

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

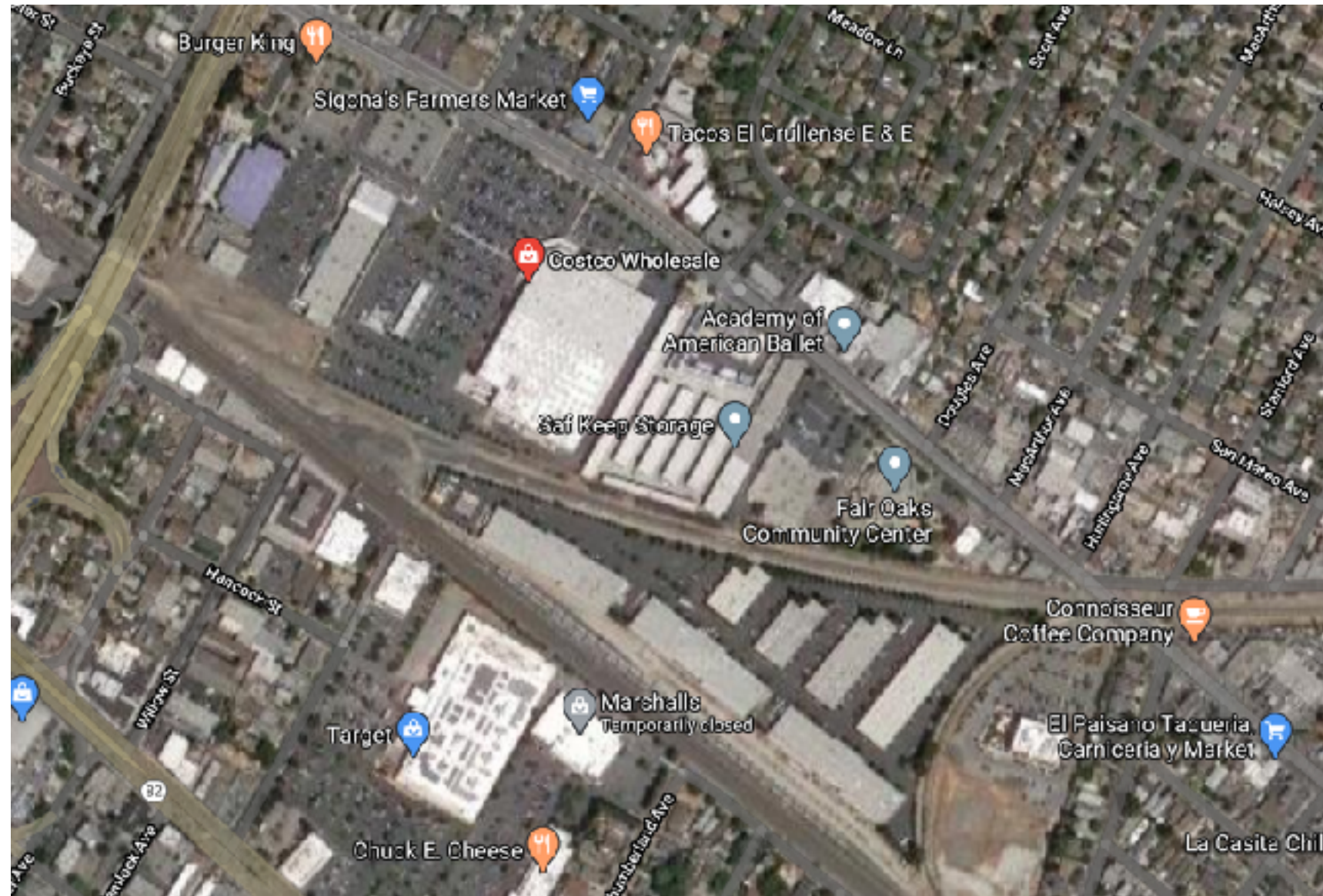
What is protein function prediction?

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

Sequence

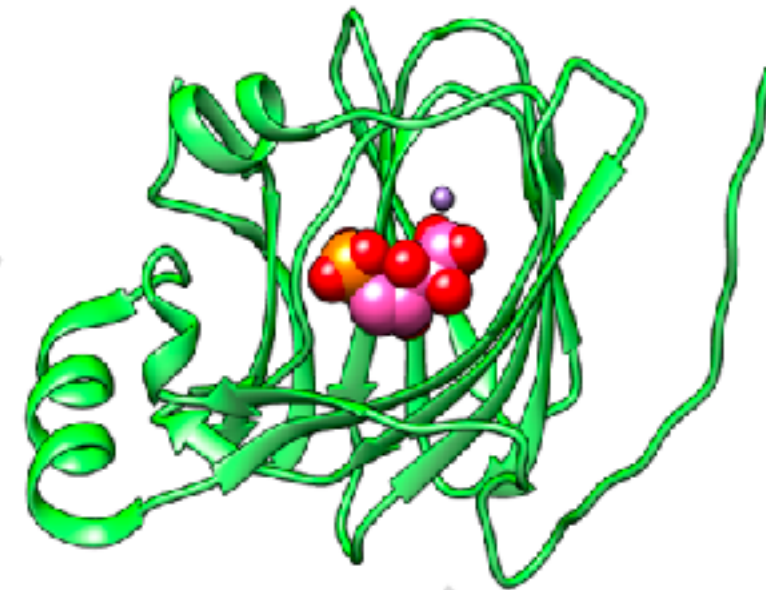
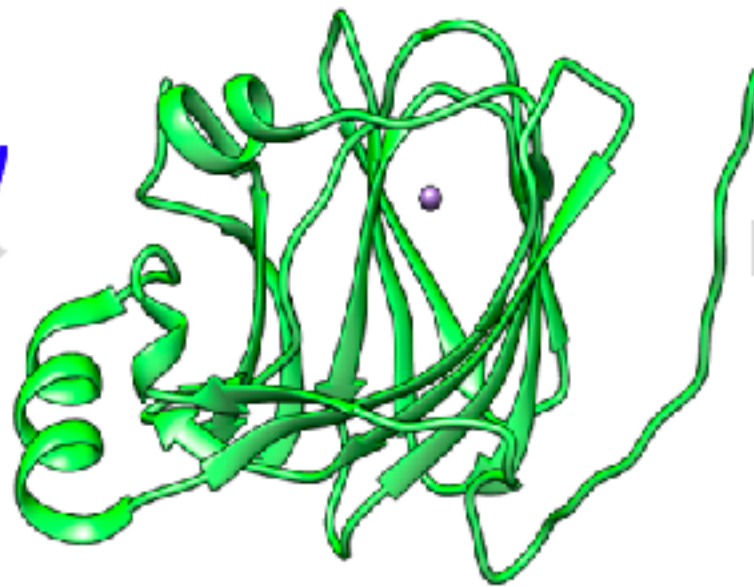
Structure

Function

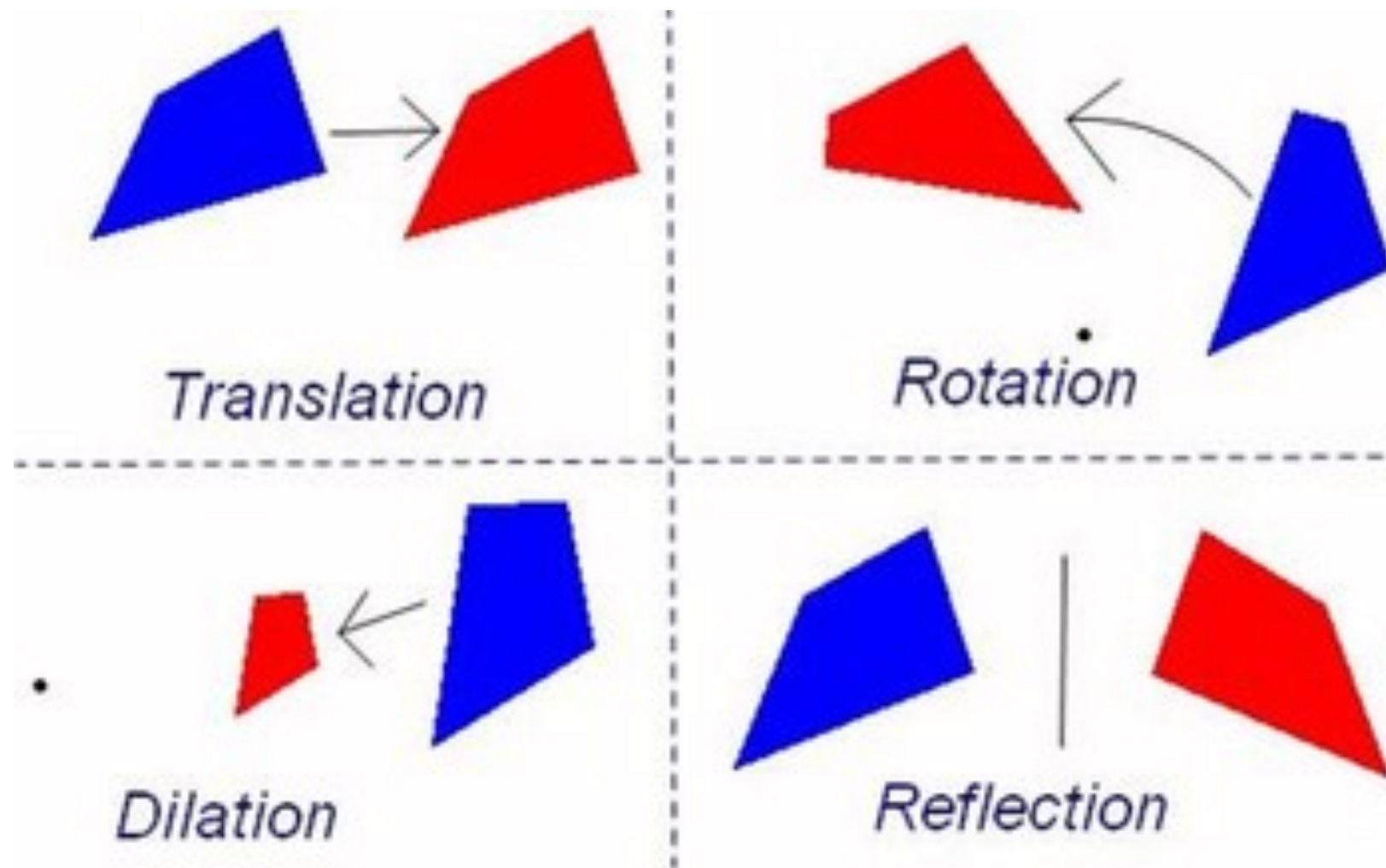
MKIAETQLFQHDI
PVVNGPYRTASGD
VWSLTITIVKIIA
EDGTIGWGETCPV
GPTYAEAHAGGAL
AALEVLASGLAGA
EALPLPLHTRMDS
LL...

Modeling

AI



What we want to maintain?



Translation equivariance: $X + g$

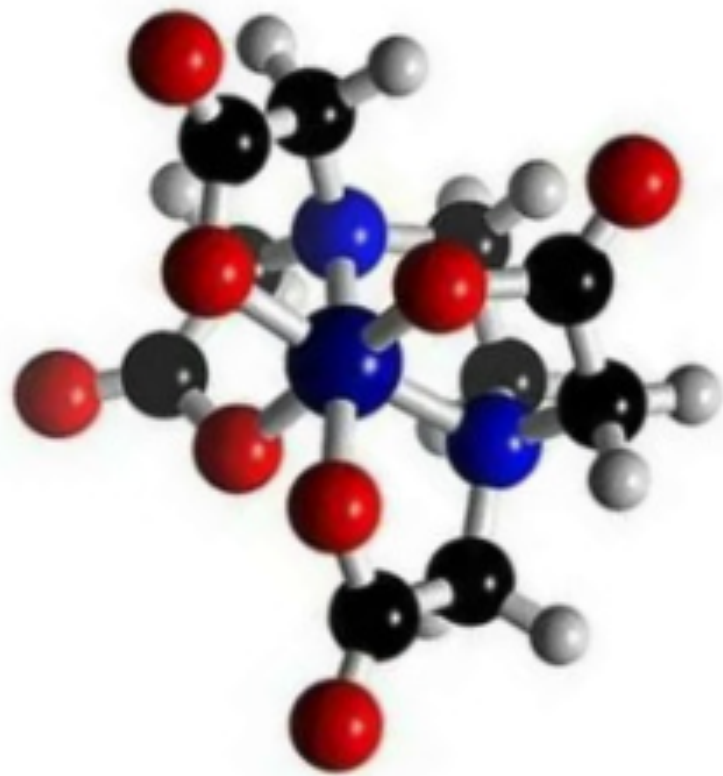
Rotation equivariance: QX

The embedding of the protein remains the same for any g and any Q .

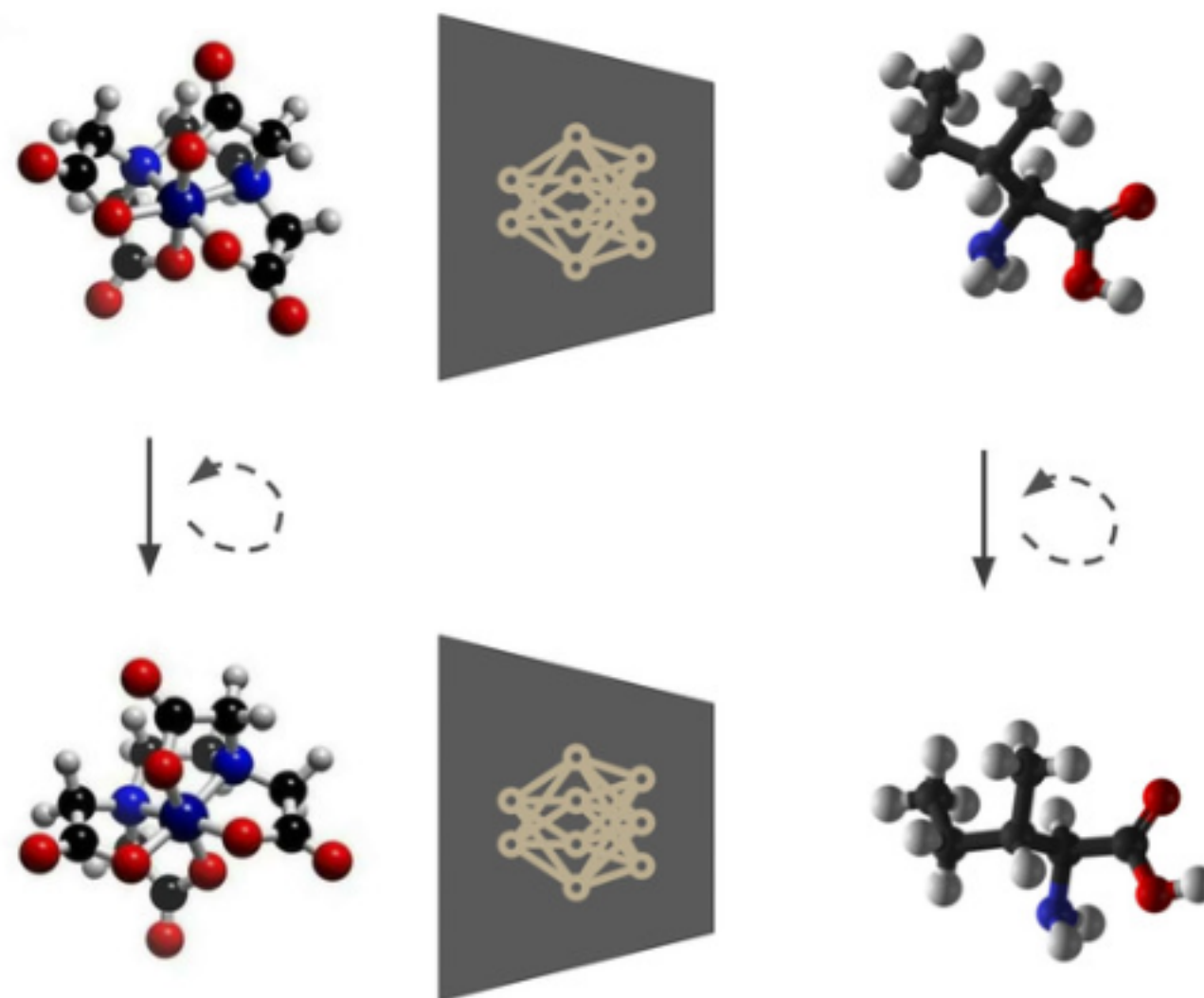
translation, rotation and reflection invariance



translation, rotation and reflection invariance



translation, rotation and reflection equivariance



Predict the binding site between a protein and a compound

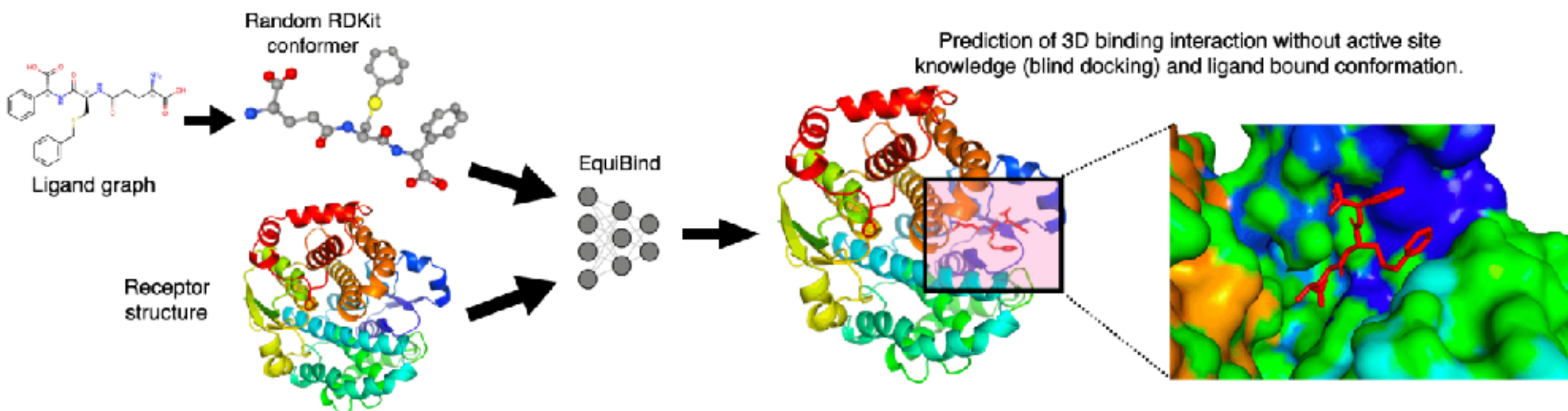


Figure 1. High-level overview of the structural drug binding problem tackled by EQUIBIND.

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 Encyclopedia 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 Encyclopedia 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 ▾

Search

[Help](#)[» Japanese](#)**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 \(Target\)](#)
[Antinfectives](#)**Organism-specific entry points****KEGG Organisms**Enter org code(s) [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 ▾

Search

[Help](#)[» Japanese](#)**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]

[Pathway](#)
[Brite](#)
[Brite table](#)
[Module](#)
[Network](#)
[KO \(Function\)](#)
[Organism](#)
[Virus](#)
[Compound](#)
[Disease \(ICD\)](#)
[Drug \(ATC\)](#)
[Drug \(Target\)](#)
[Antinfectives](#)**KEGG MEDICUS** [Health information resource](#) [[Drug labels search](#)]**Organism-specific entry points****KEGG Organisms** Enter org code(s) [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 DISEASE Database

Diseases viewed as perturbed states of the molecular system

» [Japanese](#)

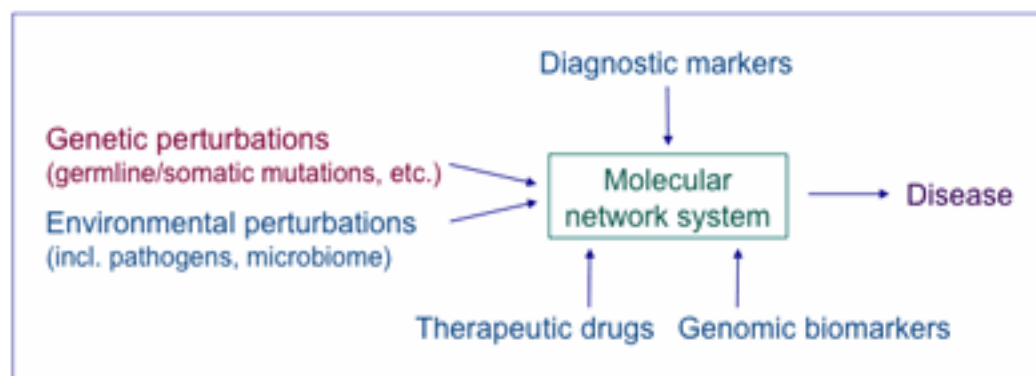
Menu **PATHWAY** **BRITE** **NETWORK** **DISEASE** **DRUG** **ENVIRON** **Pathogen** **Virus** **MEDICUS**

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

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

[Top](#)

[DISEASE \(31\)](#)
[DRUG \(116\)](#)
[DGROUP \(0\)](#)
[ENVIRON \(0\)](#)
[COMPOUND \(0\)](#)

1 to 31 of 31

Entry	Name	Description	Category	Pathway	Gene
H00242	Liddle syndrome	Liddle syndrome (LIDS) 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	(LIDS1) SCN1B [HSA:6336] [KO:K04825] (LIDS2) SCN1G [HSA:6340] [KO:K04827] (LIDS3) SCN1A [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:6336] [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) JHH [HSA:3549] [KO:K11969] (BDA1C, BDA2, BDC) GDF5 [HSA:8200] [KO:K04664] (BDA2) BMPR1B [HSA:658] [KO:K04664]

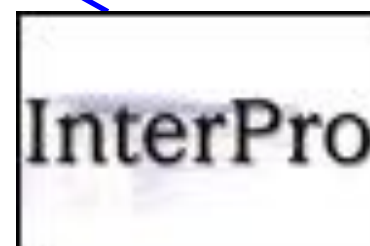
Three widely used databases for project functions

- Enzyme Commission (EC), Transporter Classification (TC)
- Kyoto Encyclopedia 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

