

# What Can an AI Competition do for You?

Iddo Friedberg  
Iowa State University  
College of Veterinary Medicine



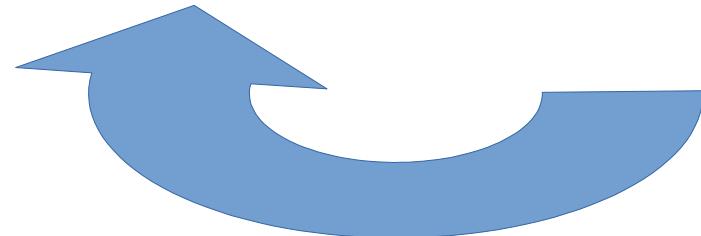
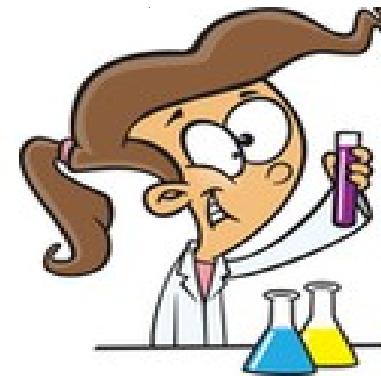
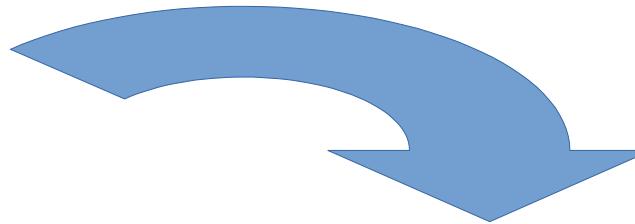
@iddux

[iddo-friedberg.net](http://iddo-friedberg.net)

# Today's Goal

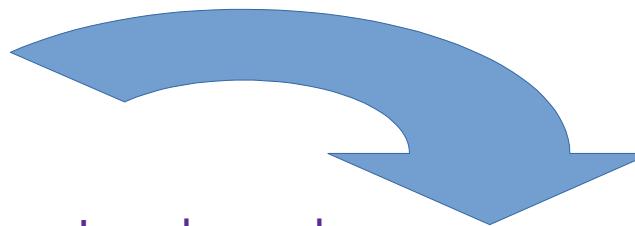


# Methods and Discovery

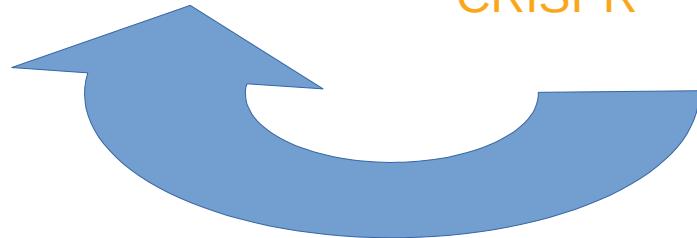


# Methods and Discovery

- Patch clamping
- Mass spec
- RNASeq
- Bioinformatics
- CRISPR



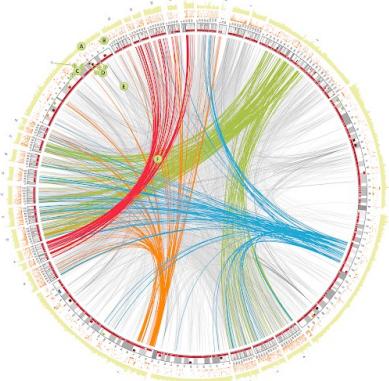
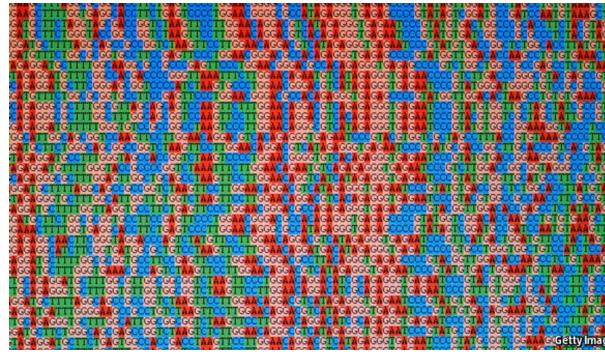
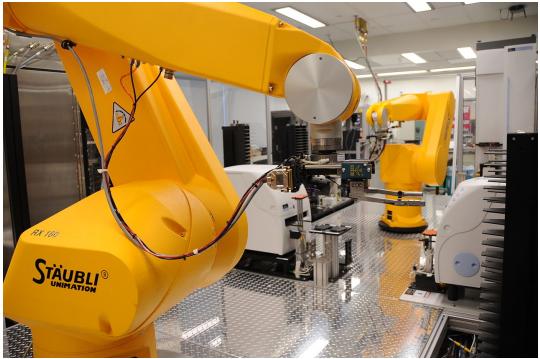
- Ion channels
- Protein expression
- Gene regulation
- Everything
- CRISPR



# Science is Symbiotic



# Genomics: More Data than Knowledge



# Predictive Models in Biology

- Data are abundant, but knowledge is lacking
  - Protein Sequence to Structure
  - Protein Sequence to Function
  - Image to Phenotype
  - Phenotype to Genotype
  - Genotype to Phenotype
  - Symptoms to Disease

# There are Many Methods to do one thing: Which is Best?



CHEMEX



EVA SOLO



MOKA POT



FRENCH PRESS



HARIO V60



AEROPRESS



VACUUM POT



PERCOLATOR



TURKISH GEZVE



DRIP BREW



ESPRESSO

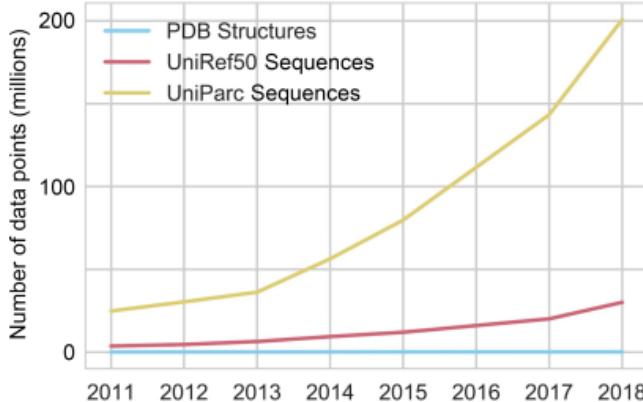


SYPHON

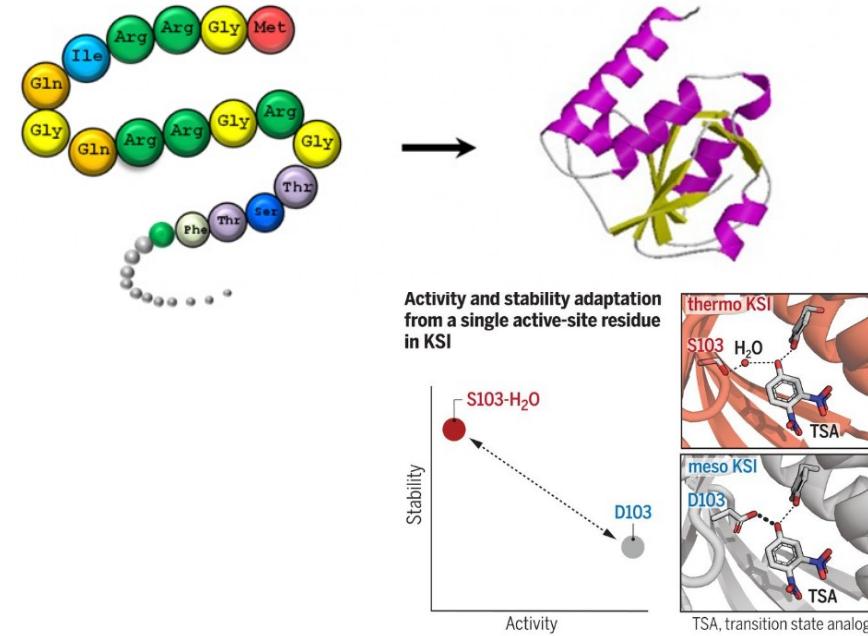
# What is “Best”?



# The Protein Structure Prediction Problem



Credit: Mohammed AlQuraishi



Pinney et al Science(2021) 371:6533

Sequence data: easy to obtain; Structure data: hard to obtain

An important problem: structure tells us how the protein functions

A hard problem: there is a large search space

# CASP: Critical Assessment of Structure Prediction

**Problem:** protein folding, or, predicting structure from sequence

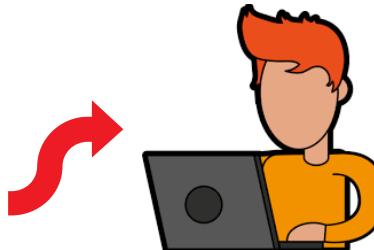
**Communities:** Structural biologists & bioinformaticians

**Challenge set:** proteins whose structures were experimentally determined.

**Assessment metrics:** Root Mean Square Deviation, GDT\_TS, etc.

```
>NP_001375270.1 enosin isoform 4 [Homo sapiens]
MPSATLHERLKTKHARPIPLGLFTINNEDEEOKNGNSRRPKAPPSVYDOKKNASSRPSAISGONN
NHSGMKPPPPRLVDDQRQLARERREEEKOLAAETVWLERERARQHKEERKKRLEEFORQKEE
RRAAEEVKRQRLEEDKERHEAVVRTMERSOKPKOKHNPHGSSLHCSPTHSAAARRLOLSPWESSIV
VNRLLTPTHSLARSKSTAALSGEAACSCPTIMPYKAHSRNNSNDRKPLVFTPPCEGSRRLLHGASSYK
KERERENVFLTSGTRRAVSPSPNKARQPARSLWLPKSCLPHLPGTPRTPSLPPGSVKAAPAVRPPS
PGNIRPVKREVKEPEKDKPEKOKVANEPSSLGKRALPVKEEATVEERTPAEPEVGAAPMAPAPAS
APAPASAPAPAPVTPAMVSAPSTVNASASVKTSGAGTDPEEATRLLAERKRLAREOREKEERERPREQE
ELERQKRELQARVAEERRTRREESSELRLEQAREKEEQDQQAERLAREEAEARQDQKEEEARV
EEAEVRDREKEKHDFREQEEERLERKKKEEMKTRRTTEADTKKTSDRQNGIAKGALTGGTTEVSALPCT
TNAPGNGKPKVGSFHVITSHOSKVTESTPDLPEKOPNENGVSQNENFEETIILP1GKPSRLDVTNSEP
EIPLNPILAFDDEGLPLPOVDGVOVQTQTAEV
```

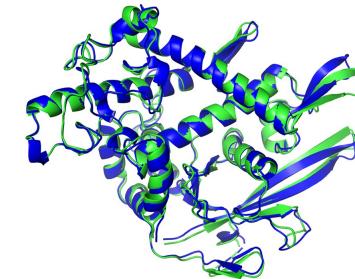
Sequence of unknown structure



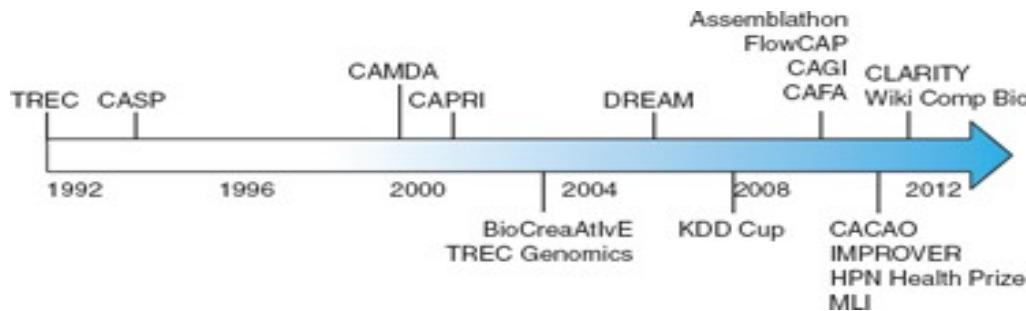
Predictions

Experimental data

Scoring



# History of Competitions



TREC: Text Retrieval Conference run by NIST

CASP: structure prediction

CAPRI: Protein Interaction

BioCreative: Text mining

DREAM: umbrella – many biomedical challenges

CAGI: genotype / phenotype

CAFA: Function

CACAO: educational

# CAFA

- Critical Assessment of Function Annotation
  - Predict function from amino-acid sequence
- **Community:** function predictors, ontologists, experimental biologists
- **Problem:** given a protein, which functions are associated with it?
- **Challenge set:** proteins that organizers know their functions, but predictors don't

# Example: CAFA

- Critical Assessment of Function Annotation
  - Predict function from amino-acid sequence
- Community: function predictors, ontologists, experimental biologists
- Problem: given a protein, which functions are associated with it?
- Challenge set: proteins that organizers know their functions, but predictors don't
- Assessment Metrics: oh, boy...

