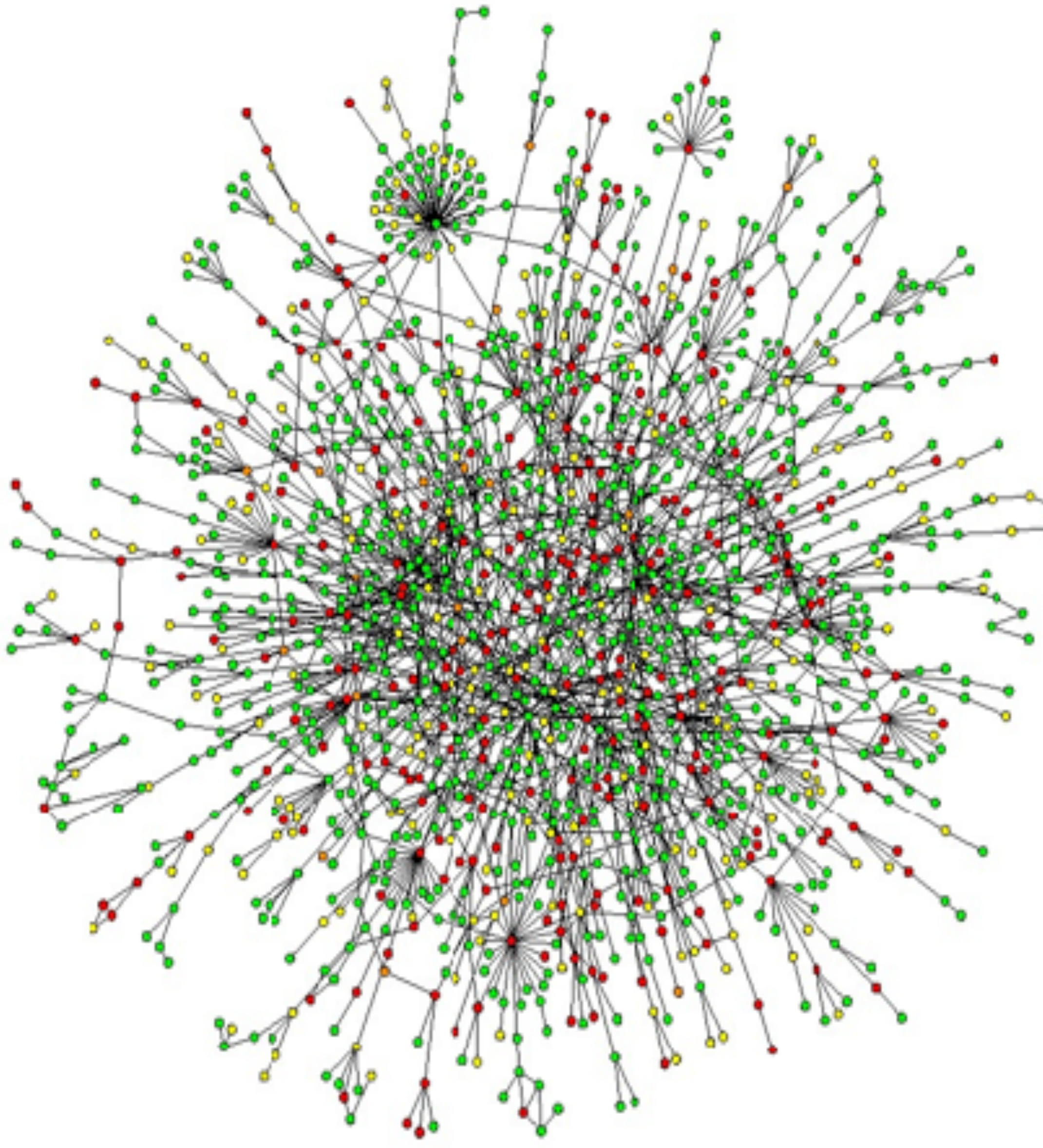
- The easiest way to infer the molecular function of an uncharacterized sequence is by finding a similar and annotated sequence
- BLAST (sequence-sequence local alignment tool) (e.g., Blast2GO)
- Problem:
  - Find similar sequence (sequence aligment)
  - Use these sequences to transfer annotation

kNN-based (fiind similar functions)

- The Function Association Matrix, describes the probability that two GO terms are associated to the same protein based on the frequency at which they co-occur in UniProt sequences.
- For example, the biological process "positive regulation of transcription, DNA-dependent" is strongly associated with the molecular function "DNA binding activity" (P(GO:0045893|GO:0003677) = 0.455).
- Predict non-observed GO terms based on observed ones

Network-Based Approach

- Protein-protein interaction network
- Closer that two nodes are in the network, the more functionally similar they will be in terms of cellular pathway or process as opposed to molecular function
- Non-neighboring proteins with similar network connectivity patterns can have similar molecular functions



Network distance is correlated to GO annotation similarity

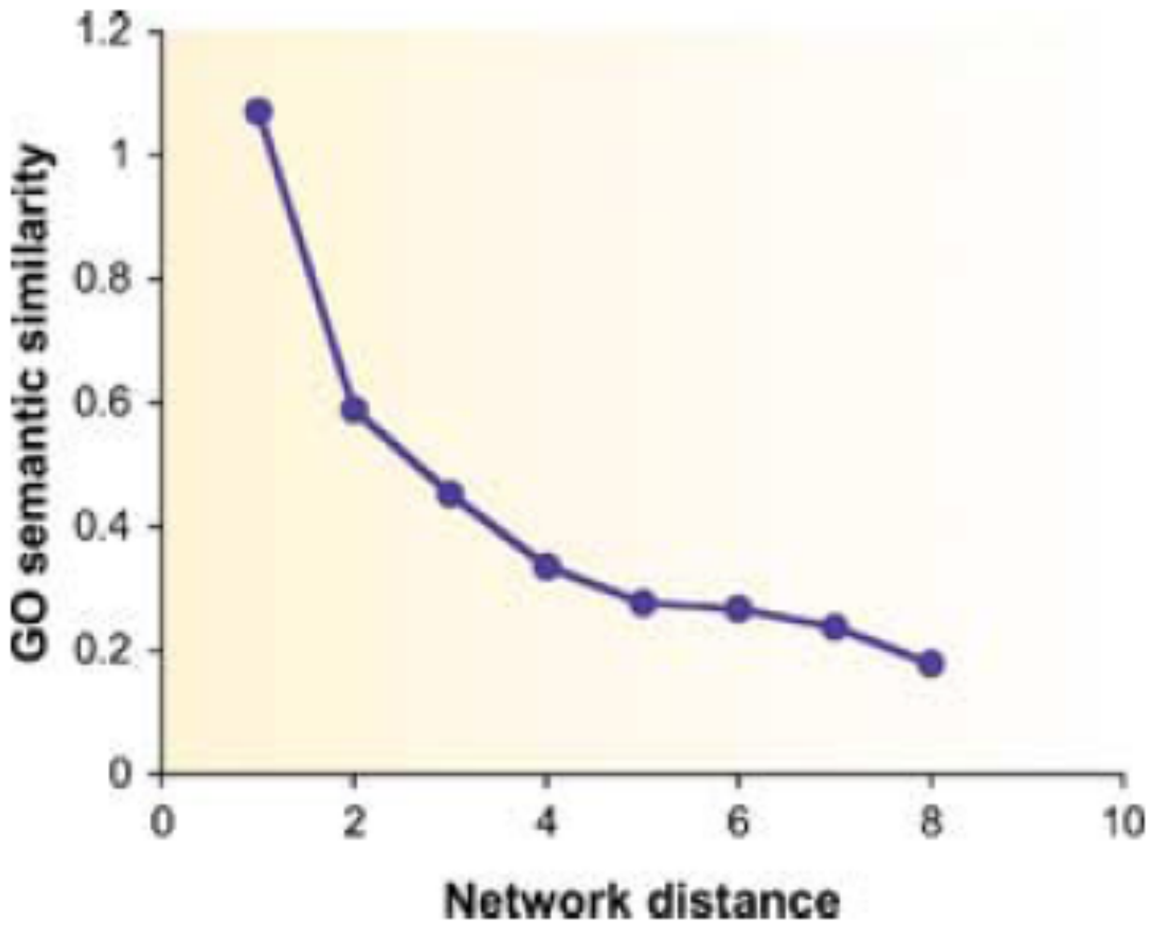


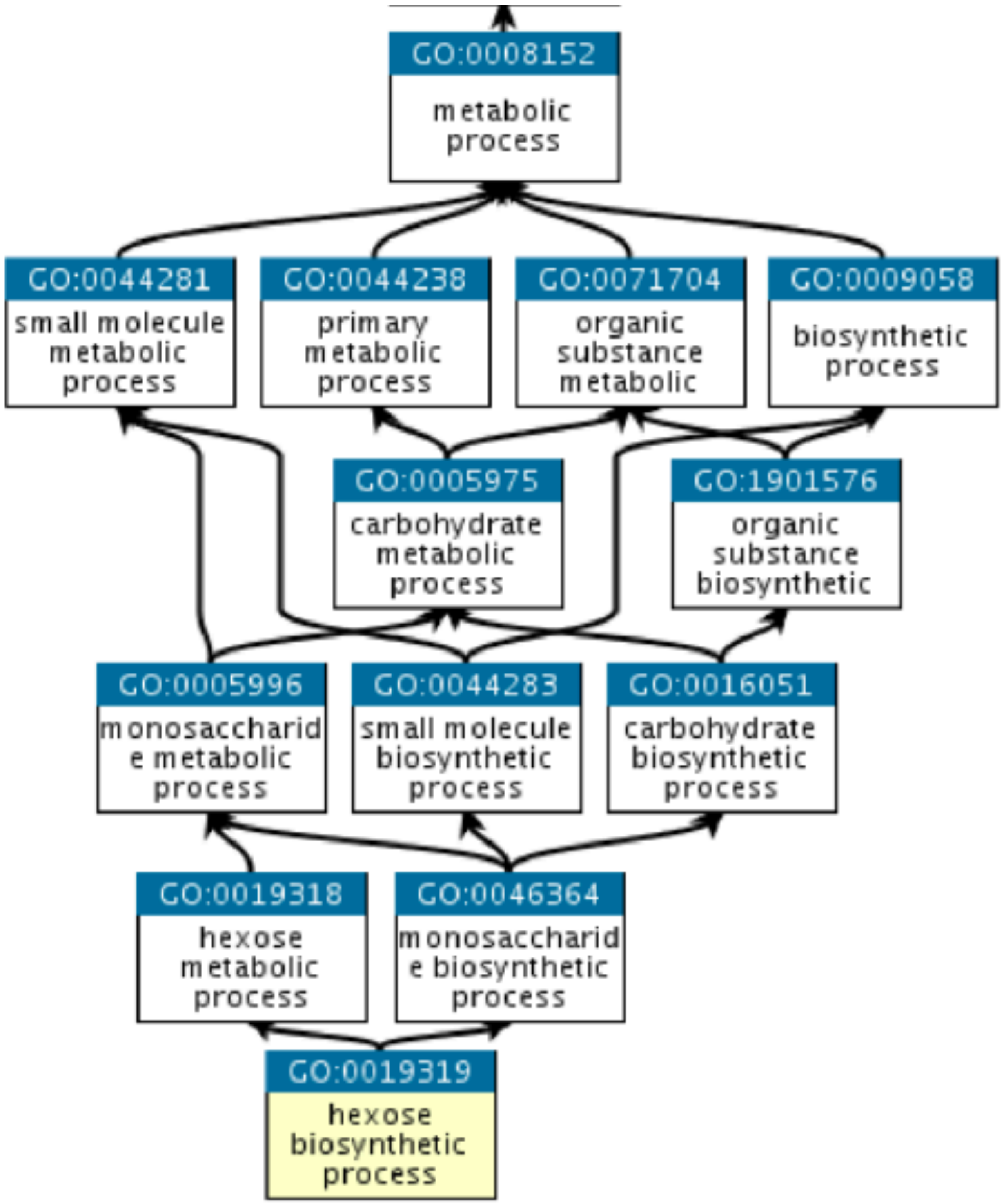
Figure 3 Correlation between protein functional distance and network distance. *X*-axis: distance in the network. *Y*-axis: average functional similarity of protein pairs that lie at the specified distance. The functional similarity of two proteins is measured using the semantic similarity of their GO categories (Lord *et al*, 2003).

Sharan et al., Molecular Systems Biology, 2007

Gene Ontology: A directed acyclic graph

Less specific

More specific



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

Label modeling

- Transfer across species
- Zero-shot/few-shot problem

Training data: How many proteins do we have annotations?

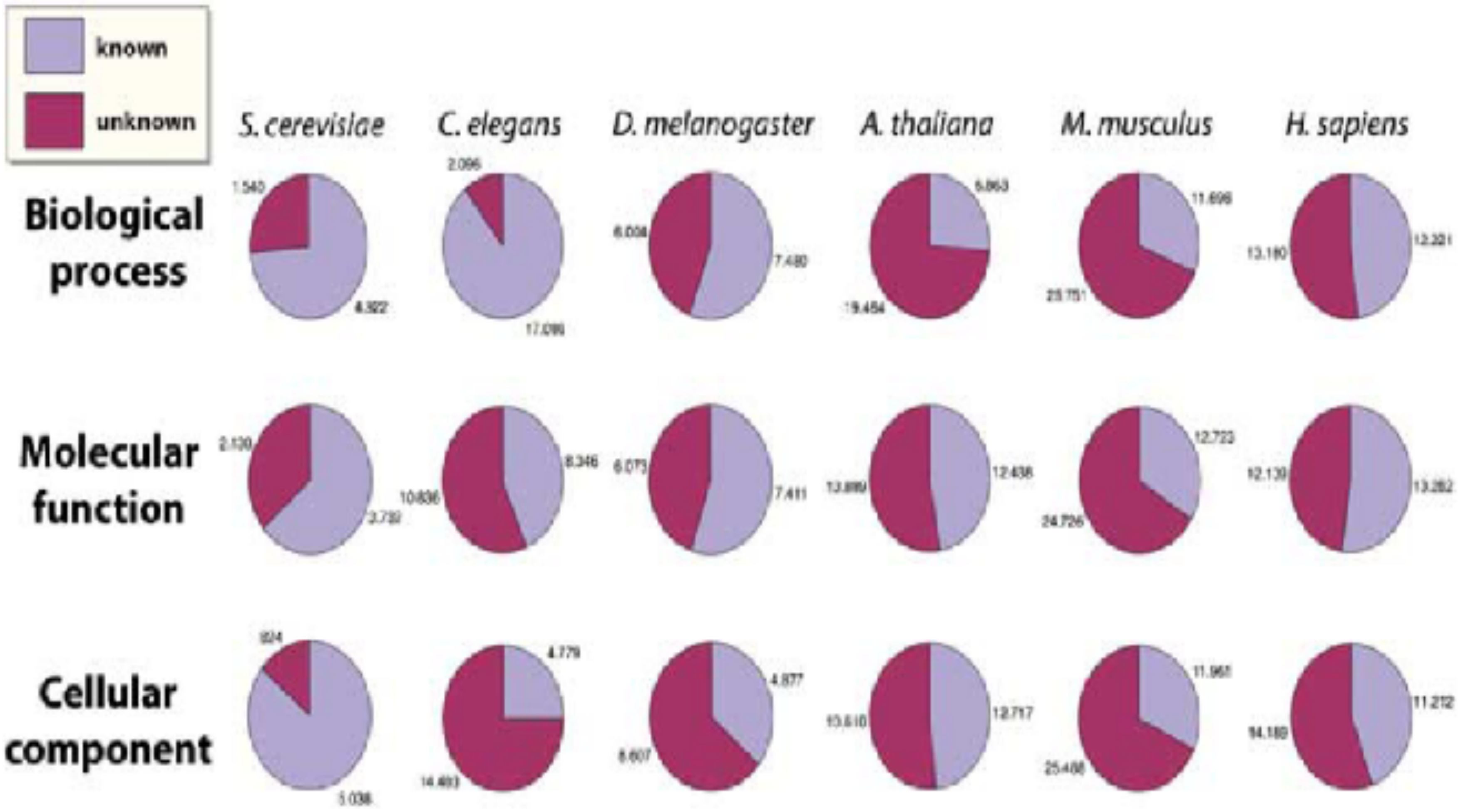
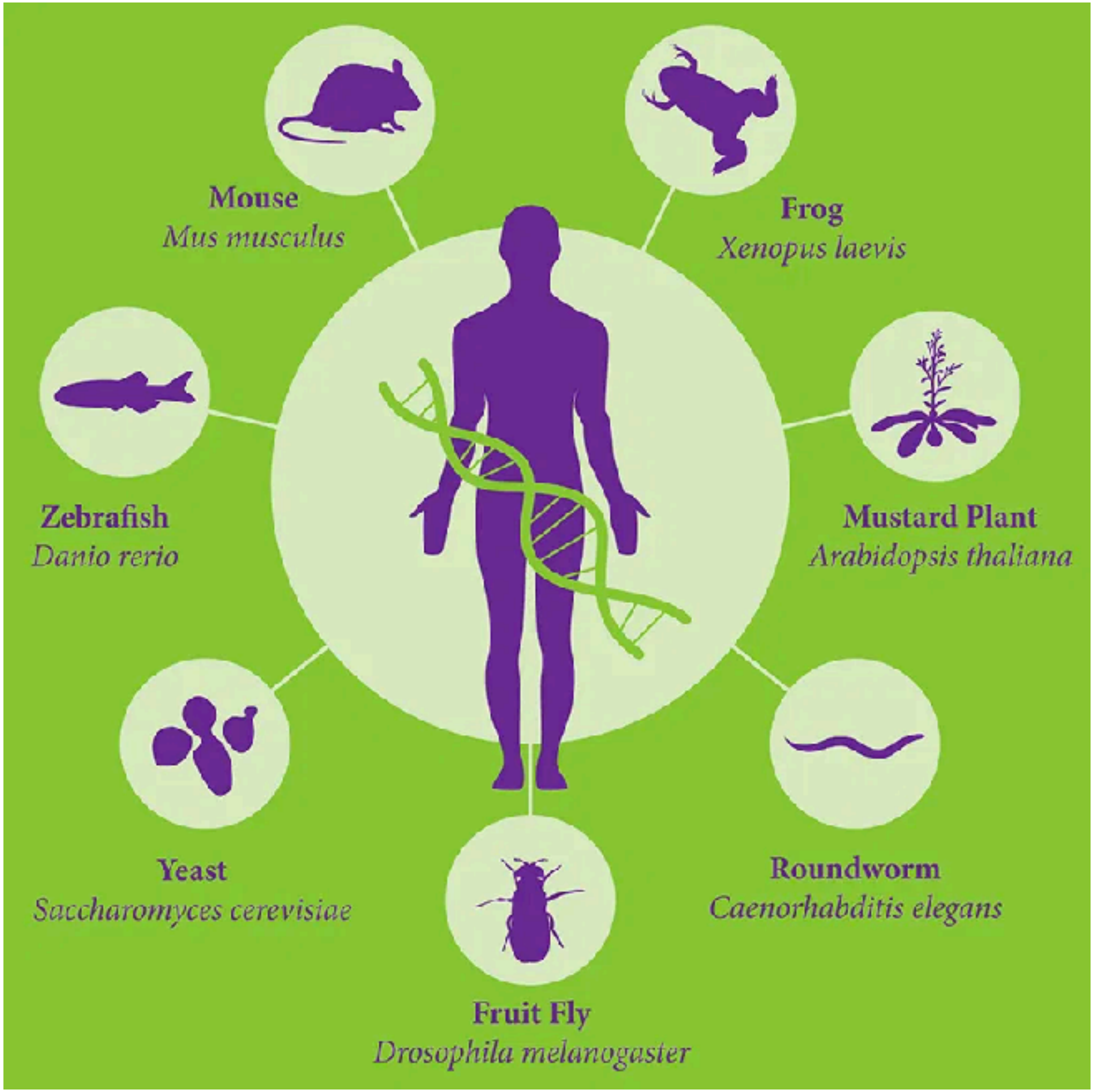