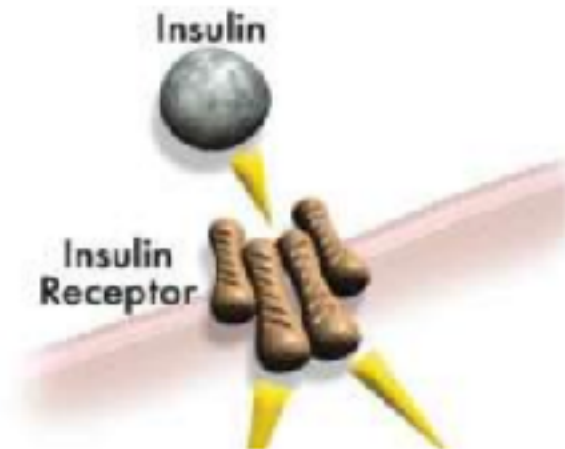


FlyBase



1. Molecular Function

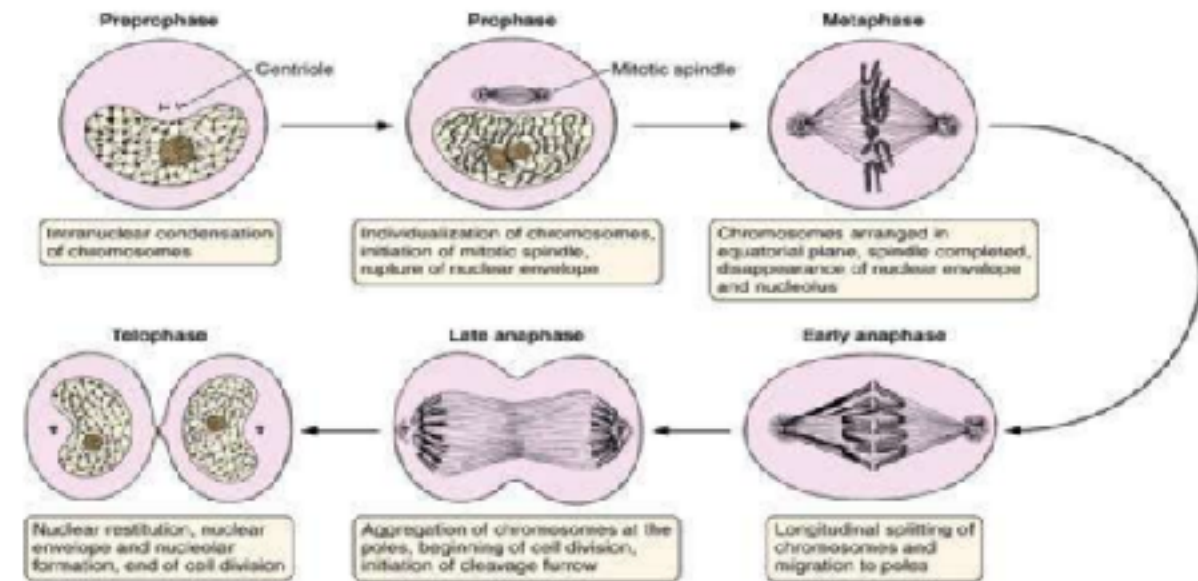
An elemental activity or task or job



- protein kinase activity
- insulin receptor activity

2. Biological Process

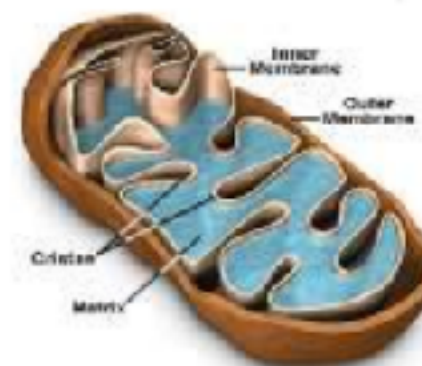
A commonly recognized series of events



- cell division

3. Cellular Component

Where a gene product is located

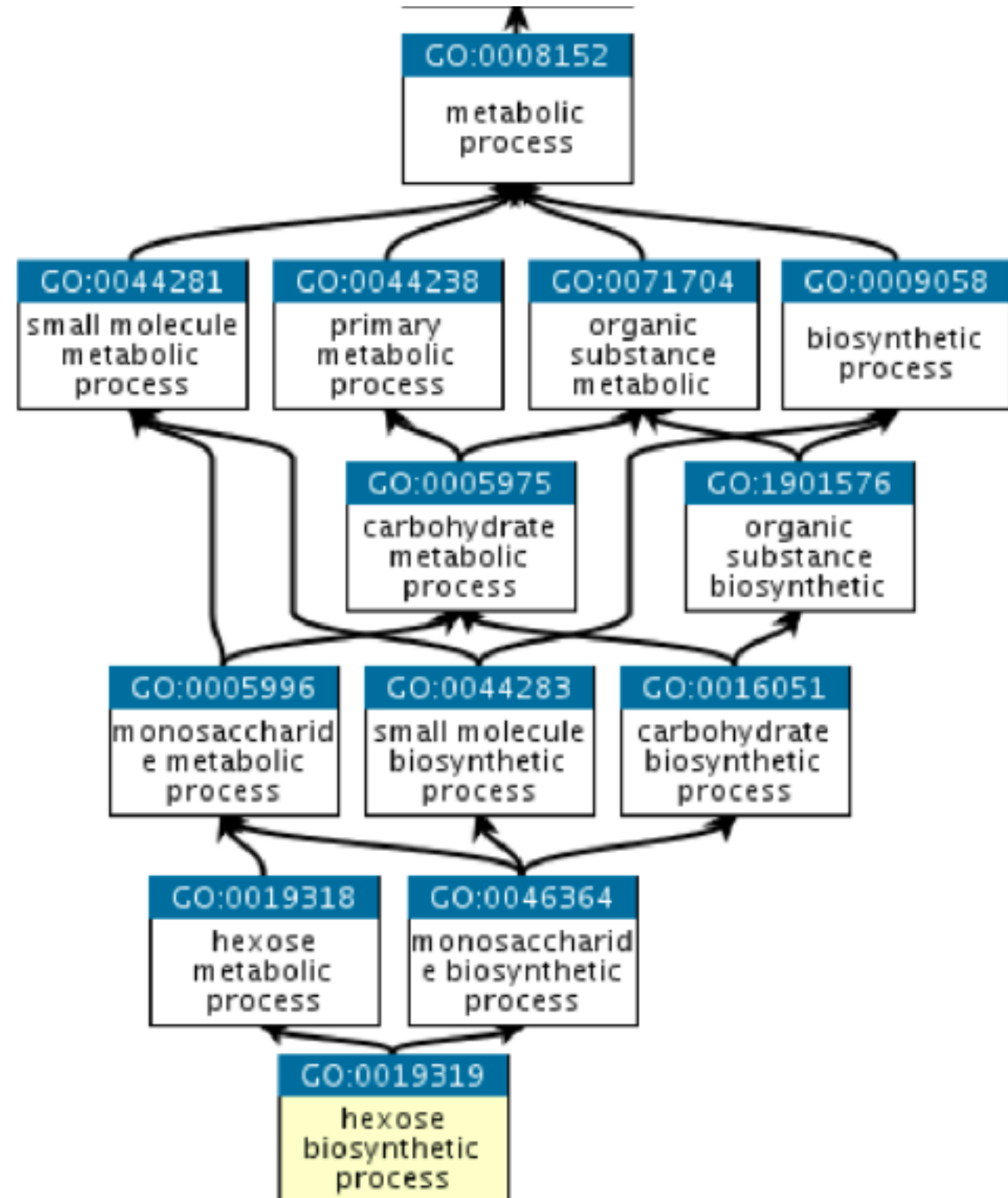


- mitochondrion
- mitochondrial matrix
- mitochondrial inner membrane

Gene Ontology: A directed acyclic graph

Less specific

More specific



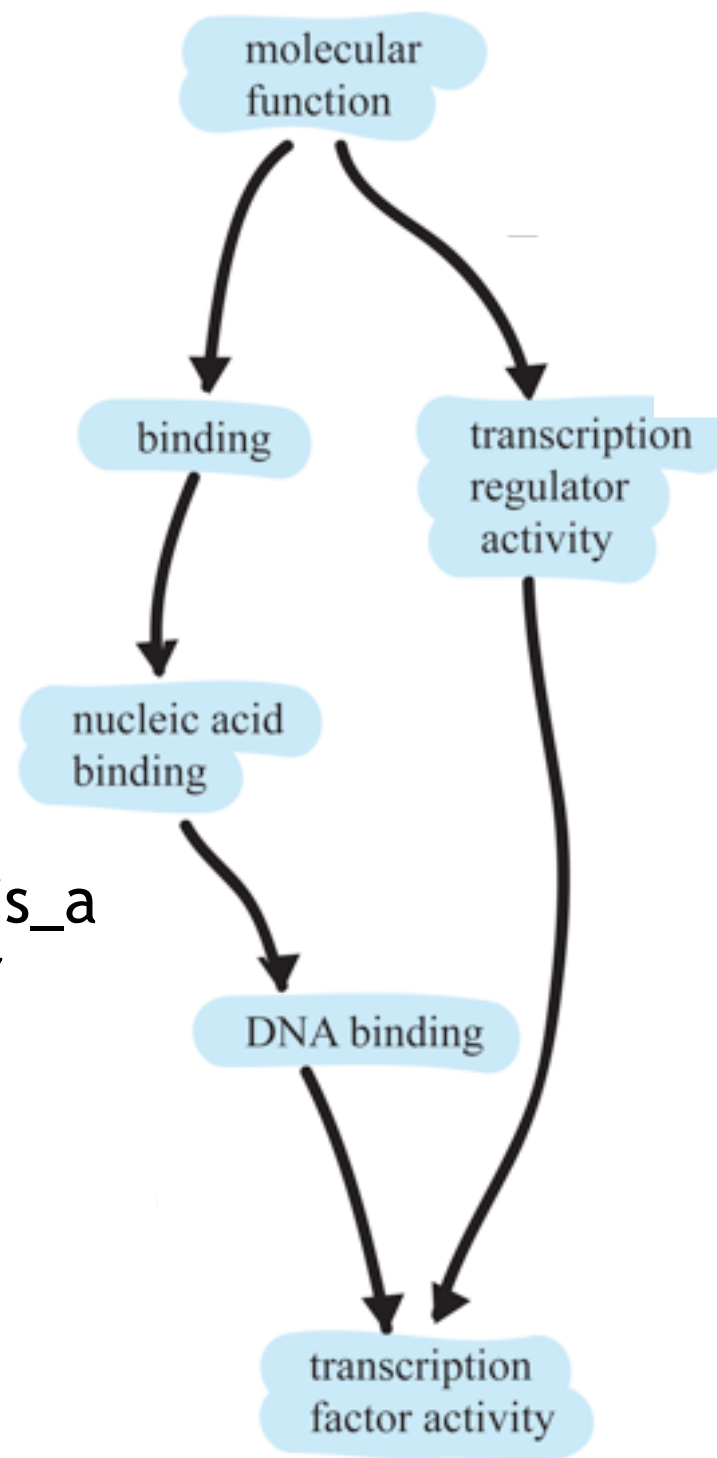
Molecular function ontology

Nucleic acid binding is a type of binding.

is_a

DNA binding is a type of nucleic acid binding.

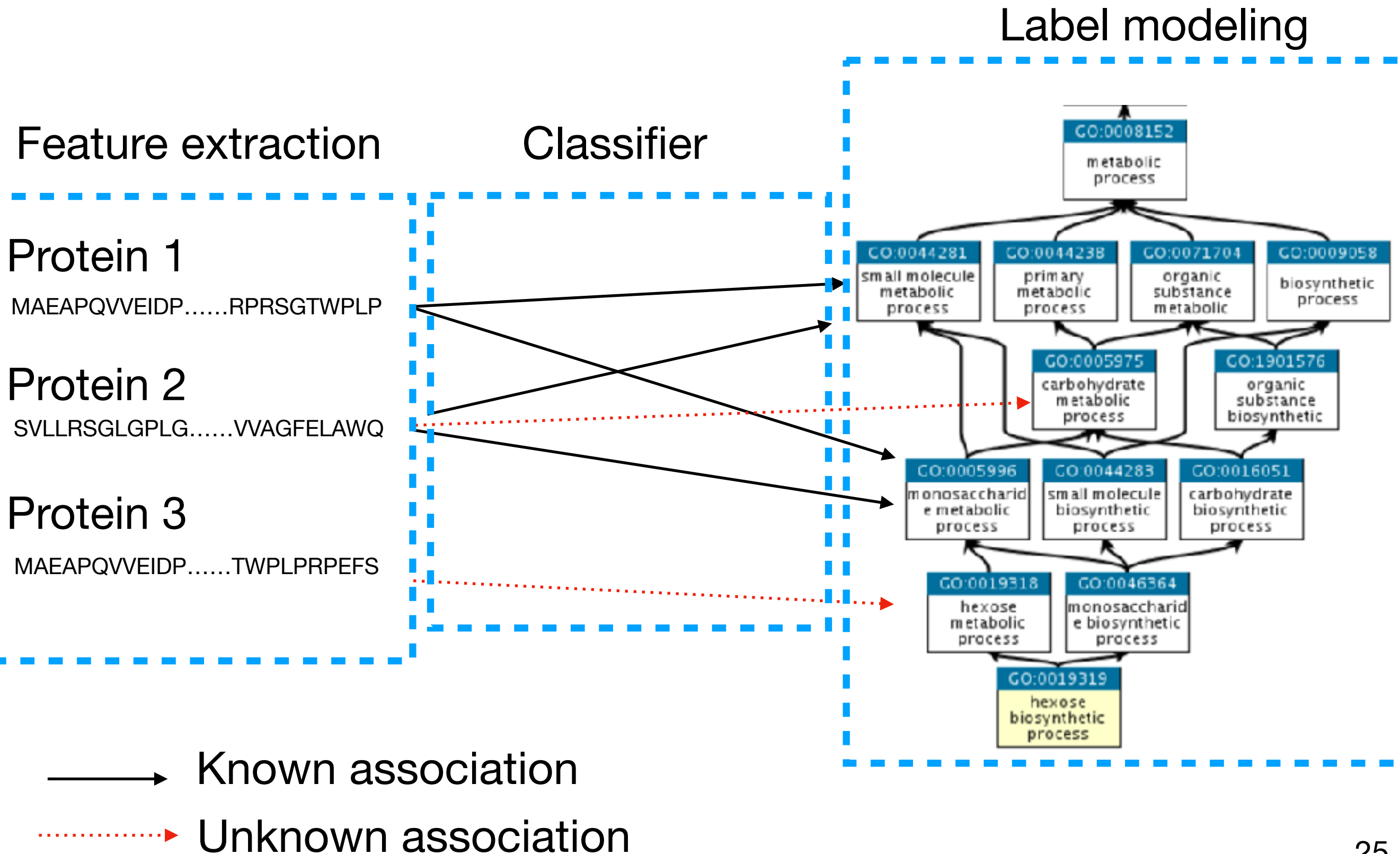
is_a



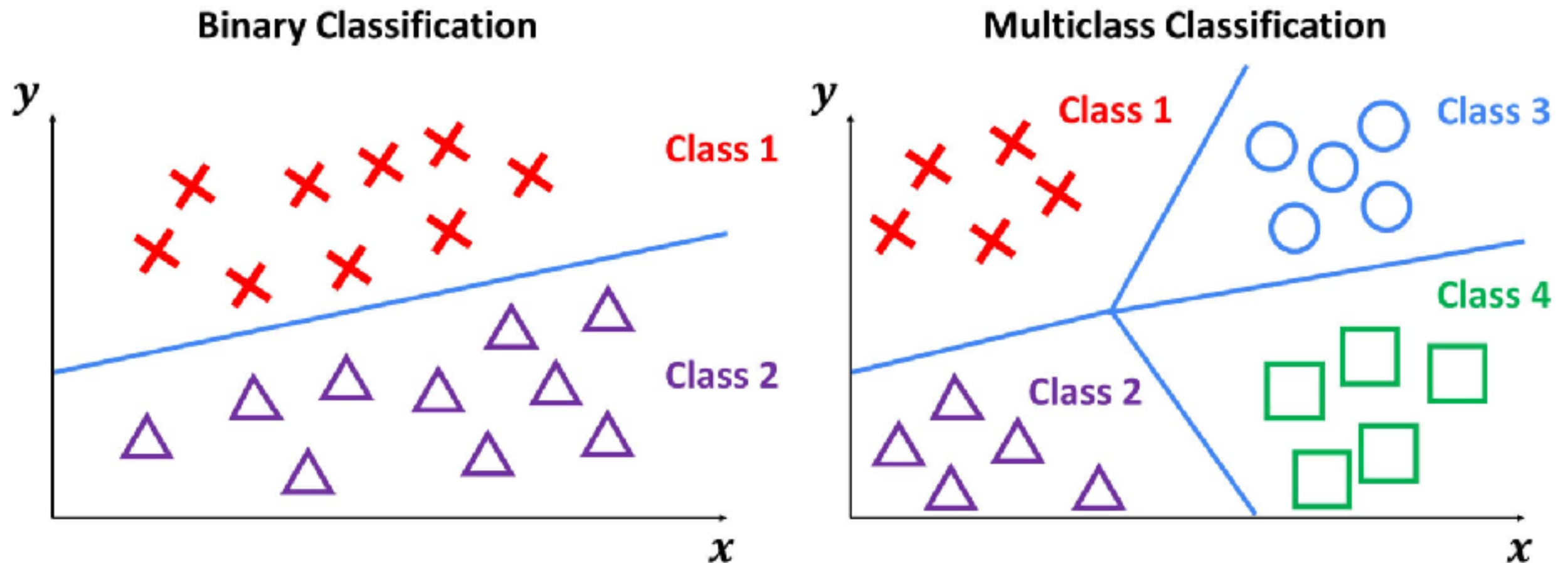
Today

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

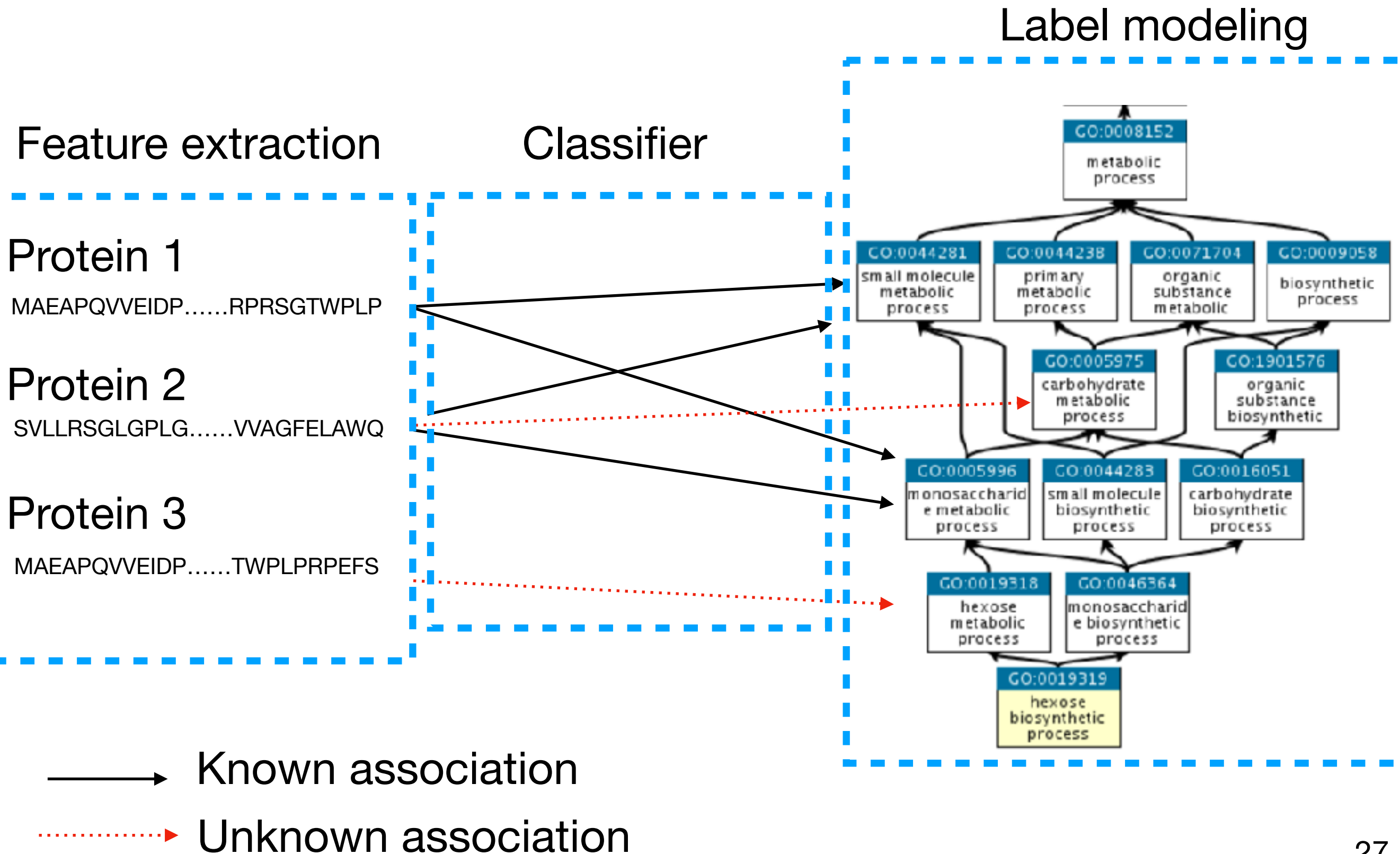
Problem setting for protein function prediction



Introduction to machine learning classification



Problem setting for protein function prediction



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	Literature_Refs	Optional DefX_Ref	Optional: public definition comment
biological_process			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	(More, Less)	(More, Less)	
☒ negative_regulation			
☒ positive_regulation			

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

Verify Input

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;

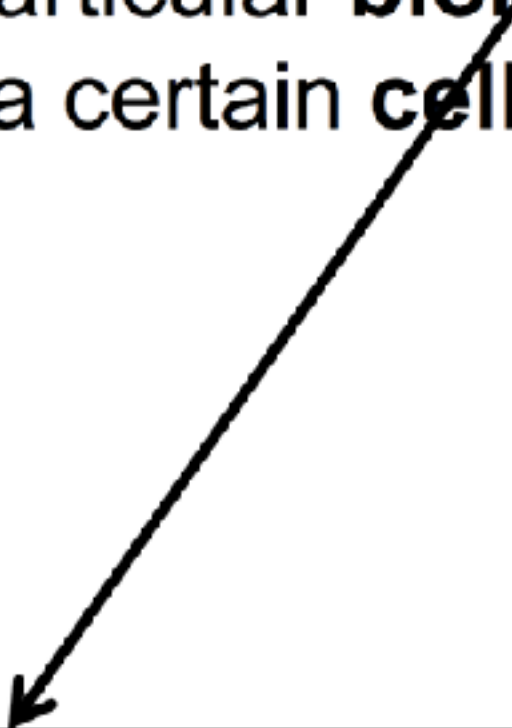


<i>Accession</i>	<i>Name</i>	<i>GO ID</i>	<i>GO term name</i>	<i>Reference</i>	<i>Evidence code</i>
P00505	GOT2	GO:0004069	aspartate transaminase activity	PMID:2731362	IDA

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**

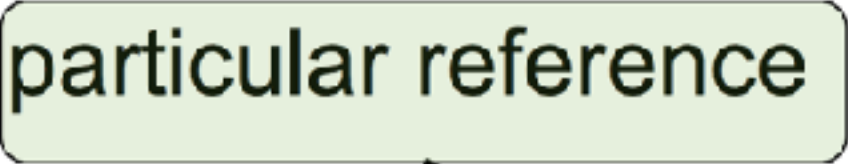


<i>Accession</i>	<i>Name</i>	<i>GO ID</i>	<i>GO term name</i>	<i>Reference</i>	<i>Evidence code</i>
P00505	GOT2	GO:0004069	aspartate transaminase activity	PMID:2731362	IDA

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



<i>Accession</i>	<i>Name</i>	<i>GO ID</i>	<i>GO term name</i>	<i>Reference</i>	<i>Evidence code</i>
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

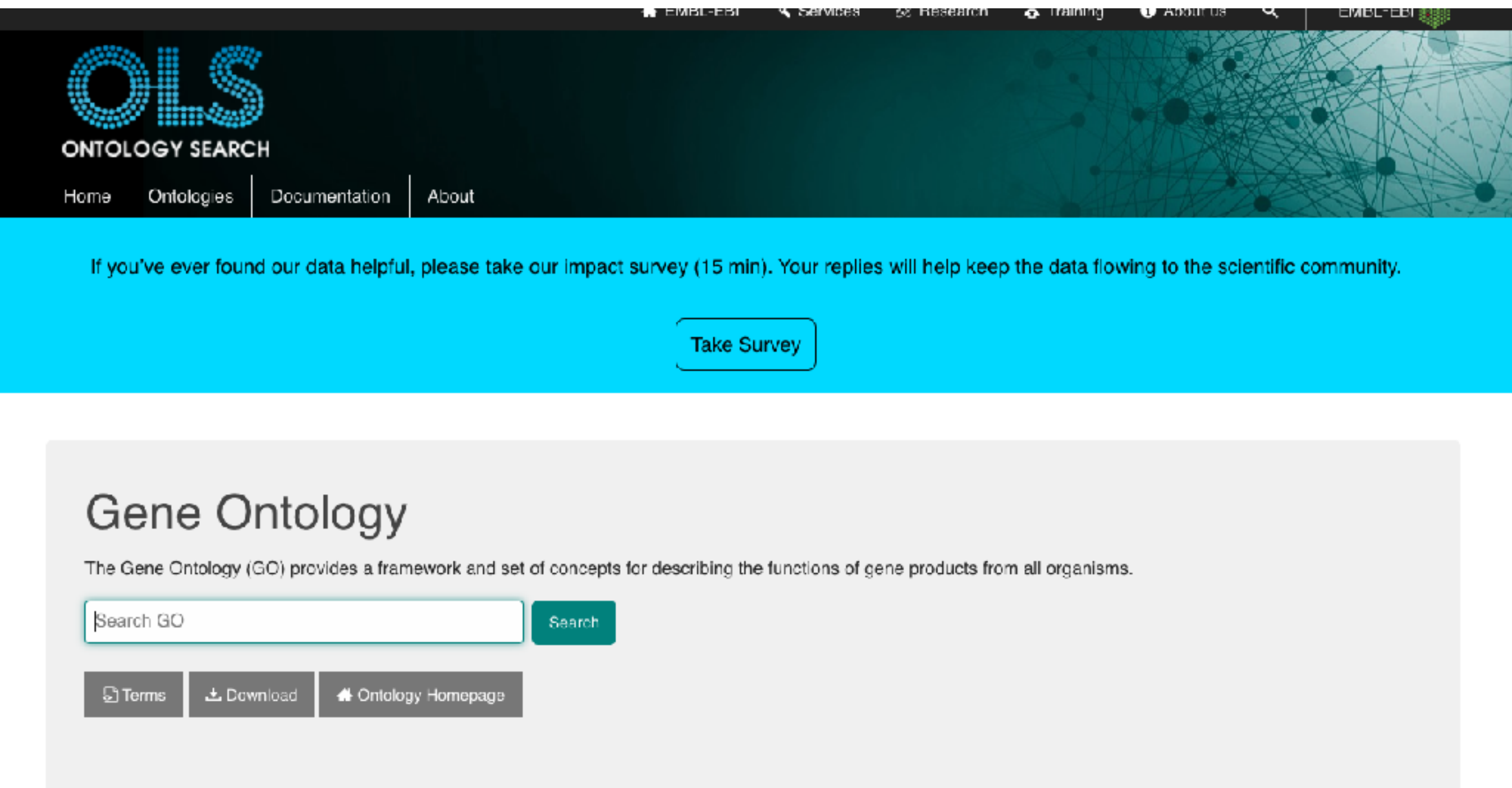


Electronic annotation

- 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



Let's take a look at this database



The screenshot shows the OLS (Ontology Lookup Service) website. At the top, there is a dark navigation bar with links for EMBL-EBI, Services, Research, Training, and About Us. Below this, the OLS logo is displayed on the left, and a network diagram is on the right. A secondary navigation bar contains links for Home, Ontologies, Documentation, and About. A prominent blue banner across the middle of the page contains a message about an impact survey and a 'Take Survey' button. Below the banner, the 'Gene Ontology' section is visible, featuring a description of GO, a search bar, and links for Terms, Download, and Ontology Homepage.

EMBL-EBI Services Research Training About Us

OLS
ONTOLOGY SEARCH

Home Ontologies Documentation About

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.