# Competition Timeline: Getting Benchmark data

>NP\_001375279.1|<seq>SLLIILSAGGAFR...A [None, Aspartic]  
MPCATANLHLKKTNAPPIPLGLFTNWEEDQONWNSRRPKADPSXVODOKKNASSPPASAISGONN  
NHSCKNPPPVILRVDDQRRLARERREERKQLAAREIVMLEREARADHYXKLEERKKRLEEGKQEE  
RKHQEEVYDQVYQVYV  
VIRLLTPHTFLARSKTAALSGAAACSP1DIMPYKAHNSRNSDRPKLFVTTPGGSSRRIHGTA5YK  
KERERENVILFLTSGRRAVSPSNPKAKQPARSLMLPKSLPHLPCTPPTPSLSPRGSVKAAFAVRPSP  
PKERERENVILFLTSGRRAVSPSNPKAKQPARSLMLPKSLPHLPCTPPTPSLSPRGSVKAAFAVRPSP  
APAPASAPAPAPAPVPTAMVSAPSPTWNASAVSKTSACTTPEAATLLEKQRRLAREOREKEERREDE  
EELKEERREDEOREKEERREDEOREKEERREDEOREKEERREDEOREKEERREDEOREKEERREDE  
EEAEVRVDEERKHFQREEQERLERKKKLEIEINRTRTEADTKTSQDQNGIAKGALTGTGVSAVPLCT  
TNAPQNCKPQGSPhVTSQH5KVTVESTDLOKPNENGVSQNEFEEIINLPIGSKPSRLDVNTSESP  
EPPLMPHLAFDQEGTLGPPLPVWVWVETDQTAEV

Sequences of unknown  
structure

Data accumulation phase

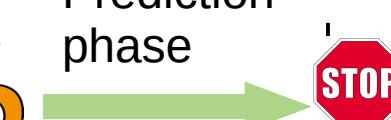


Experiments

Prediction phase



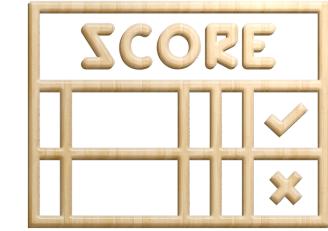
Hide results



Show results



Assess/ Score

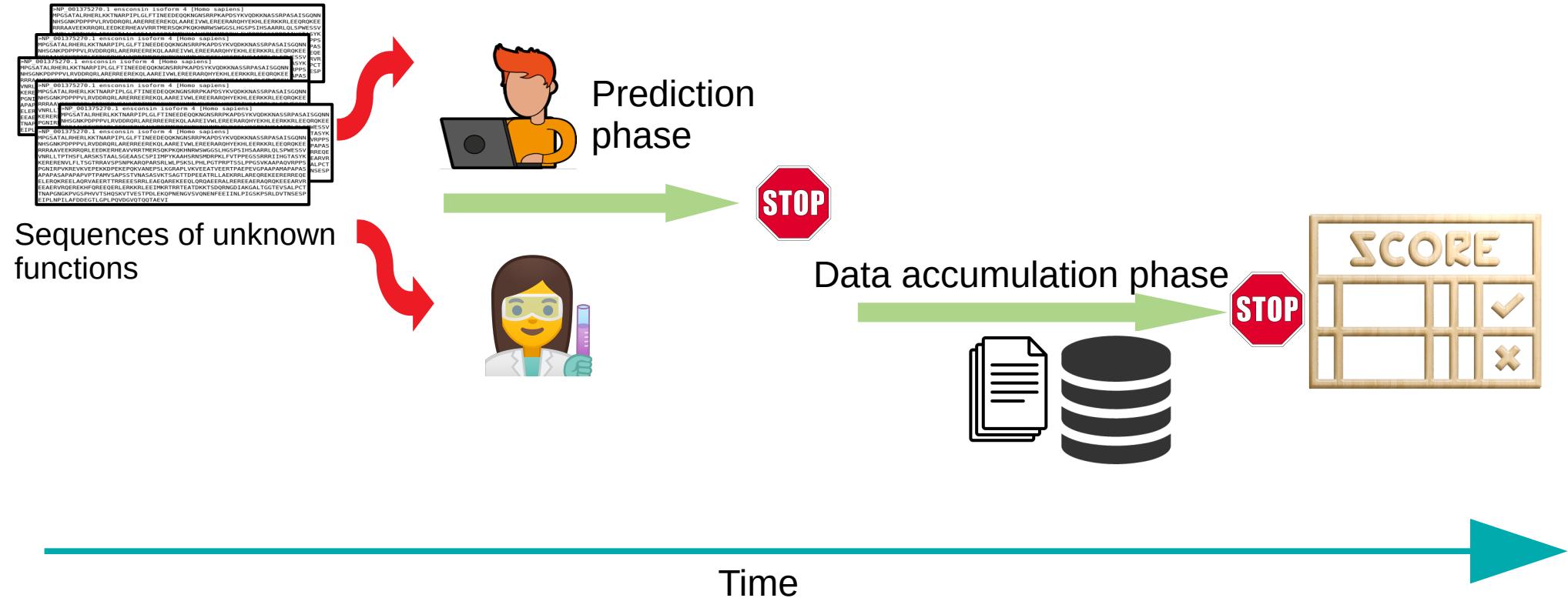


Time

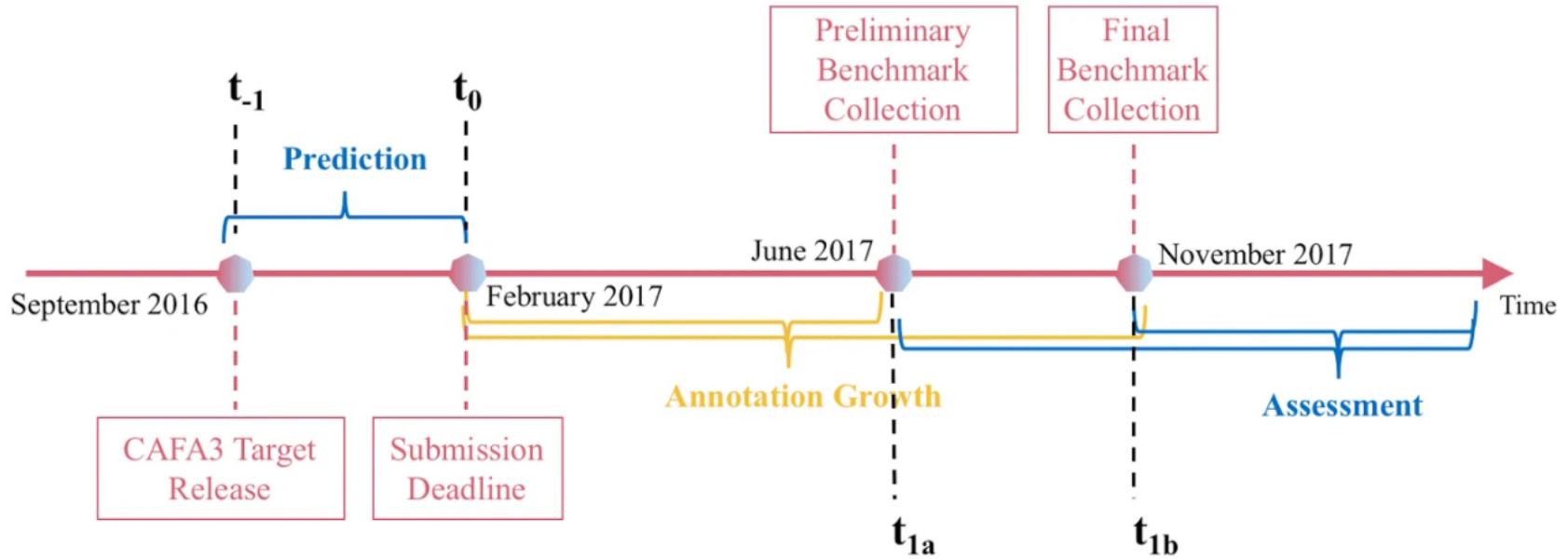


# Big Data competitions: use a time challenge

Data accumulate over time **mostly after** the prediction phase ends



# CAFA Timeline

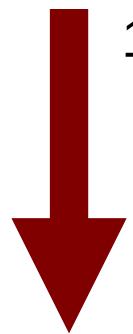


CAFA3 timeline

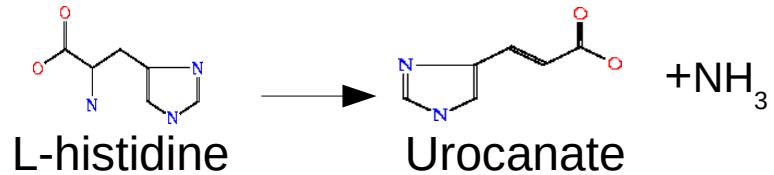
# What is Protein Function?

Histidine amino lyase  
(HAL, Histidase)

Biochemical



$10^{-9} \text{ m}$



# What is Protein Function?

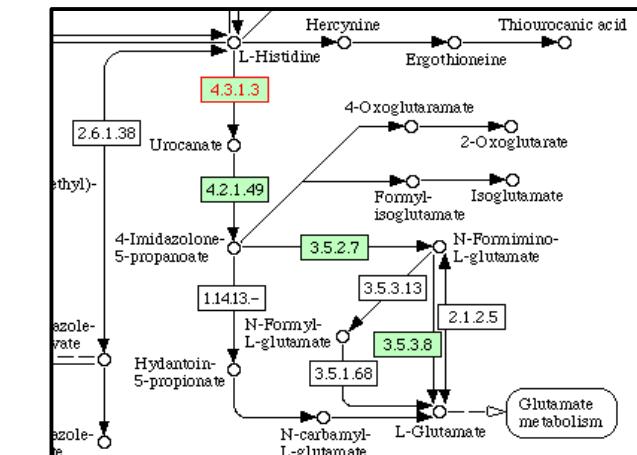
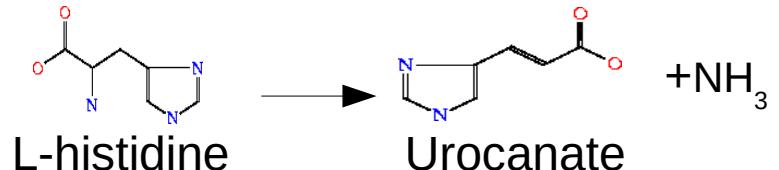
Biochemical

$10^{-9} \text{ m}$

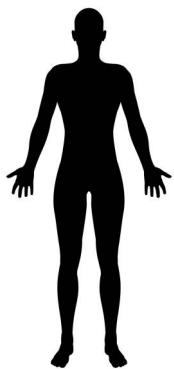
Cellular

$10^{-6} \text{ m}$

Histidine amino lyase  
(HAL, Histidase)



# What is Protein Function?



Biochemical

$10^{-9} \text{ m}$

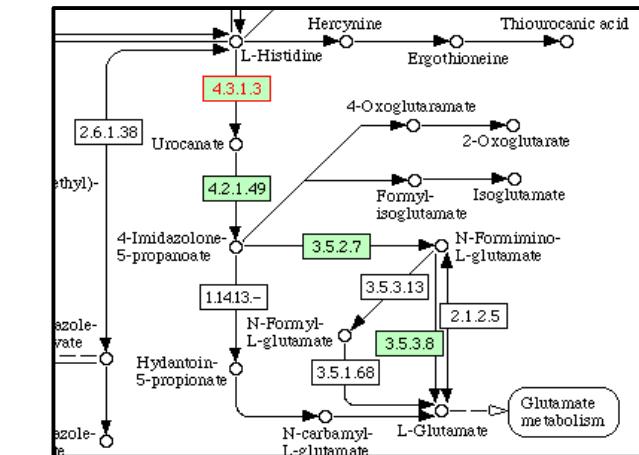
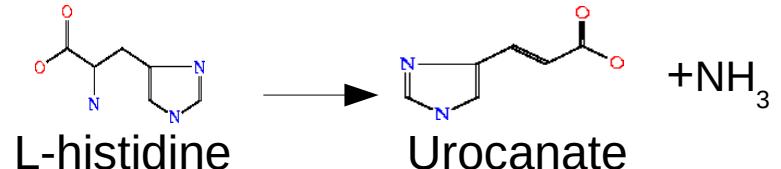
Cellular

$10^{-6} \text{ m}$

Phenotype

$10^{-1} \text{ m}$

Histidine amino lyase  
(HAL, Histidase)



Mutation: Histidinemia

(Mental impairment and speech defect)

# Natural Language isn't standard!

“HAL—which is the first enzyme in the degradation pathway of L-histidine—catalyzes the non-oxidative deamination of its substrate to trans-uropionic acid”.