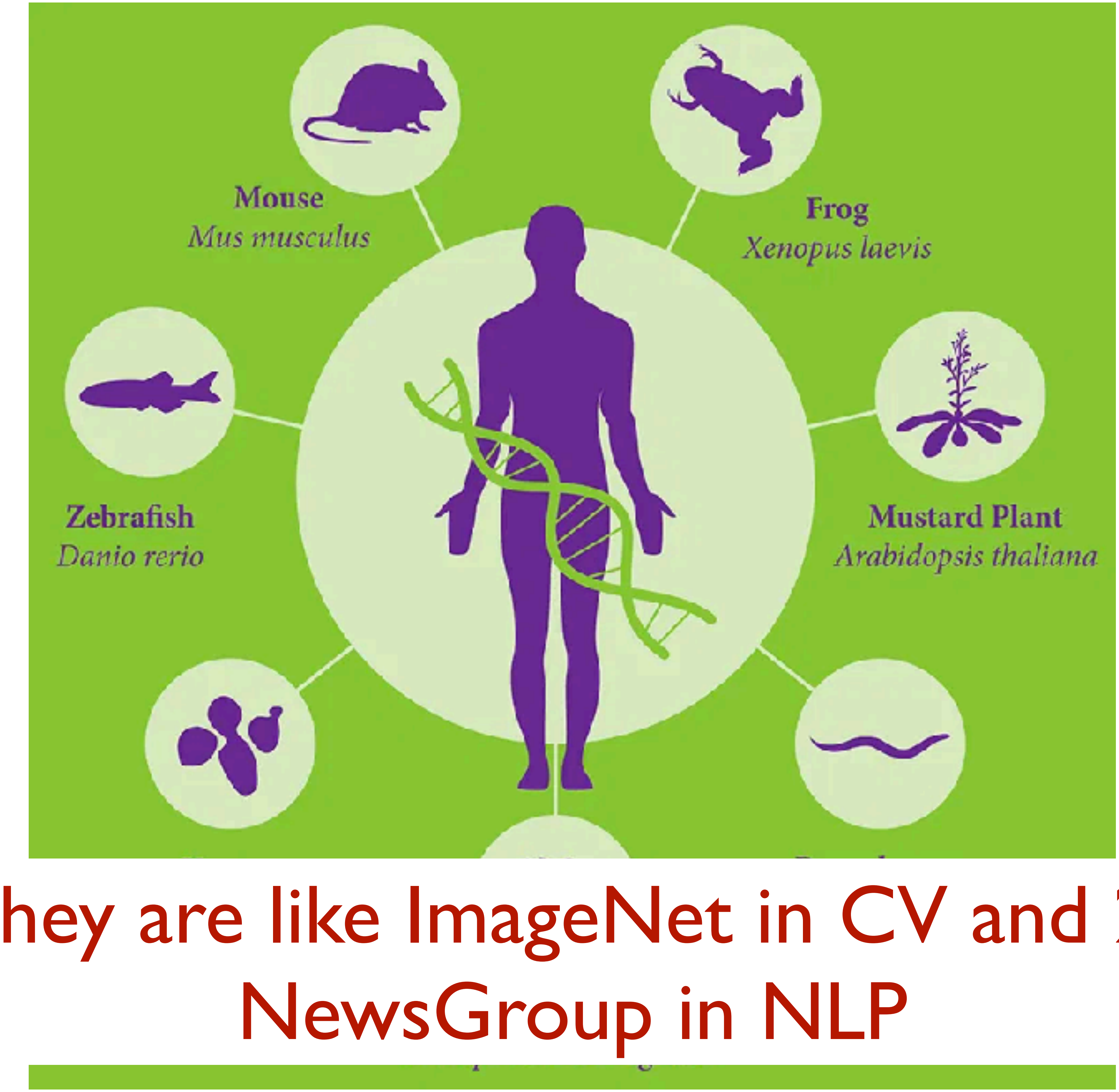
Figure 1 Extent of annotation of proteins in model species. For each species, the charts give the fractions and numbers of annotated and unannotated proteins, according to the three ontologies of the GO annotation. The numbers are based on the Entrez Gene and the WormBase databases as of September 2006.

Sharan et al., Molecular Systems Biology, 2007

Model organism

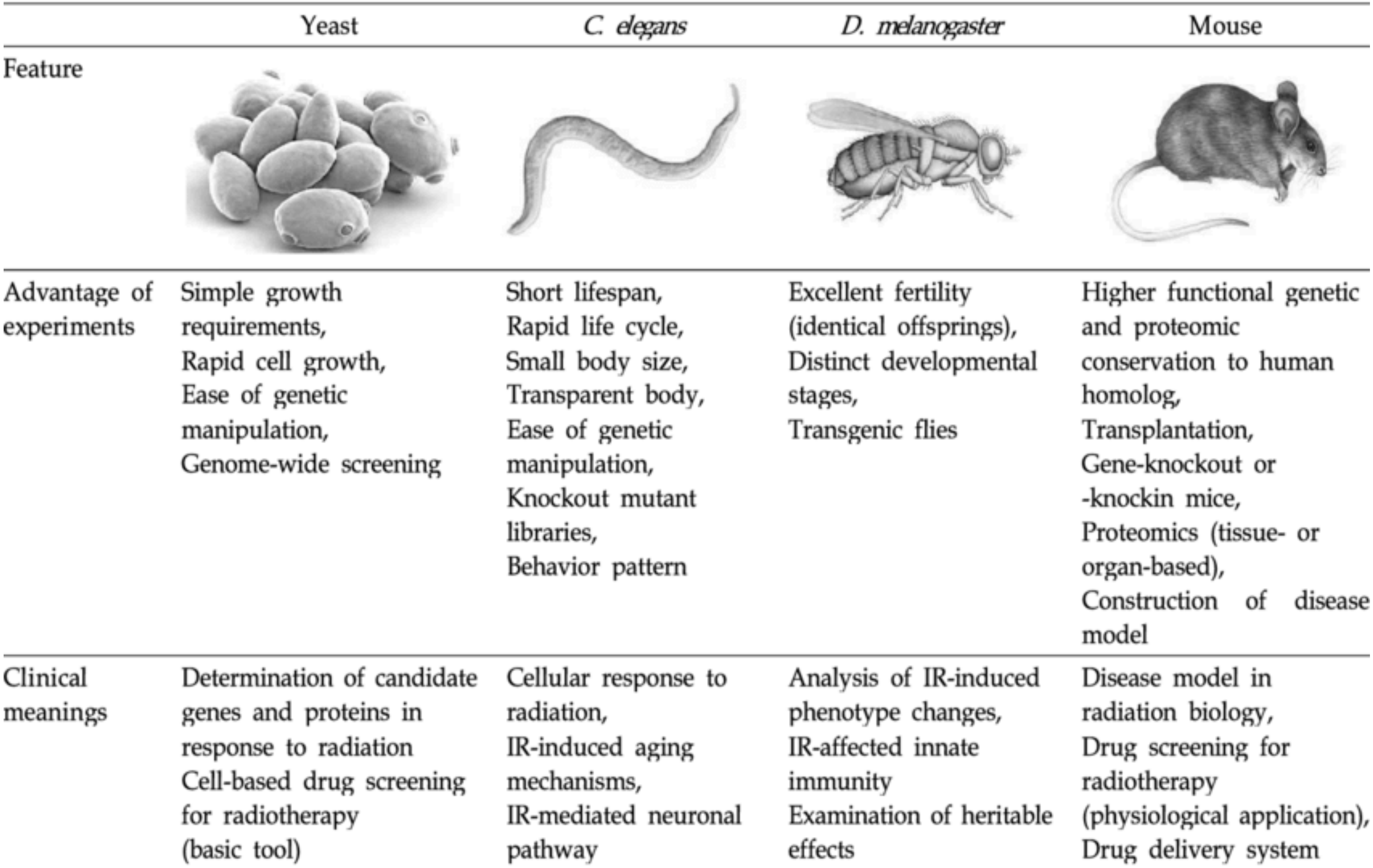


Model organism



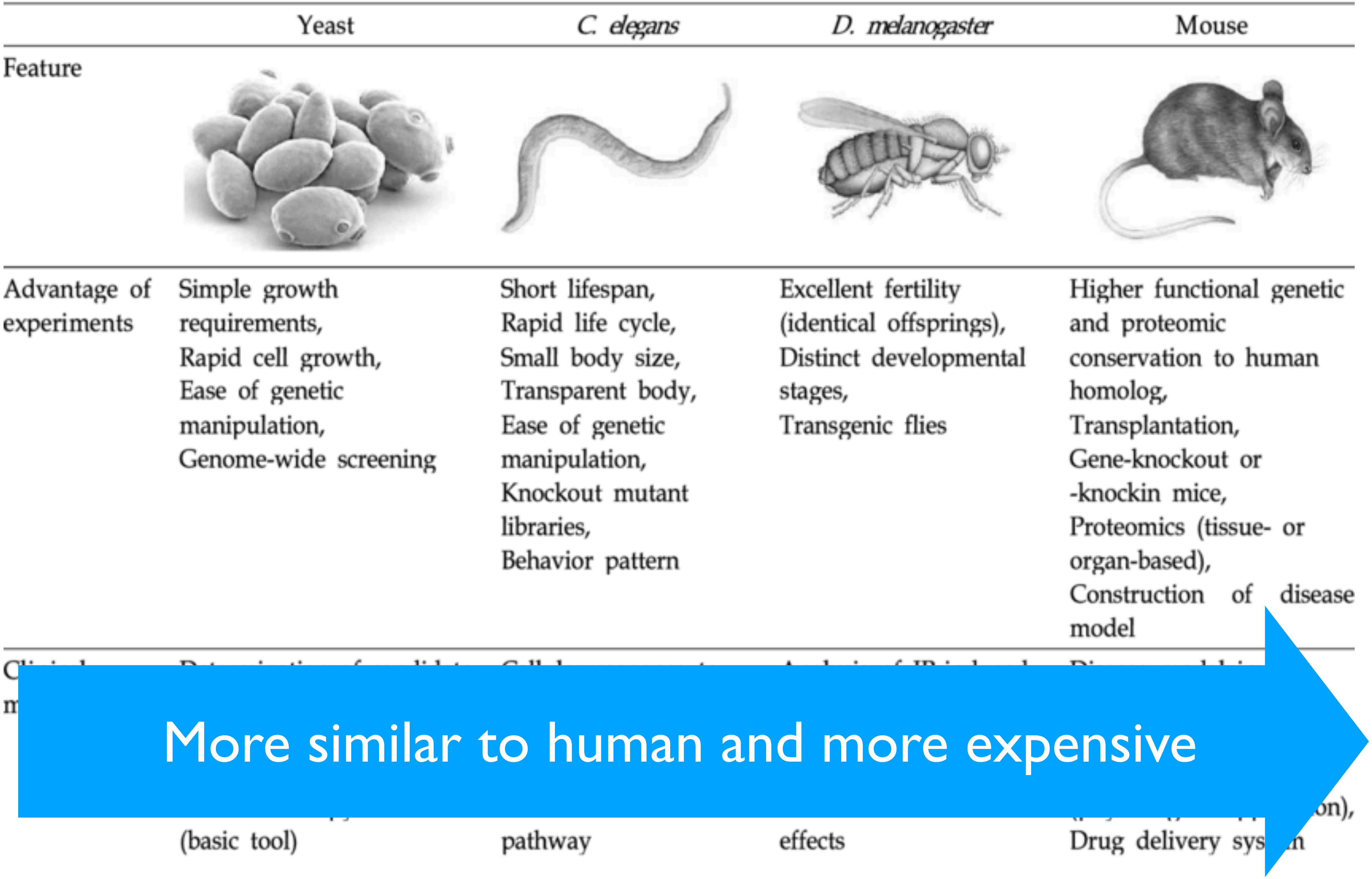
They are like ImageNet in CV and 20 NewsGroup in NLP

Current State of Function of Model Genome Annotation



	Yeast	*C. elegans*	*D. melanogaster*	Mouse
Feature				
Advantage of experiments	Simple growth requirements, Rapid cell growth, Ease of genetic manipulation, Genome-wide screening	Short lifespan, Rapid life cycle, Small body size, Transparent body, Ease of genetic manipulation, Knockout mutant libraries, Behavior pattern	Excellent fertility (identical offsprings), Distinct developmental stages, Transgenic flies	Higher functional genetic and proteomic conservation to human homolog, Transplantation, Gene-knockout or -knockin mice, Proteomics (tissue- or organ-based), Construction of disease model
Clinical meanings	Determination of candidate genes and proteins in response to radiation Cell-based drug screening for radiotherapy (basic tool)	Cellular response to radiation, IR-induced aging mechanisms, IR-mediated neuronal pathway	Analysis of IR-induced phenotype changes, IR-affected innate immunity Examination of heritable effects	Disease model in radiation biology, Drug screening for radiotherapy (physiological application), Drug delivery system

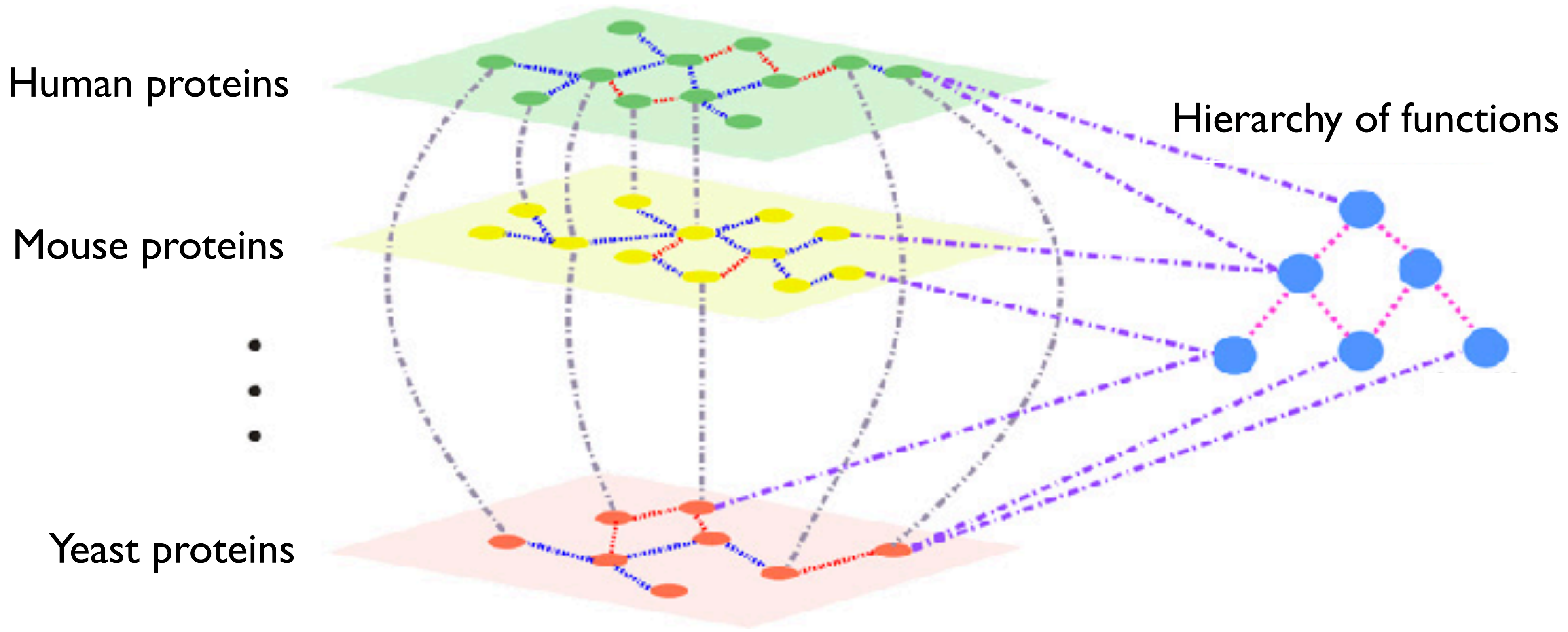
Current State of Function of Model Genome Annotation



	Yeast	*C. elegans*	*D. melanogaster*	Mouse
Feature				
Advantage of experiments	Simple growth requirements, Rapid cell growth, Ease of genetic manipulation, Genome-wide screening	Short lifespan, Rapid life cycle, Small body size, Transparent body, Ease of genetic manipulation, Knockout mutant libraries, Behavior pattern	Excellent fertility (identical offsprings), Distinct developmental stages, Transgenic flies	Higher functional genetic and proteomic conservation to human homolog, Transplantation, Gene-knockout or -knockin mice, Proteomics (tissue- or organ-based), Construction of disease model
	(basic tool)	pathway	effects	on), Drug delivery sys n