Take Survey

Gene Ontology

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

Search GO Search

Terms Download Ontology Homepage

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 state, attached to the trunk of an organism, such as a limb or a branch. [GOC:rc GOC:jid]

 Tree view

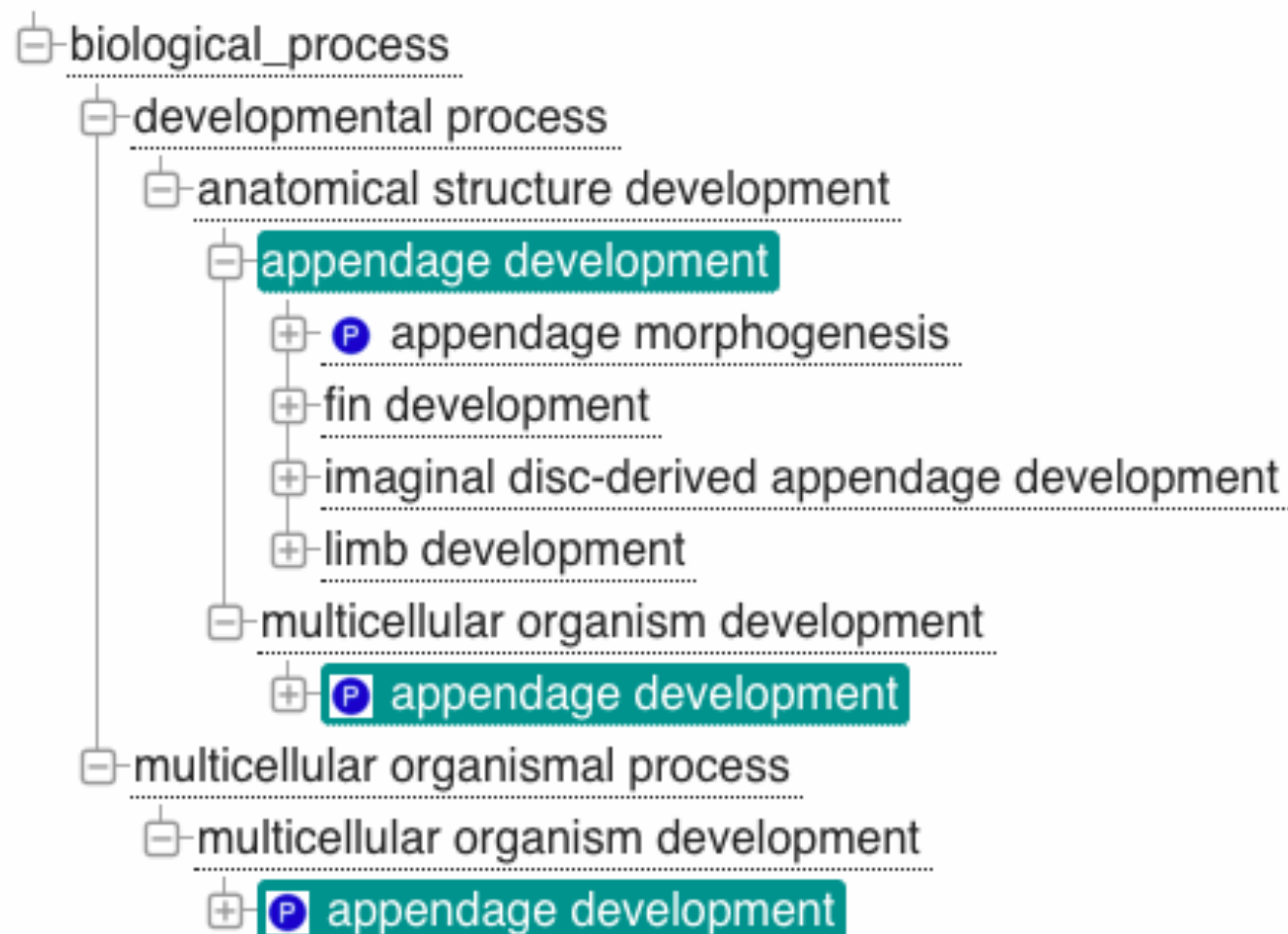
 Term mappings

 Term history

 Graph view

Reset tree

Show all siblings



[http://current.geneontology.org/products/ pages/downloads.html](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
http://ftp.ebi.ac.uk/pub/databases/GO/goa/INTPROT/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 goa_human_valid.gaf
=====
Created on Sun Jan 31 17:05:40 2021.
generated-by: PANTHER
date-generated: 2021-01-31
PANTHER version: v.15.8.
GO version: 2021-01-31.

=====
Documentation about this header can be found here: https://github.com/geneontology/go-site/blob/master/docs/gaf_validation.md

UniProtKB AQA024REG1 NUDT4B GO:0003723 GO_REF:0000043 IEA UniProtKB KW:KW-0504 F Diphasphoinositol polyphosphate phosphohydrolase NUDT4B NUDT4B protein taxon:9605 20201128 UniProt
UniProtKB AQA024REG1 NUDT4B GO:0005320 GO_REF:0000052 IDA C Diphasphoinositol polyphosphate phosphohydrolase NUDT4B NUDT4B protein taxon:9605 20201128 HPA UniProt
UniProtKB AQA024REG1 NUDT4B GO:0046872 GO_REF:0000043 IEA UniProtKB KW:KW-0470 F Diphasphoinositol polyphosphate phosphohydrolase NUDT4B NUDT4B protein taxon:9605 20201128 UniProt
UniProtKB AQA024REG1 NUDT4B GO:0052340 GO_REF:0000003 IEA EC:3.6.1.52 F Diphasphoinositol polyphosphate phosphohydrolase NUDT4B NUDT4B protein taxon:9606 20201128 UniProt
UniProtKB AQA024REG1 NUDT4B GO:0052342 GO_REF:0000003 IEA EC:3.6.1.52 F Diphasphoinositol polyphosphate phosphohydrolase NUDT4B NUDT4B protein taxon:9606 20201128 UniProt
UniProtKB AQA075B6H7 ICKV3-7 GO:0002250 GO_REF:0000043 IEA UniProtKB KW:KW-1064 P Probable non-functional immunoglobulin kappa variable 3-7 ICKV3-7 protein taxon:9606 20201128
UniProt AQA075B6H7 ICKV3-7 GO:0005306 GO_REF:0000044 IEA UniProtKB SubCell:SL-0039 C Probable non-functional immunoglobulin kappa variable 3-7 ICKV3-7 protein taxon:9606 20201128
UniProt AQA075B6H7 ICKV3-7 GO:0010314 GO_REF:0000043 IEA UniProtKB KW:KW-1200 C Probable non-functional immunoglobulin kappa variable 3-7 ICKV3-7 protein taxon:9606 20201128
UniProt AQA0750CH9 IGV1D-42 GO:0002250 GO_REF:0000043 IEA UniProtKB KW:KW-1064 F Probable non-functional immunoglobulin kappa variable 10-42 IGV1D-42 protein taxon:9605
20201128 UniProt AQA0750CH9 IGV1D-42 GO:0005080 GO_REF:0000044 IEA UniProtKB-SubCell:SL-0039 C Probable non-functional immunoglobulin kappa variable 10-42 IGV1D-42 protein taxon:9606
20201128 UniProt AQA0750CH9 IGV1D-42 GO:0019014 GO_REF:0000043 IEA UniProtKB KW:KW-1200 C Probable non-functional immunoglobulin kappa variable 10-42 IGV1D-42 protein taxon:9605
20201128 UniProt AQA0750CH9 IGLV4-59 GO:0002250 GO_REF:0000043 IEA UniProtKB KW:KW-1064 F Immunoglobulin lambda variable 4-59 IGLV4-59 protein taxon:9605 20201128 UniProt
UniProt AQA0750CH9 IGLV4-59 GO:0005080 GO_REF:0000044 IEA UniProtKB-SubCell:SL-0039 C Immunoglobulin lambda variable 4-59 IGLV4-59 protein taxon:9606 20201128
UniProt AQA0750CH9 IGLV4-59 GO:0019014 GO_REF:0000043 IEA UniProtKB KW:KW-1200 C Immunoglobulin lambda variable 4-59 IGLV4-59 protein taxon:9605 20201128 UniProt

```

- Enable transferring knowledge across species

A good dataset for ML

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

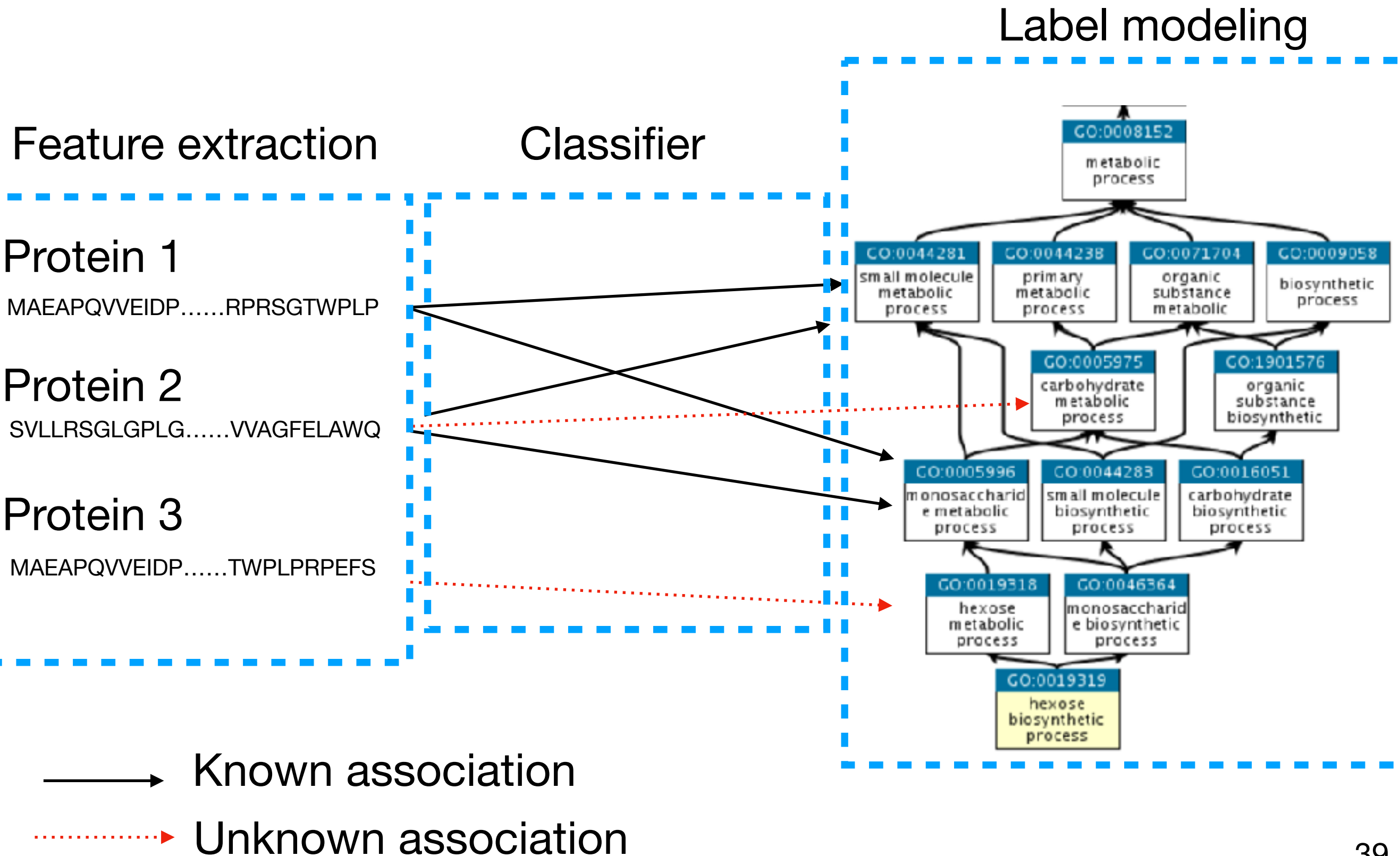
Electronic annotations	269,207,317
Manual annotations*	2,752,604

* 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



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/>

[BLAST](#) [Align](#) [Retrieve/ID mapping](#) [Peptide search](#) [SPARQL](#)

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 example:
P31946 P62258
ALBU_HUMAN
EFTU_ECOLI
2. If you need to convert to another identifier type, select the source and target type.
3. Click the Submit button.

[Help](#) [Help](#)

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

FASTA ▾

5

breast cancer type 1 susceptibility protein homolog [Rattus norvegicus]

NCBI Reference Sequence: NP_036646.1

[GenPept](#) [Identical Proteins](#) [Graphics](#)

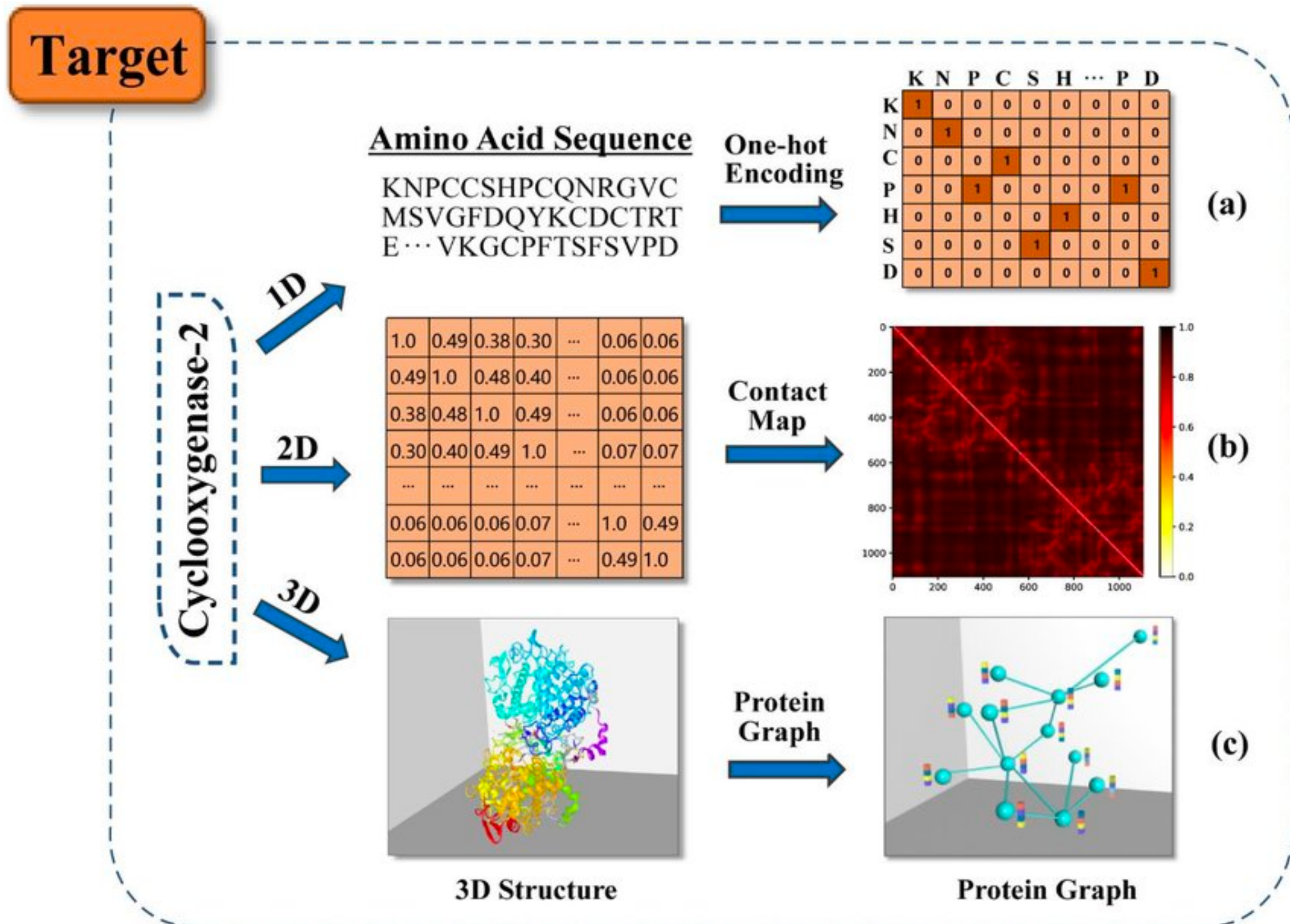
>NP_036646.1 breast cancer type 1 susceptibility protein homolog [Rattus norvegicus]