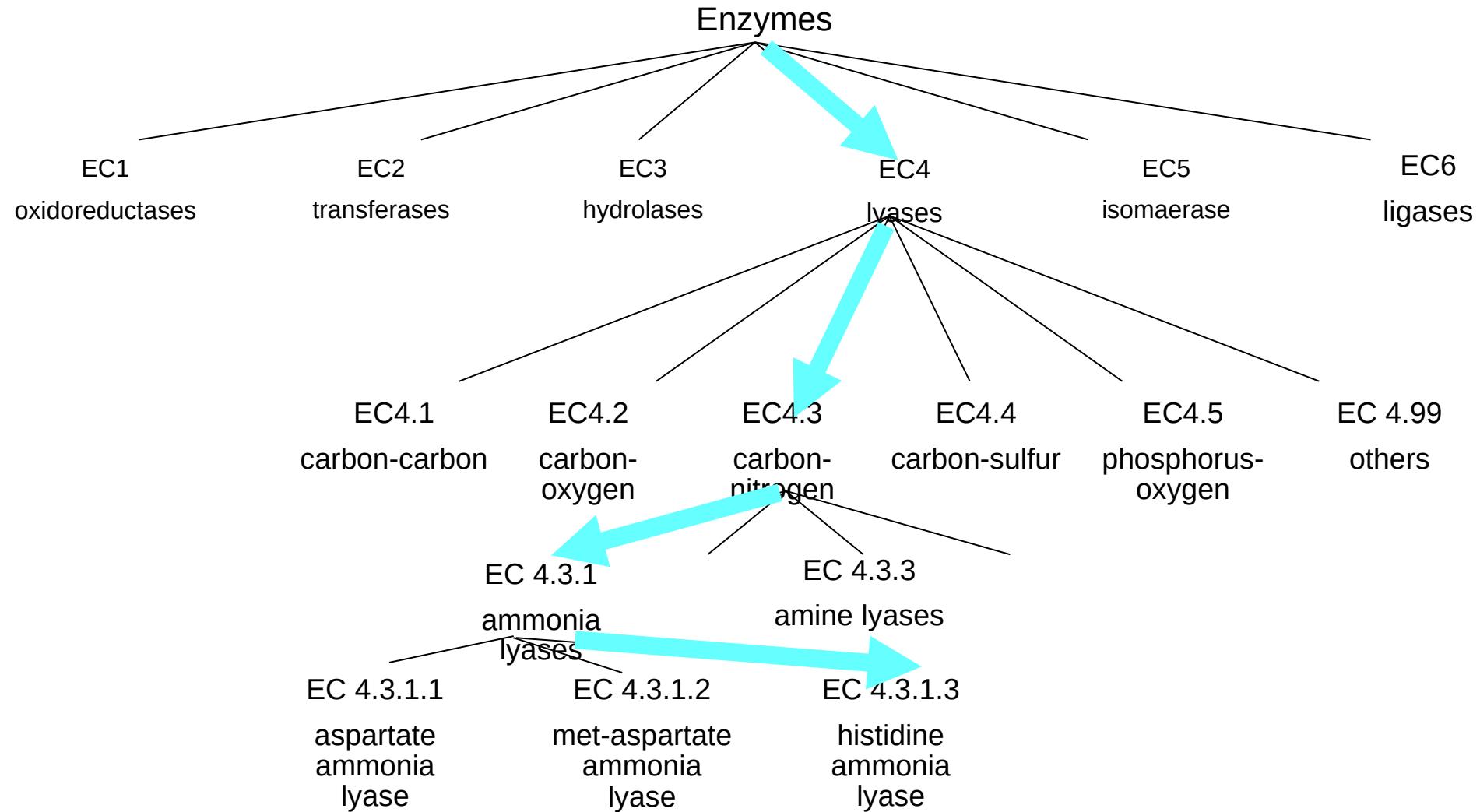
László Poppe, (2001)  
COCB

“Histidase catalyzes the elimination of the alpha-amino group of histidine using a 4-methylidene-imidazole-5-one (MIO), which is formed autocatalytically from the internal peptide segment 142Ala-Ser-Gly.”

# Too many words, synonyms, meanings...

“An **ontology** is the formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse”

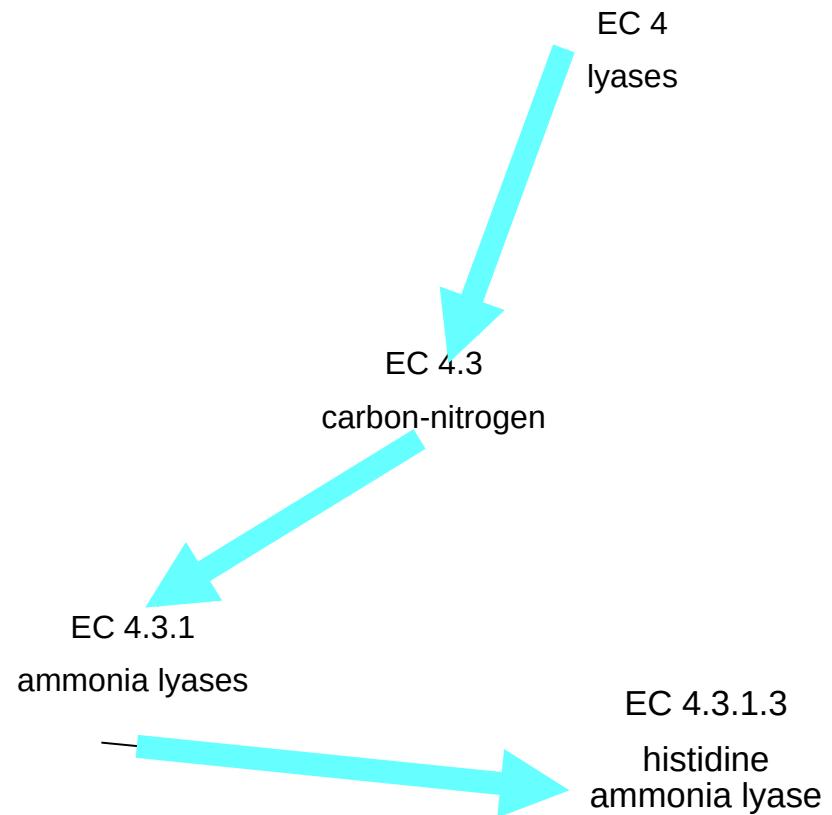
# A Simple Ontology: the Enzyme Commission Classification



# Enzyme Commission Classification

E.C. Provides a relationship description by:

- 1) Using a **controlled vocabulary**
- 2) Going from the general to the specific
- 3) Defining the scope of interest



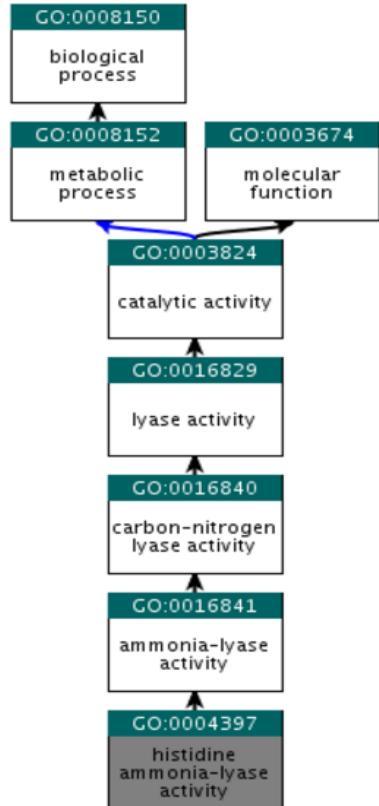
# What do we need?

- A standardized functional vocabulary
- A comprehensive vocabulary
- Capture different functional aspects
- Capture different relationship types

# The Gene Ontology (GO) Project

- A major project standardizing the representation of genes and gene product attributes.
- Major Goals:
  - 1) Develop and maintain a controlled vocabulary and relationships of gene and gene product attributes
  - 2) Annotate genes and gene products
  - 3) Provide tools to access data provided by the GO project

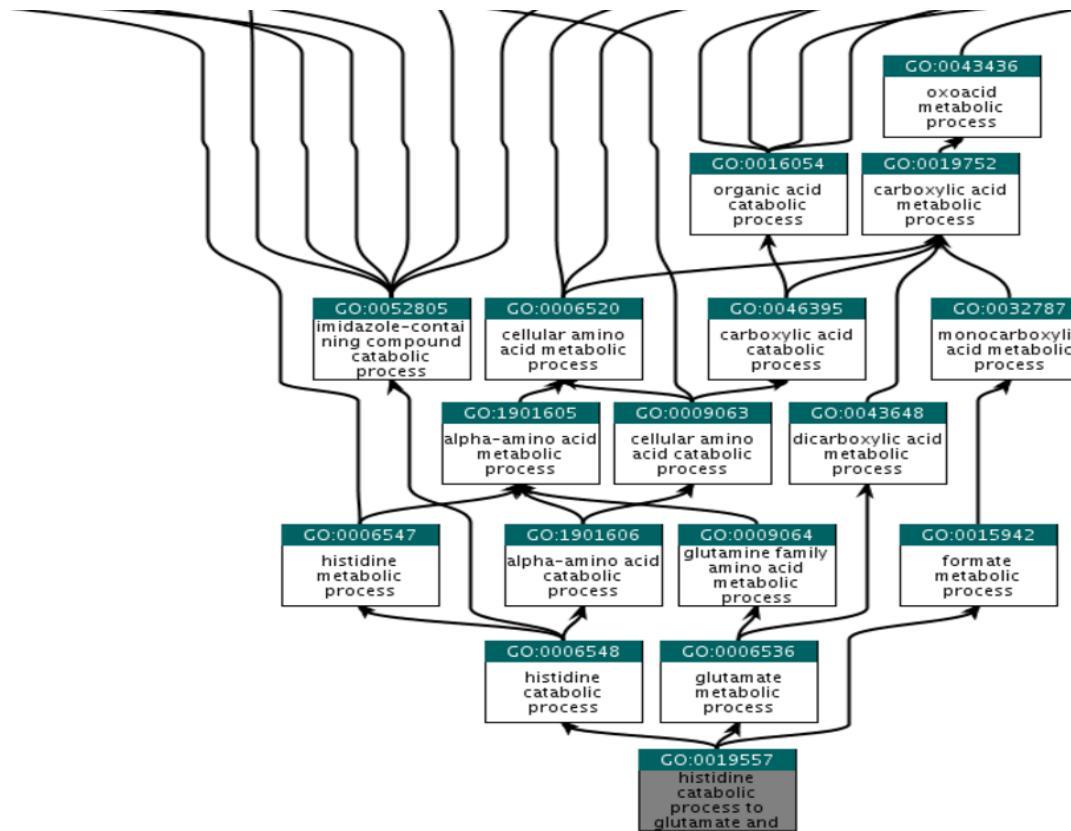
# Open the Ontology, HAL...