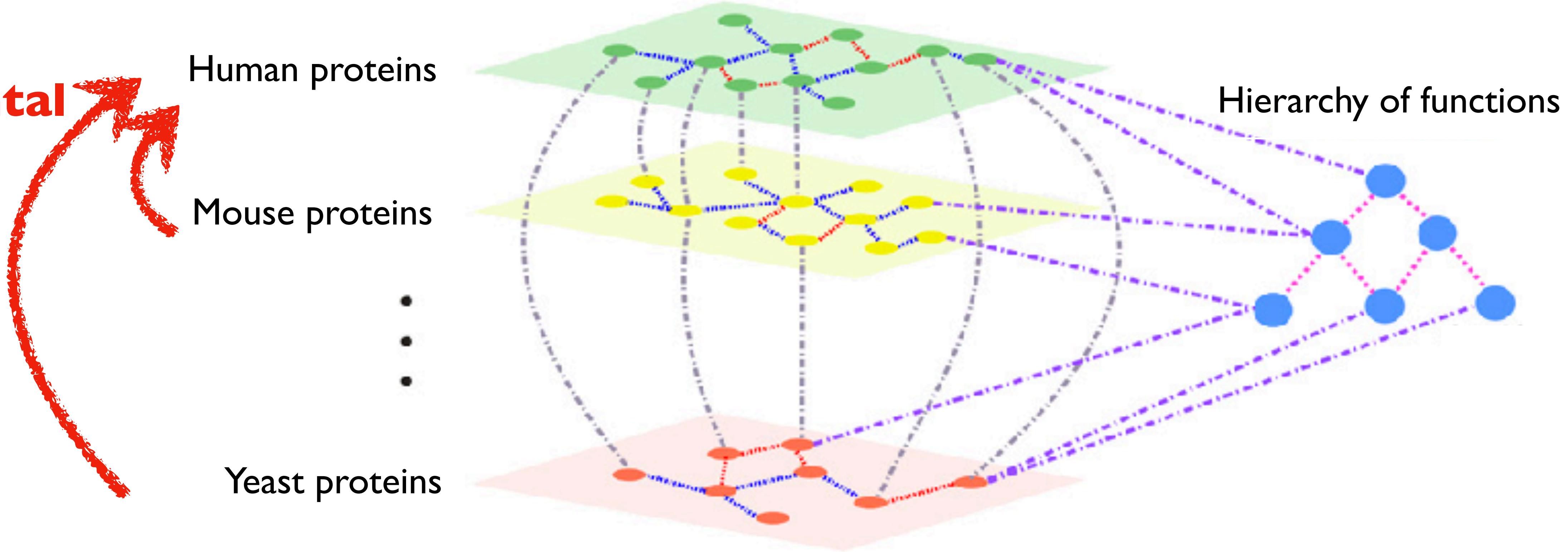
More similar to human and more expensive

transfer knowledge from other species to human



[Wang et al. PSB2017]

Transfer knowledge from other species to human



Transfer experimental results

Human proteins

Mouse proteins

Yeast proteins

Hierarchy of functions

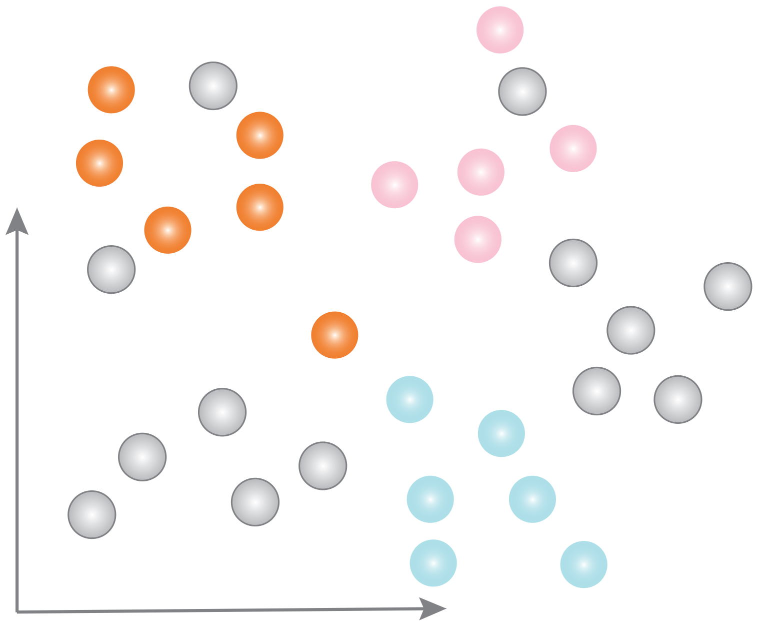
A unique heterogeneous network dataset

- Nodes: **16K** human proteins, **16K** mouse proteins, **6K** yeast proteins, **11K** fruit fly proteins, **13K** worm proteins
- Edges: **7** edge types
- Labels: **227K** protein function associations

Wang et al. PSB2017

Key challenge: novel functions

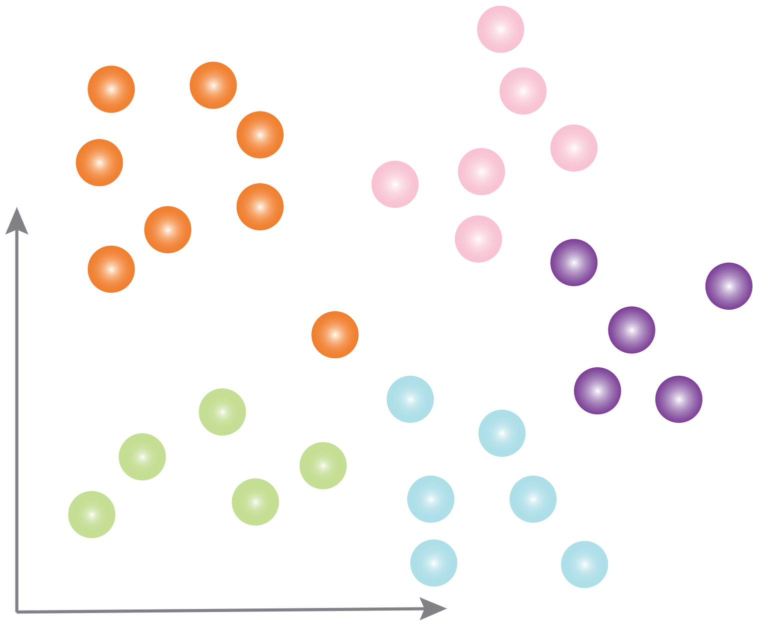
Input



● Unannotated proteins

● ● ● Annotated reference proteins

Gold standard

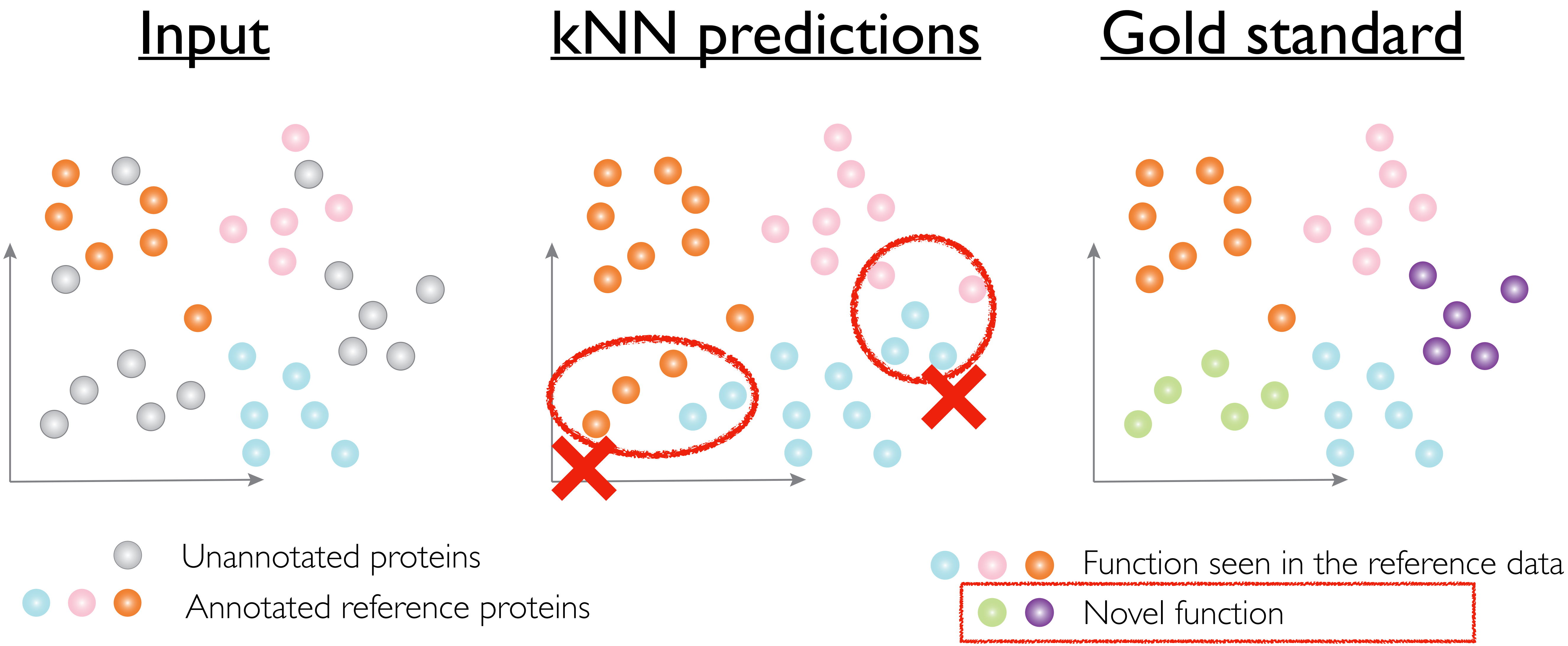


● ● ● Function seen in the reference data

● ● Novel function

How to correctly classify proteins into novel functions?

Existing methods cannot annotate novel functions



Input

kNN predictions

Gold standard

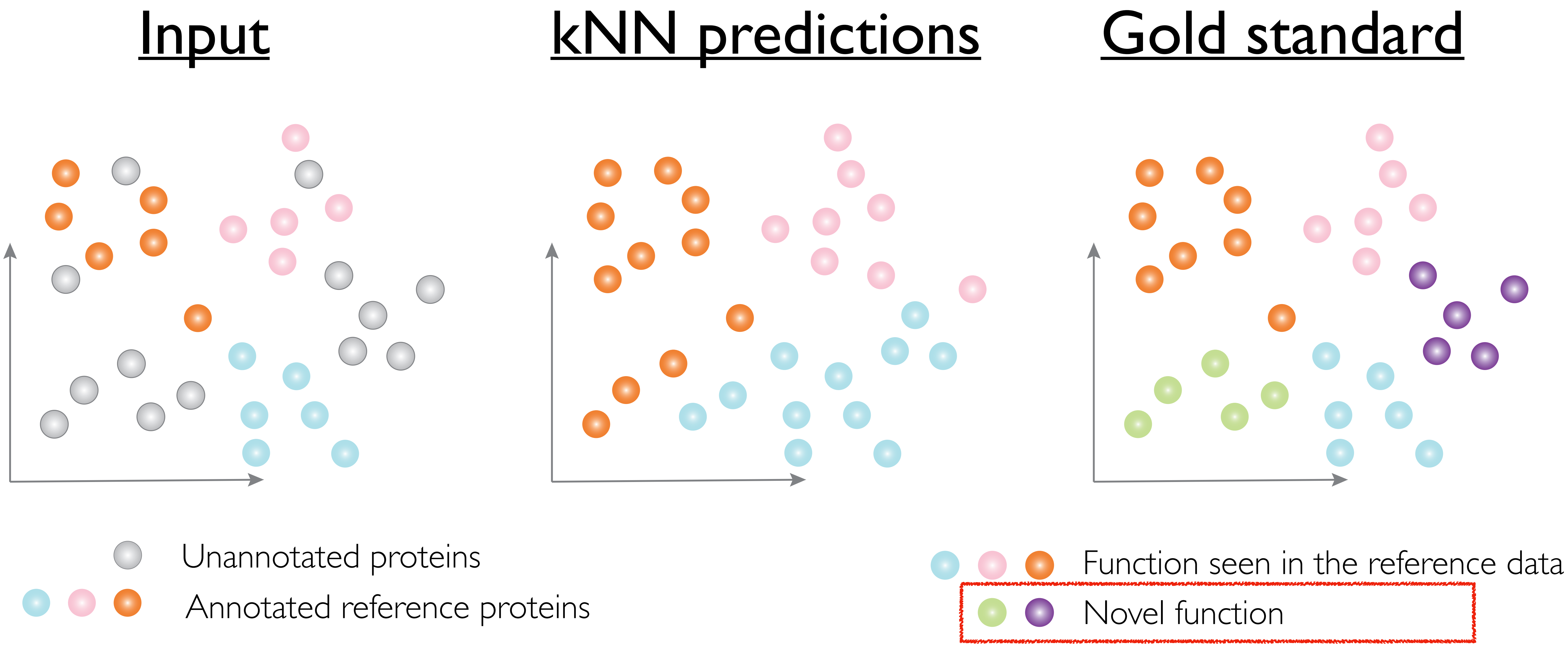
● Unannotated proteins

● ● ● Annotated reference proteins

● ● ● Function seen in the reference data

● ● Novel function

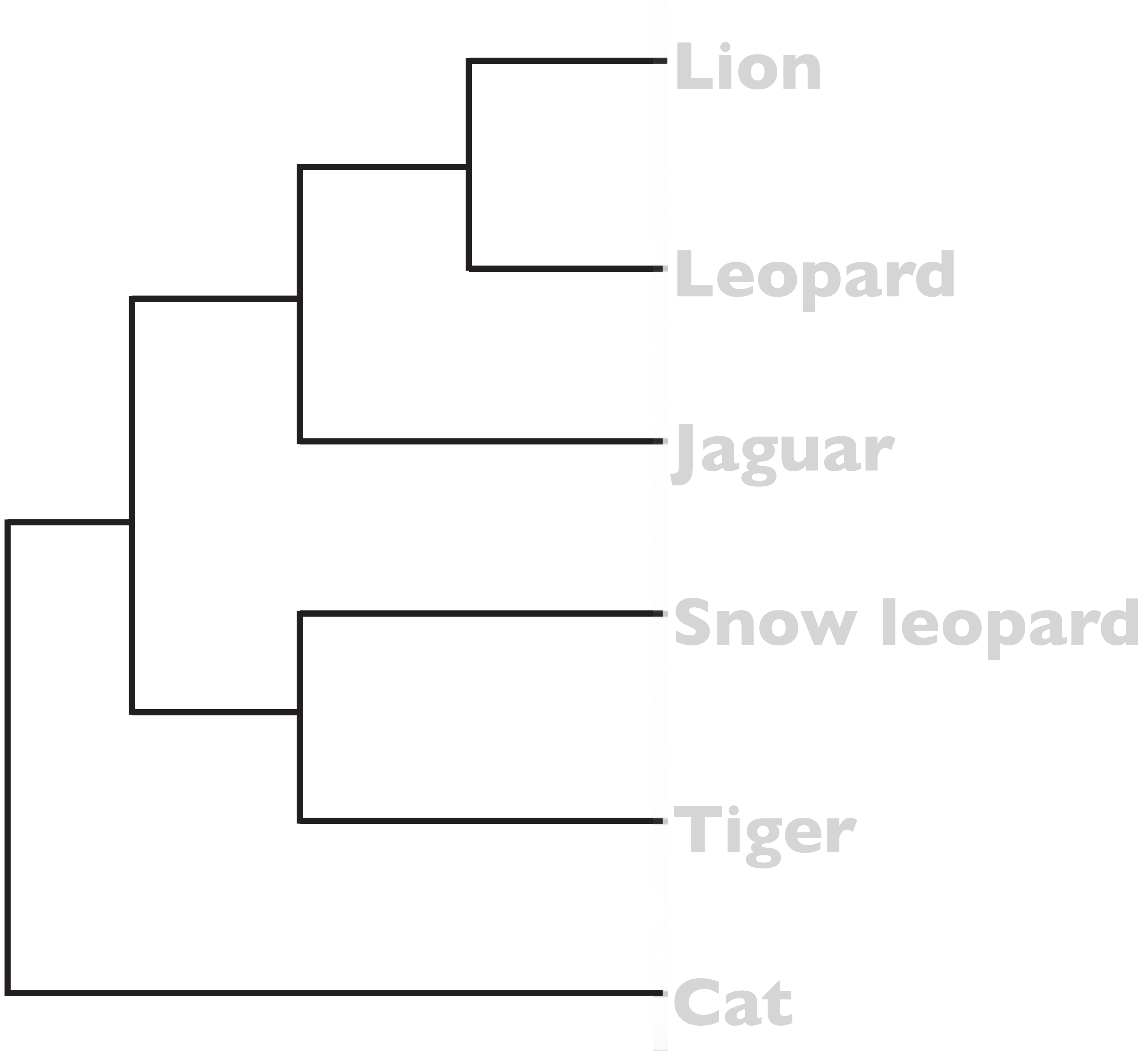
Existing methods cannot annotate novel functions


Input
kNN predictions
Gold standard
● Unannotated proteins
● ● ● Annotated reference proteins
● ● ● Function seen in the reference data
● ● Novel function

Zero-shot learning: classify samples into novel classes using side information/class attributes