```
MDLSAVRIQEVQNVLHAMQKILECPICLELIKEPVSTQCDHIFCKFCMLKLLNQKKGPSQCPLCKNEITK
RSLOGSARFSQLVEELLKIIDAFELDTGMQCANGFSFSKKKNSSSELLNEDASIIQSVGYRNRVKKLQOI
ESGSATLKDSLVSQLSNLGIVRSMKKNRQTQPQNKSVYIALESSESSEERNAPDGCSVRDQELFQIAPGG
AGDEGKLNSAKKAACDFSEGIRNIEHHQCSDKDLNPTENHATERHPEKCPRISVANVHVEPCGTDARASS
LQRGTRSLLFTEDRDLAEKAEFCDRSKQSGAAVSQQSRWADSKETCNGRPVPRTEGKADPNVDSLGRKQ
WNHPKSLCPENSGATTDPWITLNSSIQKVNEWFSRTGEMLTSDNASDRRPASNAEAAVVLEVSNEVDGC
FSSSKIDLVAPDPDNAVMTSGRDFSKPVENIINDKIFGKTYQRKGSRPHLNHVTEIIGTFTTEPQIIQ
EQPFTNKLKRKRSTCLHPEDFIKKADLTVVQRISENLNQGTDOMEPNDQAMSITSNGQENRATGNDLQRC
RNAHPIESLRKEPAFTAKAKSISNSISDLEVELNVHSSKAPKKNRLRRKSTRCVLPLEPISRNPSPPTCA
ELQIESCGSSEETKKNNSNQTTPAGHIREPQLIEDTEPAADAKKNPNEHIRKRSASDAFPPEEKL MNKAGL
LTSCSSPRKPQGPVNPSPERKGIEQLEMCOMPDNKELGDLVLGGEPGKPTPEPSEESTSVSLVPD TDYD
TQNSVSILEANTVRYARTGSVQCMTQFVASENPKELVHGSNNAGSGSECFKHPLRHELNHNQETIEMEDS
ELDTQYLQNTFQVSKRQSFALFSKLRSPQKDCTLVGARSVPSREPSPKVTSRGEQKERQGESEESHVQ
AVTVTVGLPVPCQEGKPGAVTMCADVSR LCPSSHYSRSCENGLNTTDKSGISQNSHFRQSVSPLRSSIKTD
NRKTLTEGRFEKHTERGMGNETA VQSTIHTISLNNRGDACLEASSGSVIEVHSTGENVQGLDRNRGPKV
NTVSLLDSTQPGVSKQSAPVSDKYLEIKQESKAVSADFSPCLFSDHLEKPMRSDKTFQVCSETPDDLDD
VEIQENASFGEGGITEKSAIFNGSVLRRESSRSPSPVTHASKSRSLHRGSRKLEFSEESDSTEDDLPCF
QHLLSRVSSTPELTRCSSVVTQRVPEKAKGTQAPRKSSISDCNNEVILGEASQEYQFSEDAKCSGSMFSS
QHSAALGSPANALSQDPDFNPPSKQRRHQAEENEEAFLSDKELISDHEDMAACLEEASDQEEDSIIPDSVA
SGYESEANLSEDCSQSDILTQQRATMKDNLIKQEQMAQLEAVLEQHGSQPSGHPPCLPADPCALEDL P
DPEQNRSGTAILTSKNINENPVSQNPKRACDDKSQPQPPDGLPSGDKESGMRRPSPFKSPLTSSRCSARG
HSRSLQNRNSTSQEELLQPAXLEKSCEPHNLTGRSCLPRQDLEGTYPYPESGIRLVSSRDPDSESPKVSAL
VCTAPASTSALKISQGVAGSCRSPAAGGADTAVVEIVSKIKPEVTSPKERAERDISMVVSGLTPKEVMI
VQKFAEKYRLALTDVITEETHVVIKTD AEFVCERTLKYFLGIAGGKWIVSYSWVIKSIQERKLLSVHEF
EVKGDVVTGSNHQGP RRSRESQEKLFEG LQIYCCEPFTNMPKDELERMLQLCGASVVKELPLLTRDTGAH
PIVLVQPSAWTEDNDCPDIGQLCKGRLVMWDWVLDSISVYRCRDL DAYLVQNITCGRD GSEPQDSND
```

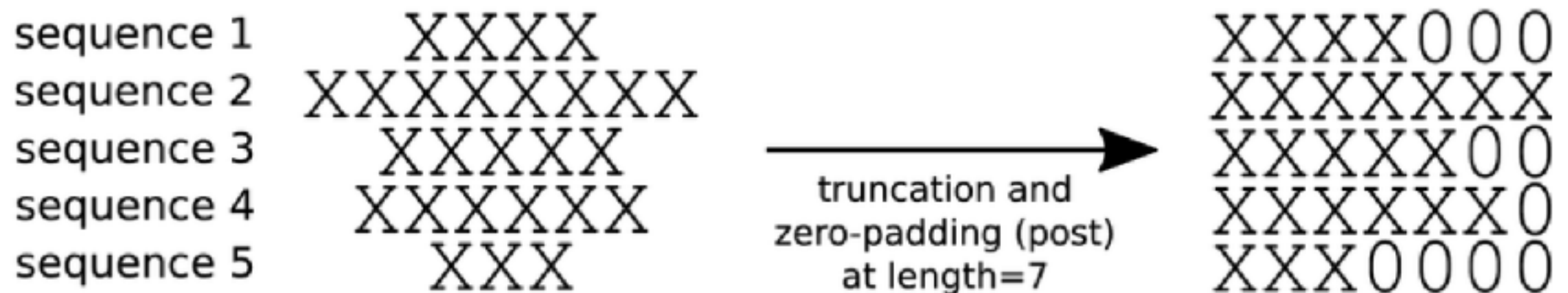
Feature extraction

- 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

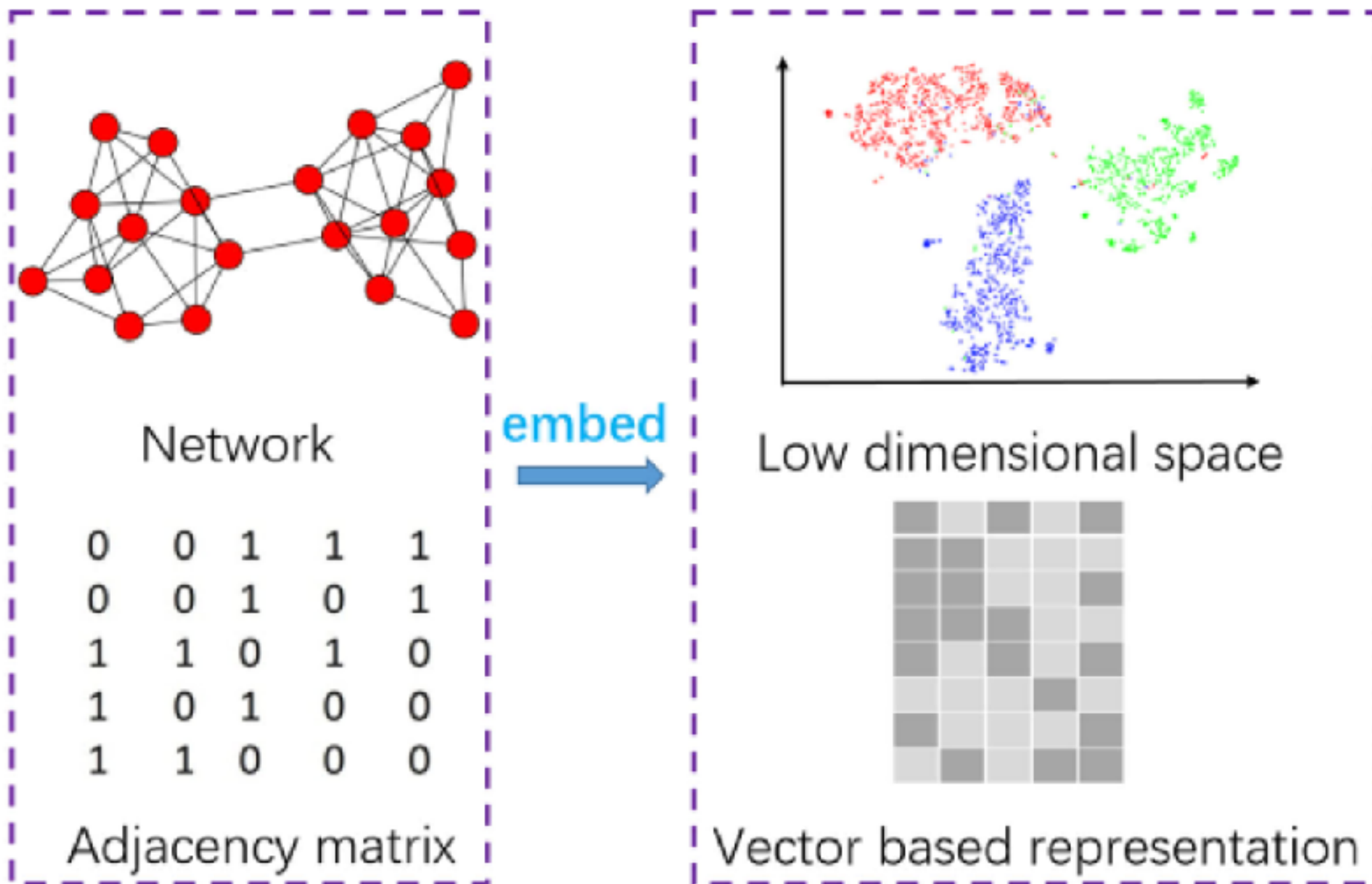
Converting proteins to numeral features



Truncation and zero-padding to have a matched length

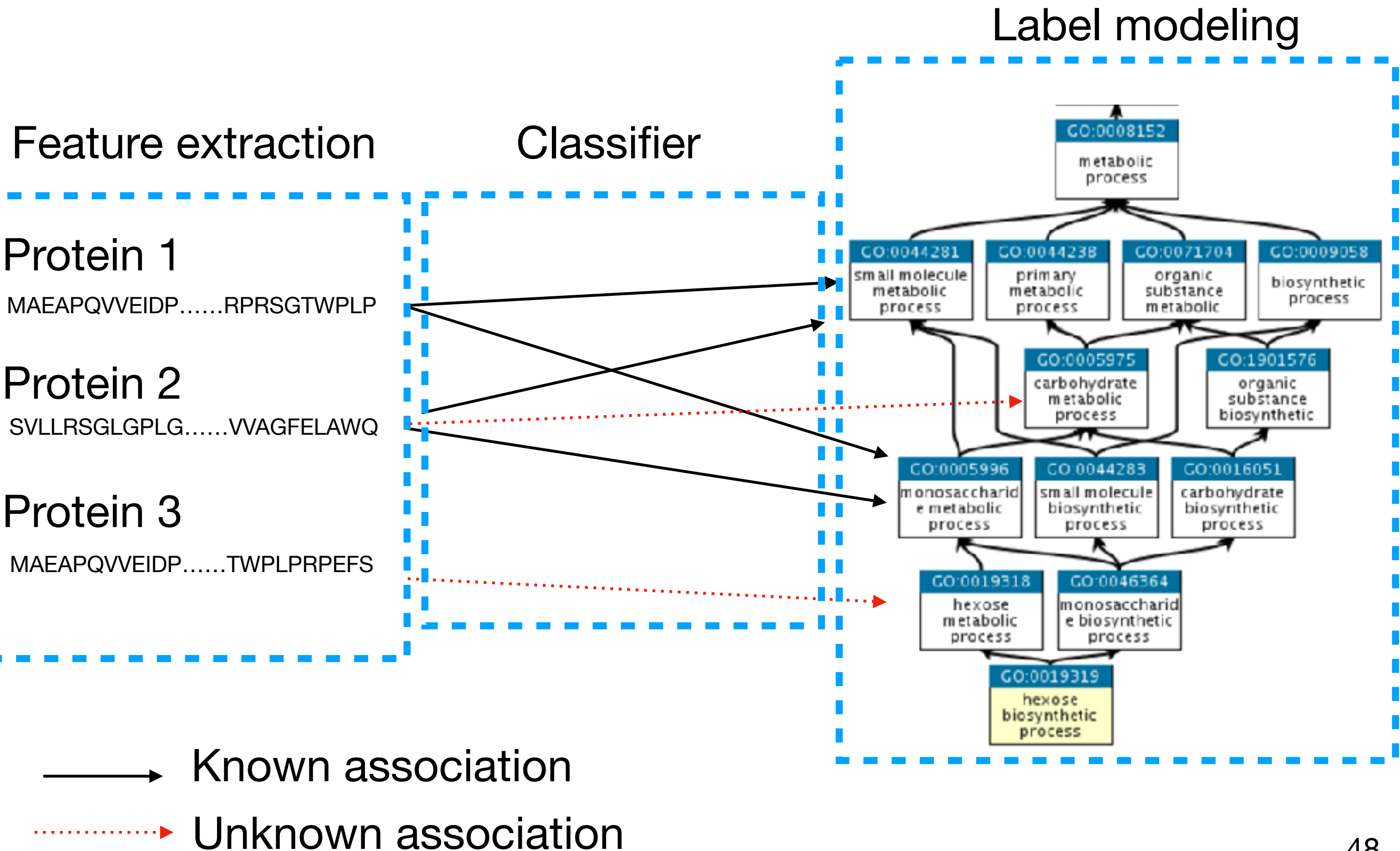