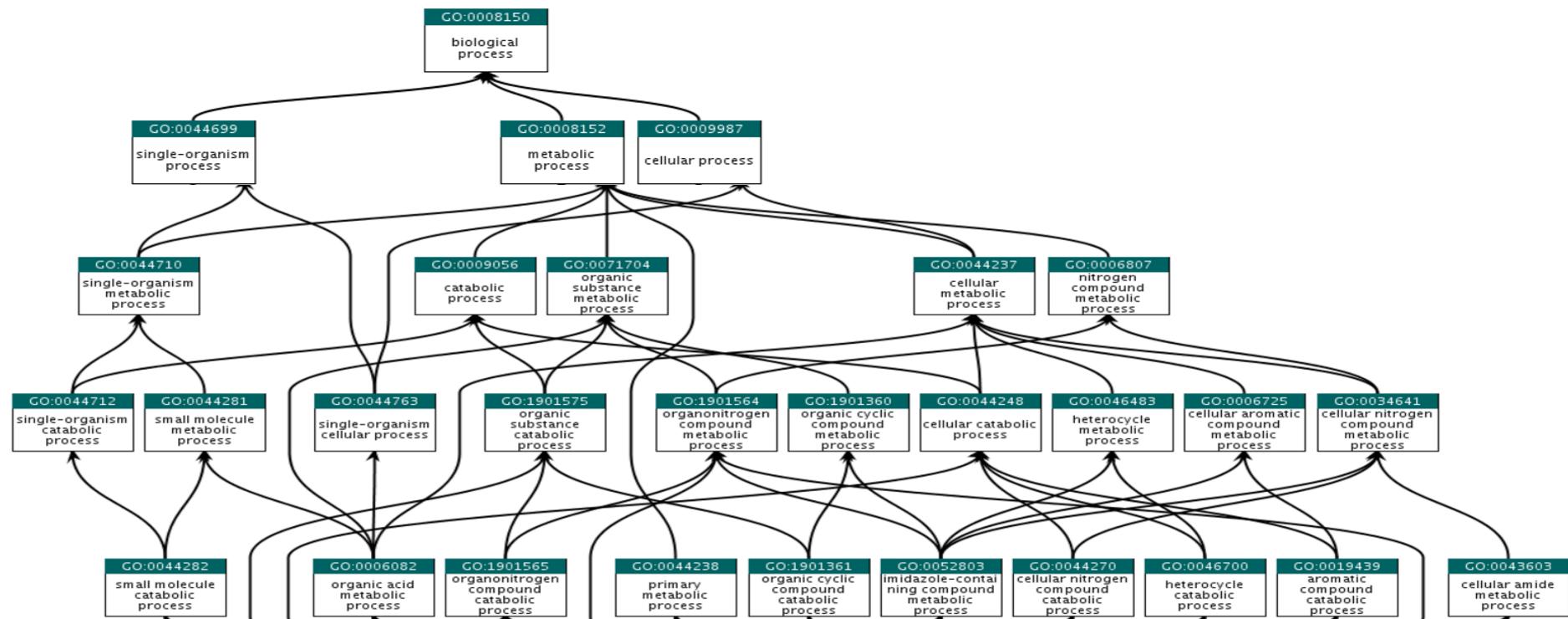
Search for GO term: [ebi.ac.uk/QuickGO](http://ebi.ac.uk/QuickGO)

Search for protein: UniProtKB

# Biological Process



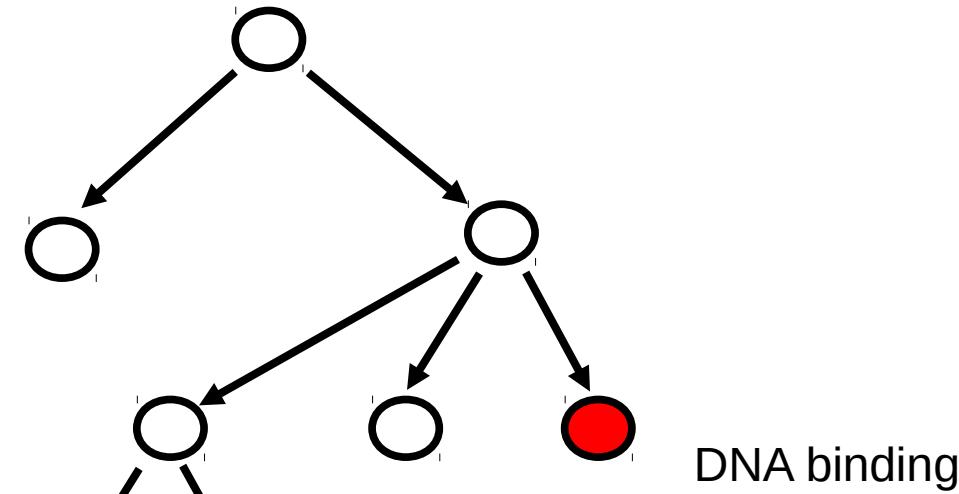
# Going up...



# Let's review

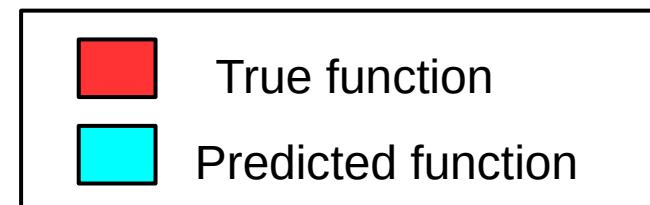
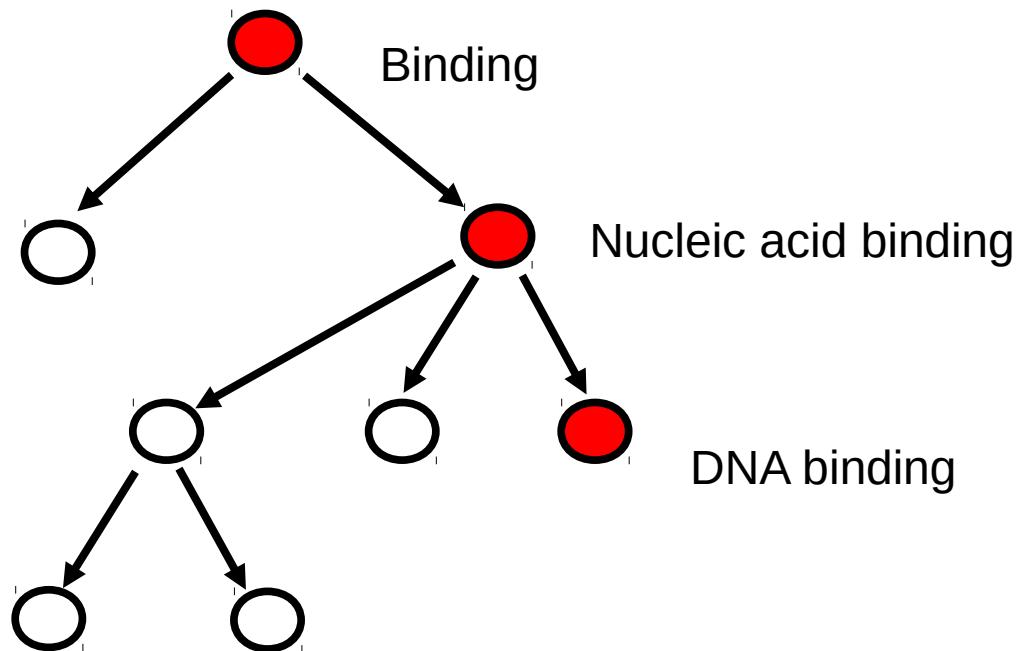
- Critical Assessments can help develop and improve methods
- Need: interesting problem, community, rules
- **Computational representation of problem and method performance**

# Assessment of Function Prediction Using Ontologies



- True function
- Predicted function

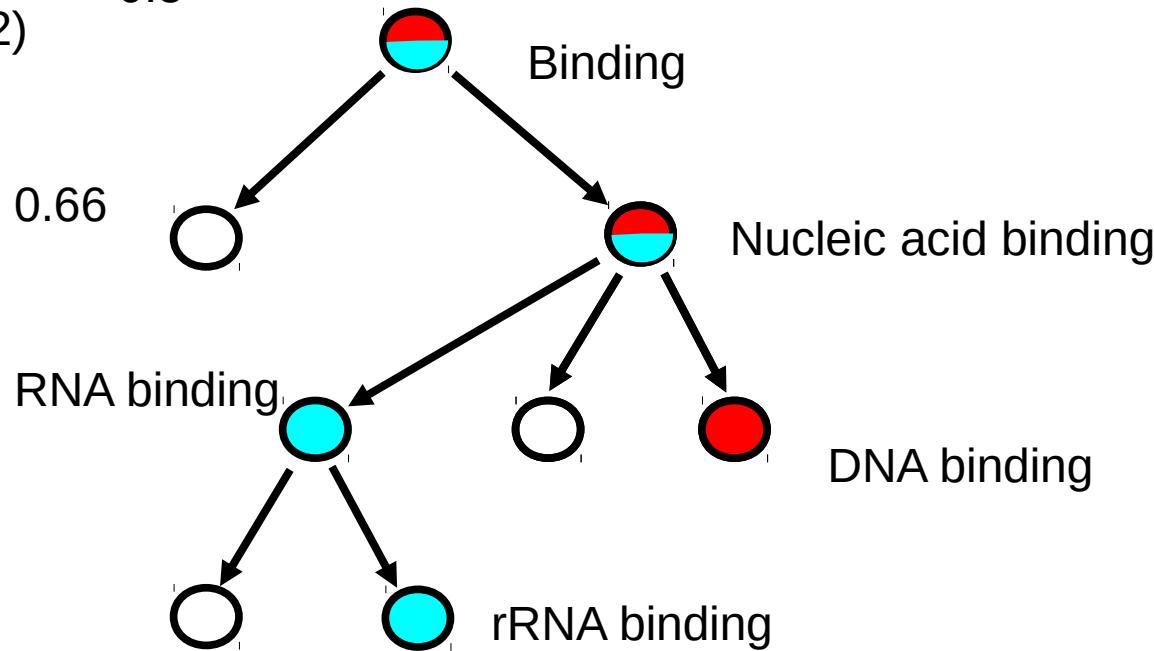
# Assessment of Function Prediction Using Ontologies