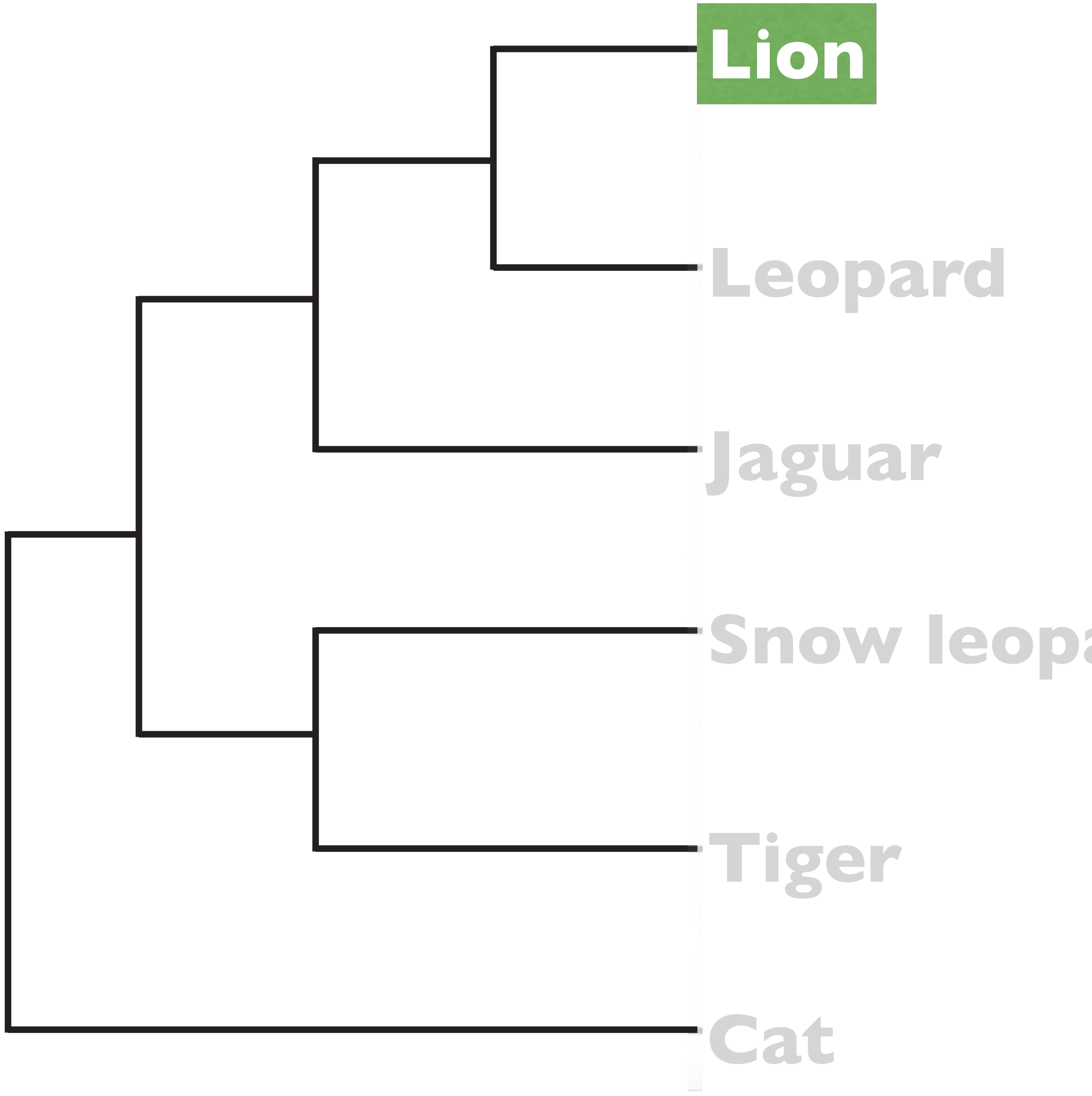
ONTOLOGICAL CLASSIFICATION OF UNSEEN ANIMALS

Ontology of great cats



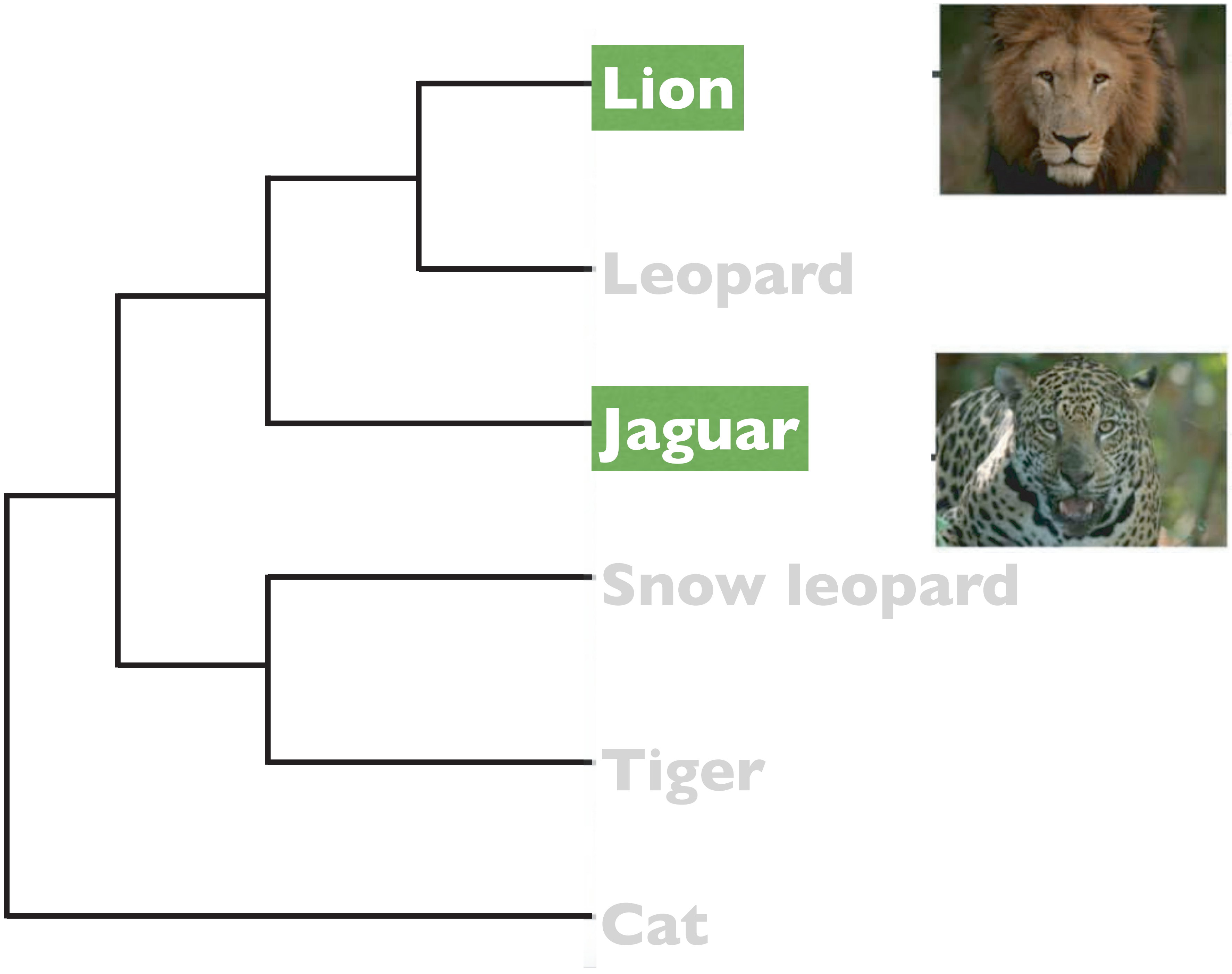
ONTOLOGICAL CLASSIFICATION OF UNSEEN ANIMALS