Protein protein network



Classifier

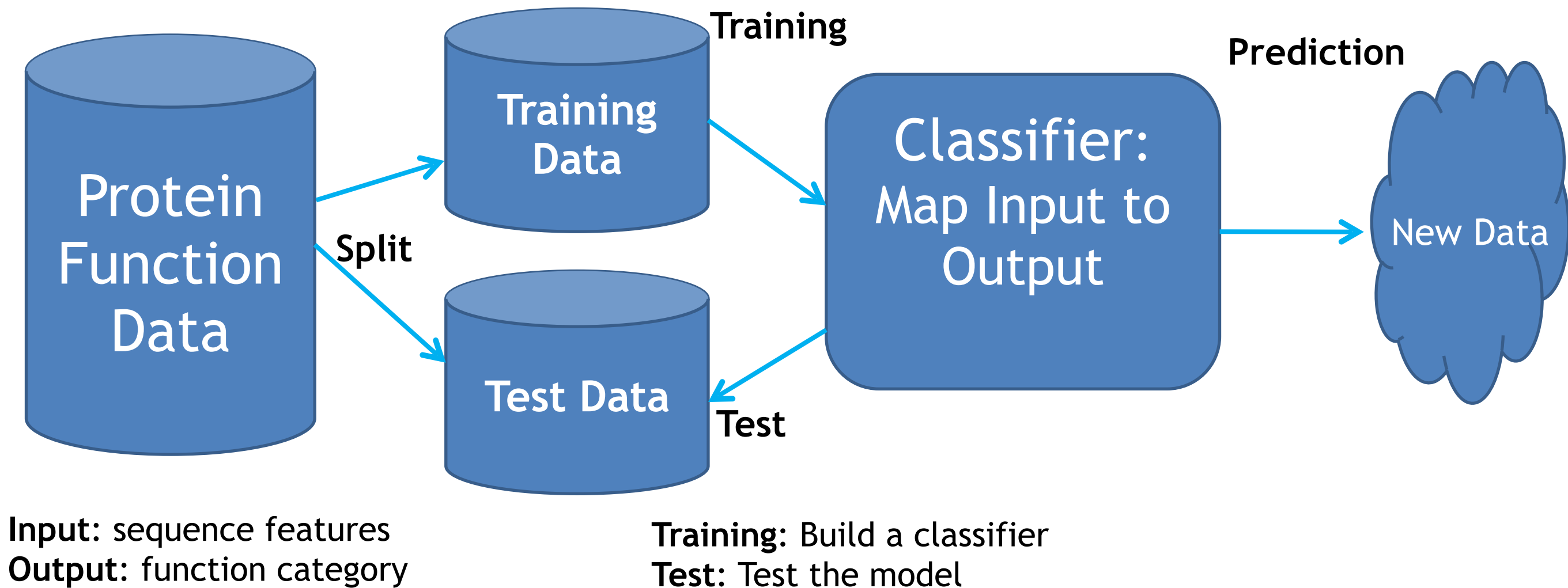
Problem setting for protein function prediction



Problem setting

- Input:
 - Features: sequence of known proteins
 - Known annotations: $\langle \text{Gene Ontology } i, \text{protein } j \rangle$
 - Label graph: gene ontology graph
- Output:
 - Unknown annotations: Should we annotate protein k to gene ontology q ?

Data Driven Machine Learning Approach



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

kNN-based (find similar proteins)

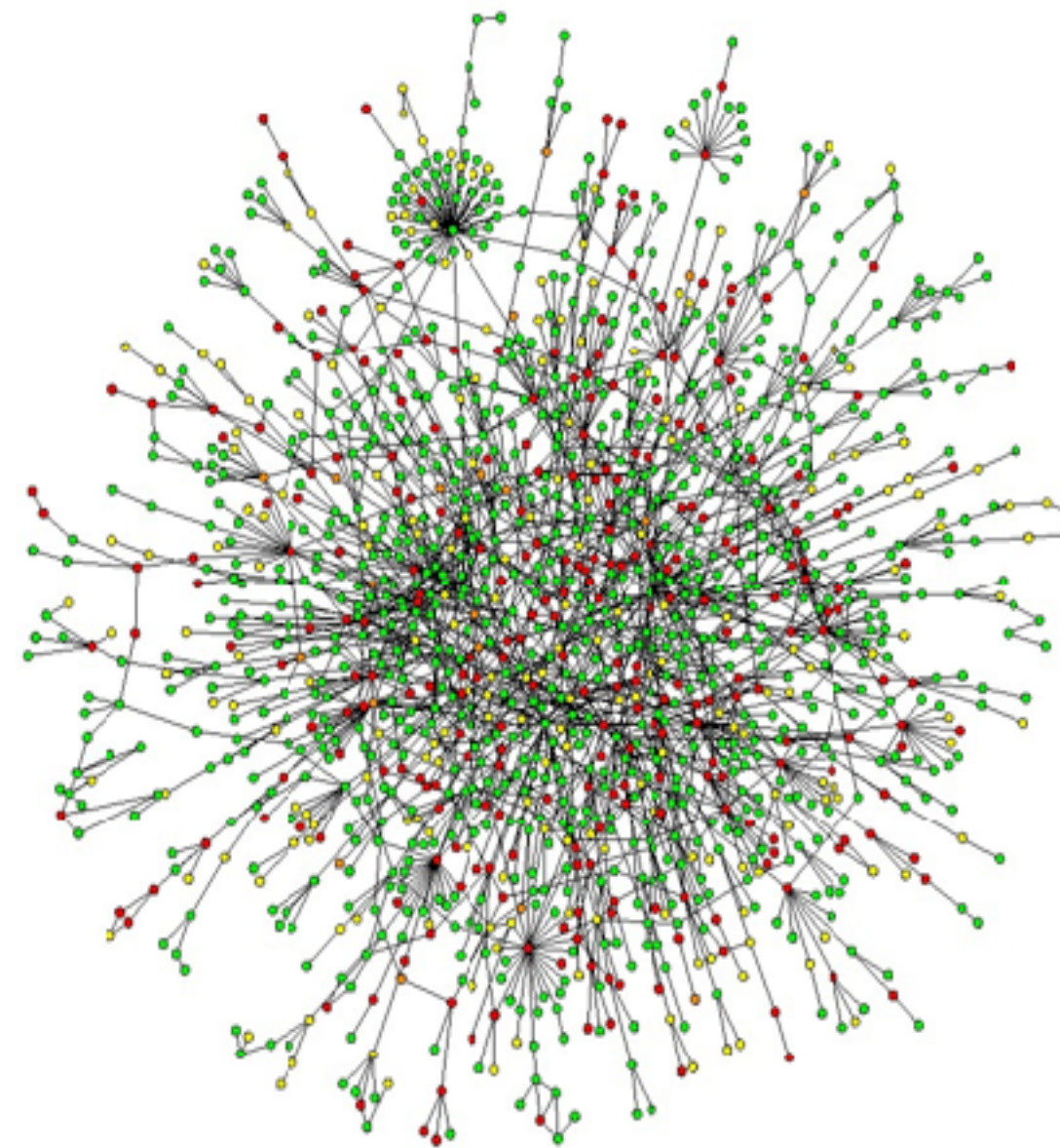
- The easiest way to infer the molecular function of an uncharacterized sequence is by finding a similar and annotated sequence
- BLAST (sequence-sequence local alignment tool) (e.g., Blast2GO)
- Problem:
 - Find similar sequence (sequence alignment)
 - Use these sequences to transfer annotation

kNN-based (find 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” ($P(\text{GO:0045893}|\text{GO:0003677}) = 0.455$).
- 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

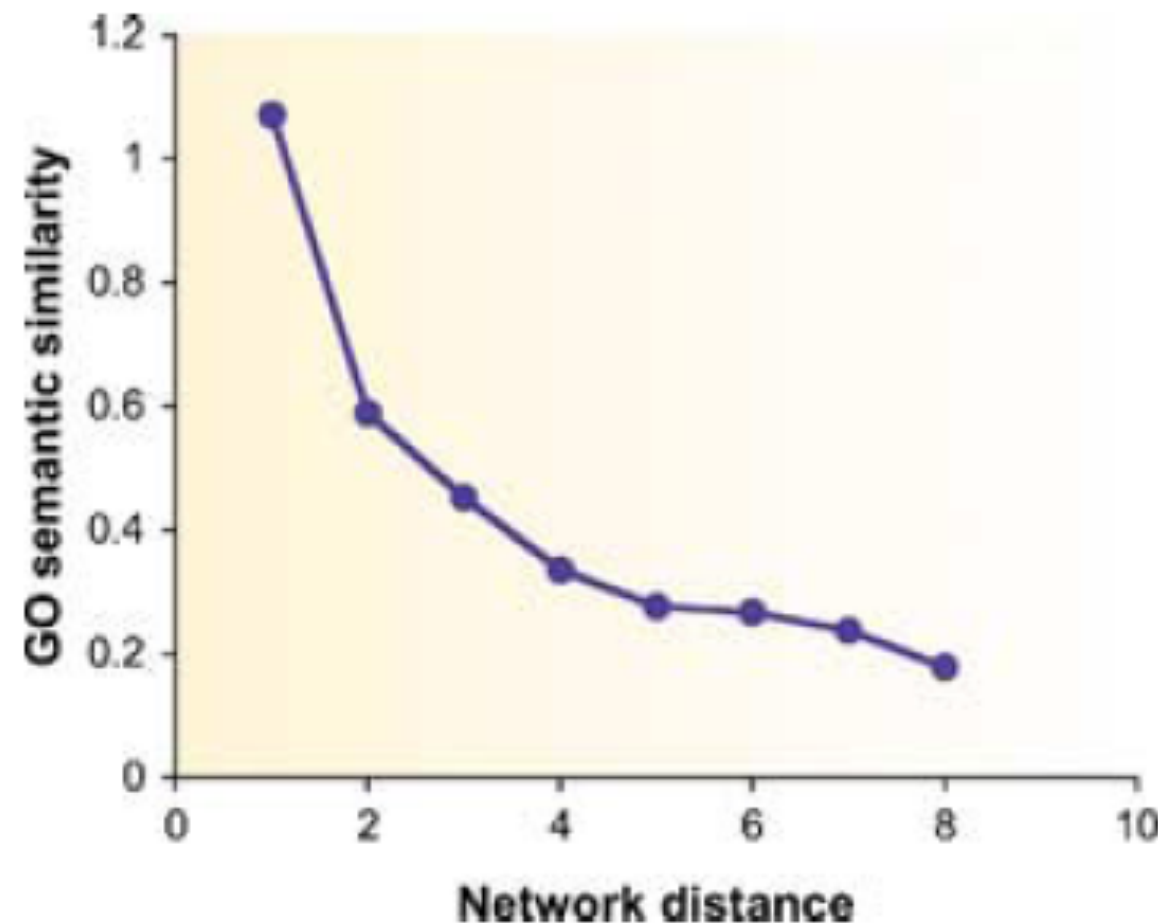
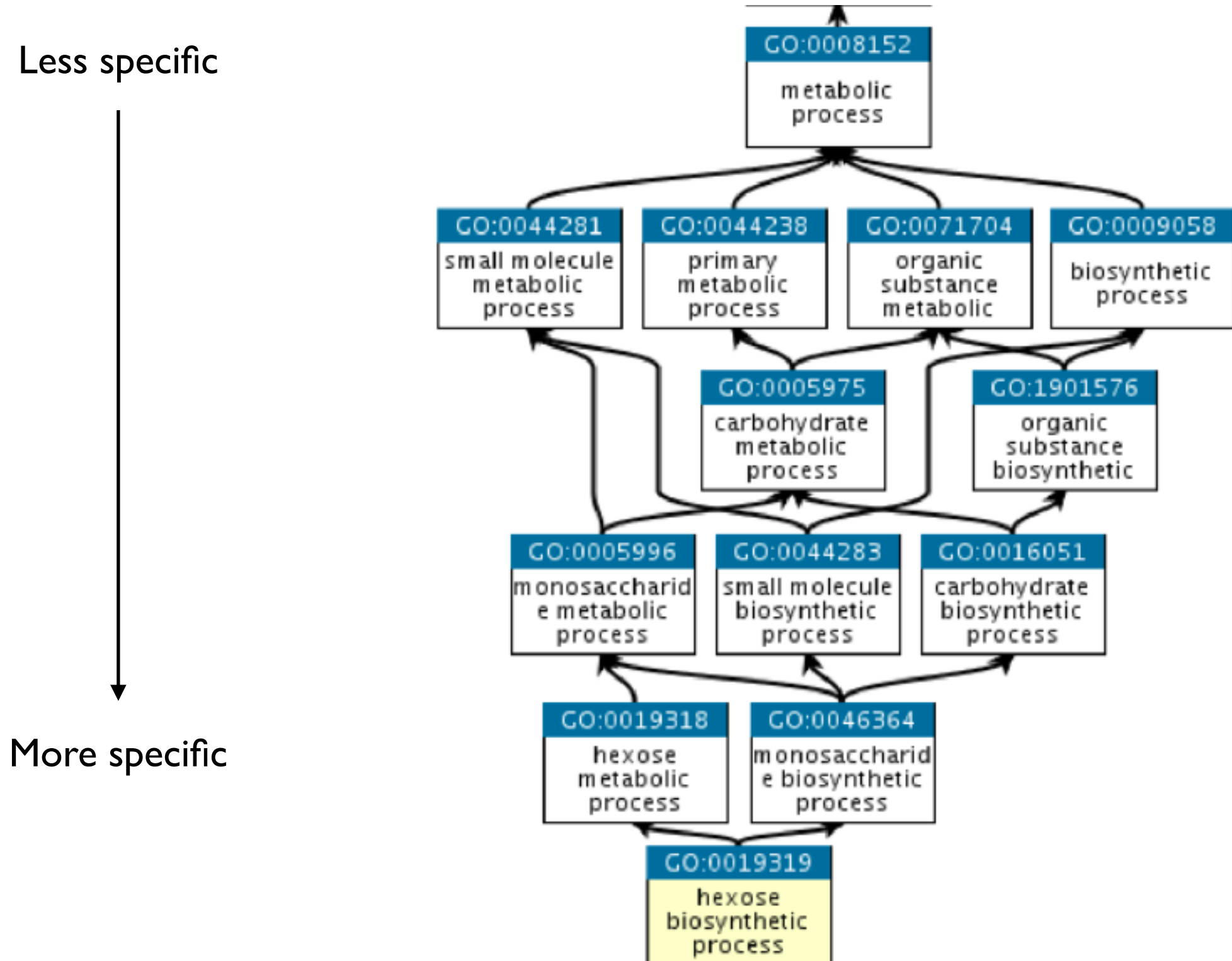


Figure 3 Correlation between protein functional distance and network distance. *X*-axis: distance in the network. *Y*-axis: average functional similarity of protein pairs that lie at the specified distance. The functional similarity of two proteins is measured using the semantic similarity of their GO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

Gene Ontology: A directed acyclic graph



Label modeling

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

Training data: How many proteins do we have annotations?

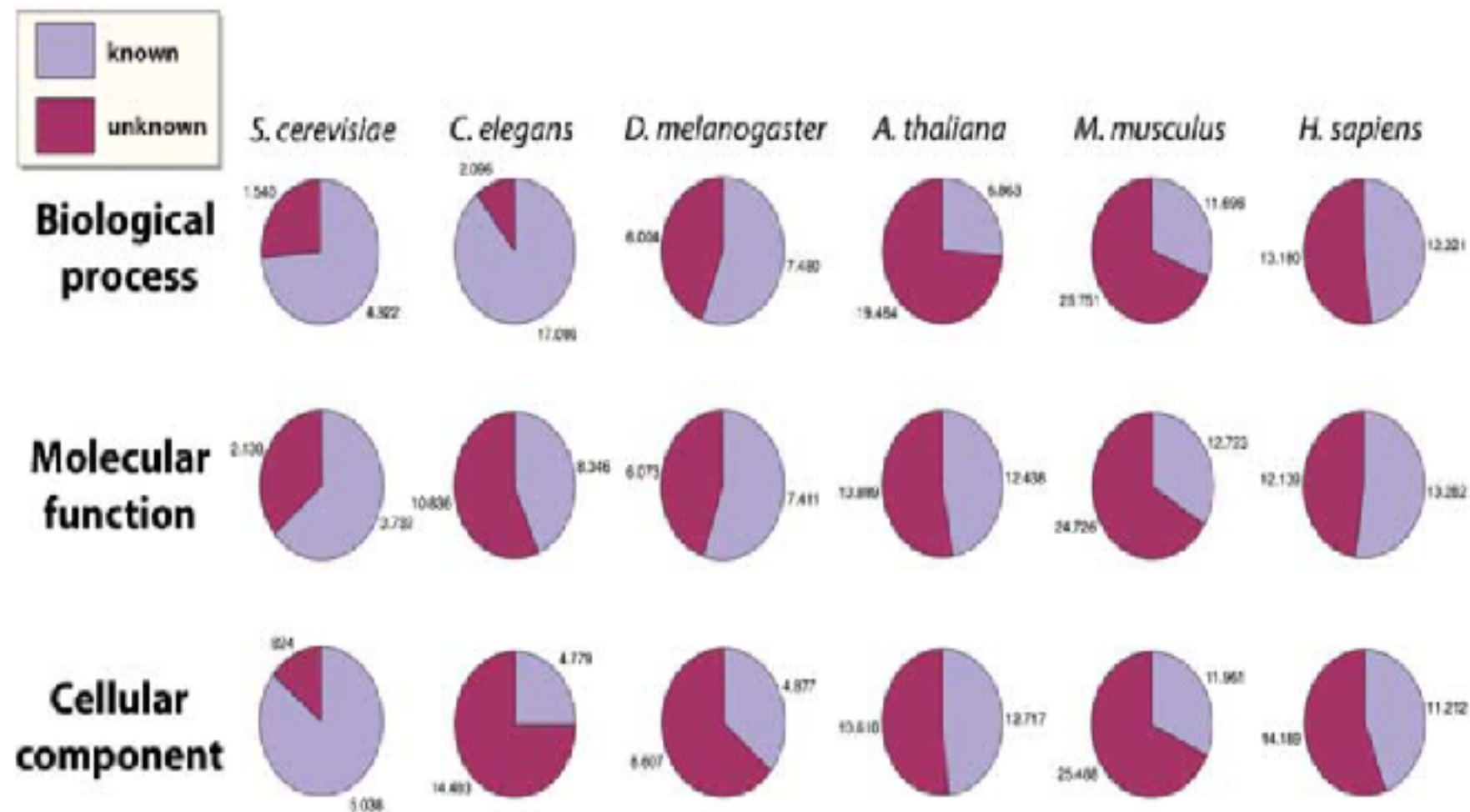
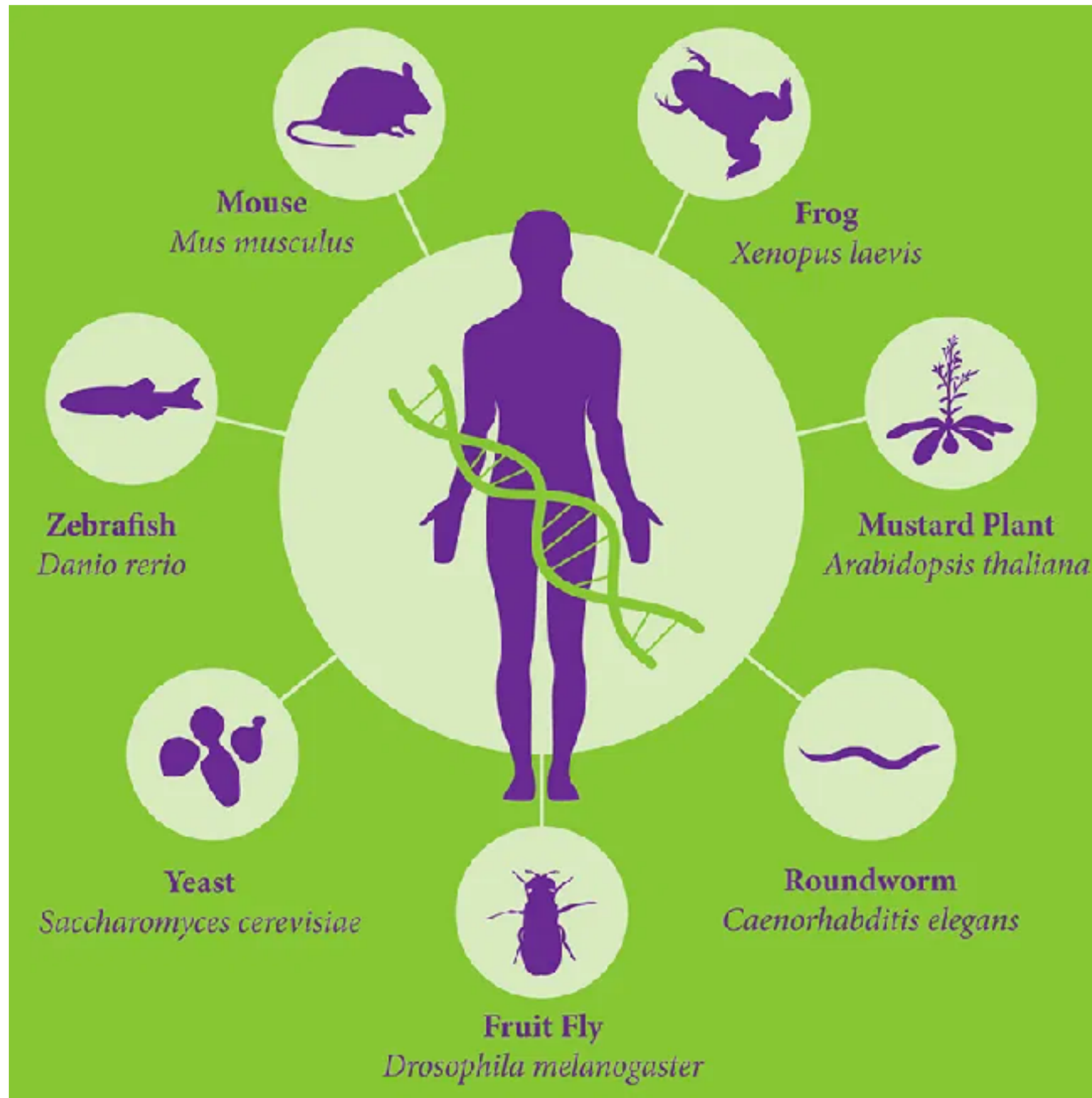


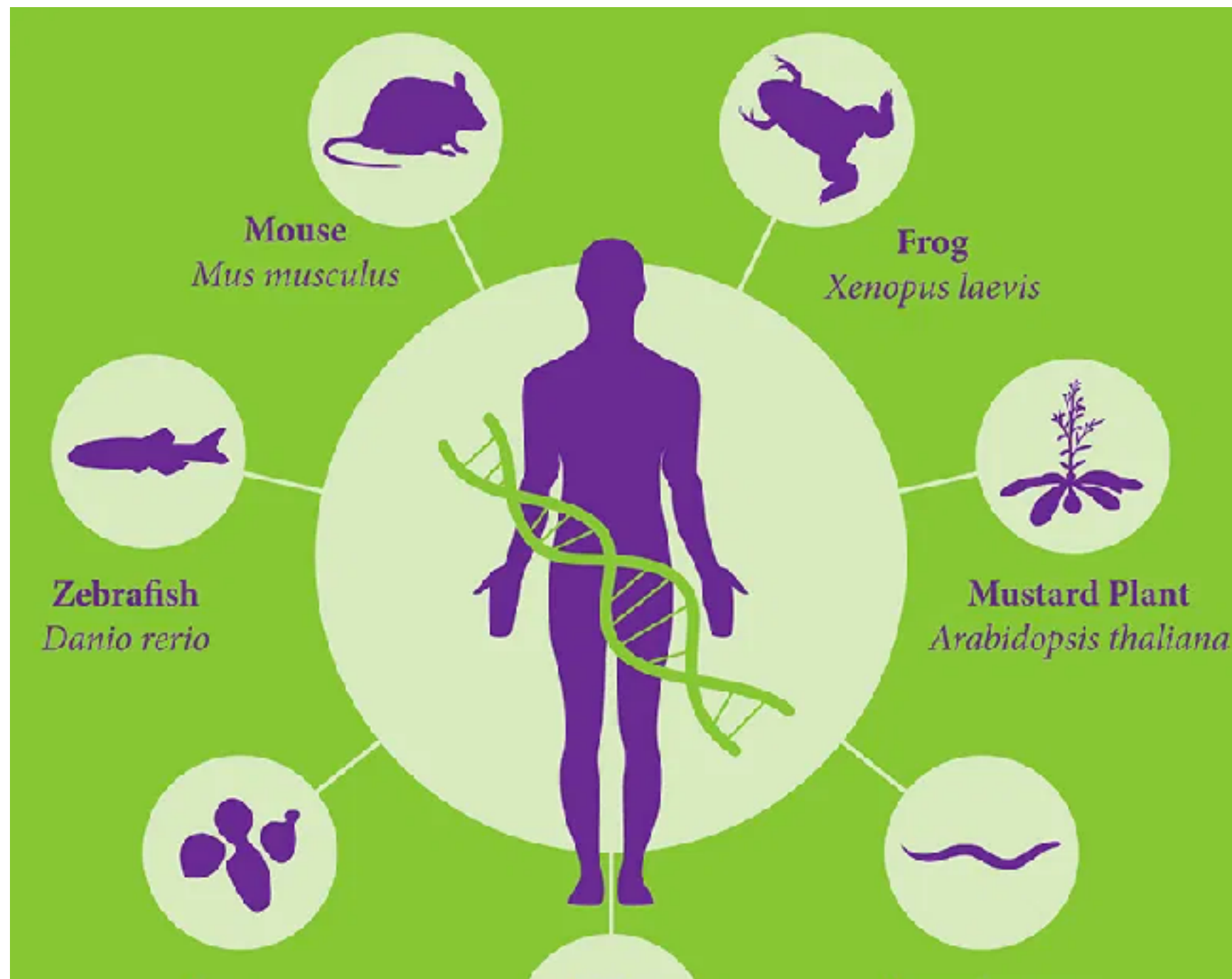
Figure 1 Extent of annotation of proteins in model species. For each species, the charts give the fractions and numbers of annotated and unannotated proteins, according to the three ontologies of the GO annotation. The numbers are based on the Entrez Gene and the WormBase databases as of September 2006.

Sharan et al., Molecular Systems Biology, 2007