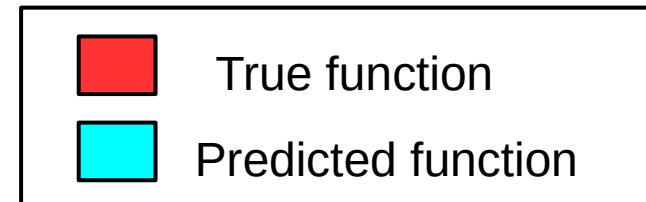
# Assessment of Function Prediction Using Ontologies

$$\text{Precision} = \frac{\text{TP}(2)}{\text{TP}(2)+\text{FP}(2)} = 0.5$$

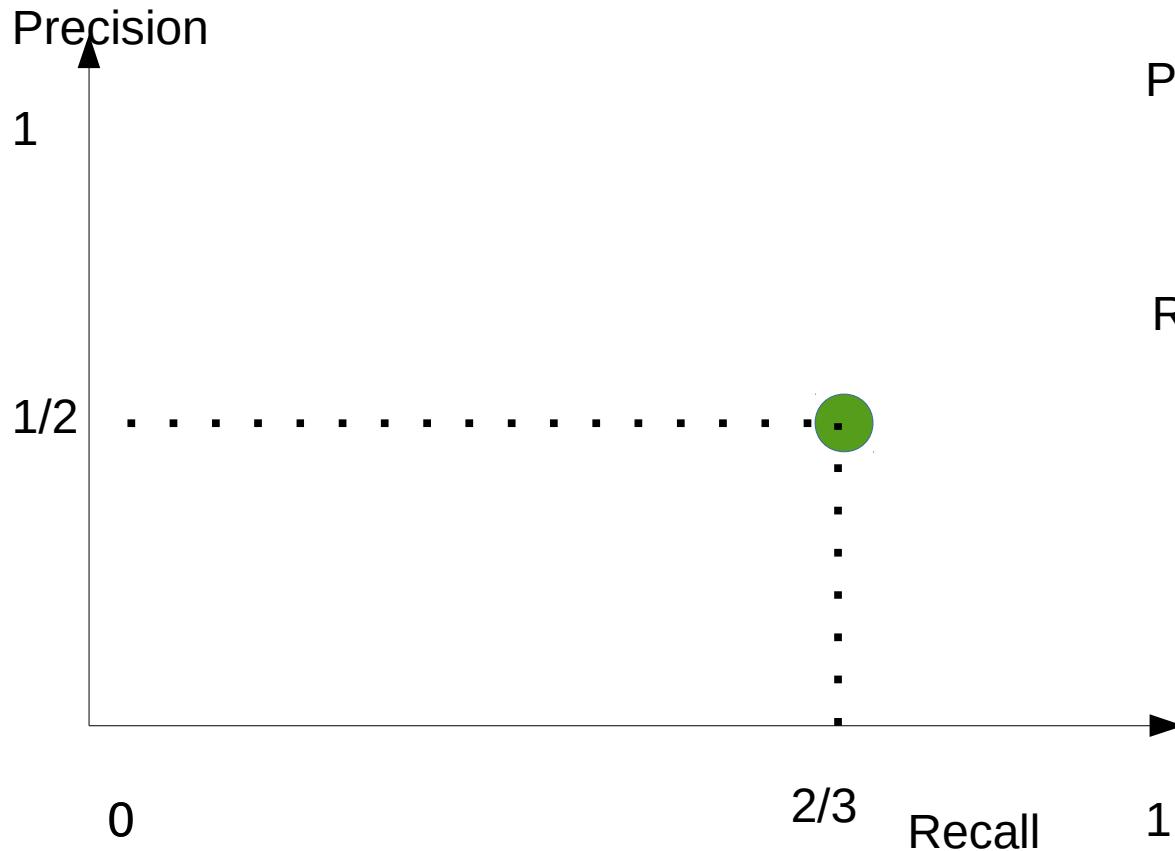
$$\text{Recall} = \frac{\text{TP}(2)}{\text{TP}(2)+\text{FN}(1)} = 0.66$$



- True Positives : 2
- False Positives: 2
- False Negatives: 1



# Precision Recall

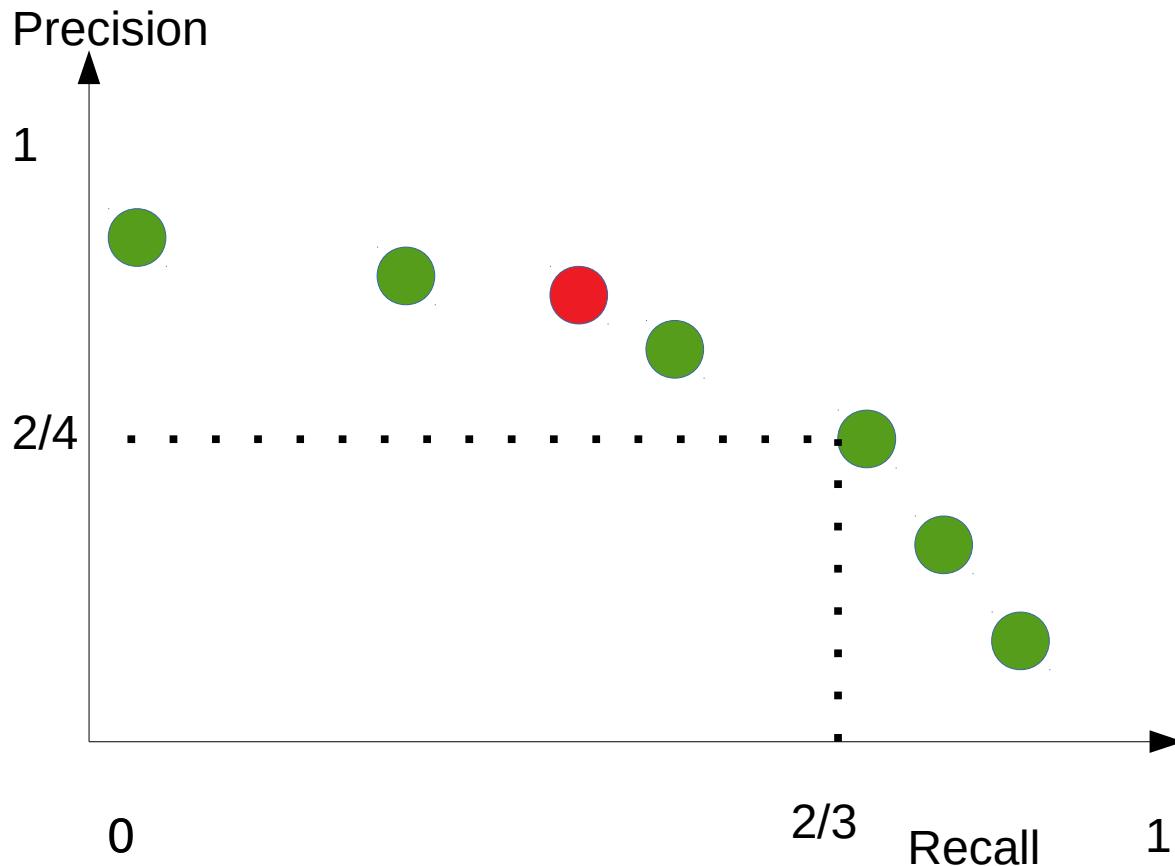


$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Protein ID	GO term	Confidence
AUTHOR ZZZ		
MODEL 1		
KEYWORDS sequence alignment.		
T96060020120	GO:0008270	0.80
T96060020120	GO:0003700	0.80
T96060020120	GO:0006351	0.80
T96060020119	GO:0005730	0.01
T96060020119	GO:0003676	0.07
T96060020119	GO:0005622	0.07
T96060020119	GO:0046872	0.07
T96060020118	GO:0008270	0.75
T96060020118	GO:0006351	0.68
T96060020118	GO:0003677	0.67
T96060020118	GO:0005634	0.67
T96060020118	GO:0006355	0.55
T96060020118	GO:0003700	0.34

# Precision Recall Curve



$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

$$F1 = \frac{\text{Pr} \times \text{Rc}}{\text{Pr} + \text{Rc}}$$

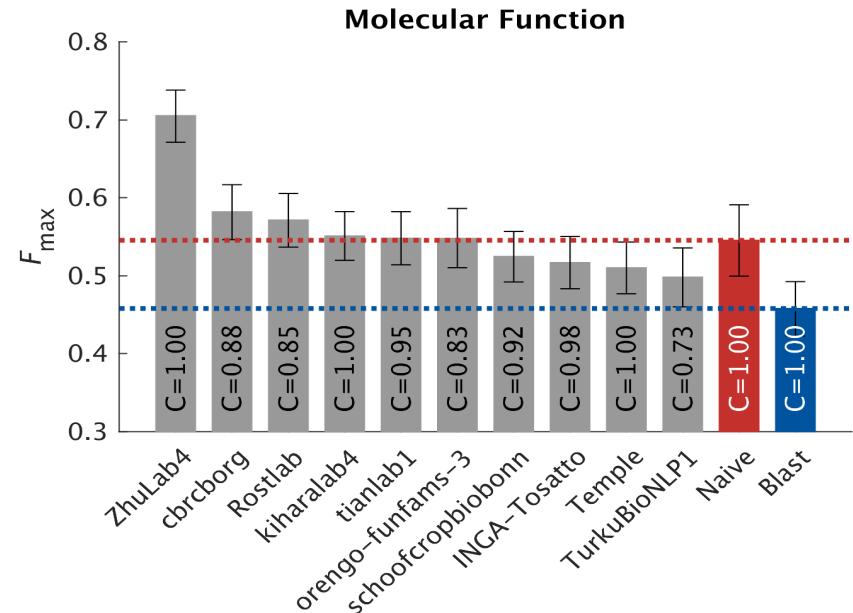
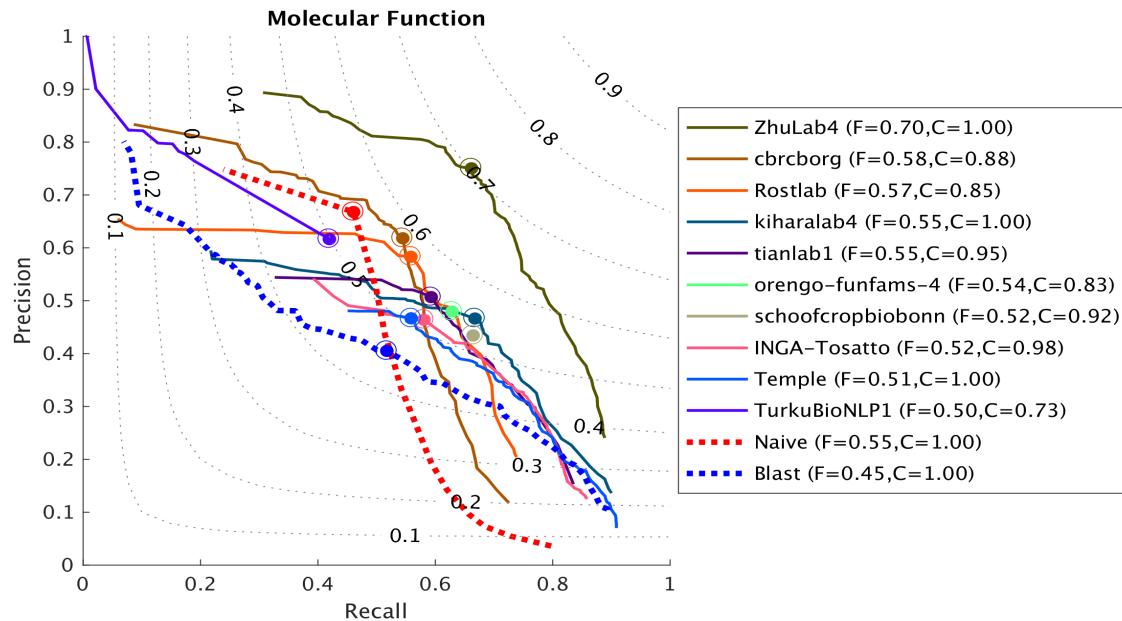
# Molecular Function