Ontology of great cats



ONTOLOGICAL CLASSIFICATION OF UNSEEN ANIMALS

Ontology of great cats



Lion

Leopard

Jaguar

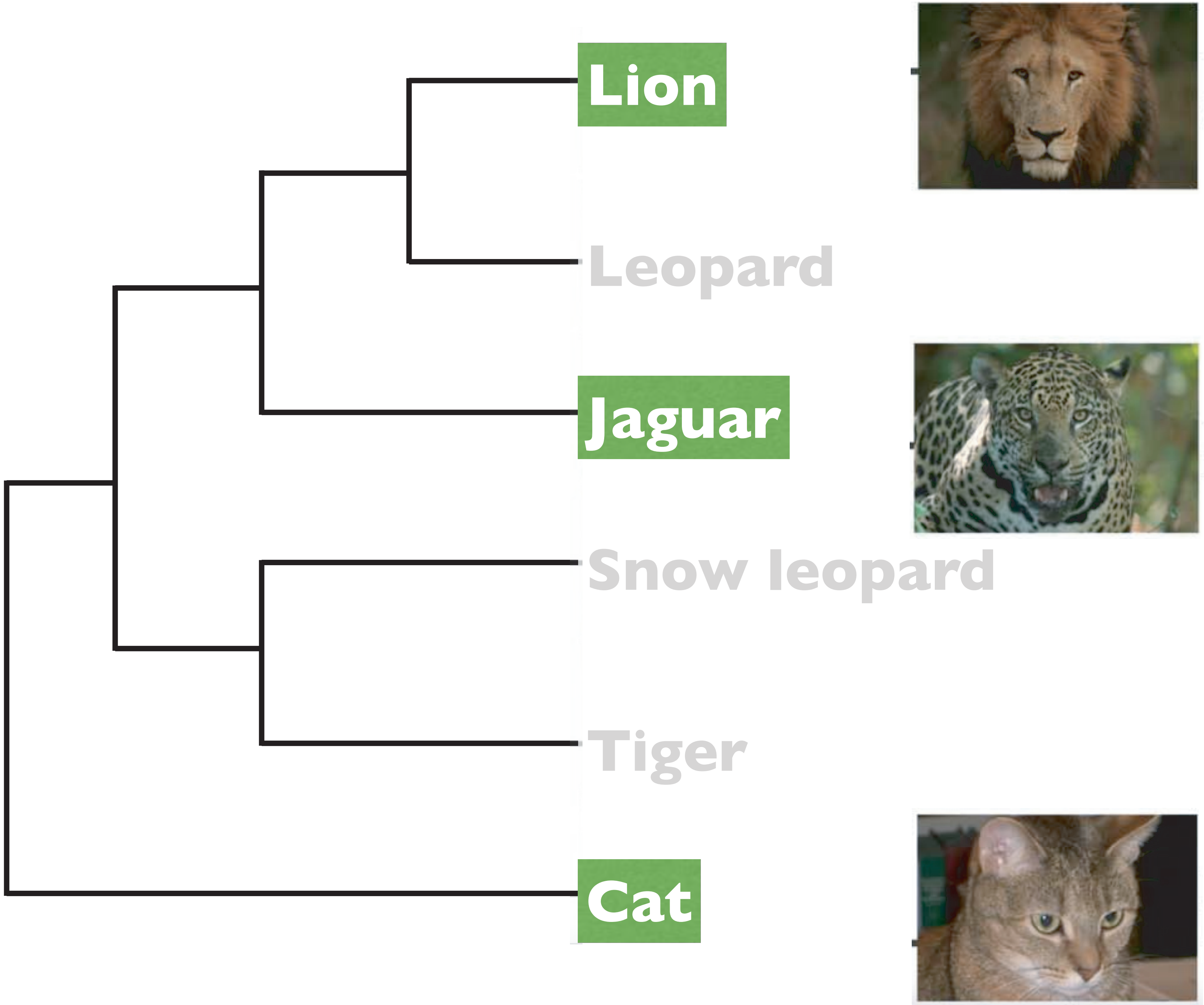
Snow leopard

Tiger

Cat

ONTOLOGICAL CLASSIFICATION OF UNSEEN ANIMALS

Ontology of great cats



Lion

Leopard

Jaguar

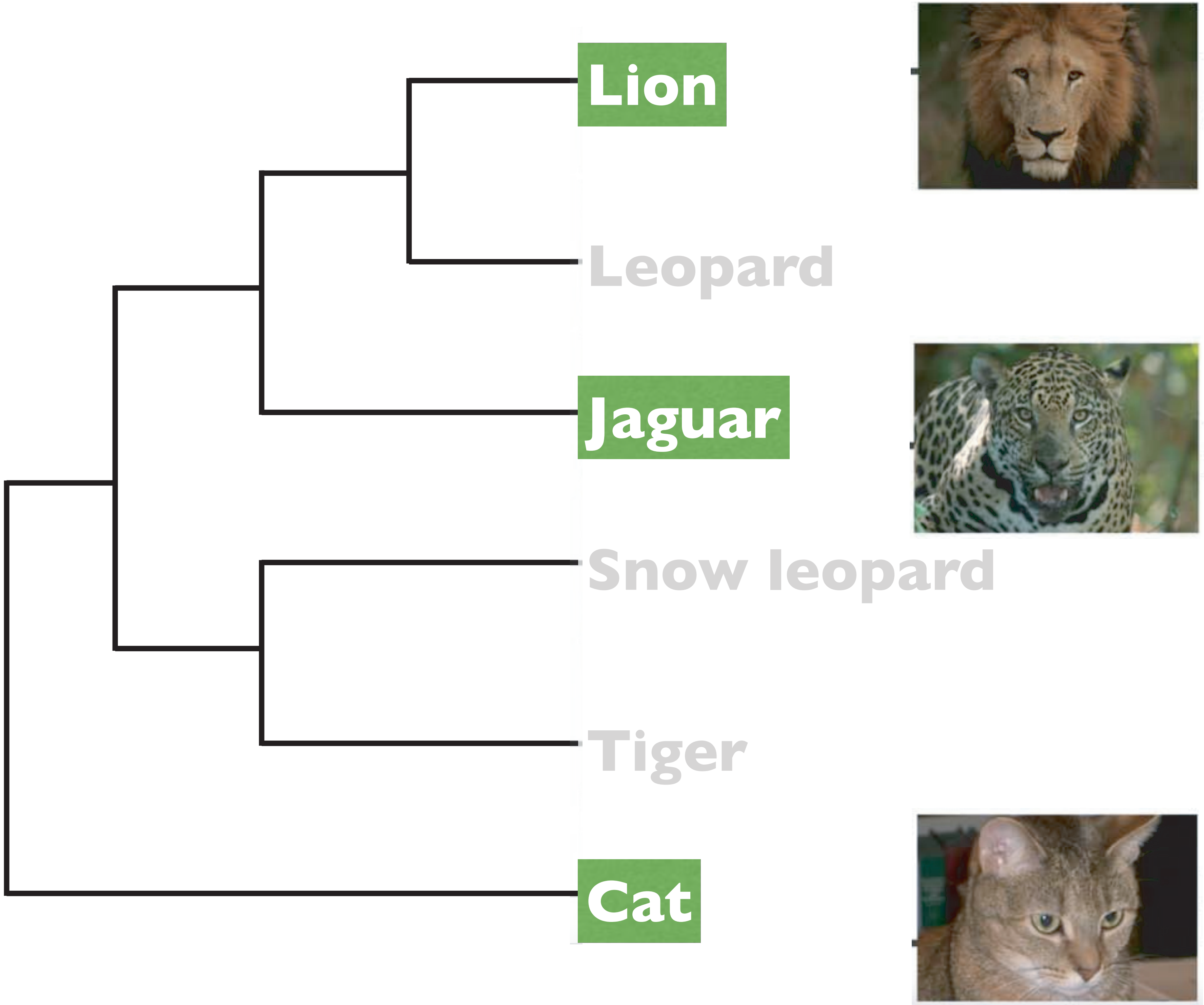
Snow leopard

Tiger

Cat

ONTOLOGICAL CLASSIFICATION OF UNSEEN ANIMALS

Ontology of great cats



Lion

Leopard

Jaguar

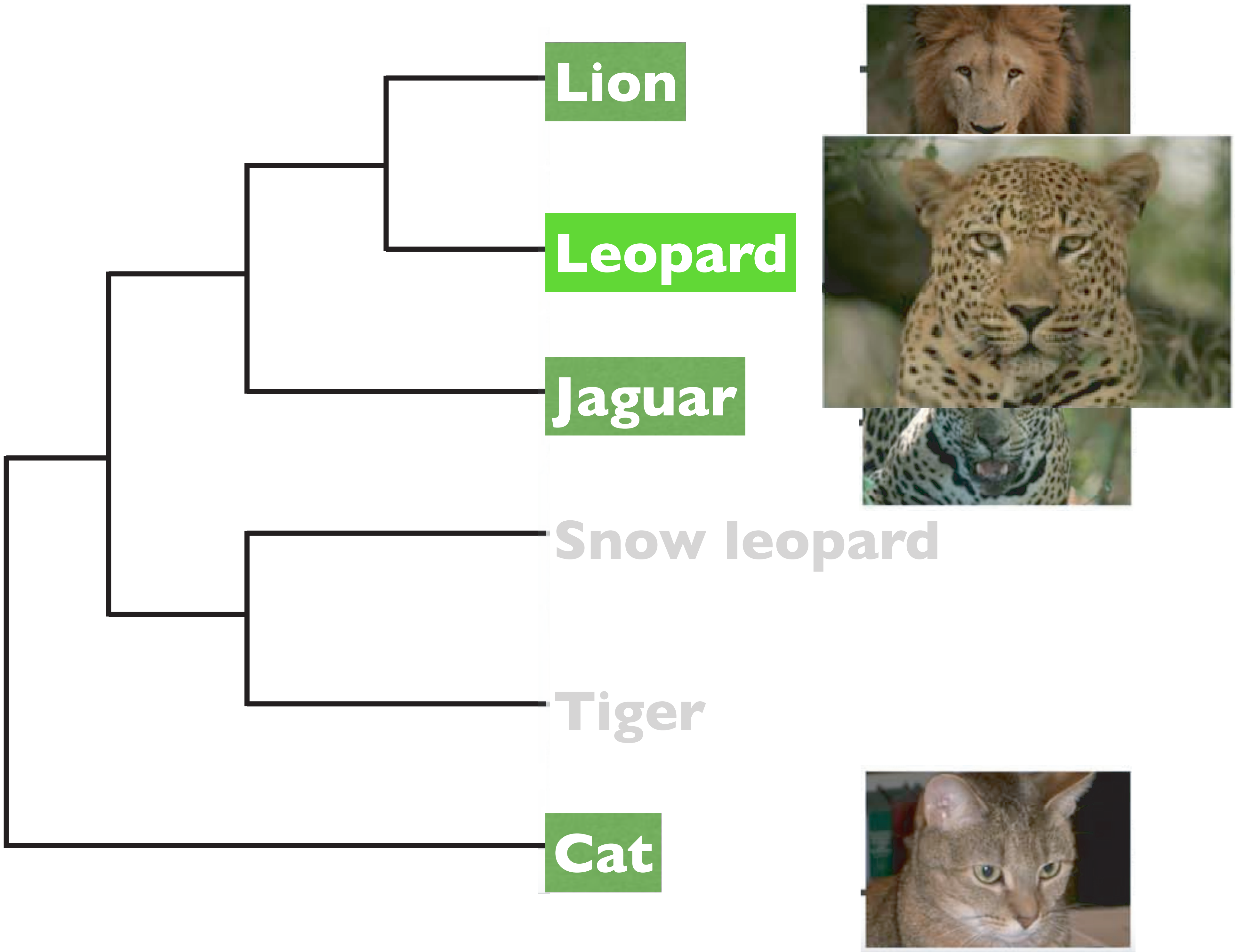
Snow leopard

Tiger

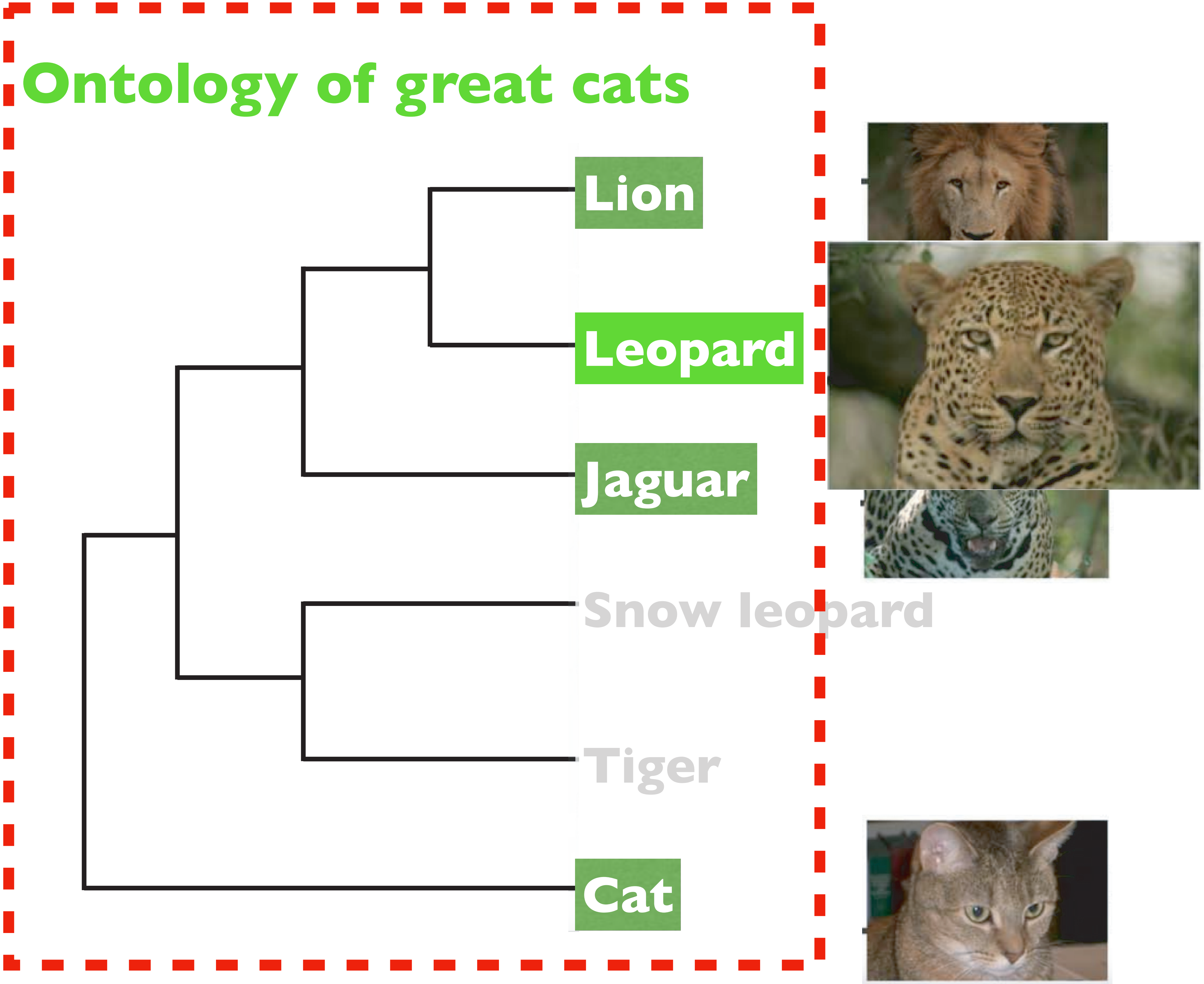
Cat

ONTOLOGICAL CLASSIFICATION OF UNSEEN ANIMALS

Ontology of great cats

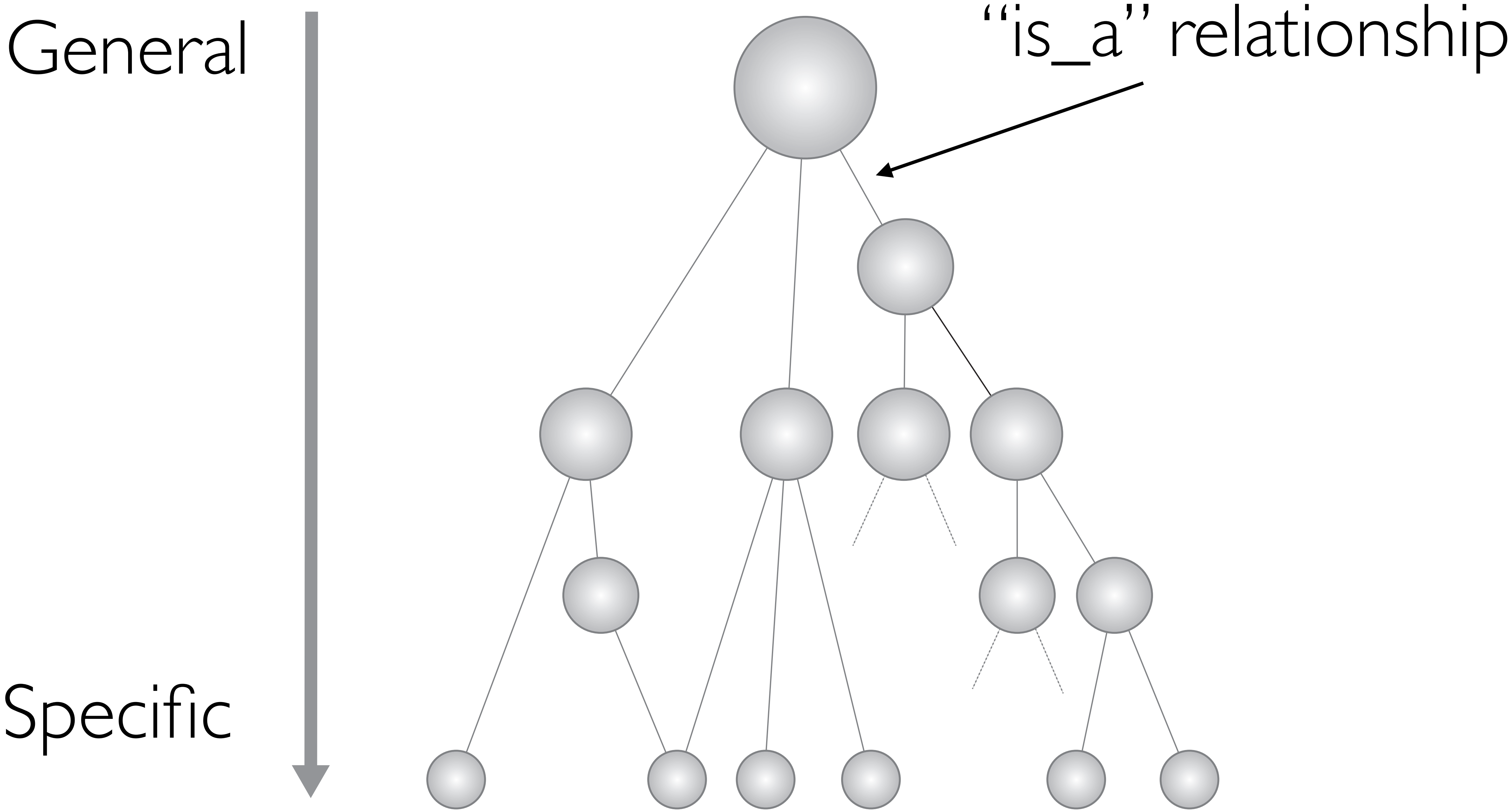


ONTOLOGICAL CLASSIFICATION OF UNSEEN ANIMALS



Ontology of great cats

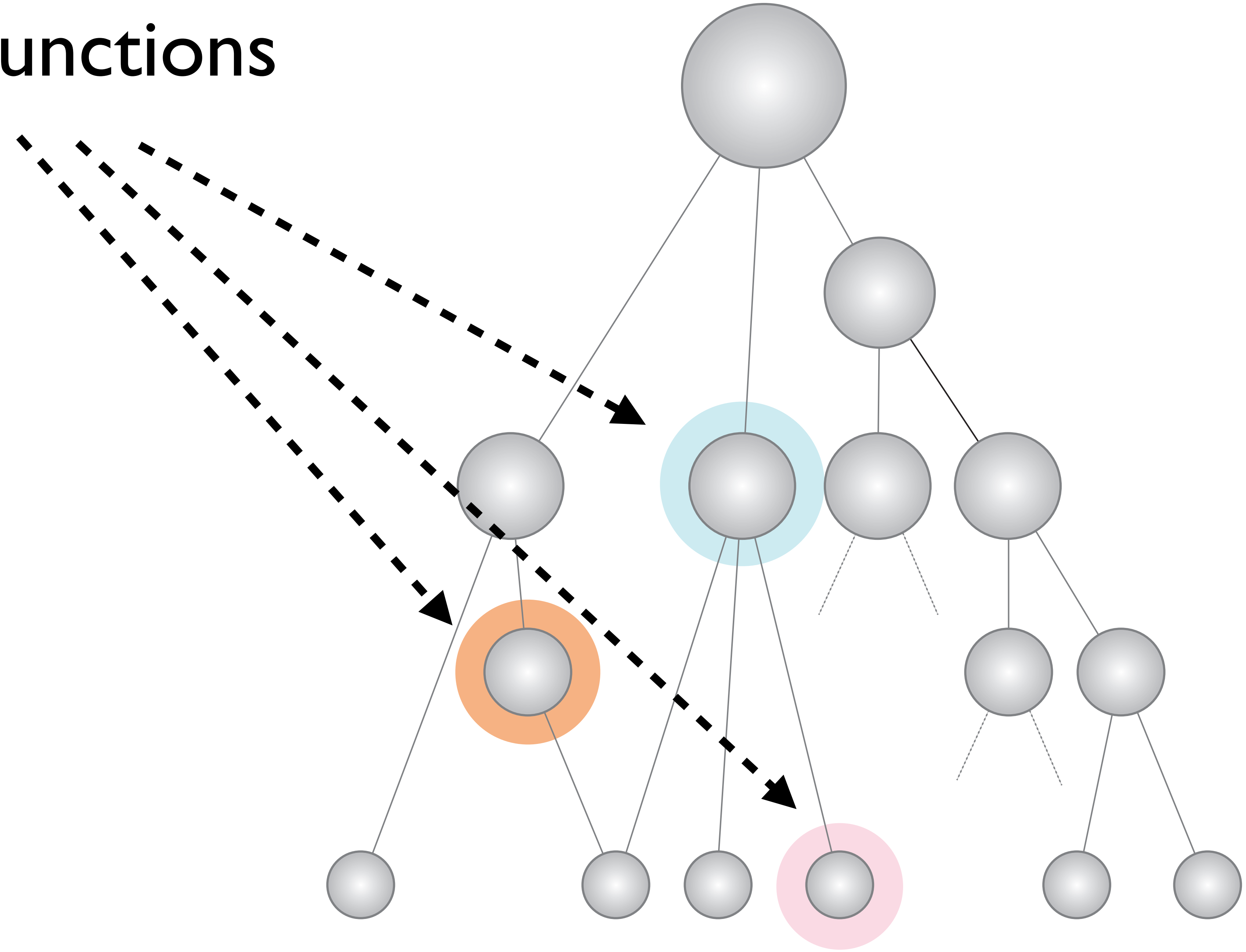
solution: use gene ontology as side information



Each node is a function. 23k functions in total.

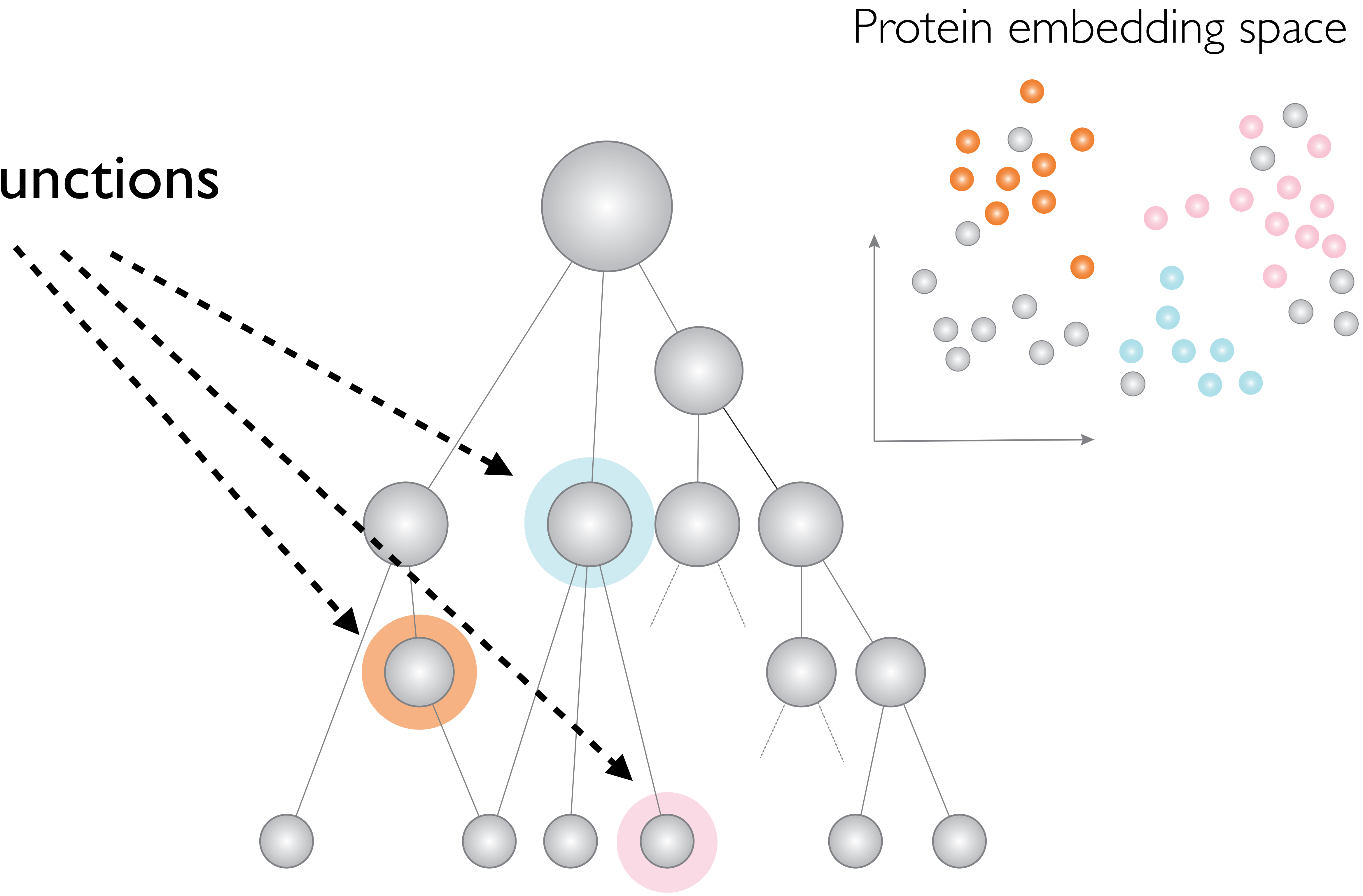
solution: use gene ontology as side information

Seen functions



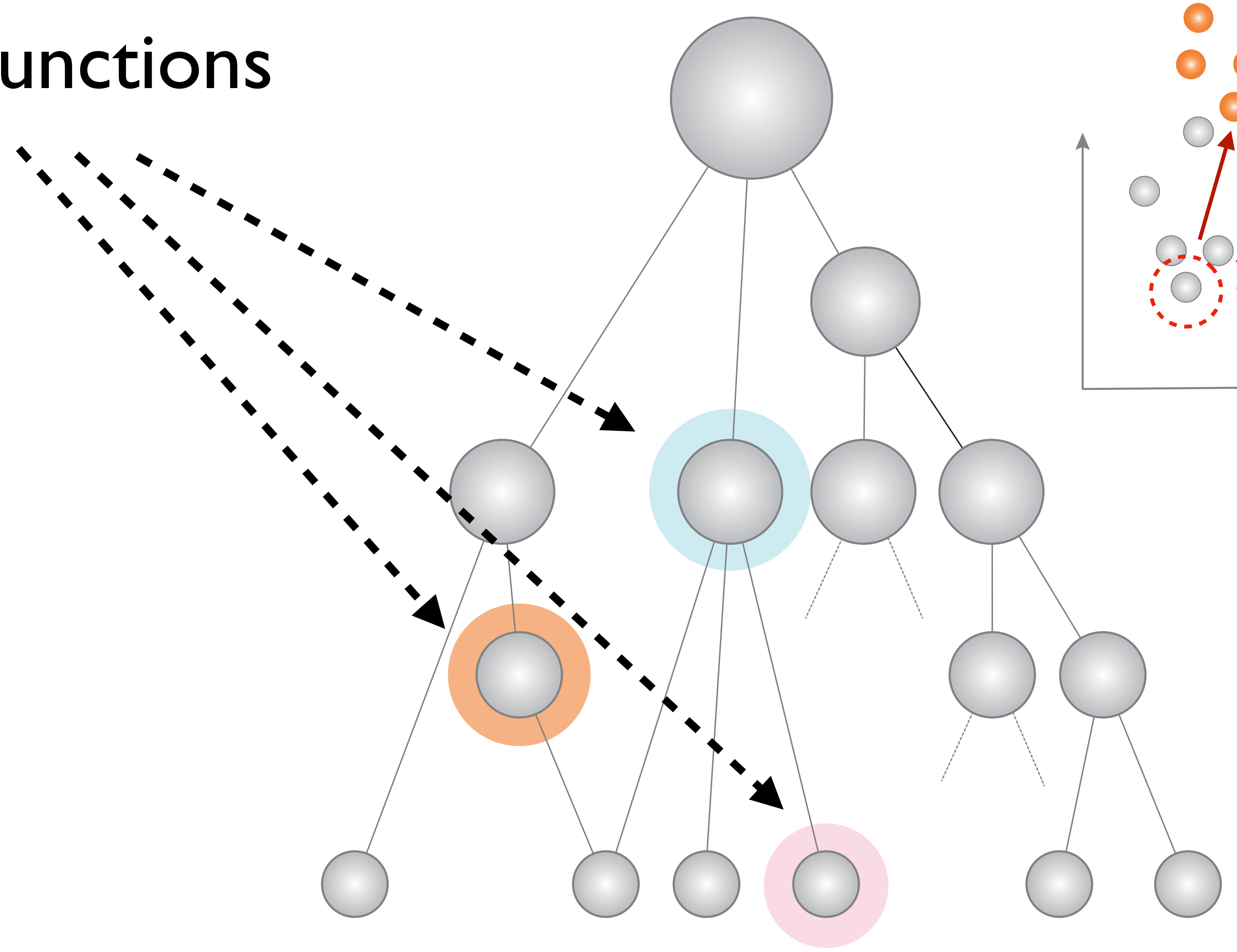
solution: use gene ontology as side information