Model organism

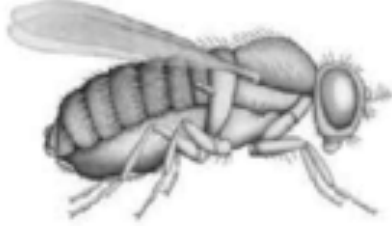


Model organism

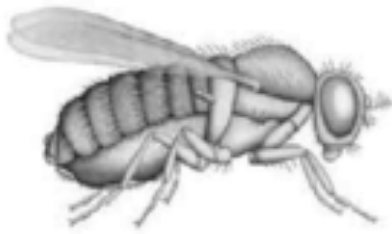


They are like ImageNet in CV and 20
NewsGroup in NLP

Current State of Function of Model Genome Annotation

	Yeast	<i>C. elegans</i>	<i>D. melanogaster</i>	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	<i>C. elegans</i>	<i>D. melanogaster</i>	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

More similar to human and more expensive

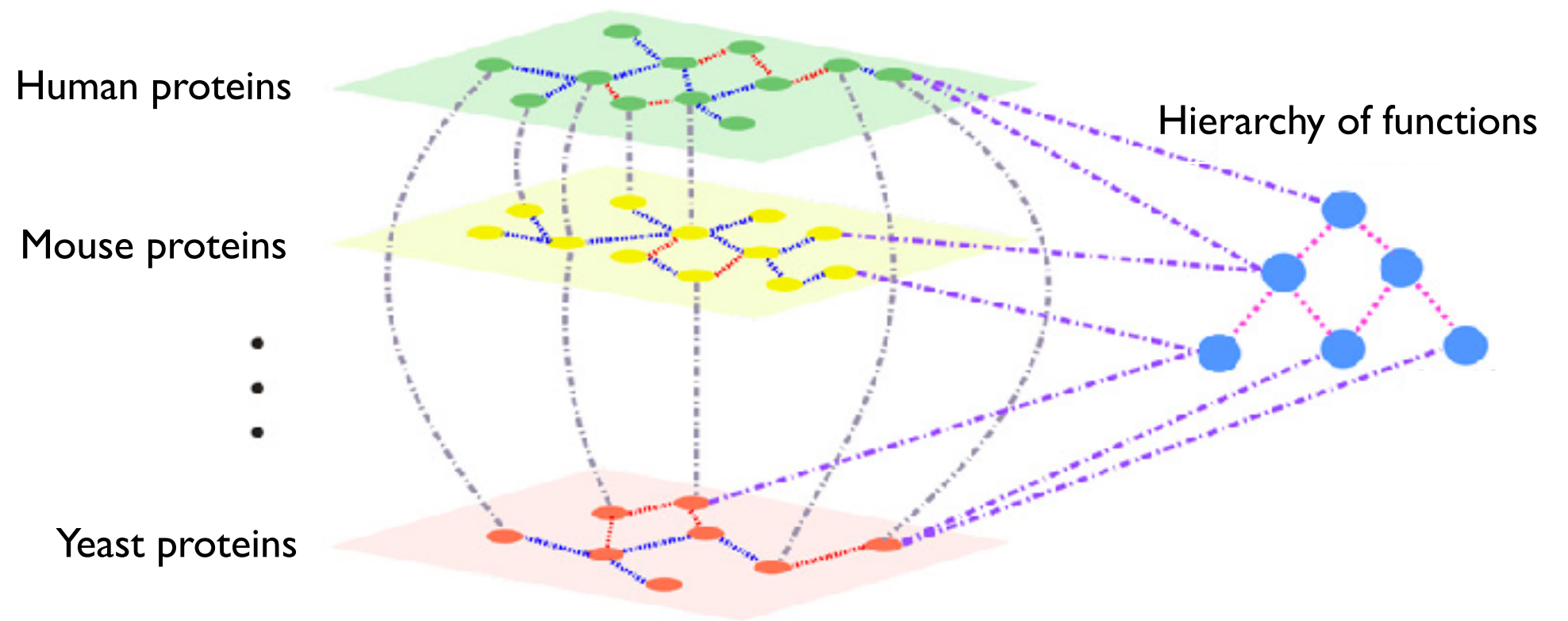
(basic tool)

pathway

effects

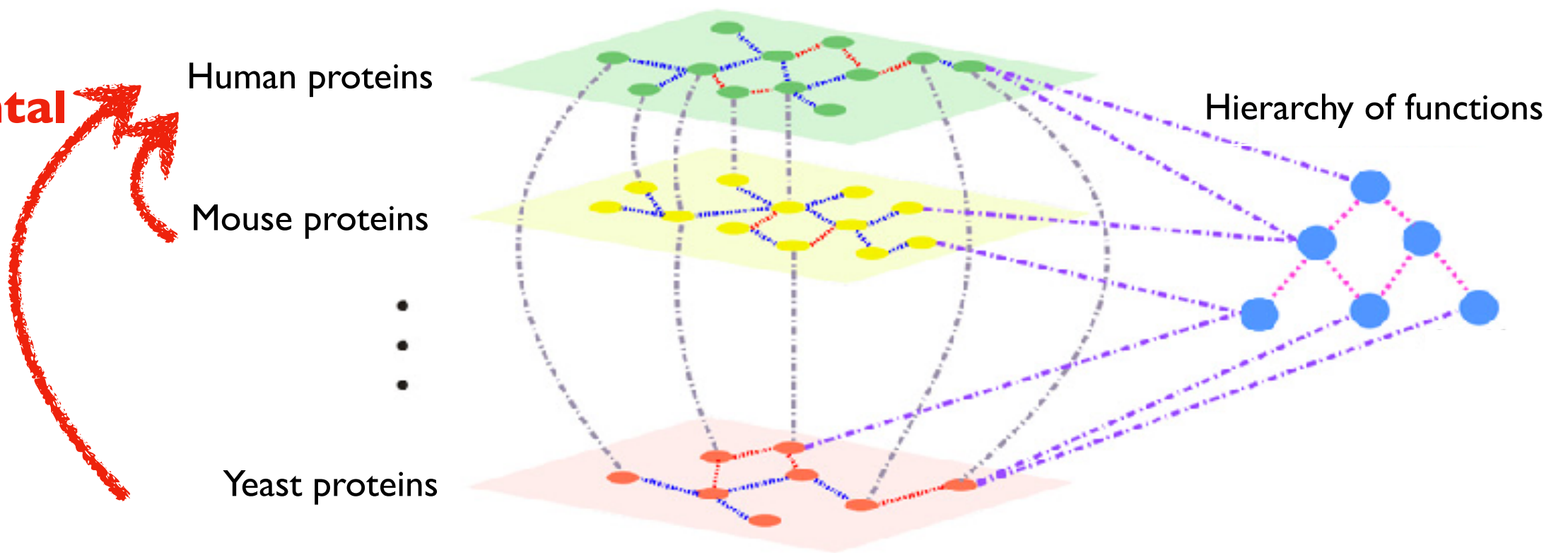
Drug delivery system

transfer knowledge from other species to human



Transfer knowledge from other species to human

**Transfer
experimental
results**

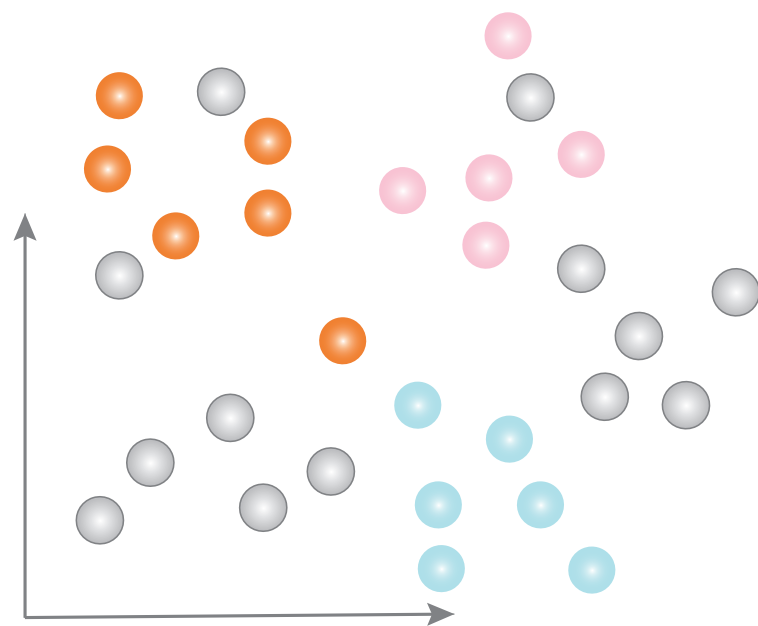


A unique heterogeneous network dataset

- **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

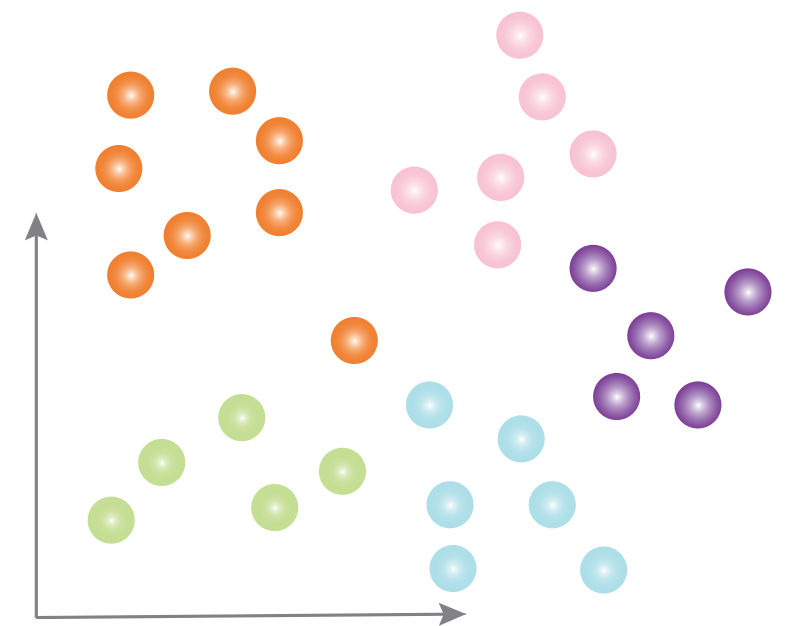
Key challenge: novel functions

Input



● Unannotated proteins
● ● Annotated reference proteins

Gold standard

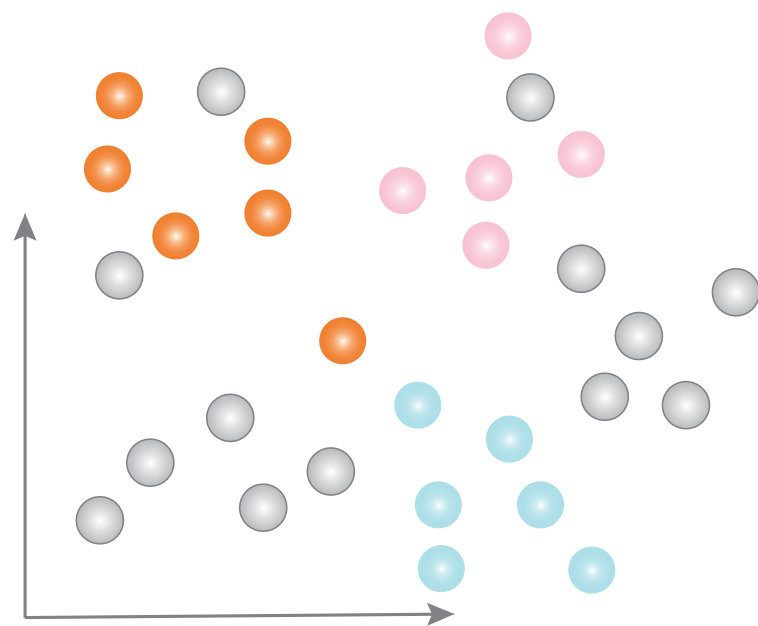


● ● ● Function seen in the reference data
● ● Novel function

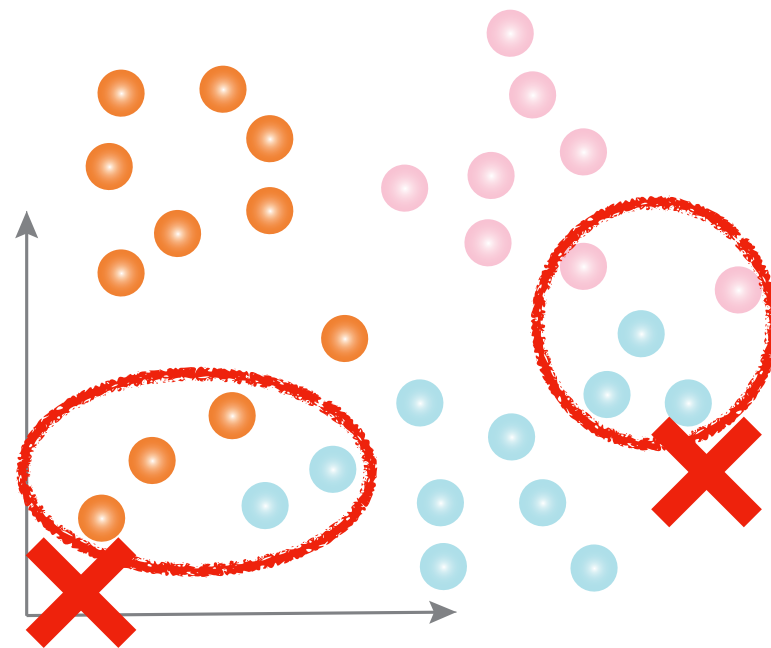
How to correctly classify proteins into novel functions?

Existing methods cannot annotate novel functions

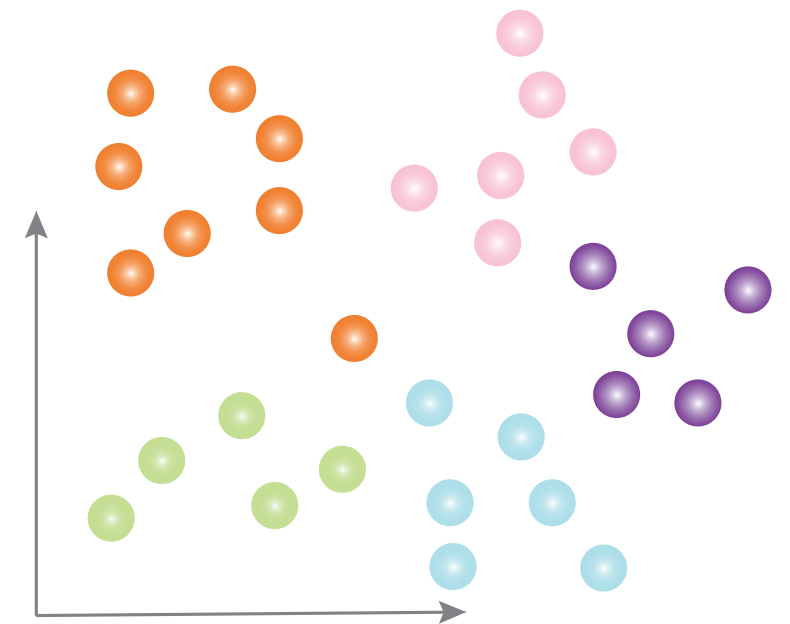
Input



kNN predictions



Gold standard

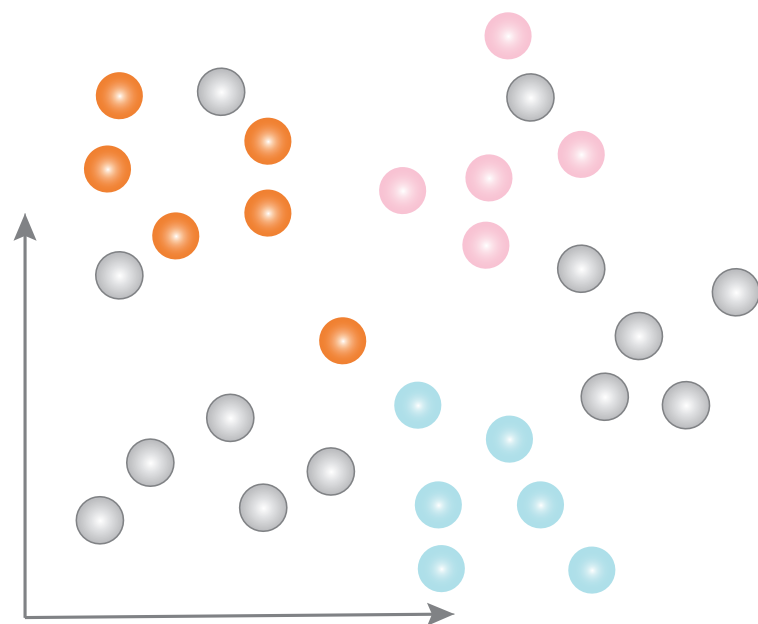


● Unannotated proteins
● ● Annotated reference proteins

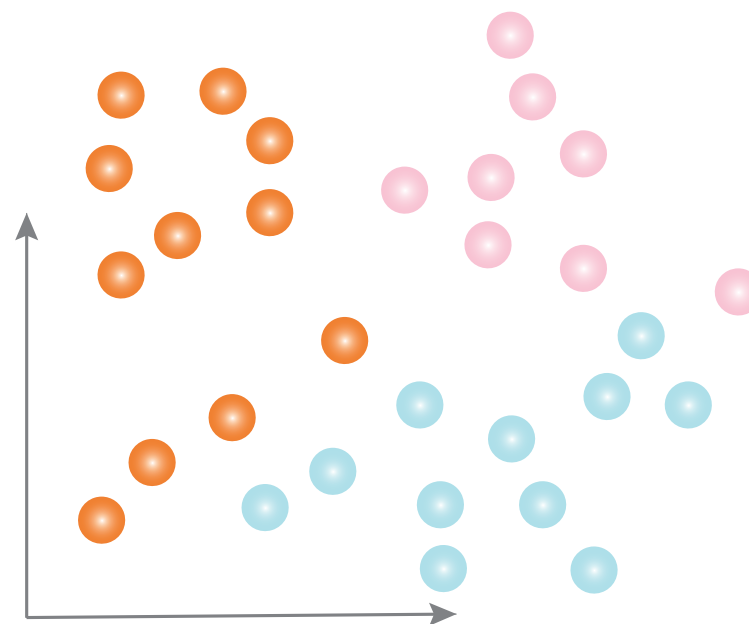
● ● ● Function seen in the reference data
● ● Novel function

Existing methods cannot annotate novel functions

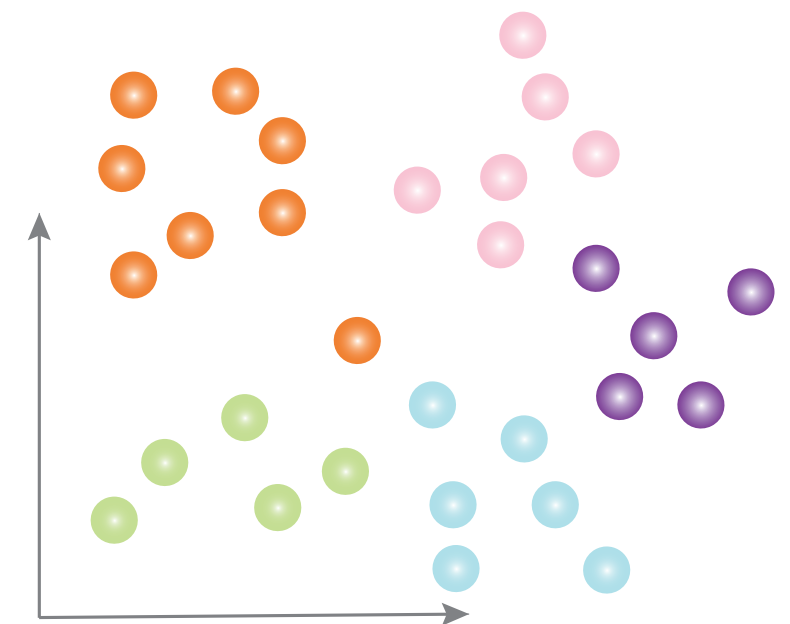
Input



kNN predictions



Gold standard



● Unannotated proteins
● Annotated reference proteins

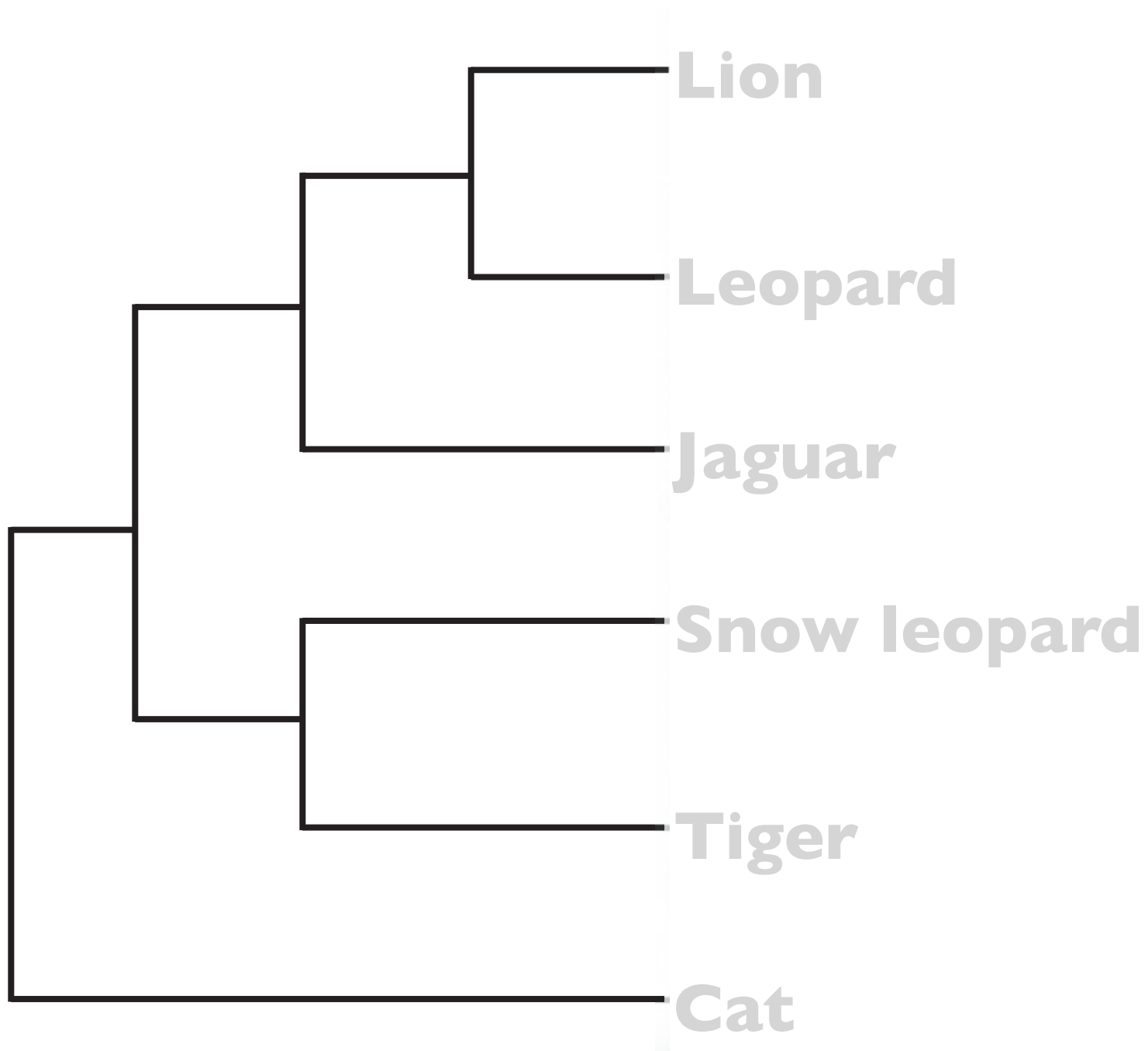
● ● ● Function seen in the reference data
● ● Novel function

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



ONTOLOGICAL CLASSIFICATION OF UNSEEN ANIMALS

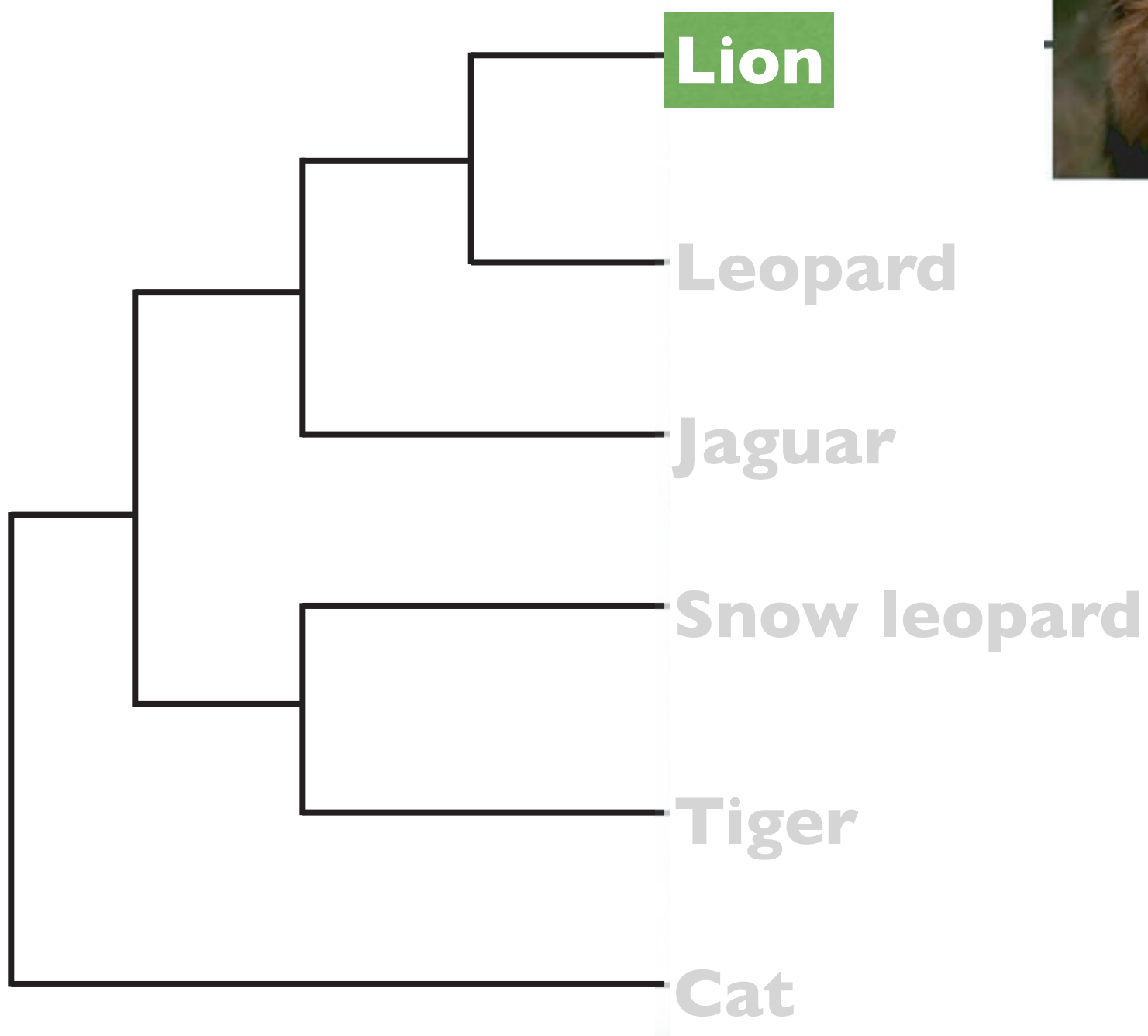
Ontology of great cats





ONTOLOGICAL CLASSIFICATION OF UNSEEN ANIMALS

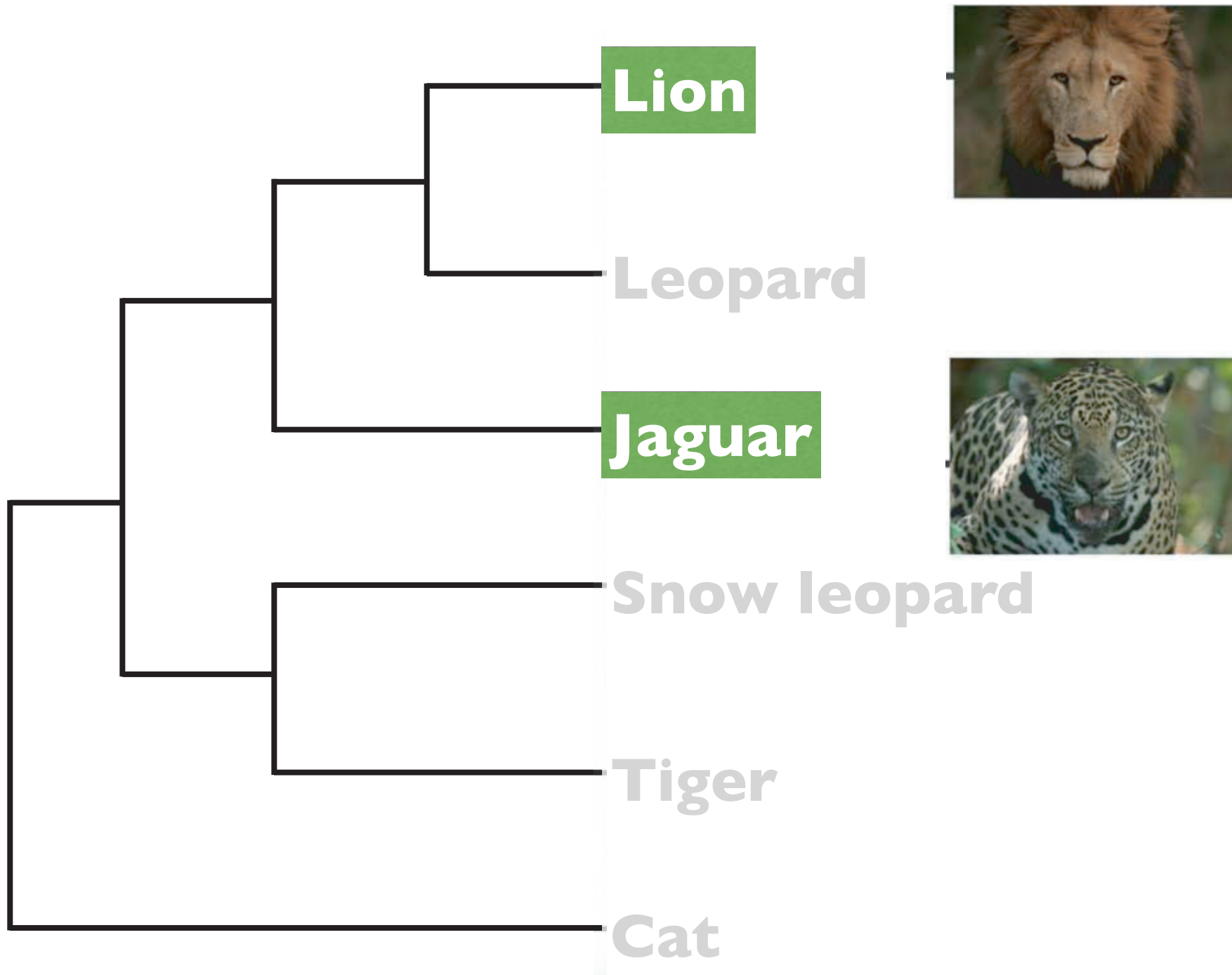