Benchmark: **all species**

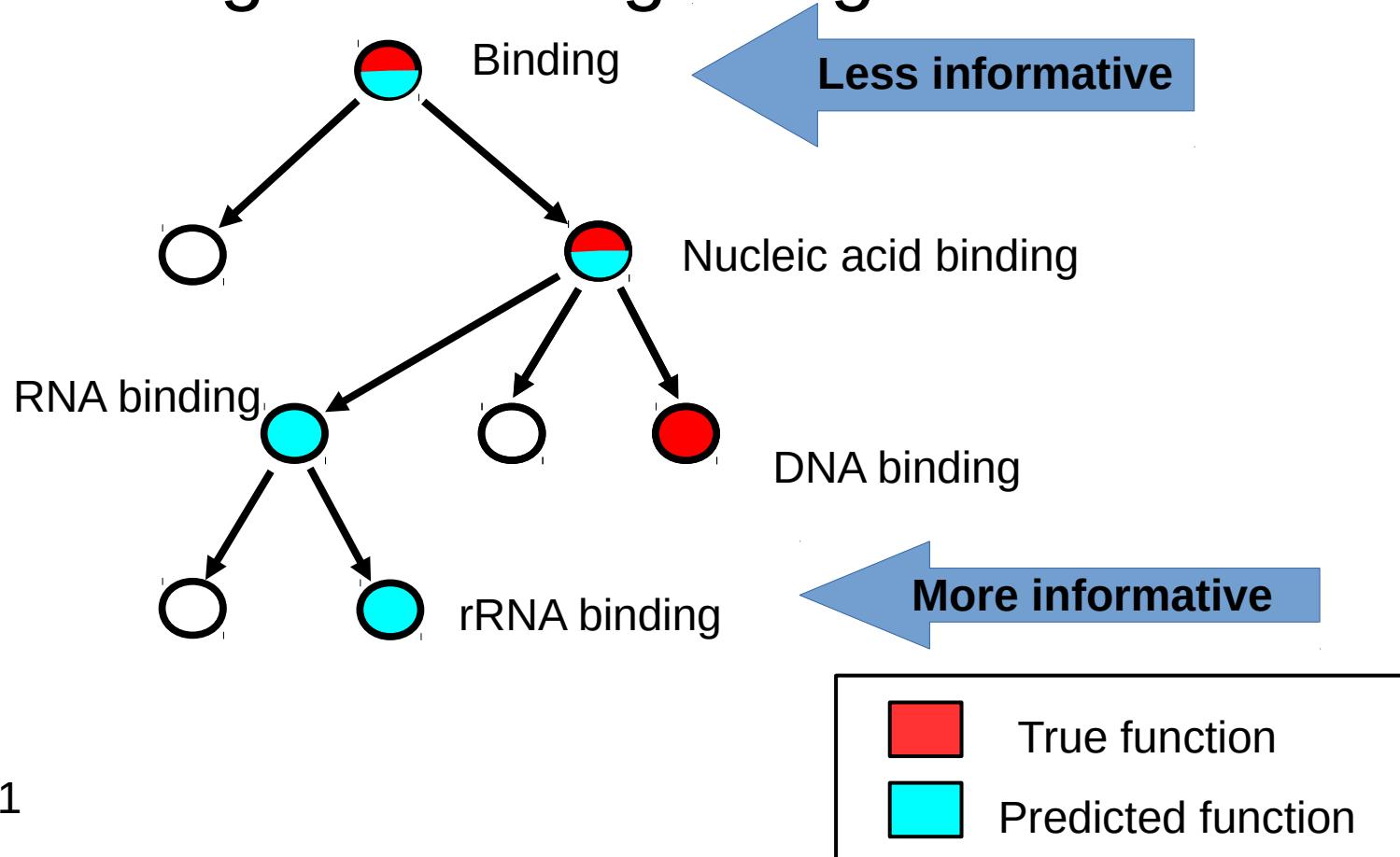
Type: **no knowledge**

Mode: **full**

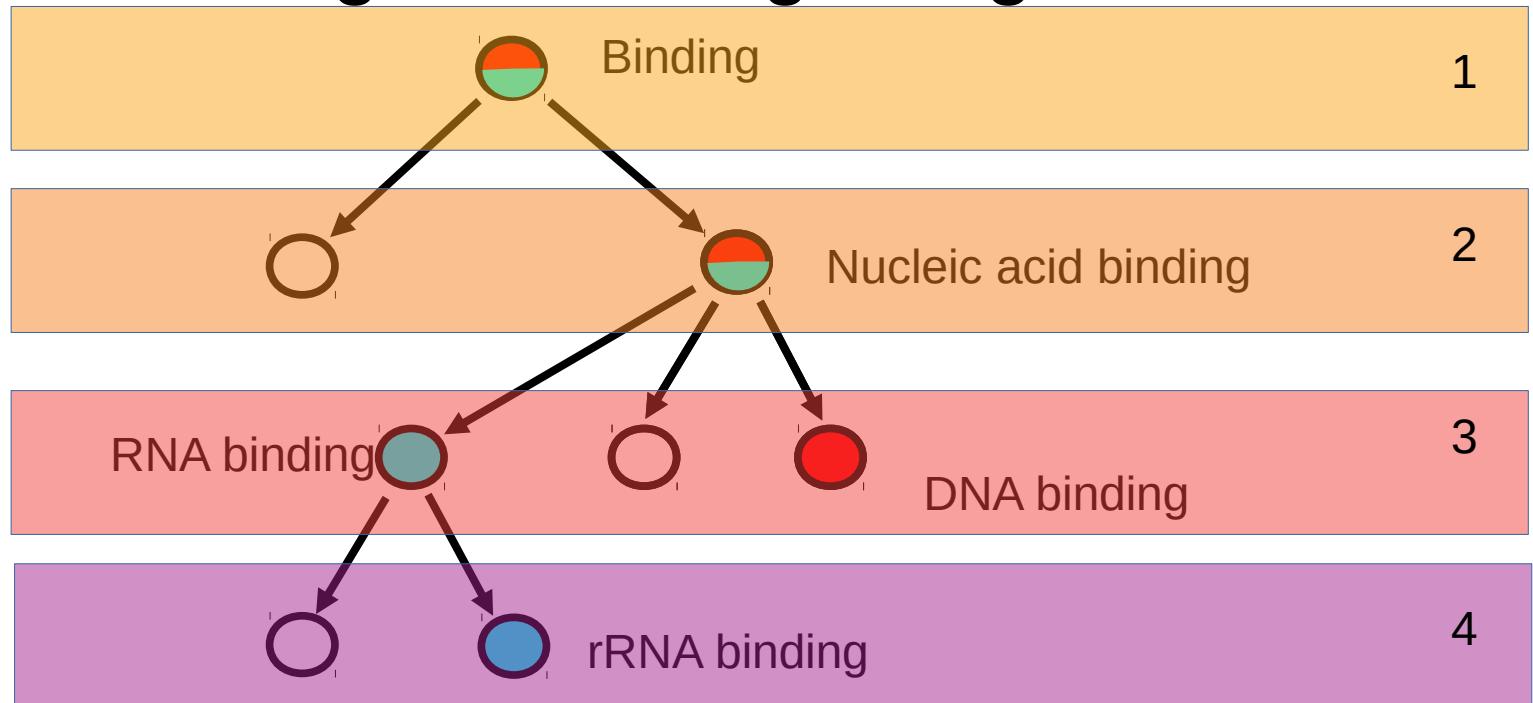
Metric: **F-max**



# Assessment of Function Prediction Using Ontologies: Adding Weight



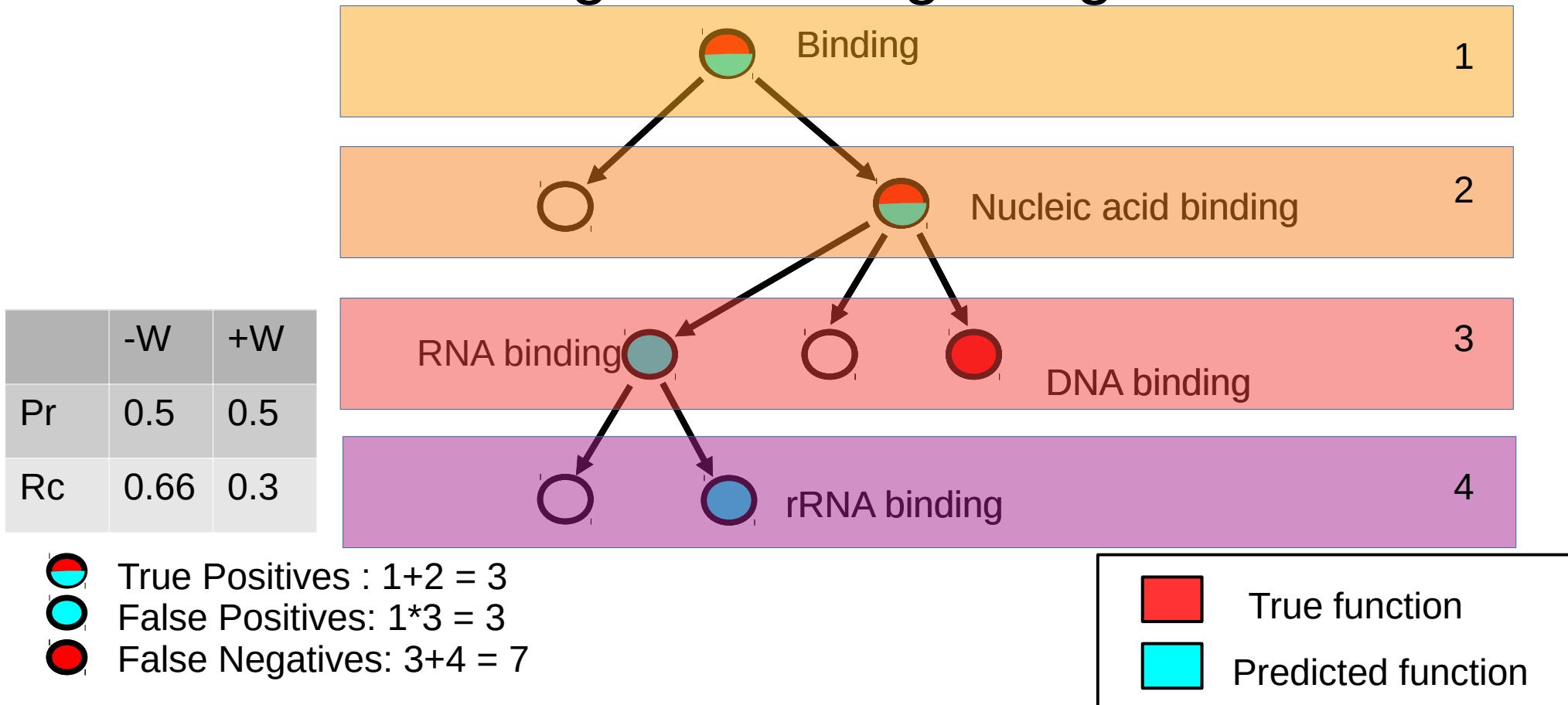
# Assessment of Function Prediction Using Ontologies: Adding Weight



True Positives :  $1+2 = 3$   
False Positives:  $1*3 = 3$   
False Negatives:  $3+4 = 7$

<span style="color:red">■</span>	True function
<span style="color:cyan">■</span>	Predicted function

# Assessment of Function Prediction Using Ontologies: Adding Weights



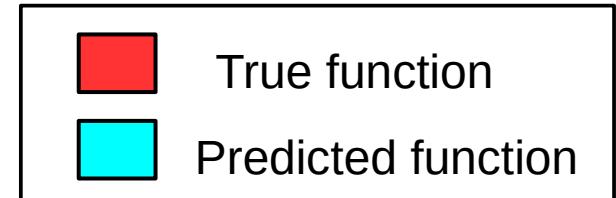
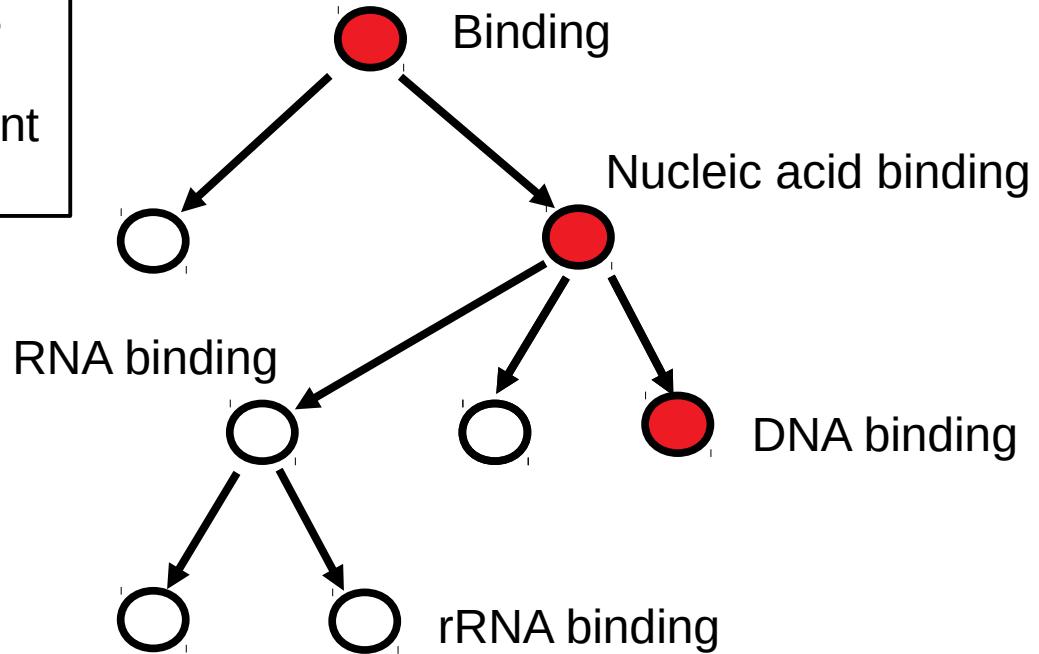
# Assessment of Function Prediction Using Ontologies: Information Theory

$$i(T) = \log \frac{1}{\Pr(T)}$$

$$\Pr(T) = \prod_{v \in T} \Pr(v | \mathcal{P}(v))$$

$$\begin{aligned} i(T) &= \log \frac{1}{\prod_{v \in T} \Pr(v | \mathcal{P}(v))} \\ &= \sum_{v \in T} \log \frac{1}{\Pr(v | \mathcal{P}(v))} \\ &= \sum_{v \in T} ia(v) \end{aligned}$$