Seen functions

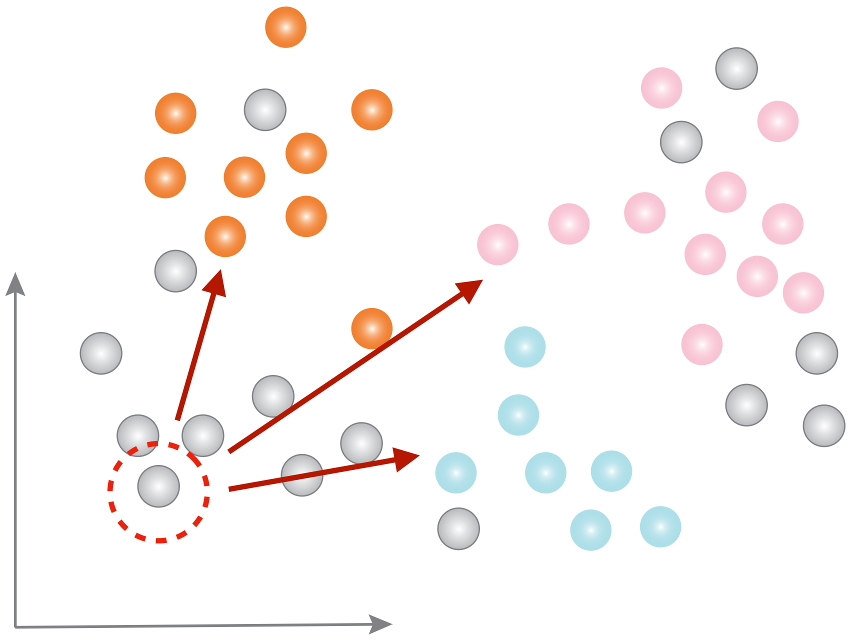


solution: use gene ontology as side information

Seen functions

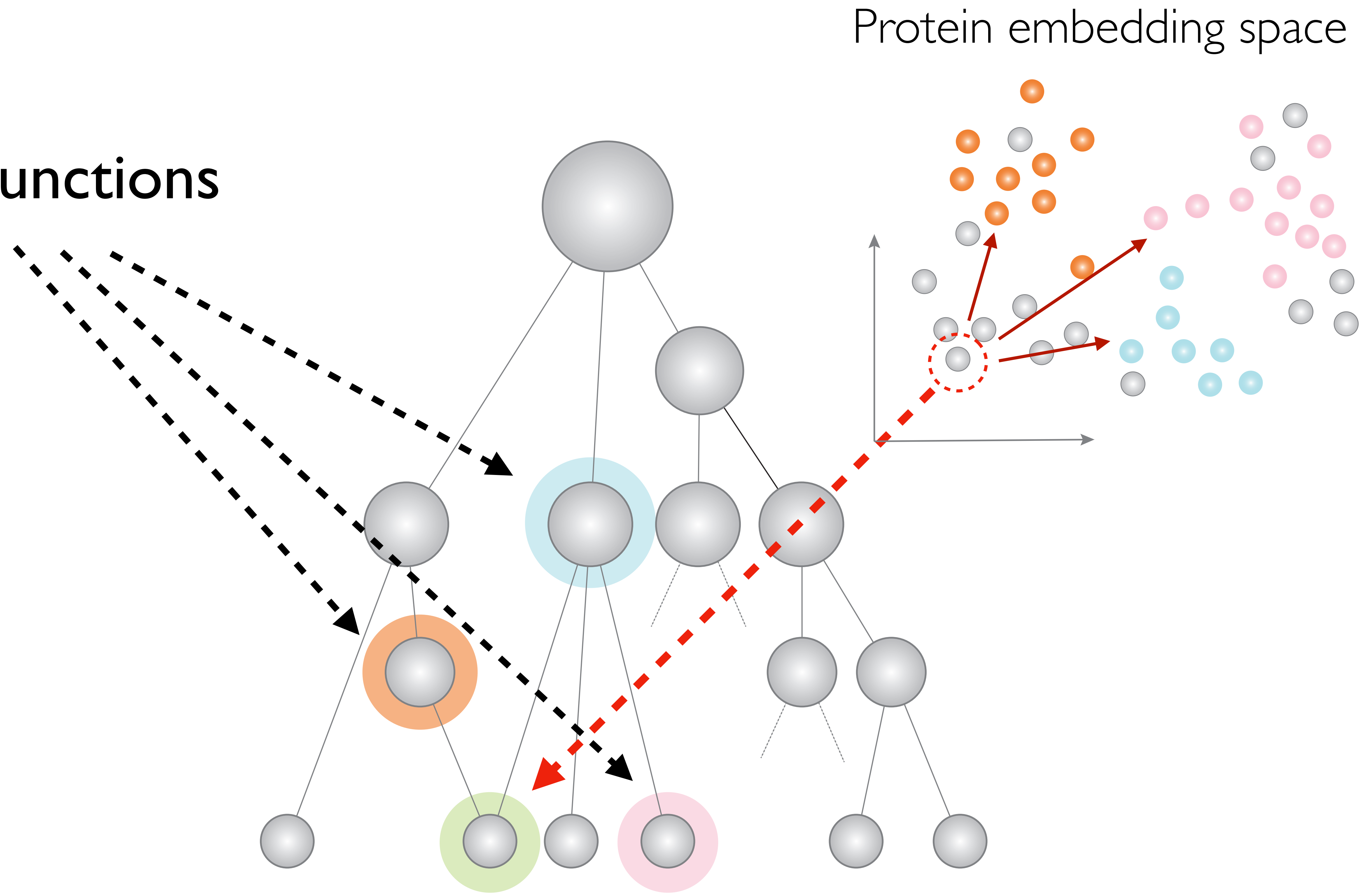


Protein embedding space



solution: use gene ontology as side information

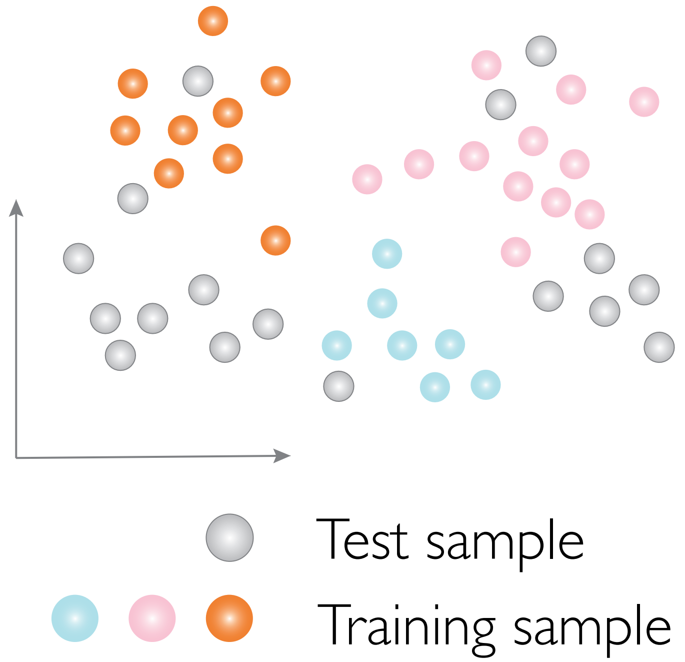
Seen functions



How to classify a test sample using class embeddings?

Input:

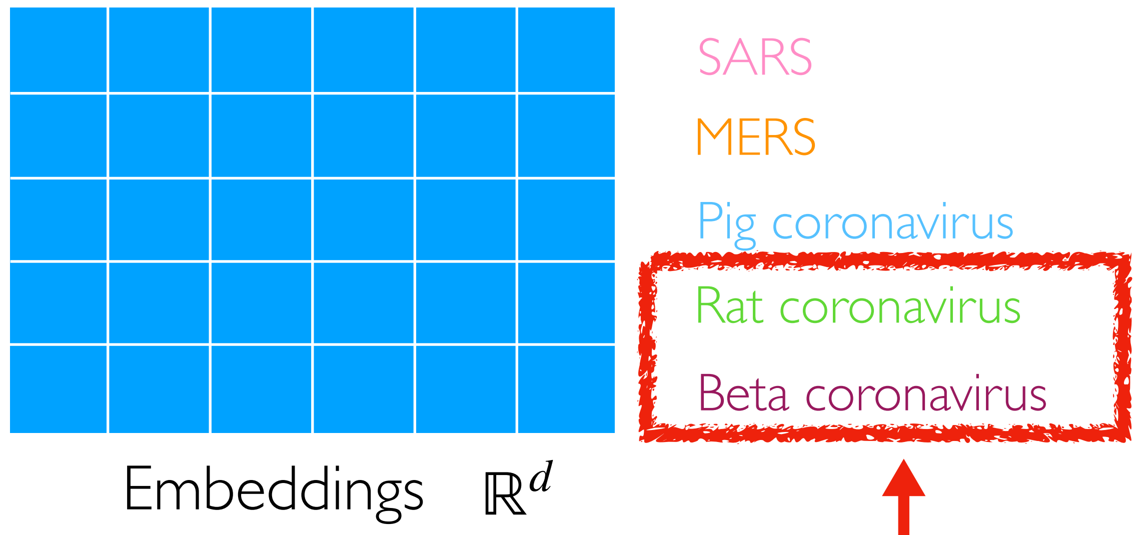
Sample features
(e.g., patient symptom lists)



● Test sample

● ● ● Training sample

Class embeddings
(e.g., disease embeddings)



SARS

MERS

Pig coronavirus

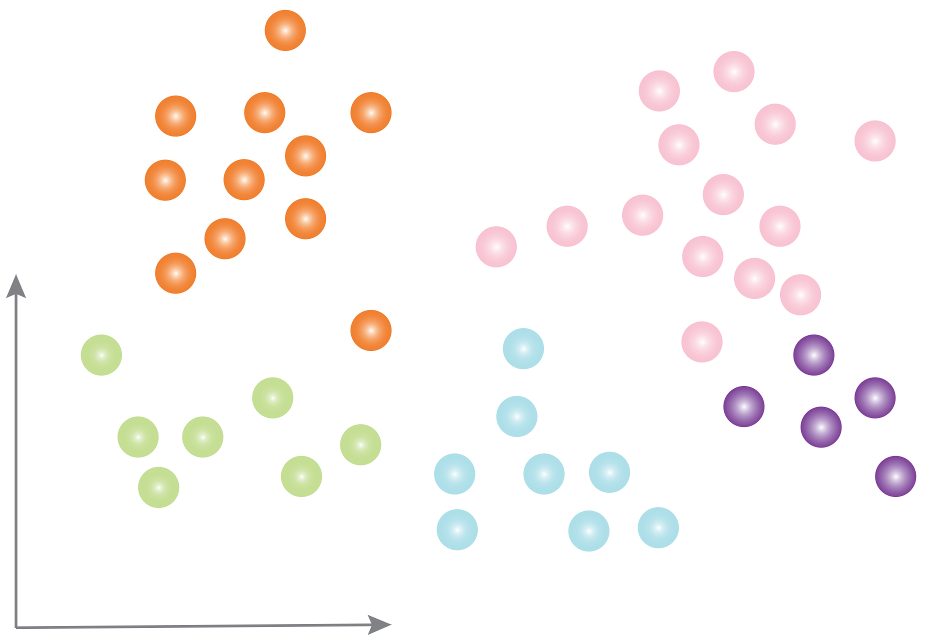
Rat coronavirus

Beta coronavirus

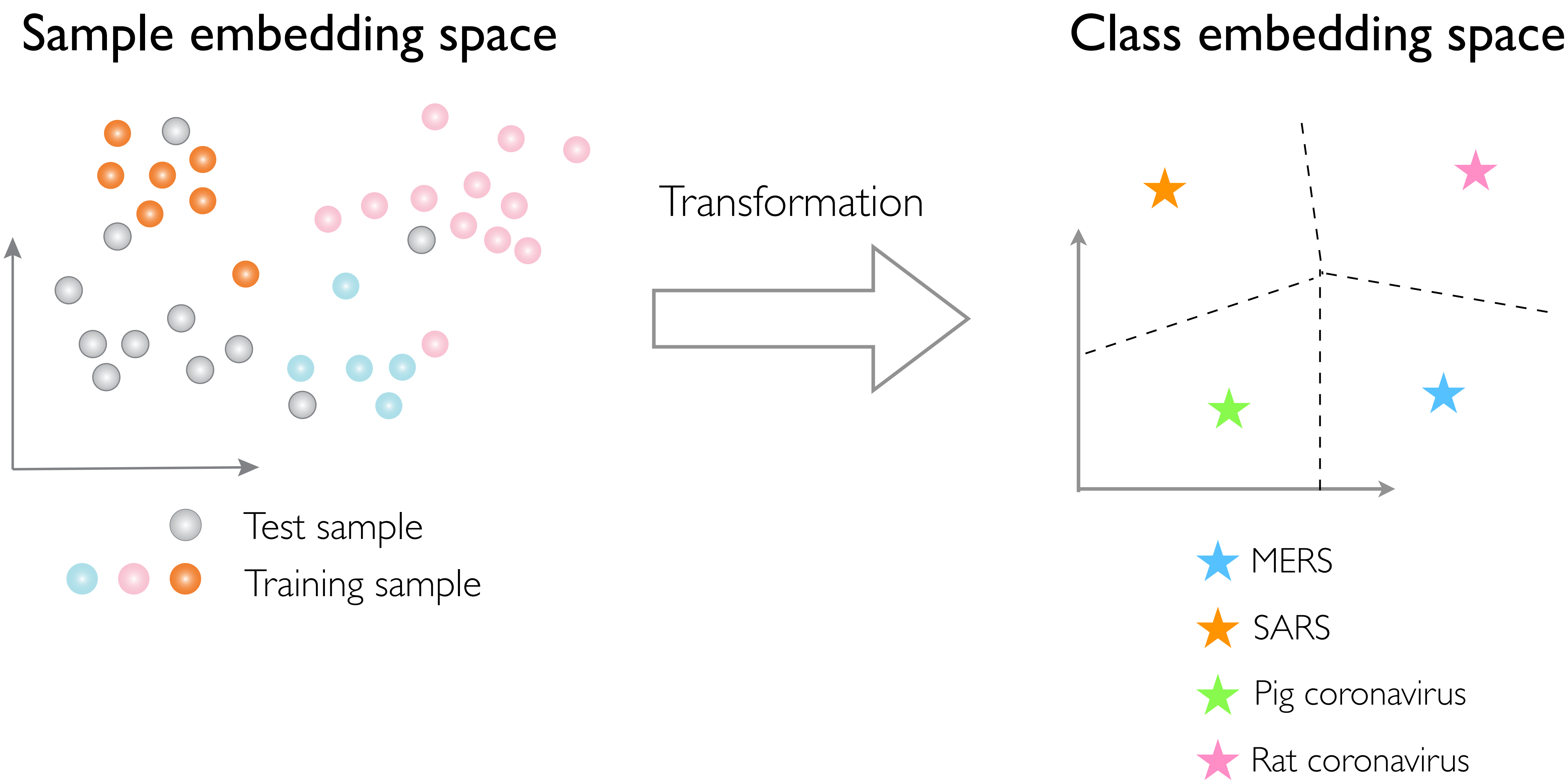
Embeddings $\mathbb{R}^d$

Never-before-seen class

Desired output:



Training: find a transformation that projects each training sample close to the embedding of its class