Ontology of great cats





ONTOLOGICAL CLASSIFICATION OF UNSEEN ANIMALS

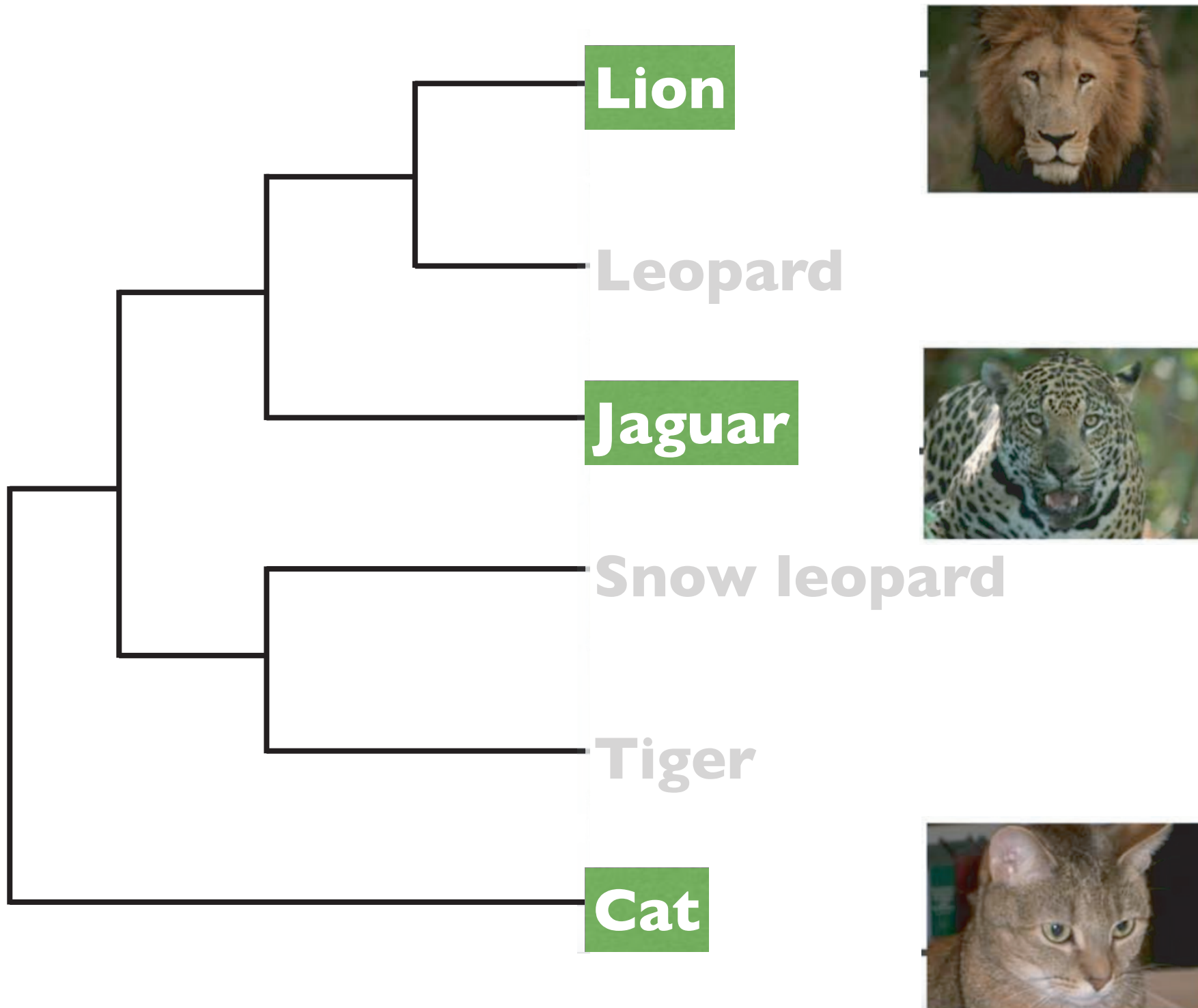
Ontology of great cats





ONTOLOGICAL CLASSIFICATION OF UNSEEN ANIMALS

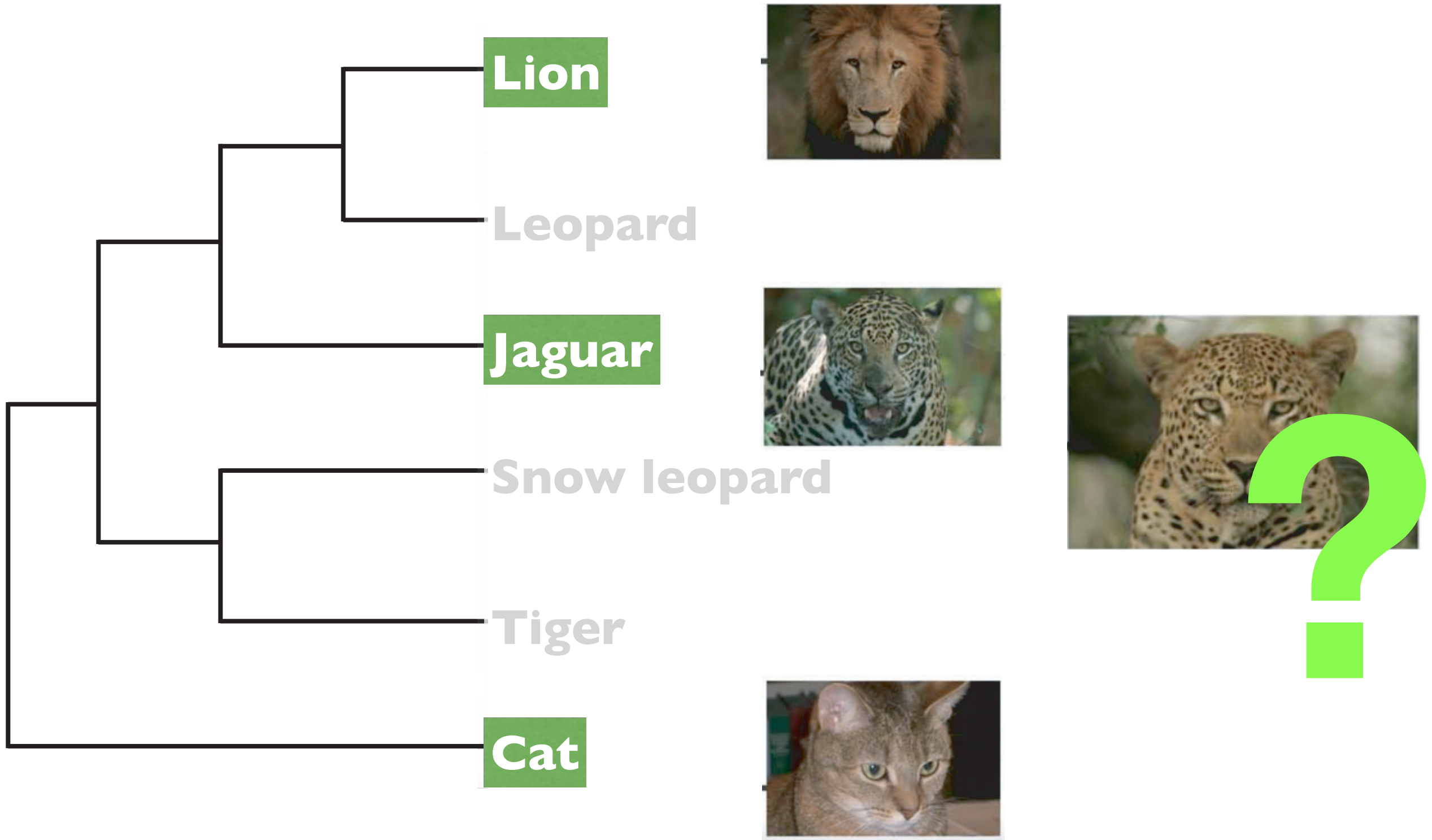
Ontology of great cats





ONTOLOGICAL CLASSIFICATION OF UNSEEN ANIMALS

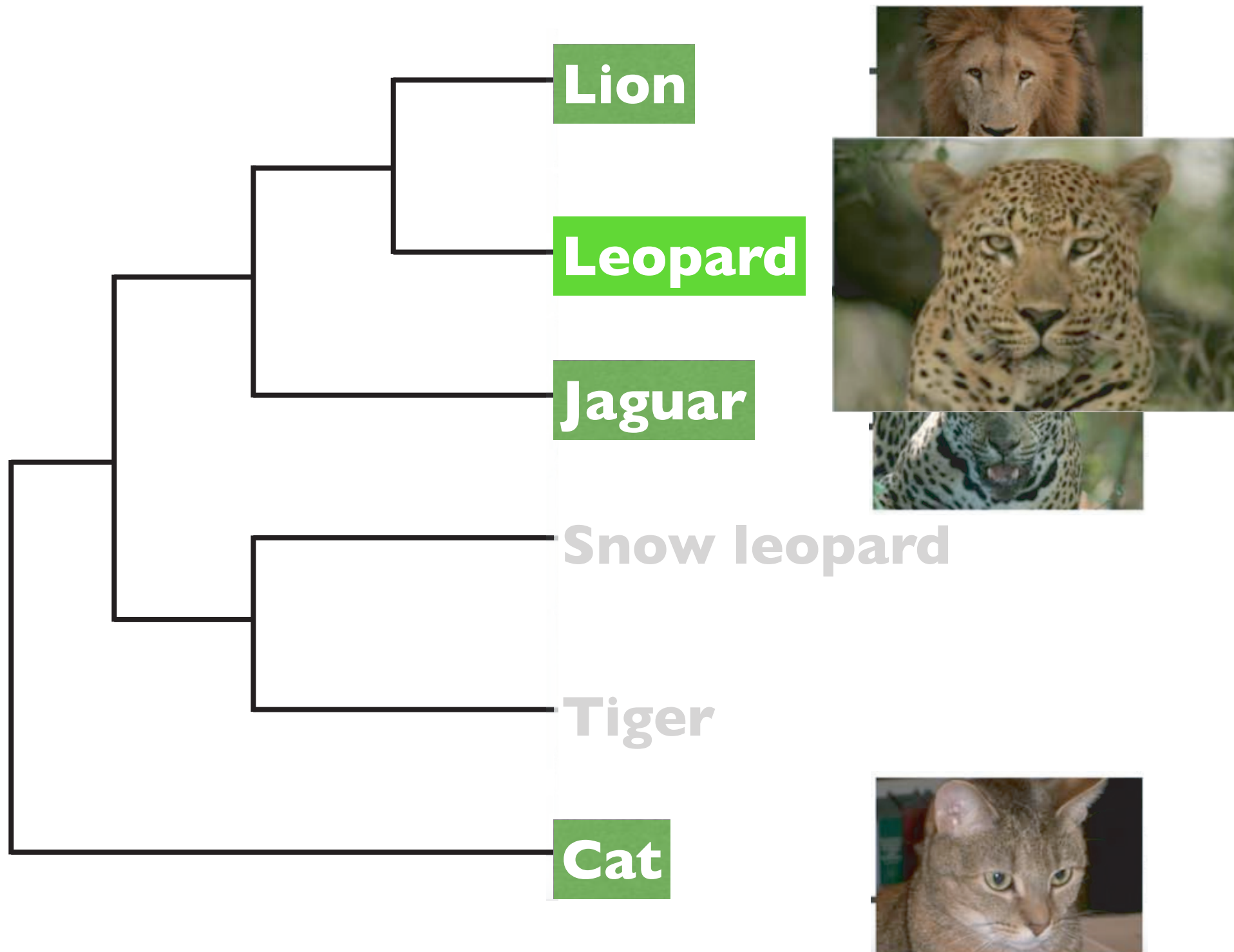
Ontology of great cats





ONTOLOGICAL CLASSIFICATION OF UNSEEN ANIMALS

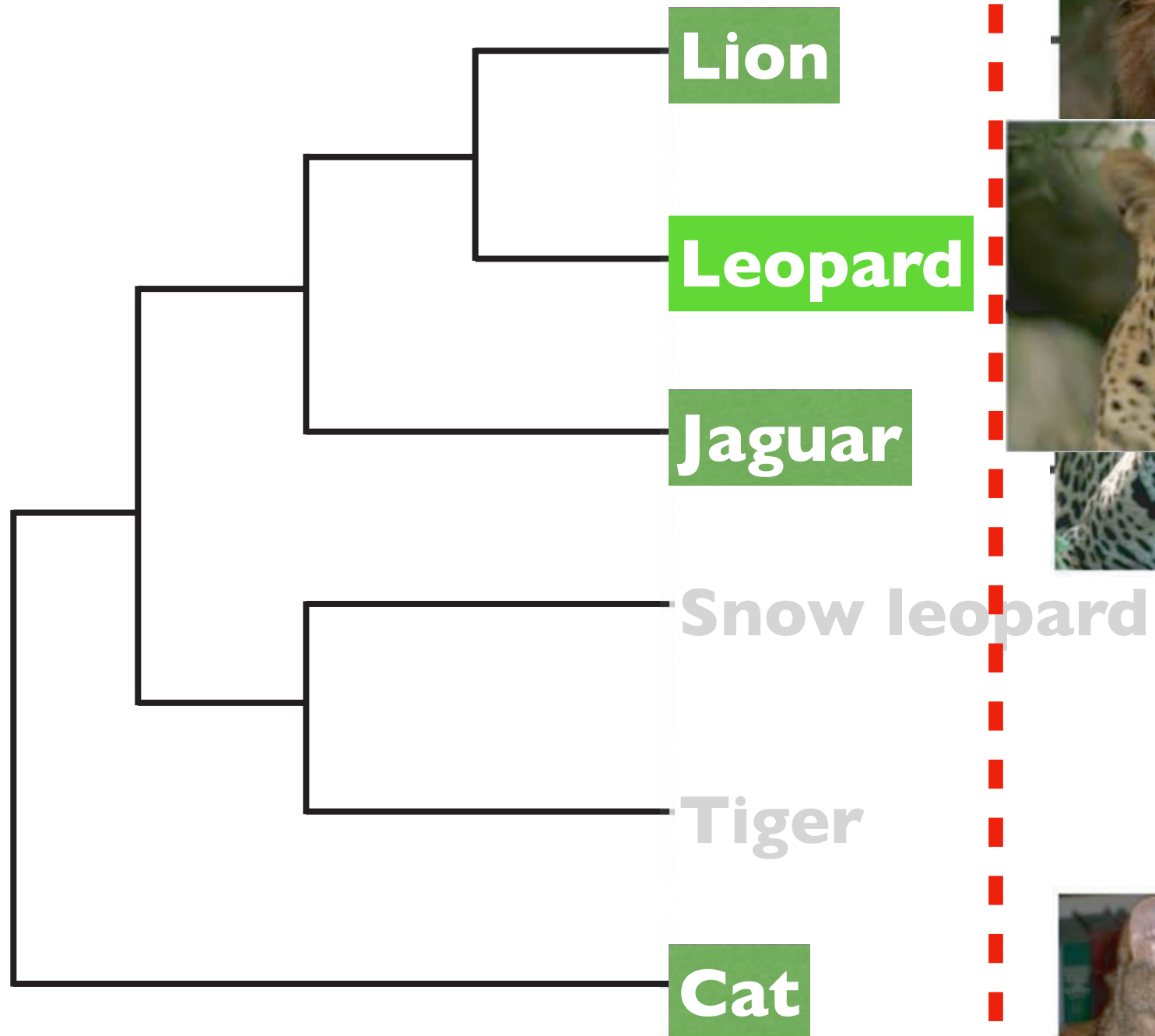
Ontology of great cats



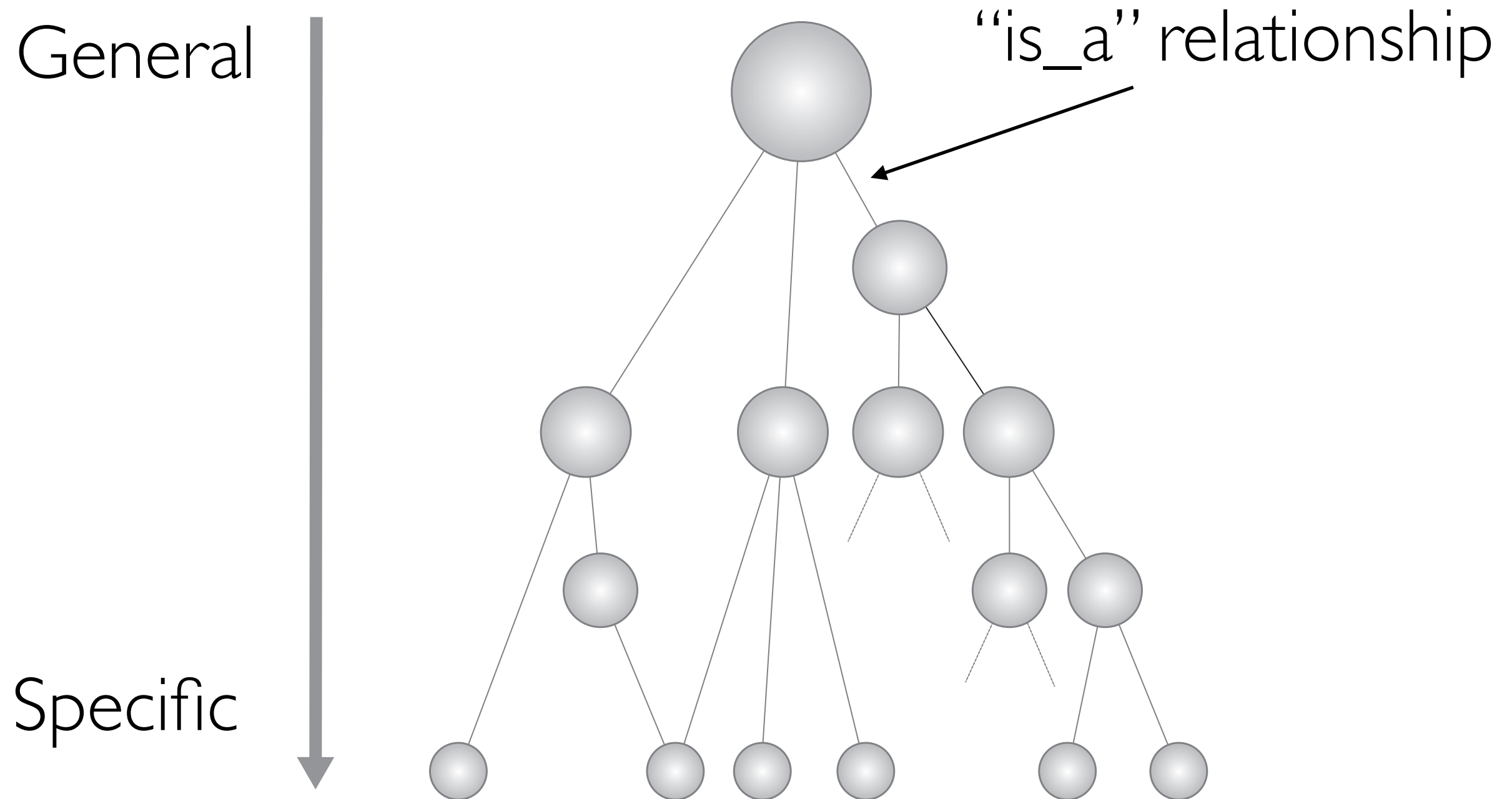


ONTOLOGICAL CLASSIFICATION OF UNSEEN ANIMALS

Ontology of great cats



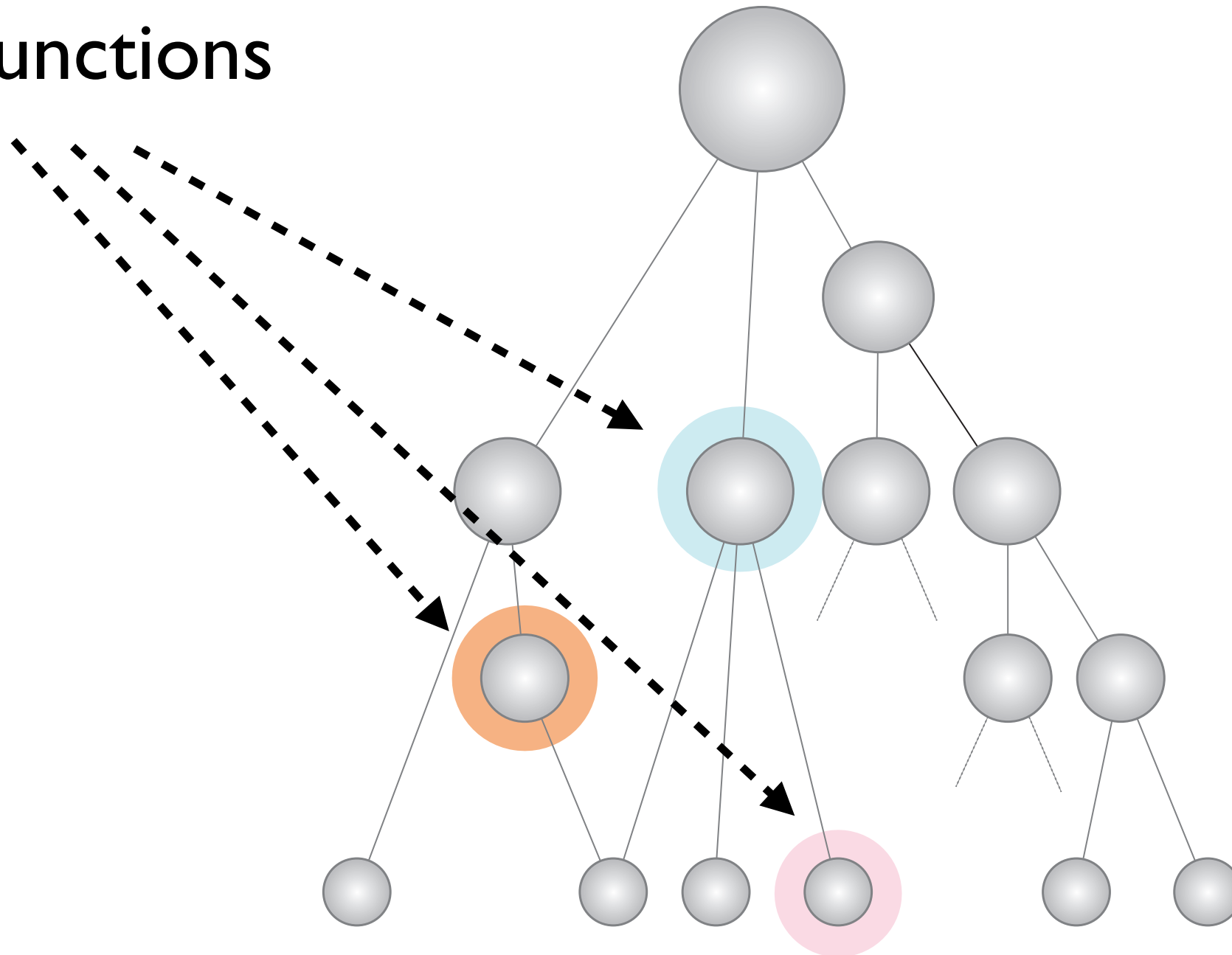
solution: use gene ontology as side information



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

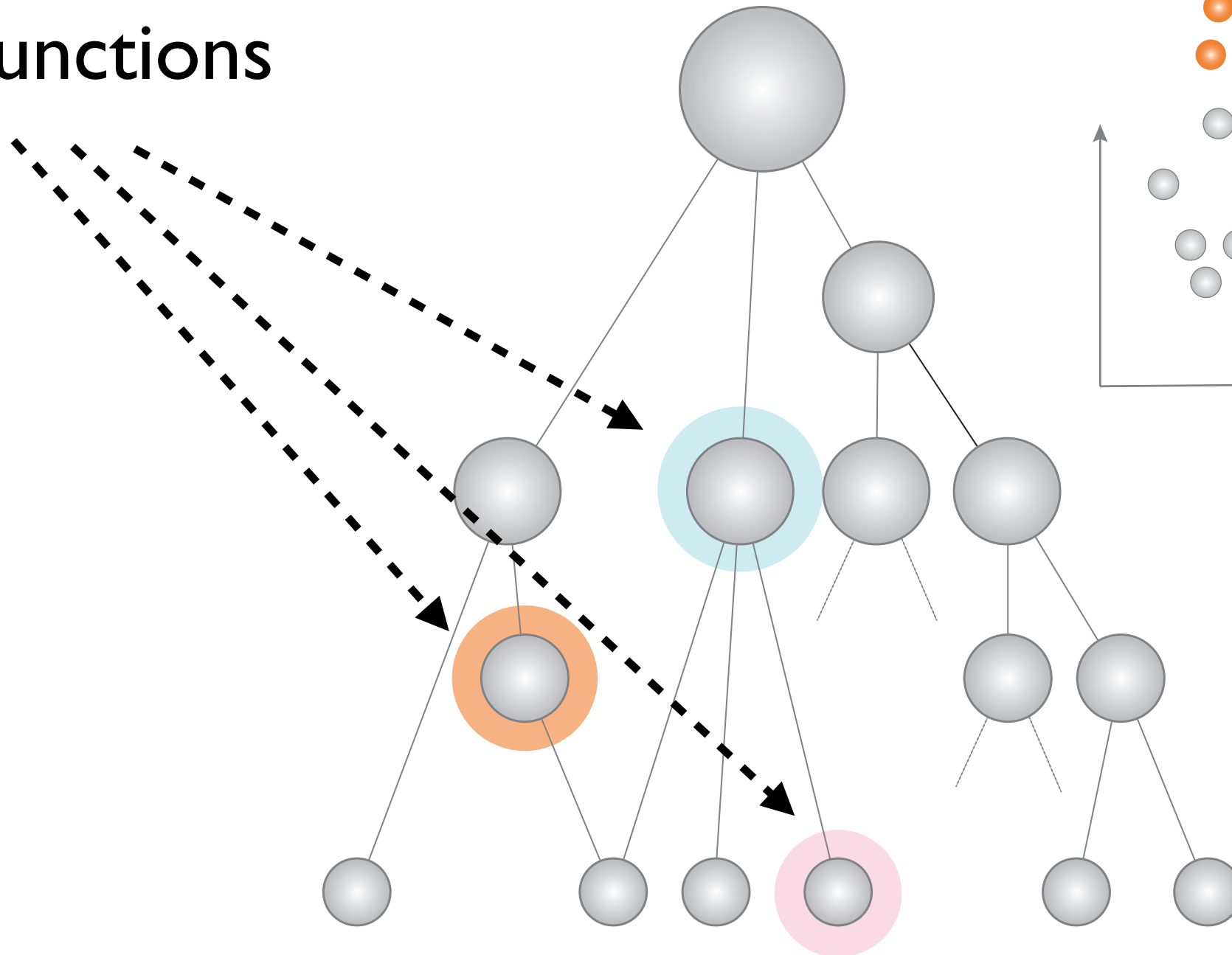
solution: use gene ontology as side information

Seen functions

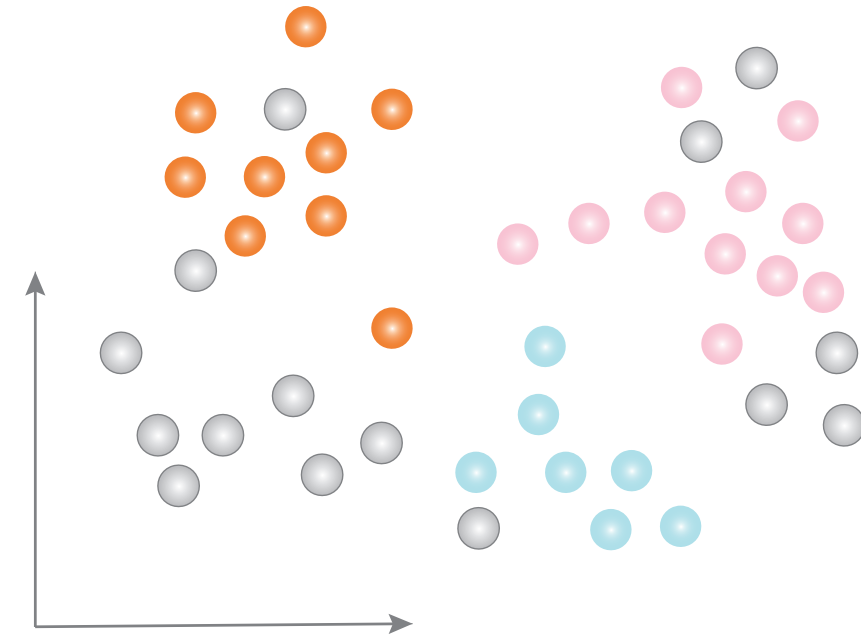


solution: use gene ontology as side information

Seen functions

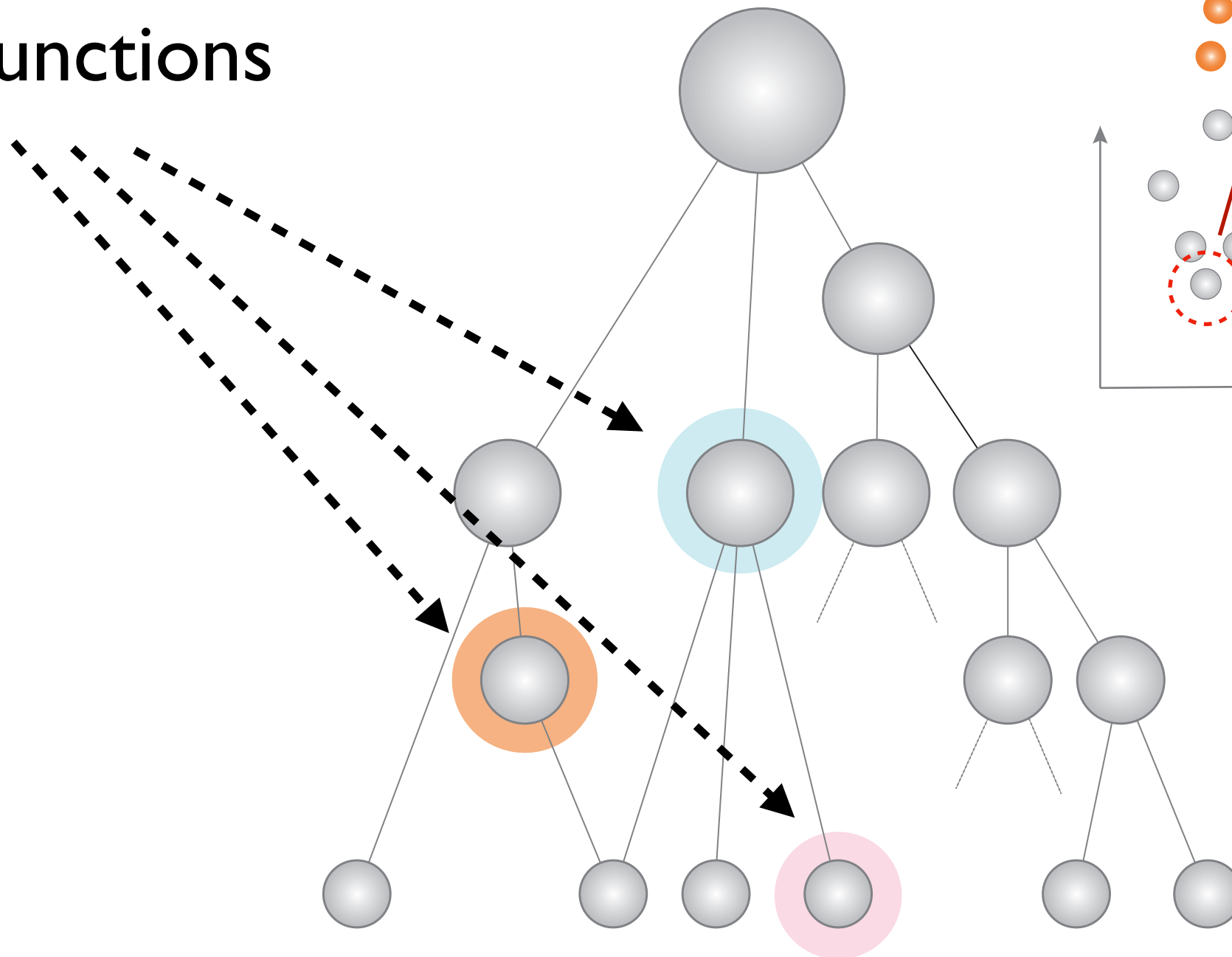


Protein embedding space

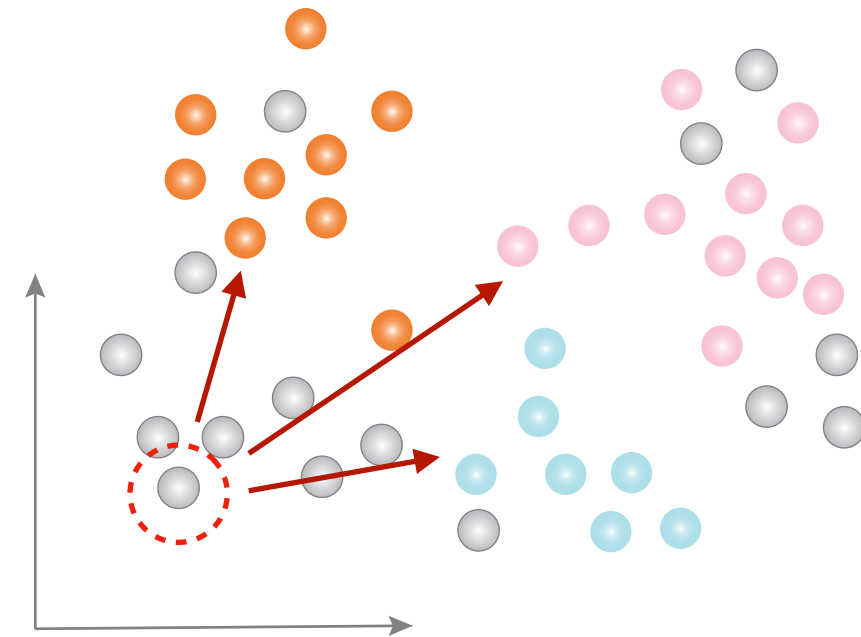


solution: use gene ontology as side information

Seen functions



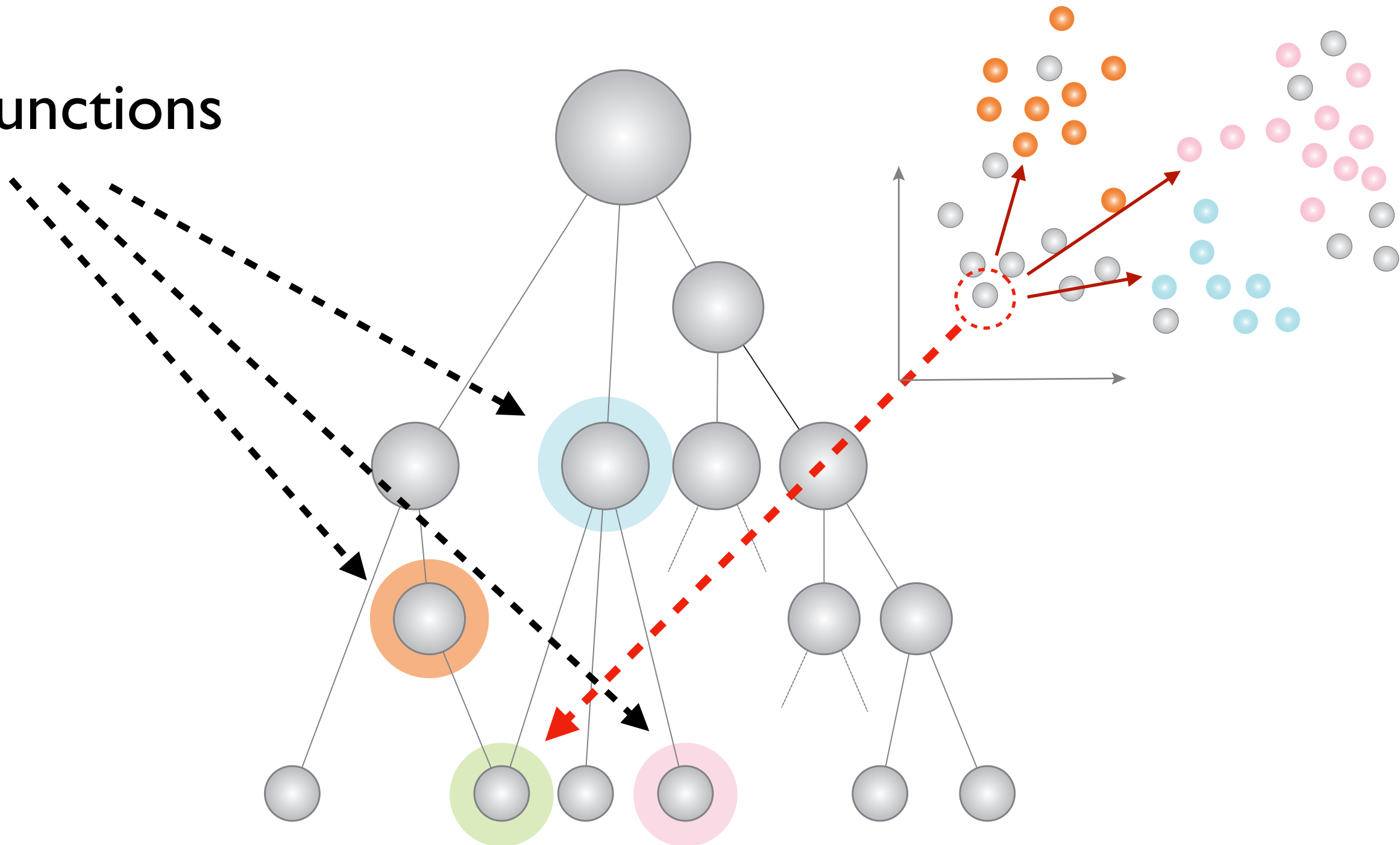
Protein embedding space



solution: use gene ontology as side information

Seen functions

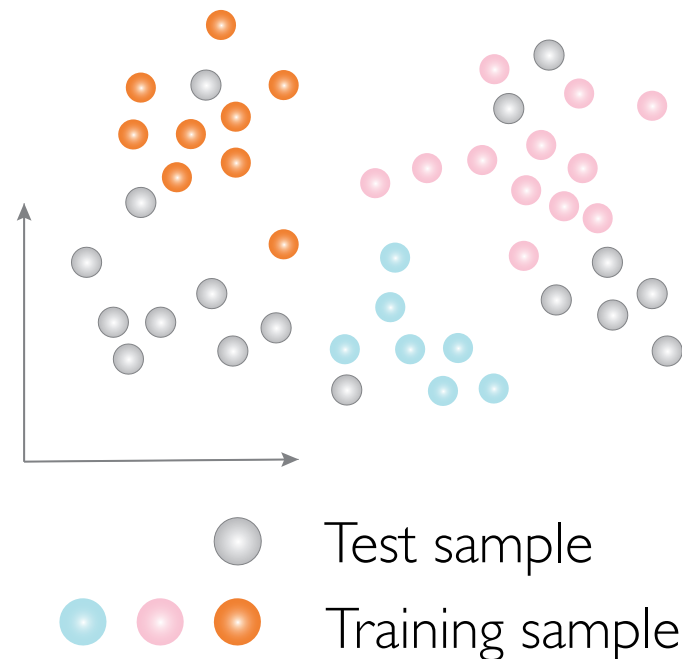
Protein embedding space



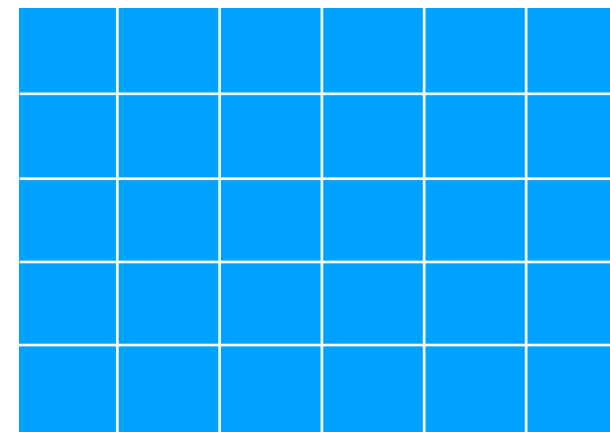
How to classify a test sample using class embeddings?

Input:

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



Class embeddings
(e.g., disease embeddings)



Embeddings $\mathbb{R}^d$

SARS

MERS

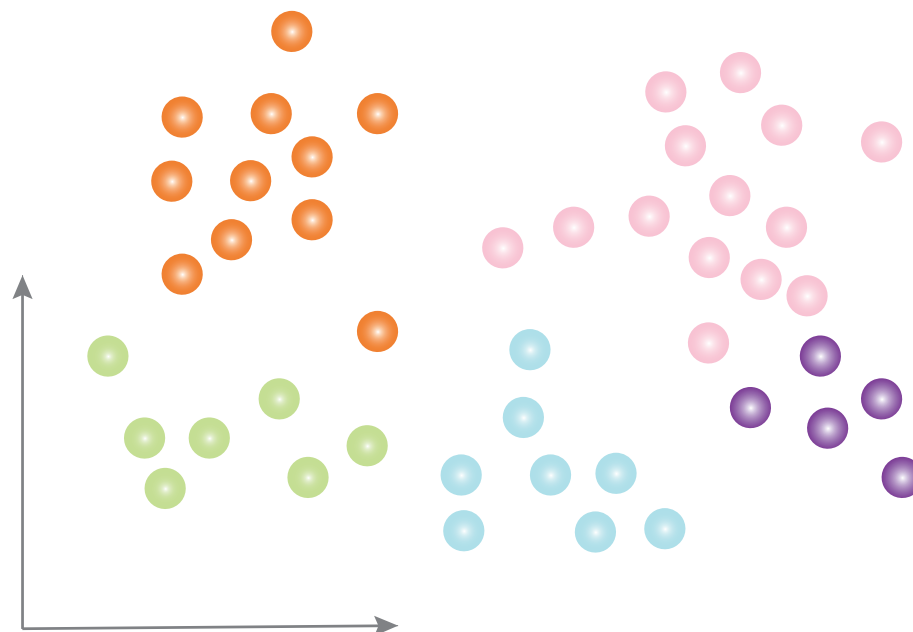
Pig coronavirus

Rat coronavirus

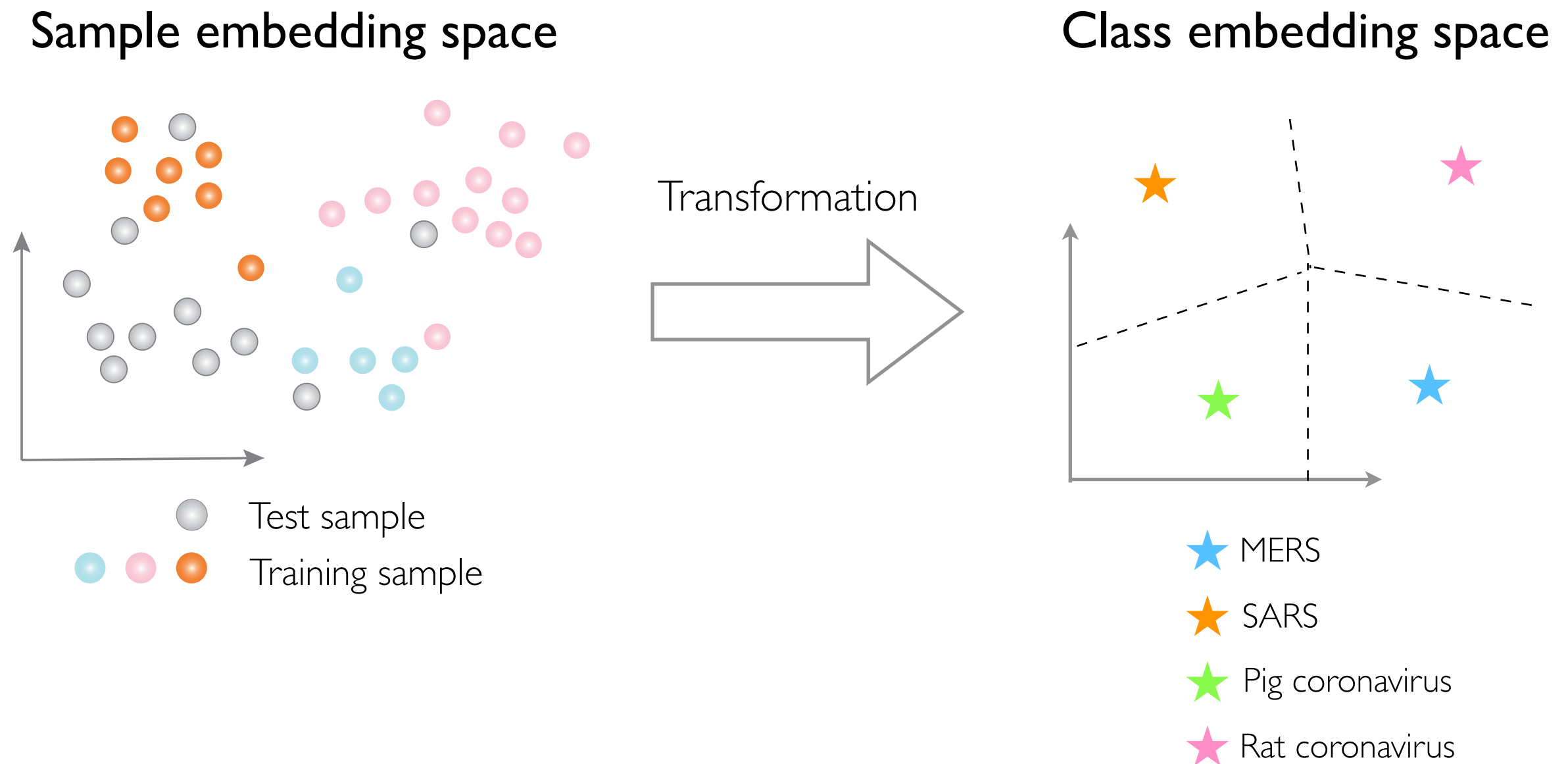
Beta coronavirus

Never-before-seen class

Desired output:

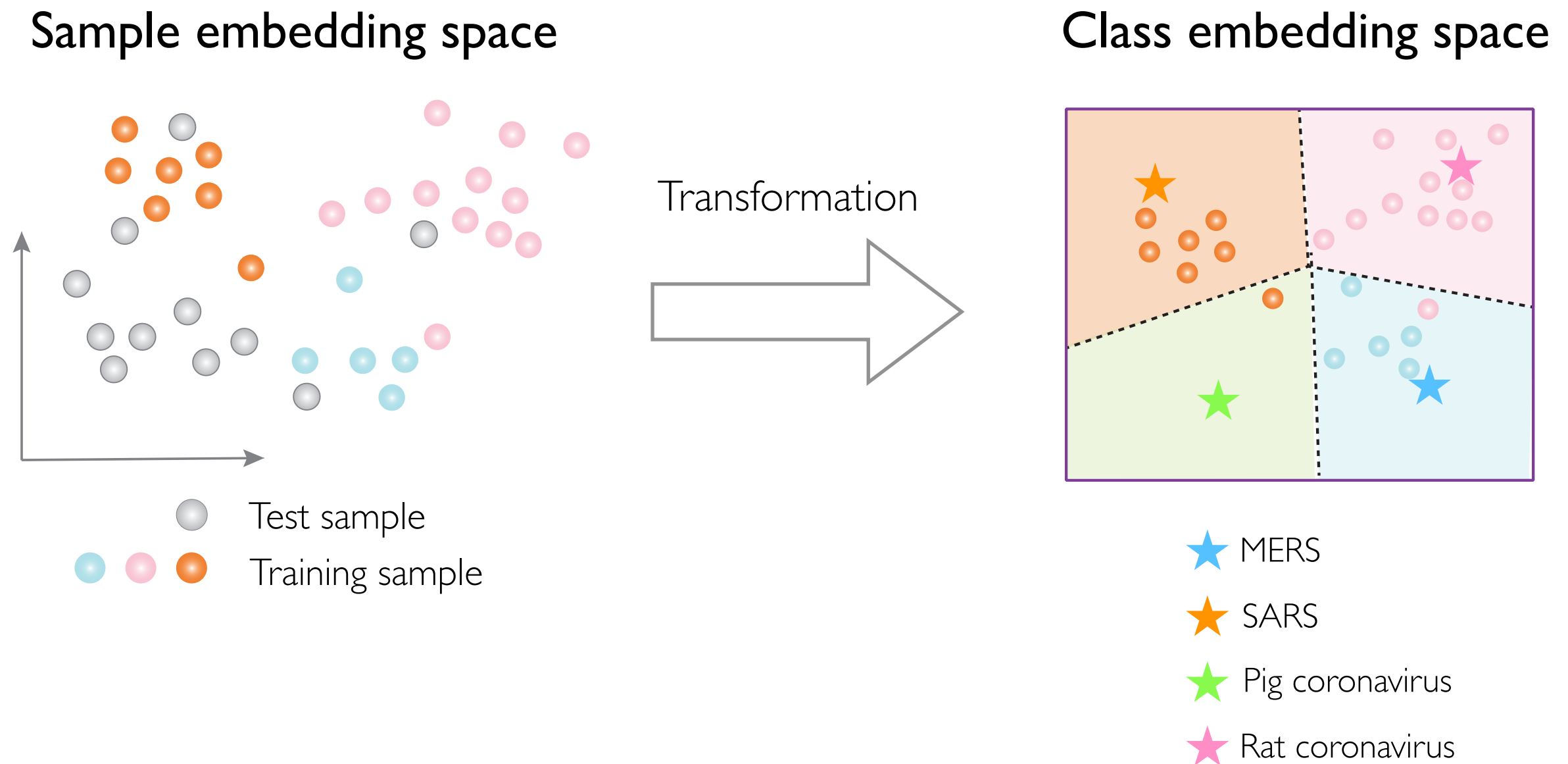


Training: find a transformation that projects each training sample close to the embedding of its class



Draw boundaries according to the midpoint between class embeddings

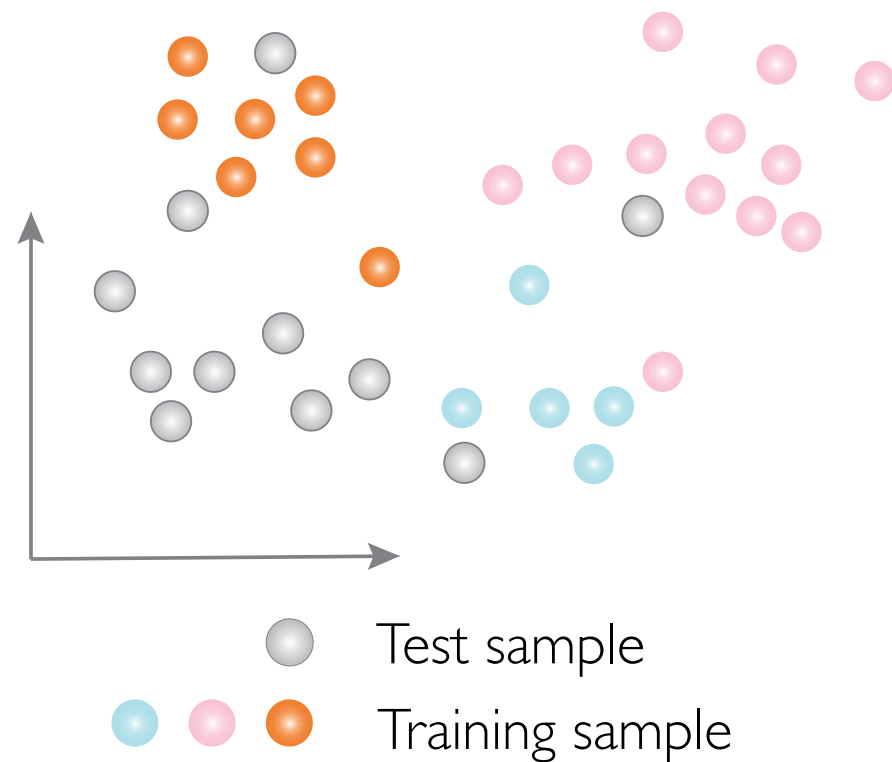
Training: find a transformation that projects each training sample close to the embedding of its class