$T$ : set of true nodes  
 $v$ : node  
 $i$ : information content  
 $ia$ : accumulated  $i$



# Assessment of Function Prediction Using Ontologies: Information Theory



Misinformation  
(accumulated IC of FP):

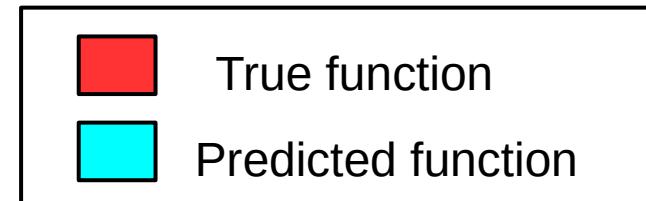
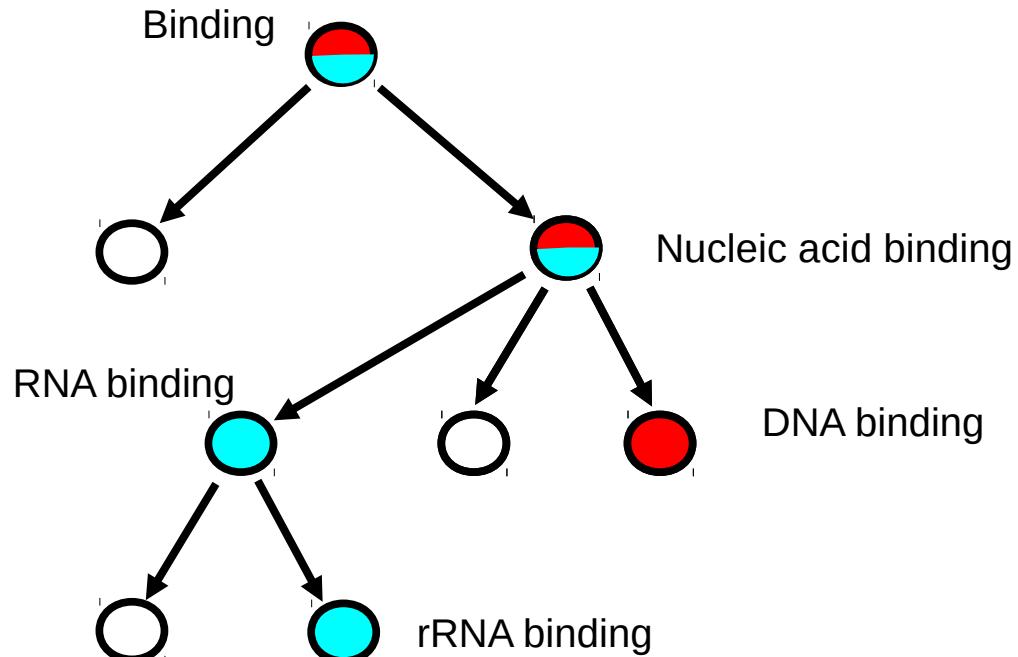
$$mi(P, T) = \sum_{v \in P - T} ia(v)$$



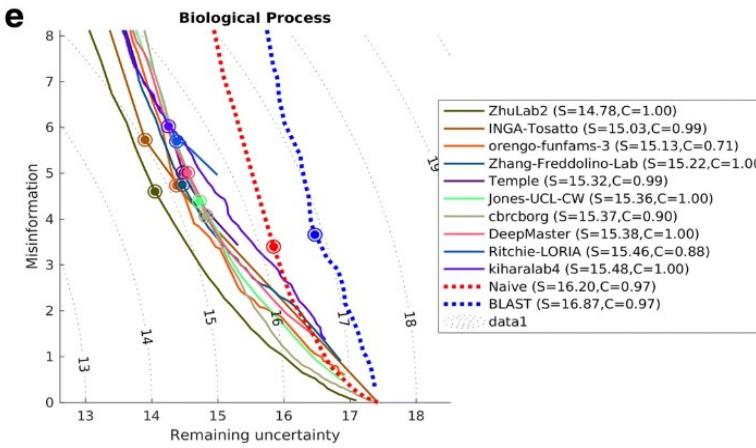
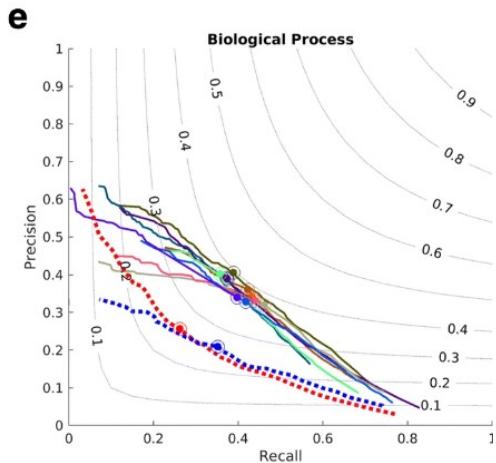
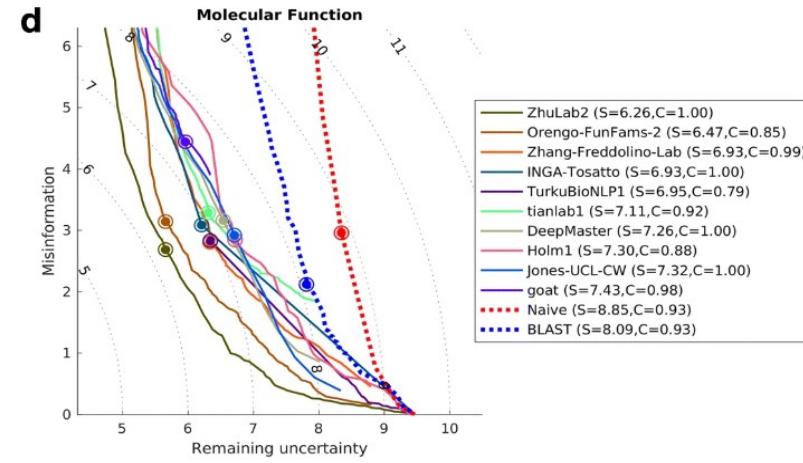
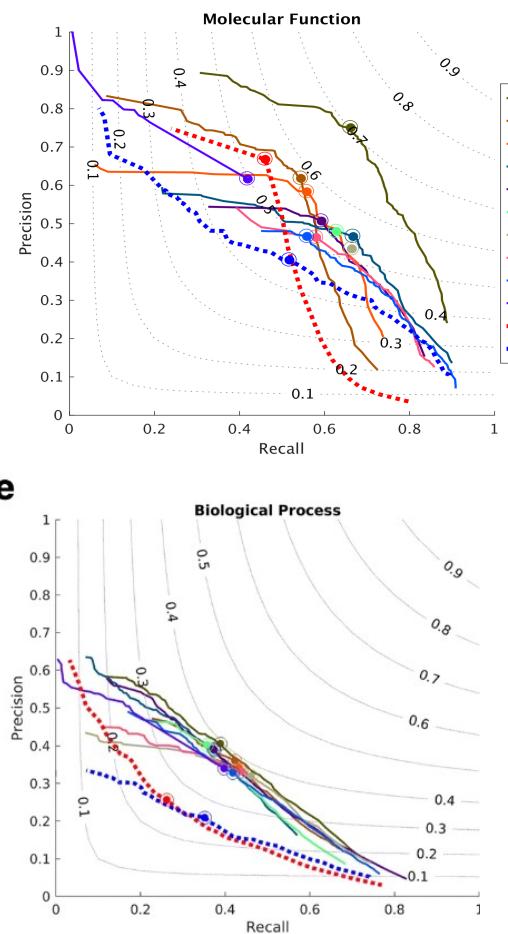
Remaining uncertainty  
(false negatives):

$$ru(P, T) = \sum_{v \in T - P} ia(v)$$

Goal: minimize ru and mi



# Precision / Recall vs. ru / mi

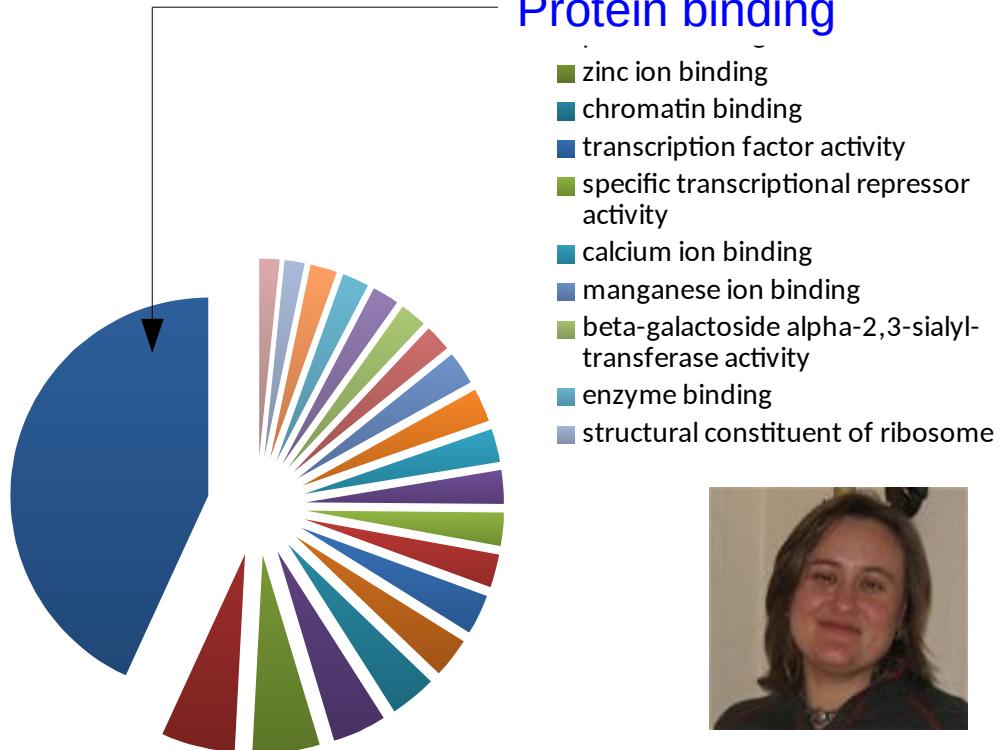


# About Assessment Metrics

- One metric rarely captures all information
- Multiple metrics are advised
- Learn and update as the competition matures