Sample embedding space

Class embedding space

Transformation

● Test sample

● ● ● Training sample

★ MERS

★ SARS

★ Pig coronavirus

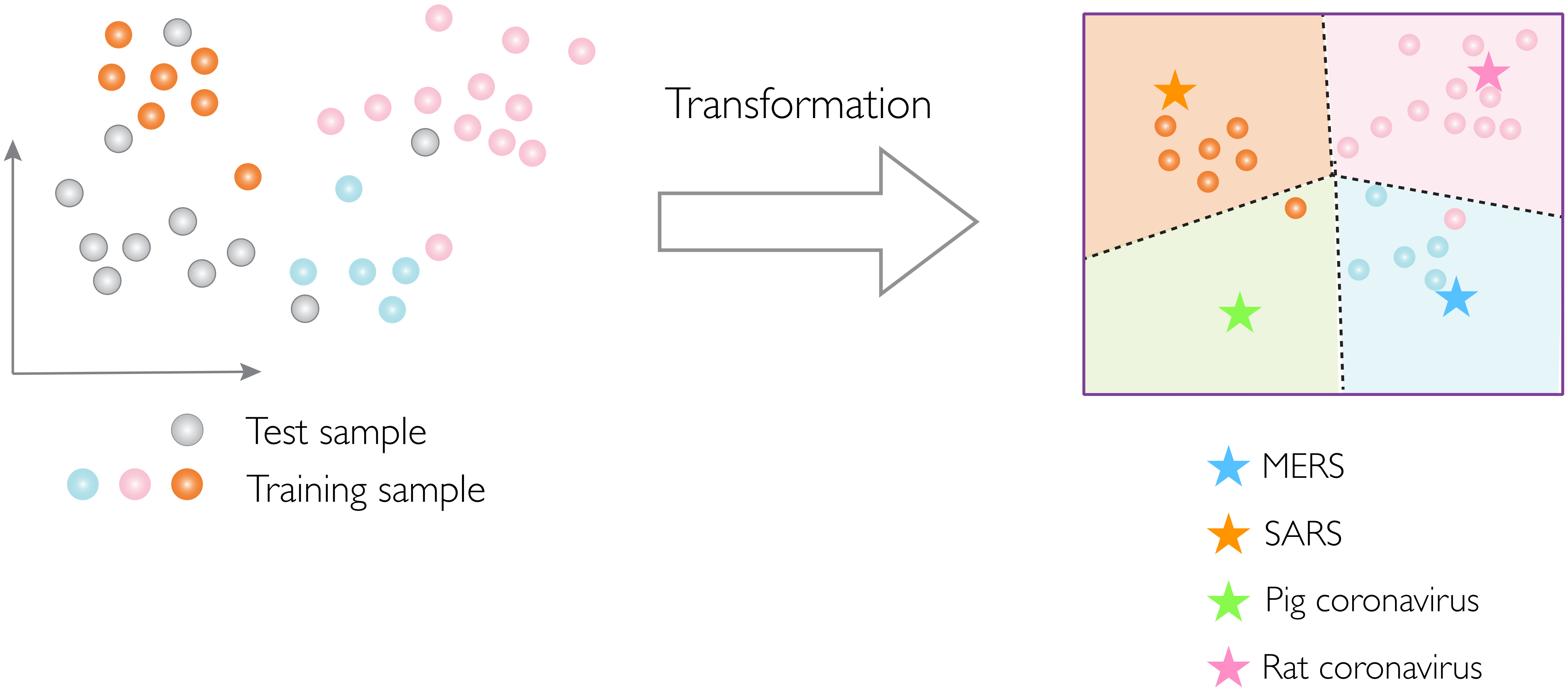
★ Rat coronavirus

Draw boundaries according to the midpoint between class embeddings

Training: find a transformation that projects each training sample close to the embedding of its class

Sample embedding space

Class embedding space

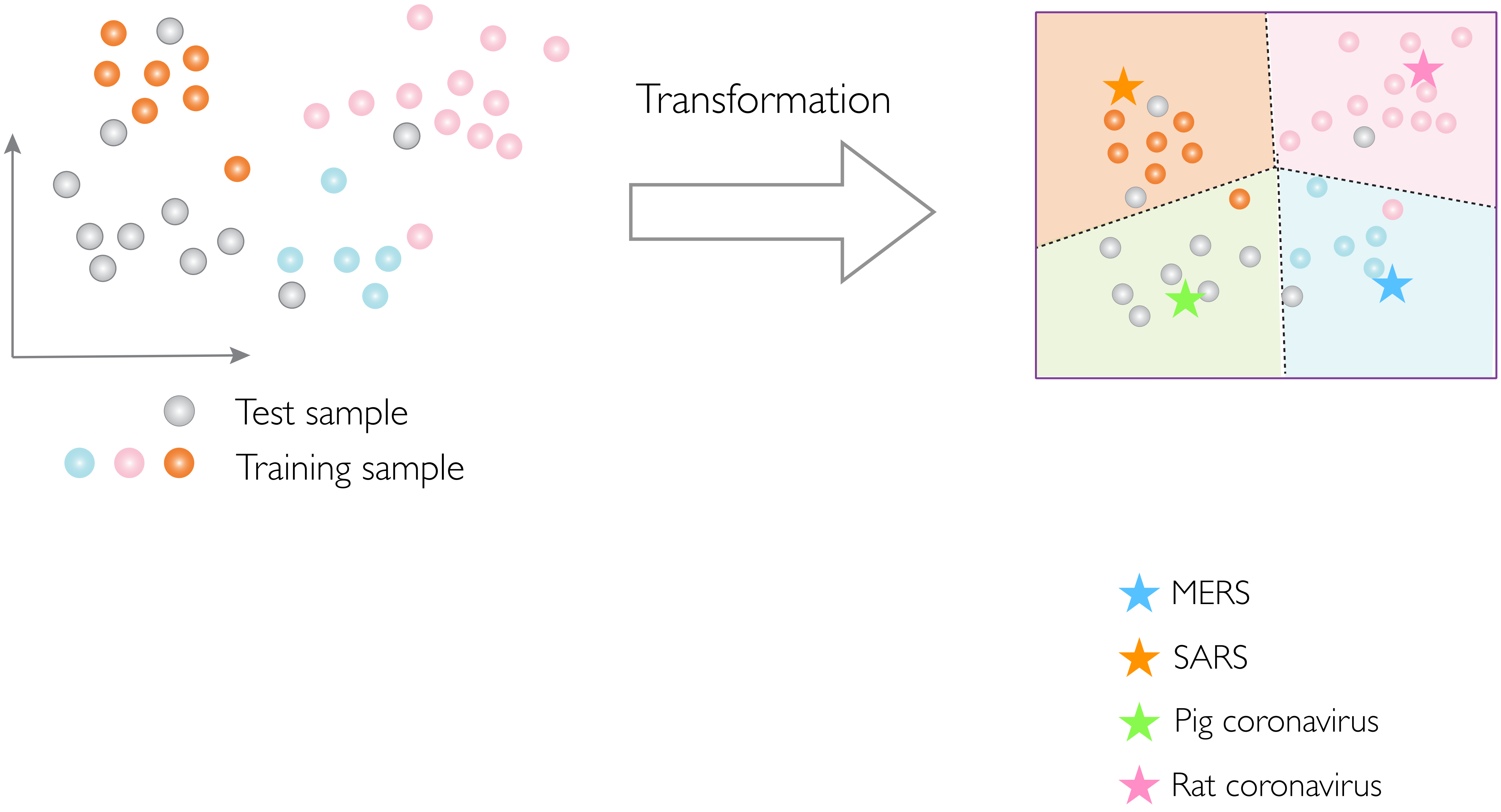


Draw boundaries according to the midpoint between class embeddings

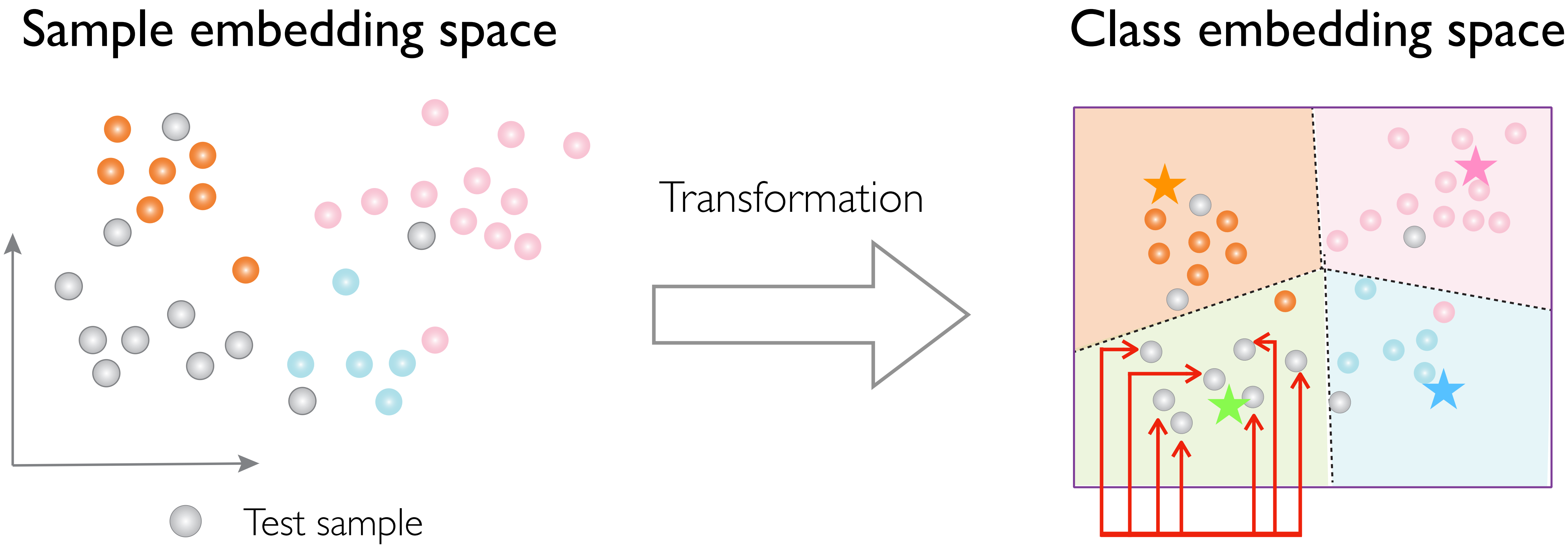
Test: project test samples using the same transformation

Sample embedding space

Class embedding space



Test: classify samples to the nearest class



Sample embedding space

Class embedding space

● Test sample

● ● ● Training sample

Key contribution: these test samples are classified into never-before-seen class

★ MERS

★ SARS

★ Pig coronavirus

★ Rat coronavirus

The Math:
use class embeddings to classify samples

Training stage:

Find transformation $\boldsymbol{W}$ that maximizes $\hat{y}_{ij}$ if training sample $\boldsymbol{i}$ belong to class $\boldsymbol{j}$ (i.e., $y_{ij} = 1$).

$$\hat{y}_{ij} = \frac{e^{\boldsymbol{f}_i \boldsymbol{W} \boldsymbol{x}_j^{\mathrm{T}}}}{\sum_k e^{\boldsymbol{f}_i \boldsymbol{W} \boldsymbol{x}_k^{\mathrm{T}}}}$$

Feature of training sample $\boldsymbol{i}$ (input)

Class embedding of class $\boldsymbol{j}$ (input)

Could be other neural network architectures (parameter)

Loss function:

$$\min_{\boldsymbol{W}} -\sum_{i=1}^{m}\sum_{j=1}^{c} y_{ij}\log \hat{y}_{ij}$$

Test stage:

Classify to the nearest class.

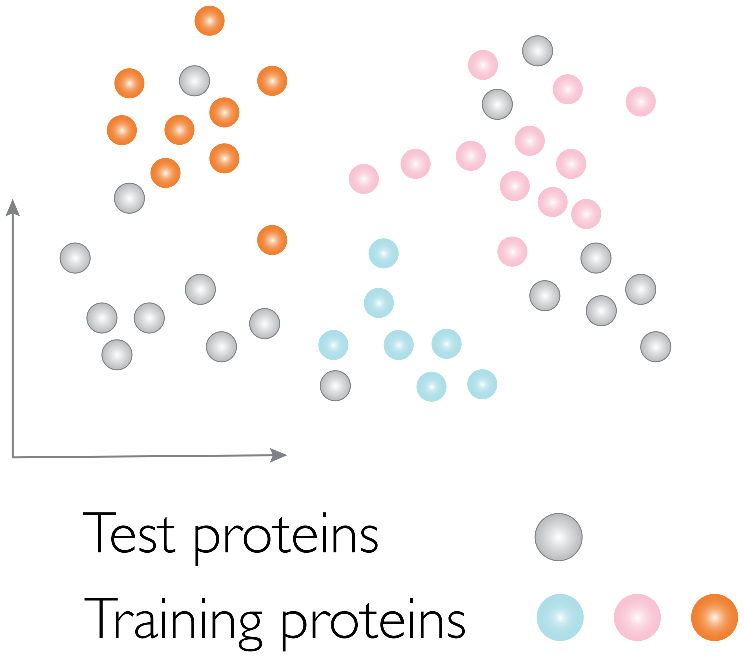
$$\mathbf{Pr}(z\,|\,j) = \frac{e^{z\boldsymbol{W}\boldsymbol{x}_j^{\mathrm{T}}}}{\sum_k e^{z\boldsymbol{W}\boldsymbol{x}_k^{\mathrm{T}}}}$$

Feature of the test sample

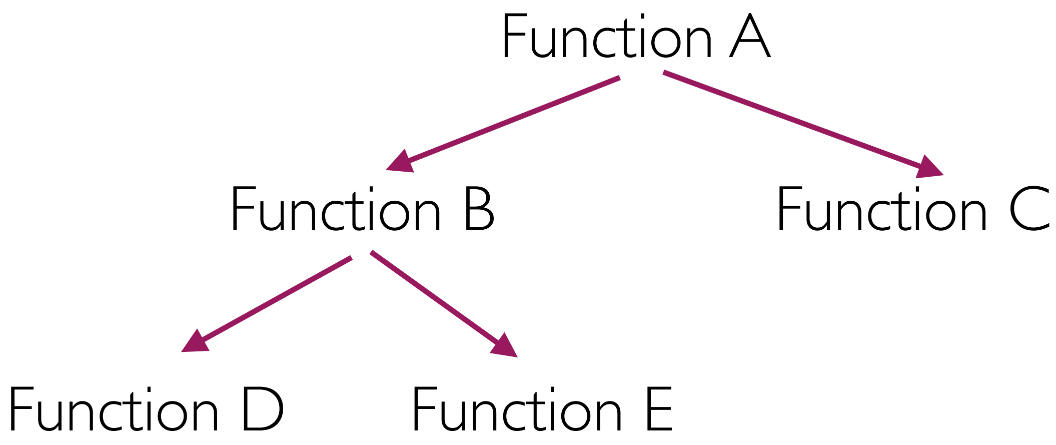
Experimental setting: classify proteins into functions

Input:

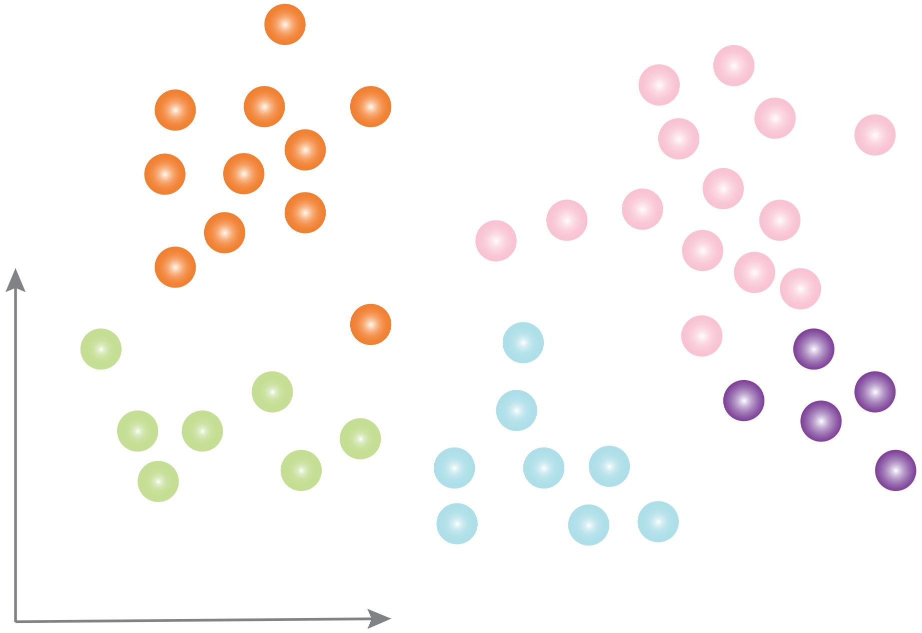
Feature representation of 60k proteins from 5 species



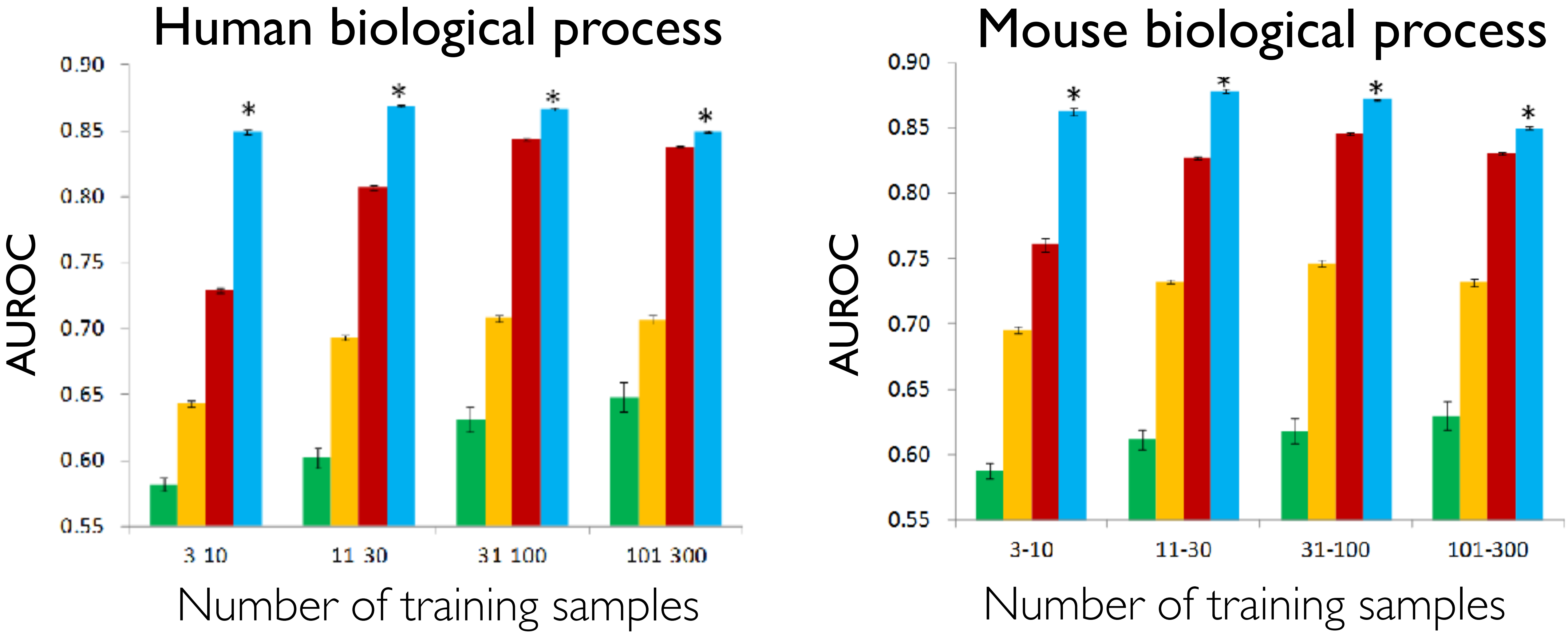
Hierarchy of 13k protein functions



Desired output:



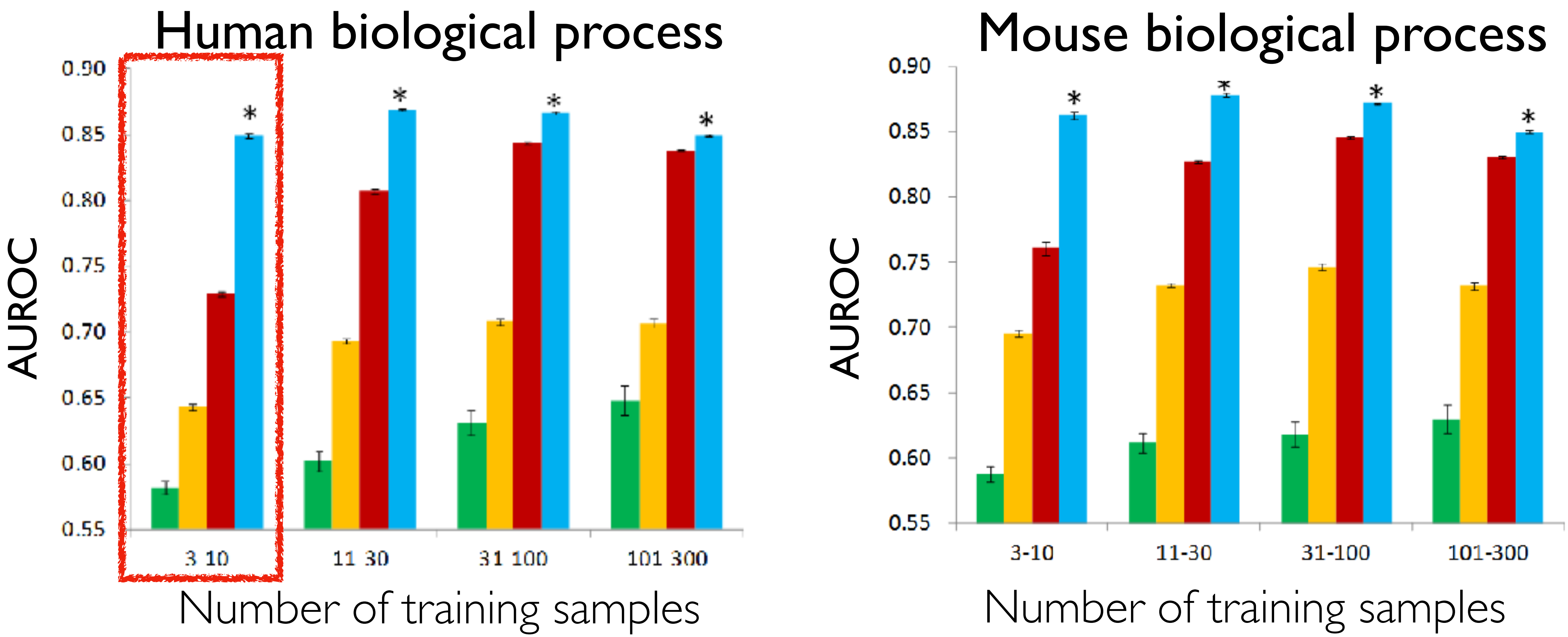
Significant improvement in few-shot classes on all five species



- 🟦 **Our method [Wang et al. 2015]**
- 🟩 Protein network embeddings without using class Hierarchy [Cho et al. 2015]
- 🟨 Heterogeneous network integration [Mostafavi and Morris, 2010]
- 🟥 Hierarchical classification of class Hierarchy [Sokolov and Ben-Hur, 2010]

Wang et al. Bioinformatics, ISMB2015 (best student paper candidate)

Significant improvement in few-shot classes on all five species



16% improvement in classes with very few samples

- 🟦 **Our method [Wang et al. 2015]**
- 🟩 Protein network embeddings without using class Hierarchy [Cho et al. 2015]
- 🟨 Heterogeneous network integration [Mostafavi and Morris, 2010]
- 🟥 Hierarchical classification of class Hierarchy [Sokolov and Ben-Hur, 2010]

Wang et al. Bioinformatics, ISMB2015 (best student paper candidate)

Tools and Resources

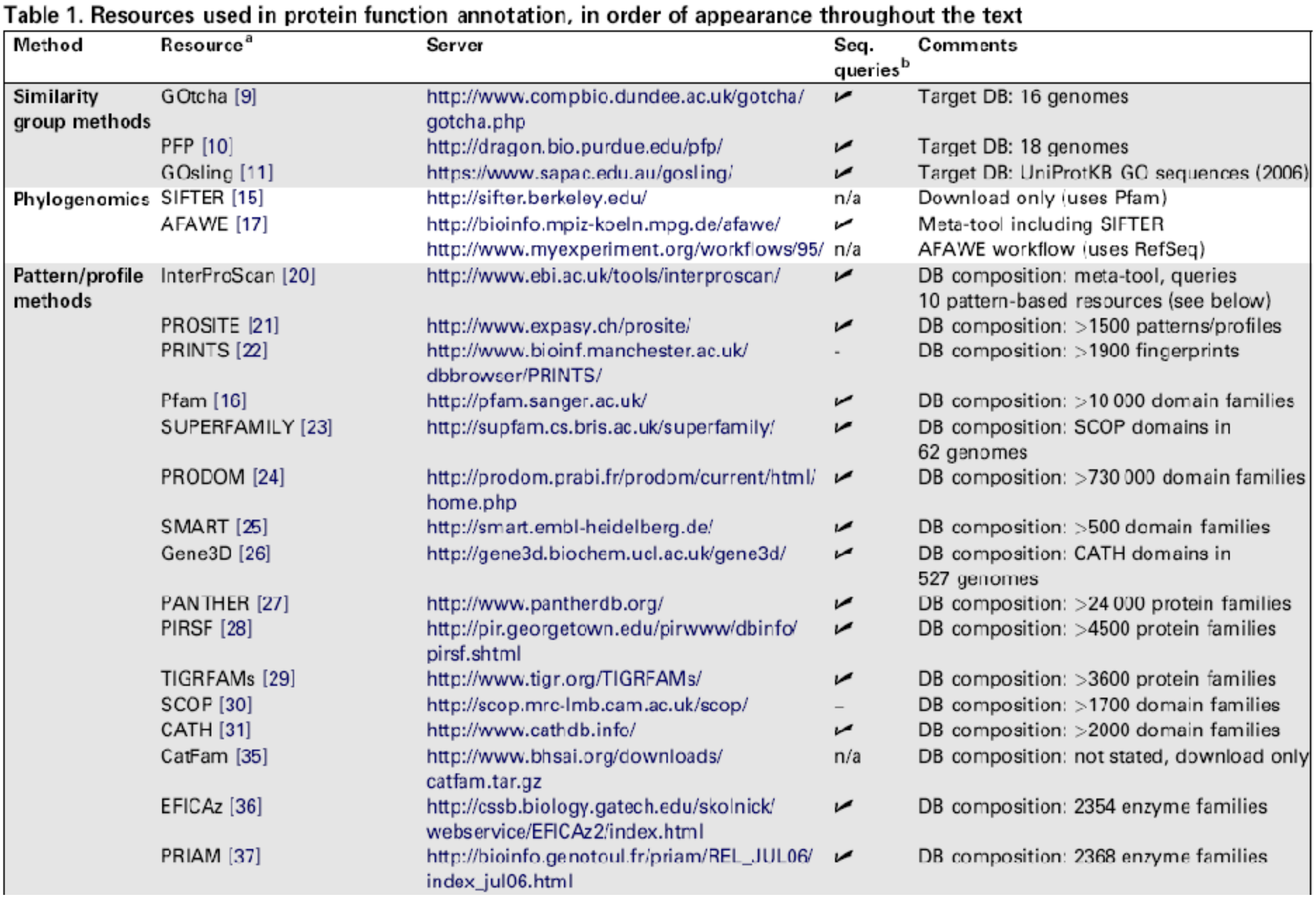
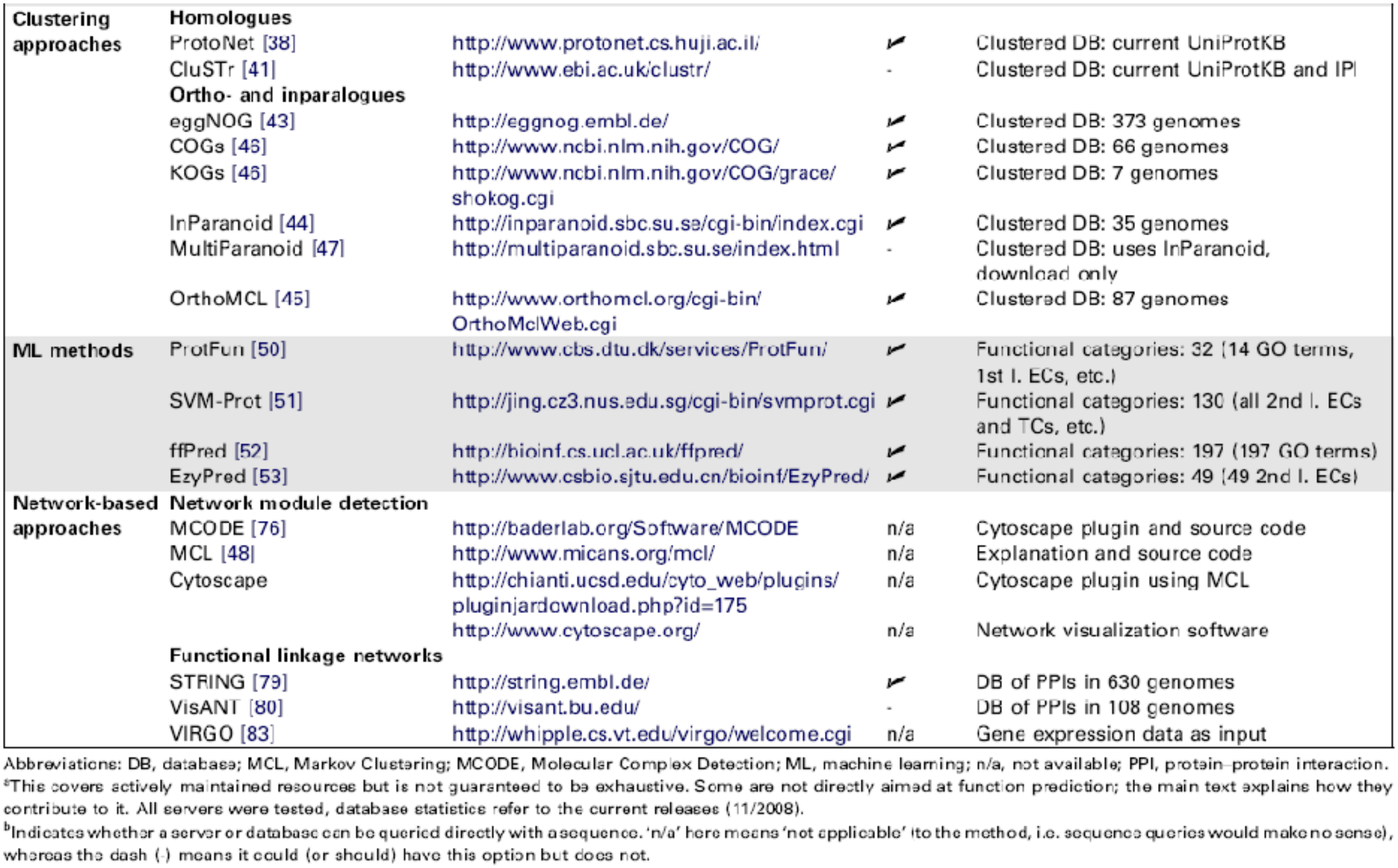


Table 1. Resources used in protein function annotation, in order of appearance throughout the text

Method	Resource[a]	Server	Seq. queries[b]	Comments
Similarity group methods	GOtcha [9]	http://www.compbio.dundee.ac.uk/gotcha/gotcha.php	✓	Target DB: 16 genomes
	PFP [10]	http://dragon.bio.purdue.edu/pfp/	✓	Target DB: 18 genomes
	GOsling [11]	https://www.sapac.edu.au/gosling/	✓	Target DB: UniProtKB GO sequences (2006)
Phylogenomics	SIFTER [15]	http://sifter.berkeley.edu/	n/a	Download only (uses Pfam)
	AFAWE [17]	http://bioinfo.mpiz-koeln.mpg.de/afawe/	✓	Meta-tool including SIFTER
		http://www.myexperiment.org/workflows/95/	n/a	AFAWE workflow (uses RefSeq)
Pattern/profile methods	InterProScan [20]	http://www.ebi.ac.uk/tools/interproscan/	✓	DB composition: meta-tool, queries 10 pattern-based resources (see below)
	PROSITE [21]	http://www.expasy.ch/prosite/	✓	DB composition: >1500 patterns/profiles
	PRINTS [22]	http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/	-	DB composition: >1900 fingerprints
	Pfam [16]	http://pfam.sanger.ac.uk/	✓	DB composition: >10 000 domain families
	SUPERFAMILY [23]	http://supfam.cs.bris.ac.uk/superfamily/	✓	DB composition: SCOP domains in 62 genomes
	PRODOM [24]	http://prodom.prabi.fr/prodom/current/html/home.php	✓	DB composition: >730 000 domain families
	SMART [25]	http://smart.embl-heidelberg.de/	✓	DB composition: >500 domain families
	Gene3D [26]	http://gene3d.biochem.ucl.ac.uk/gene3d/	✓	DB composition: CATH domains in 527 genomes
	PANTHER [27]	http://www.pantherdb.org/	✓	DB composition: >24 000 protein families
	PIRSF [28]	http://pir.georgetown.edu/pirwww/dbinfo/pirsf.shtml	✓	DB composition: >4500 protein families
	TIGRFAMs [29]	http://www.tigr.org/TIGRFAMs/	✓	DB composition: >3600 protein families
	SCOP [30]	http://scop.mrc-lmb.cam.ac.uk/scop/	–	DB composition: >1700 domain families
	CATH [31]	http://www.cathdb.info/	✓	DB composition: >2000 domain families
	CatFam [35]	http://www.bhsai.org/downloads/catfam.tar.gz	n/a	DB composition: not stated, download only
	EFICAz [36]	http://cssb.biology.gatech.edu/skolnick/webservice/EFICAz2/index.html	✓	DB composition: 2354 enzyme families
	PRIAM [37]	http://bioinfo.genotoul.fr/priam/REL_JUL06/index_jul06.html	✓	DB composition: 2368 enzyme families

Tools and Resources



Clustering approaches	**Homologues**			
	ProtoNet [38]	http://www.protonet.cs.huji.ac.il/	✓	Clustered DB: current UniProtKB
	CluSTr [41]	http://www.ebi.ac.uk/clustr/	-	Clustered DB: current UniProtKB and IPI
	Ortho- and inparalogues			
	eggNOG [43]	http://eggnog.embl.de/	✓	Clustered DB: 373 genomes
	COGs [46]	http://www.ncbi.nlm.nih.gov/COG/	✓	Clustered DB: 66 genomes
	KOGs [46]	http://www.ncbi.nlm.nih.gov/COG/grace/shokog.cgi	✓	Clustered DB: 7 genomes
	InParanoid [44]	http://inparanoid.sbc.su.se/cgi-bin/index.cgi	✓	Clustered DB: 35 genomes
	MultiParanoid [47]	http://multiparanoid.sbc.su.se/index.html	-	Clustered DB: uses InParanoid, download only
	OrthoMCL [45]	http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi	✓	Clustered DB: 87 genomes
ML methods	ProtFun [50]	http://www.cbs.dtu.dk/services/ProtFun/	✓	Functional categories: 32 (14 GO terms, 1st l. ECs, etc.)
	SVM-Prot [51]	http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi	✓	Functional categories: 130 (all 2nd l. ECs and TCs, etc.)
	ffPred [52]	http://bioinf.cs.ucl.ac.uk/ffpred/	✓	Functional categories: 197 (197 GO terms)
	EzyPred [53]	http://www.csbio.sjtu.edu.cn/bioinf/EzyPred/	✓	Functional categories: 49 (49 2nd l. ECs)
Network-based approaches	**Network module detection**			
	MCODE [76]	http://baderlab.org/Software/MCODE	n/a	Cytoscape plugin and source code
	MCL [48]	http://www.micans.org/mcl/	n/a	Explanation and source code
	Cytoscape	http://chianti.ucsd.edu/cyto_web/plugins/pluginjardownload.php?id=175	n/a	Cytoscape plugin using MCL
		http://www.cytoscape.org/	n/a	Network visualization software
	Functional linkage networks			
	STRING [79]	http://string.embl.de/	✓	DB of PPIs in 630 genomes
	VisANT [80]	http://visant.bu.edu/	-	DB of PPIs in 108 genomes
	VIRGO [83]	http://whipple.cs.vt.edu/virgo/welcome.cgi	n/a	Gene expression data as input

Abbreviations: DB, database; MCL, Markov Clustering; MCODE, Molecular Complex Detection; ML, machine learning; n/a, not available; PPI, protein–protein interaction.

[a] This covers actively maintained resources but is not guaranteed to be exhaustive. Some are not directly aimed at function prediction; the main text explains how they contribute to it. All servers were tested, database statistics refer to the current releases (11/2008).

[b] Indicates whether a server or database can be queried directly with a sequence. 'n/a' here means 'not applicable' (to the method, i.e. sequence queries would make no sense), whereas the dash (-) means it could (or should) have this option but does not.

Conclusion

- Sequence alignment is the foundation of protein function prediction
- Important databases and tools
  - KEGG, GO
  - Gene name mapping
  - NCBI reference sequence
  - CAFA

Acknowledgement

- Part of the slides are from
  - Dr. Jianlin Cheng's lecture on Analysis and Prediction of Protein Function
  - http://calla.rnet.missouri.edu/cheng/cheng_research.html
  - EMBL-EBI industry workshop 2016
  - https://www.ebi.ac.uk/about/events/2016/embl-ebi-biocuration-2016