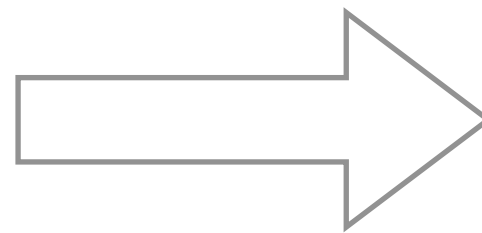
Draw boundaries according to the midpoint between class embeddings

Test: project test samples using the same transformation

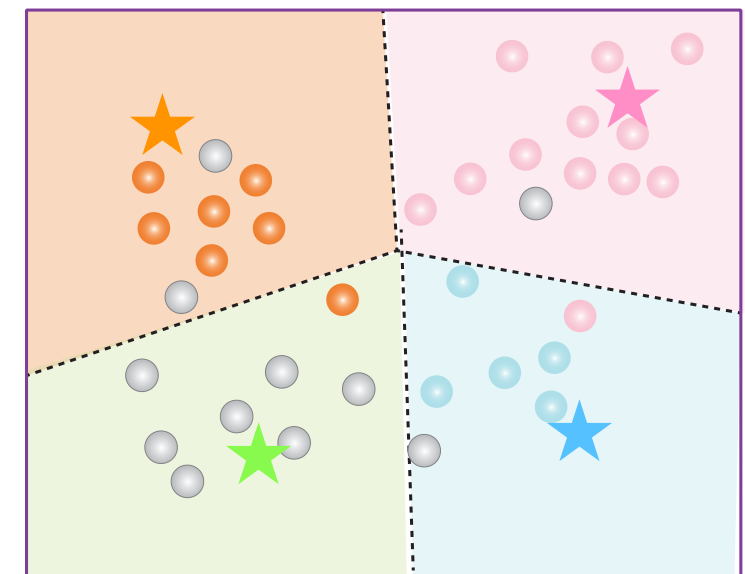
Sample embedding space



Transformation



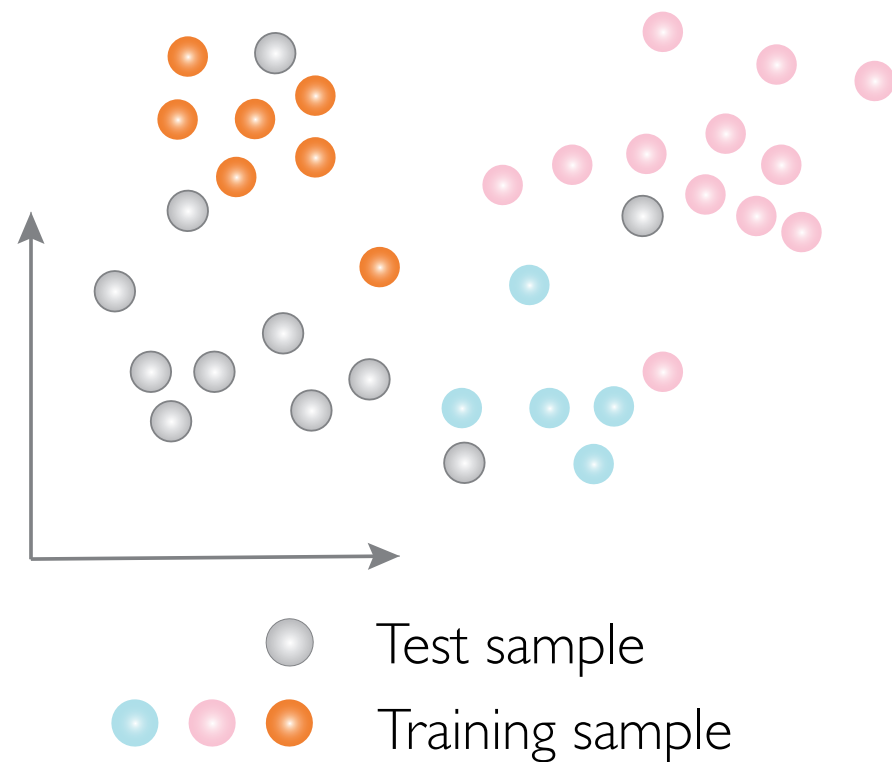
Class embedding space



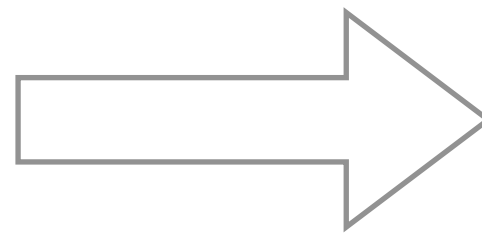
- ★ MERS
- ★ SARS
- ★ Pig coronavirus
- ★ Rat coronavirus

Test: classify samples to the nearest class

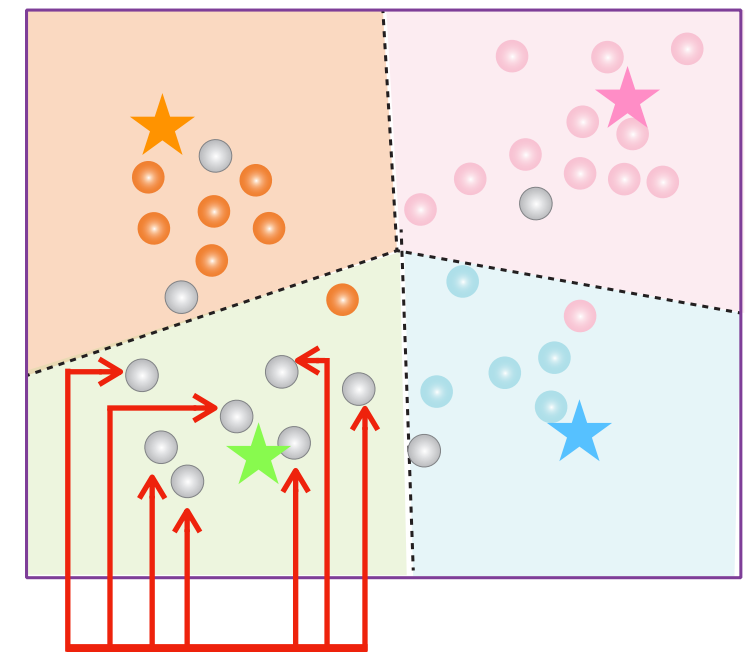
Sample embedding space



Transformation



Class embedding space



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

Feature of training sample i (input)

Class embedding of class j (input)

Could be other neural network architectures (parameter)

$$\hat{y}_{ij} = \frac{e^{f_i^T \mathbf{W} \mathbf{x}_j}}{\sum_k e^{f_i^T \mathbf{W} \mathbf{x}_k}}$$

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

Test stage:

Classify to the nearest class.

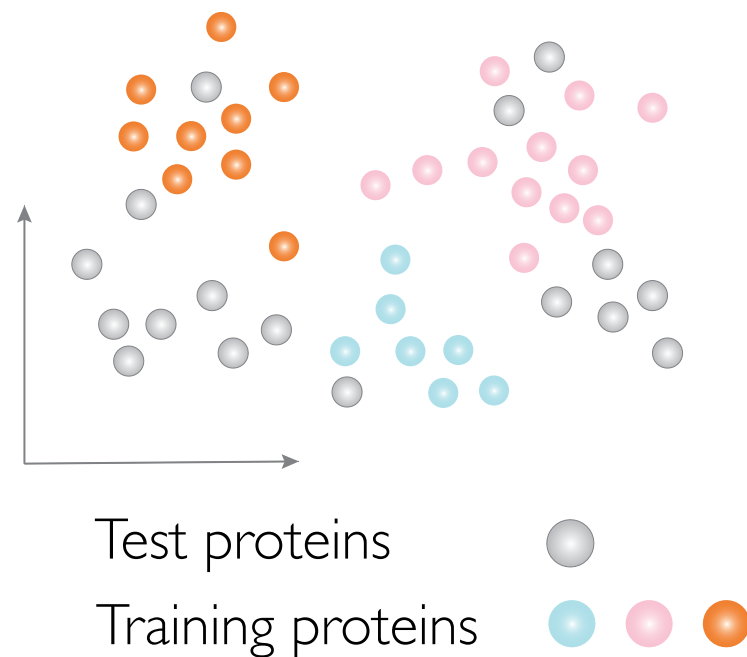
Feature of the test sample

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

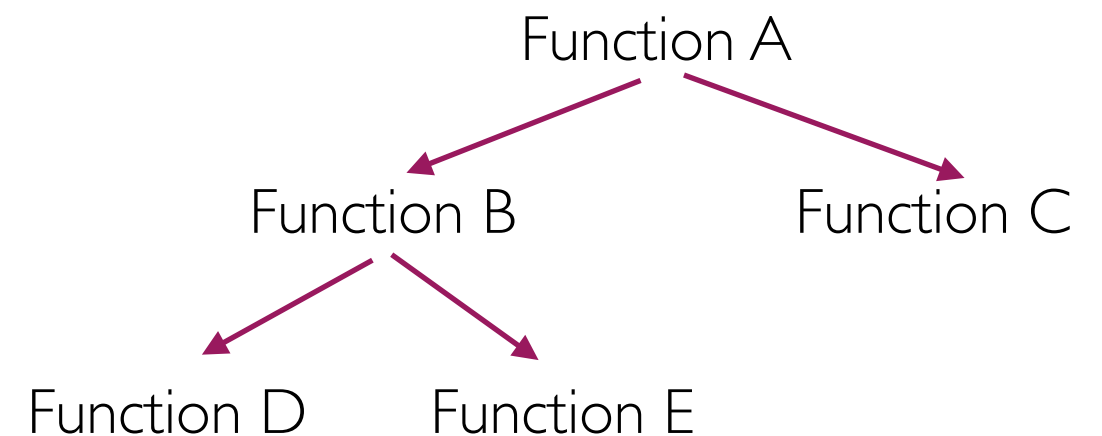
Experimental setting: classify proteins into functions

Input:

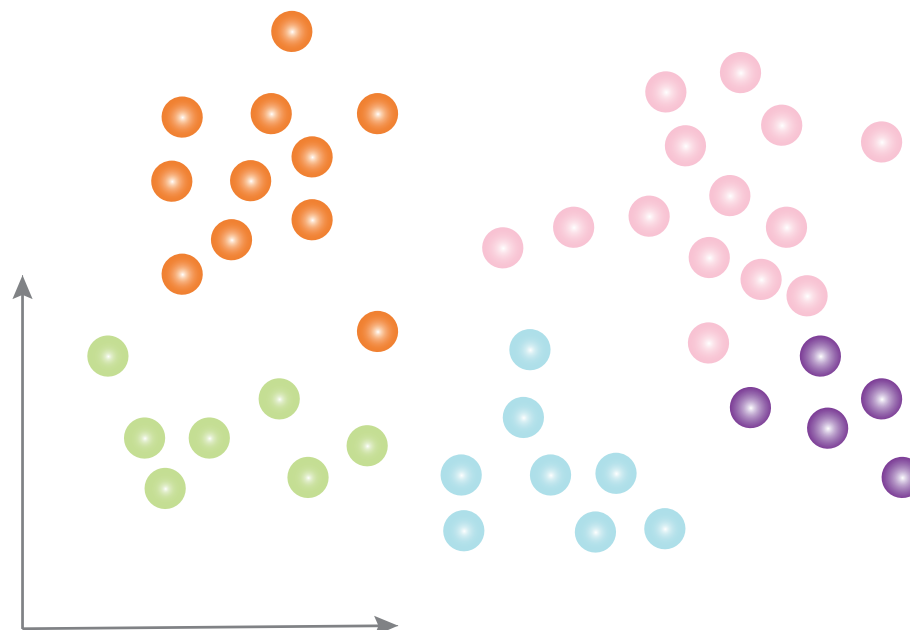
Feature representation of 60k
proteins from 5 species



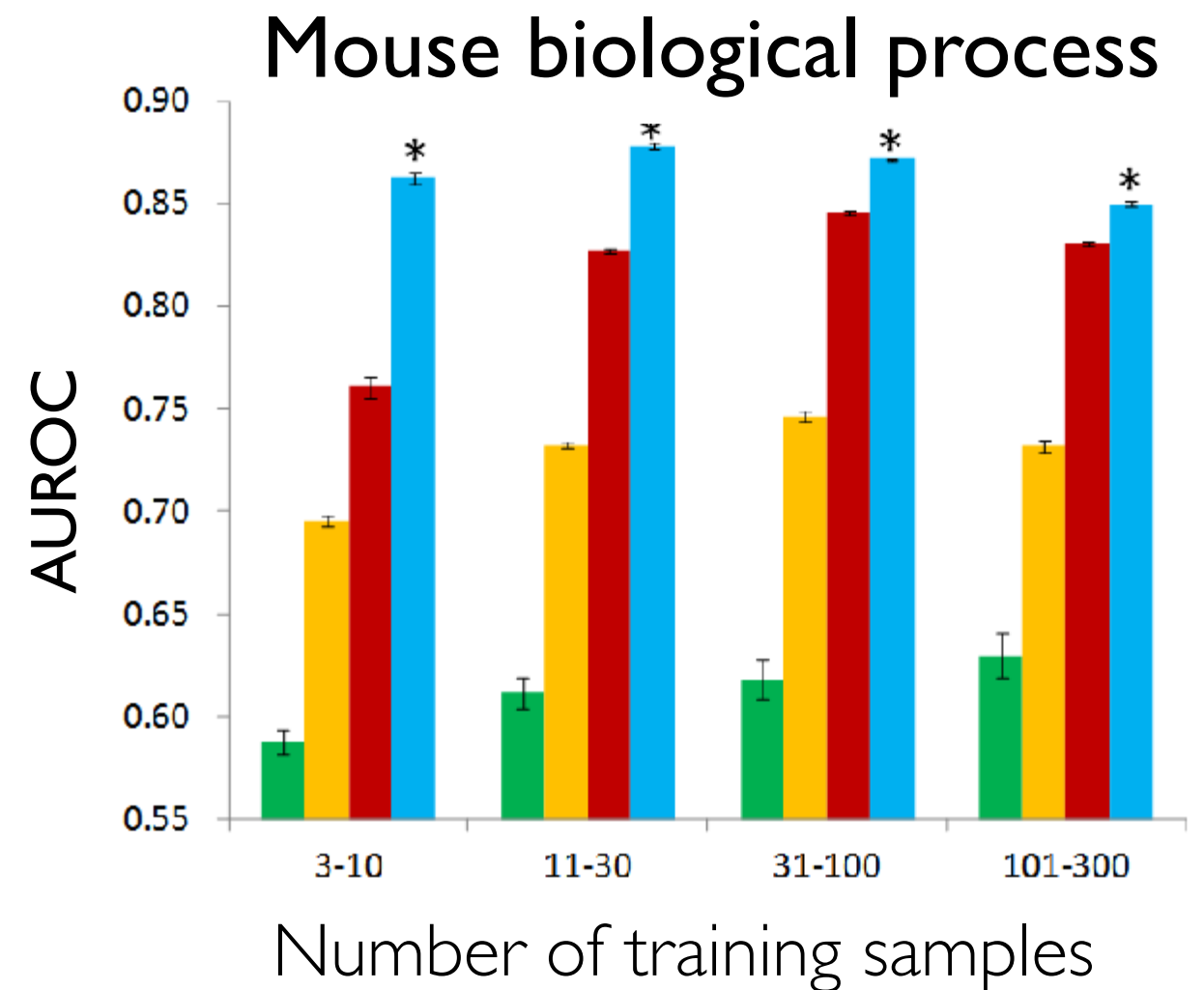
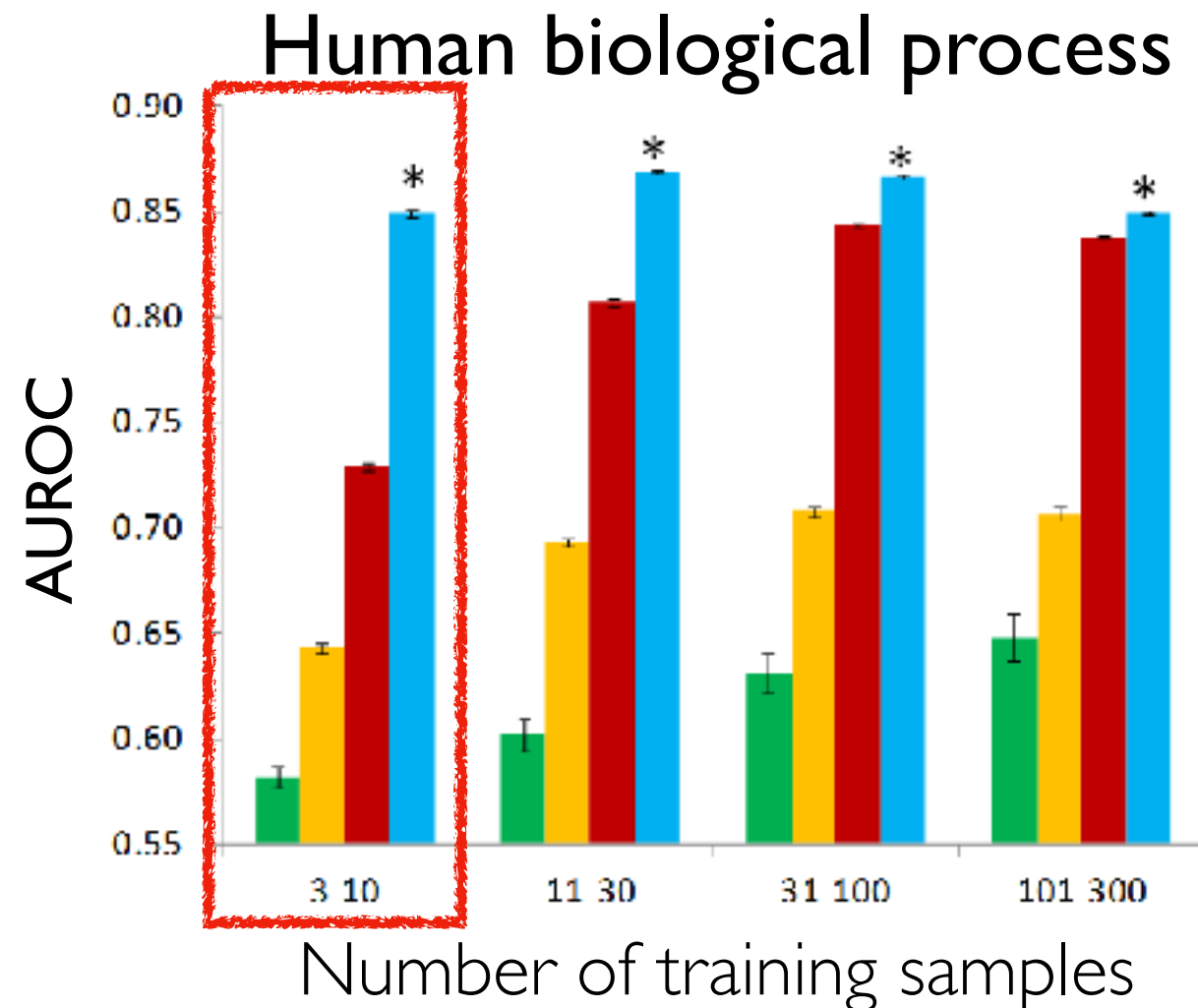
Hierarchy of 13k protein functions



Desired output:



Significant improvement in few-shot classes on all five species



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]

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://egglog.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 I. ECs, etc.)
	SVM-Prot [51]	http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi	✓	Functional categories: 130 (all 2nd I. 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 I. 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 http://www.cytoscape.org/	n/a	Cytoscape plugin using MCL 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.

^aThis 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).

^bIndicates 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>