# Databases are biased...

## Protein binding



- zinc ion binding
- chromatin binding
- transcription factor activity
- specific transcriptional repressor activity
- calcium ion binding
- manganese ion binding
- beta-galactoside alpha-2,3-sialyl-transferase activity
- enzyme binding
- structural constituent of ribosome



Alexandra Schnoes



David Ream

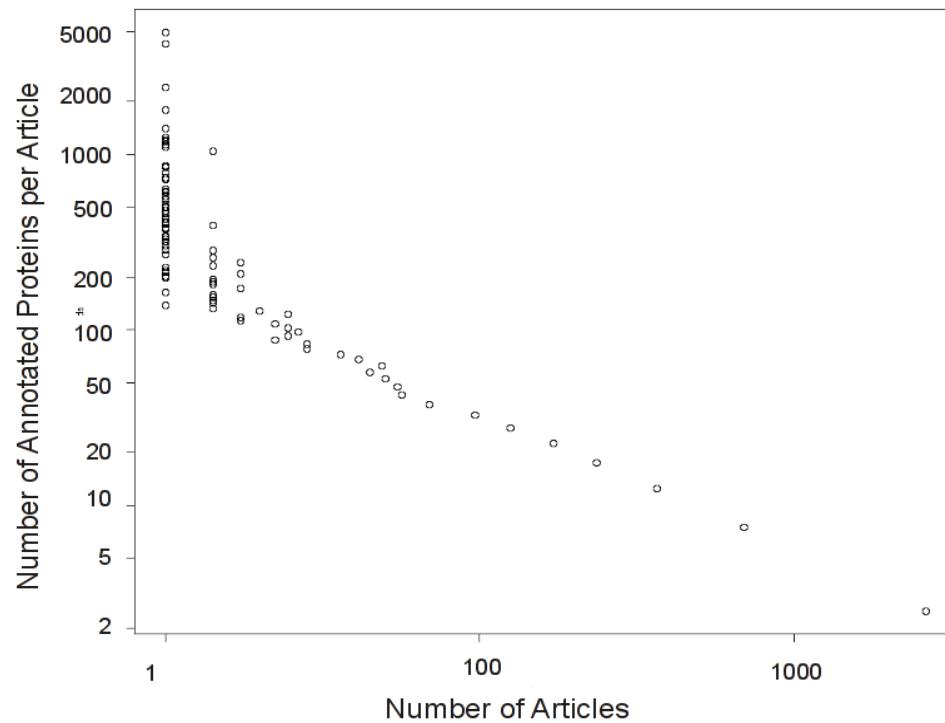


Alexander Thorman

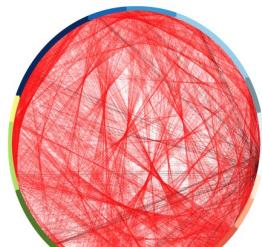
# Best Function Prediction Program (confidential source code)

```
def best_predictor_ever(inseq):  
    print "protein binding"
```

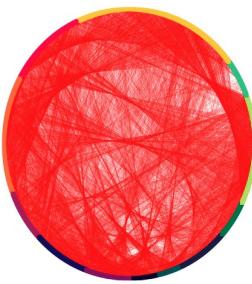
# Annotations per article



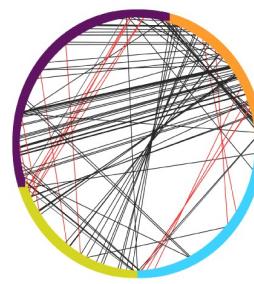
# Annotation redundancy



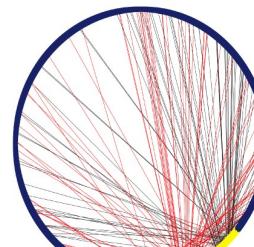
*A. Thaliana* (8879)



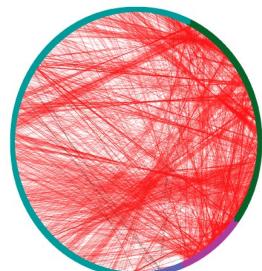
*C. elegans* (8416)



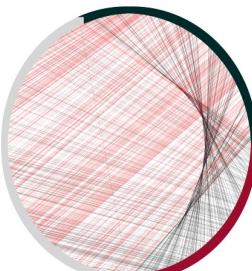
*D. melanogaster* (1217)



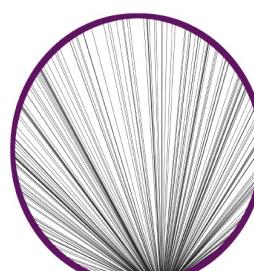
*H. sapiens* (5593)



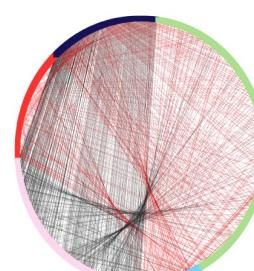
*M. musculus* (4220)



*M. tuberculosis* (2351)



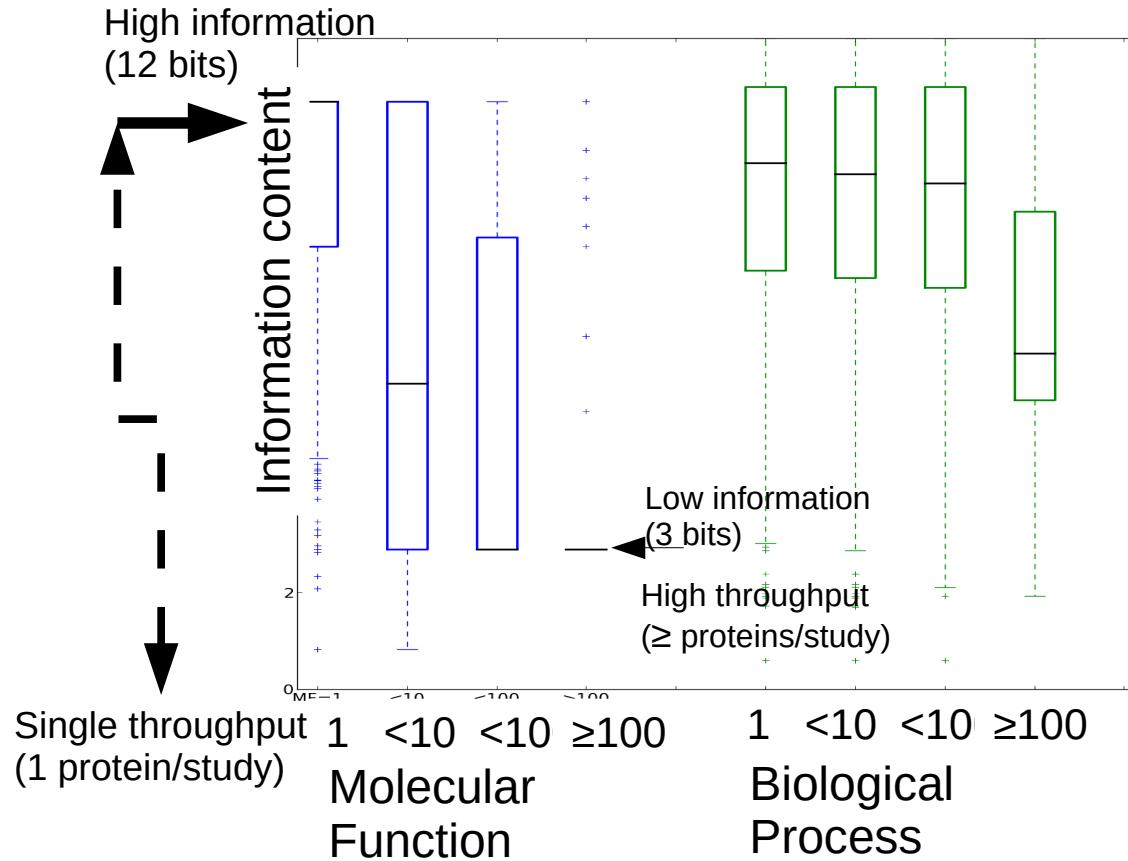
*S. pombe* (4502)



*S. cerevisiae* (3542)

— Different Ontology

— Same Ontology



Schnoes et al (2013)

# High throughput Experiments

## The Bad

- Bias our knowledge
- Bias priors for function prediction programs
- Are less informative than low-throughput experiments

## The Good

- Exclusively annotate genes otherwise unknown
- Fewer \$\$\$
- Fast results
- Consistency

DIALOGUE FOR REVERSE ENGINEERING ASSESSMENT AND METHODS



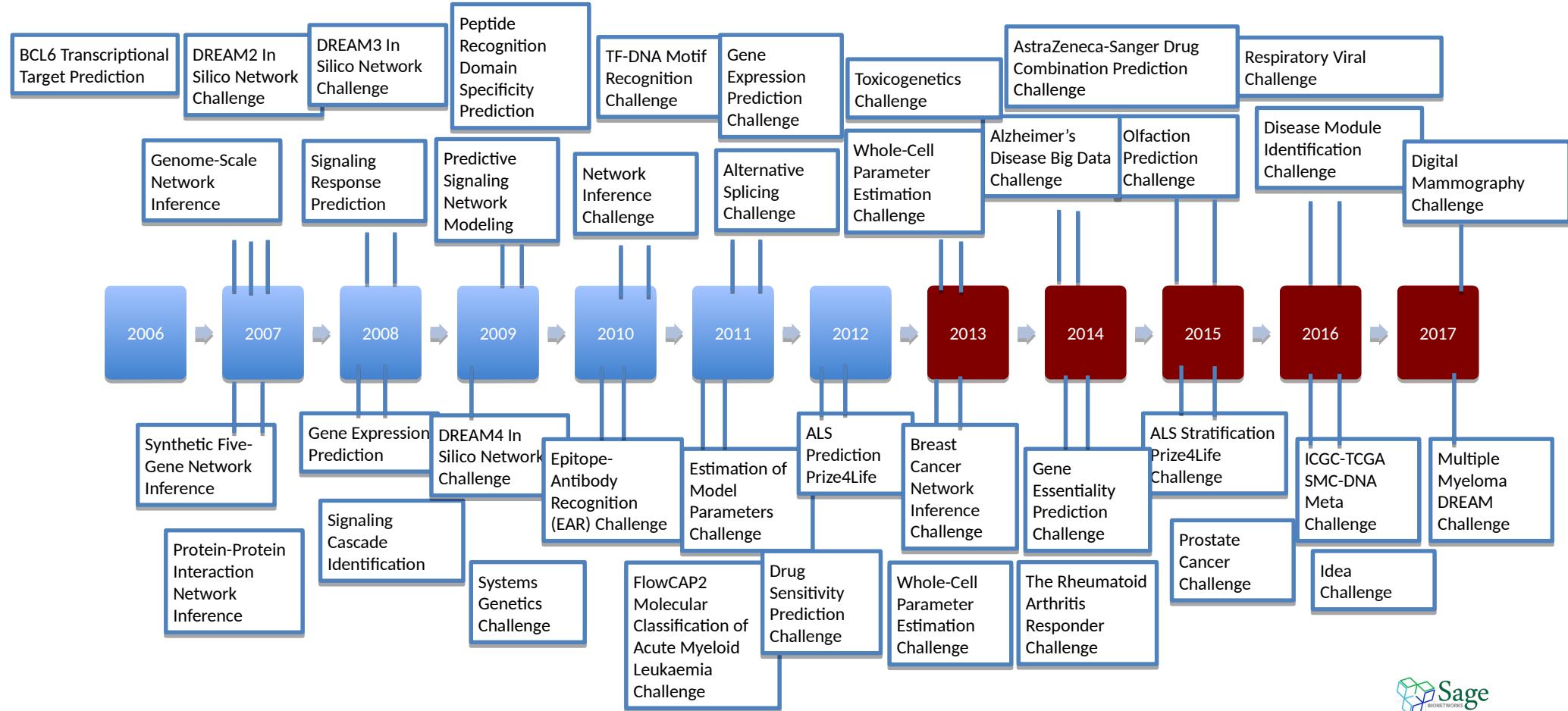
Crowdsourcing innovation in biomedicine

# DREAM Challenges

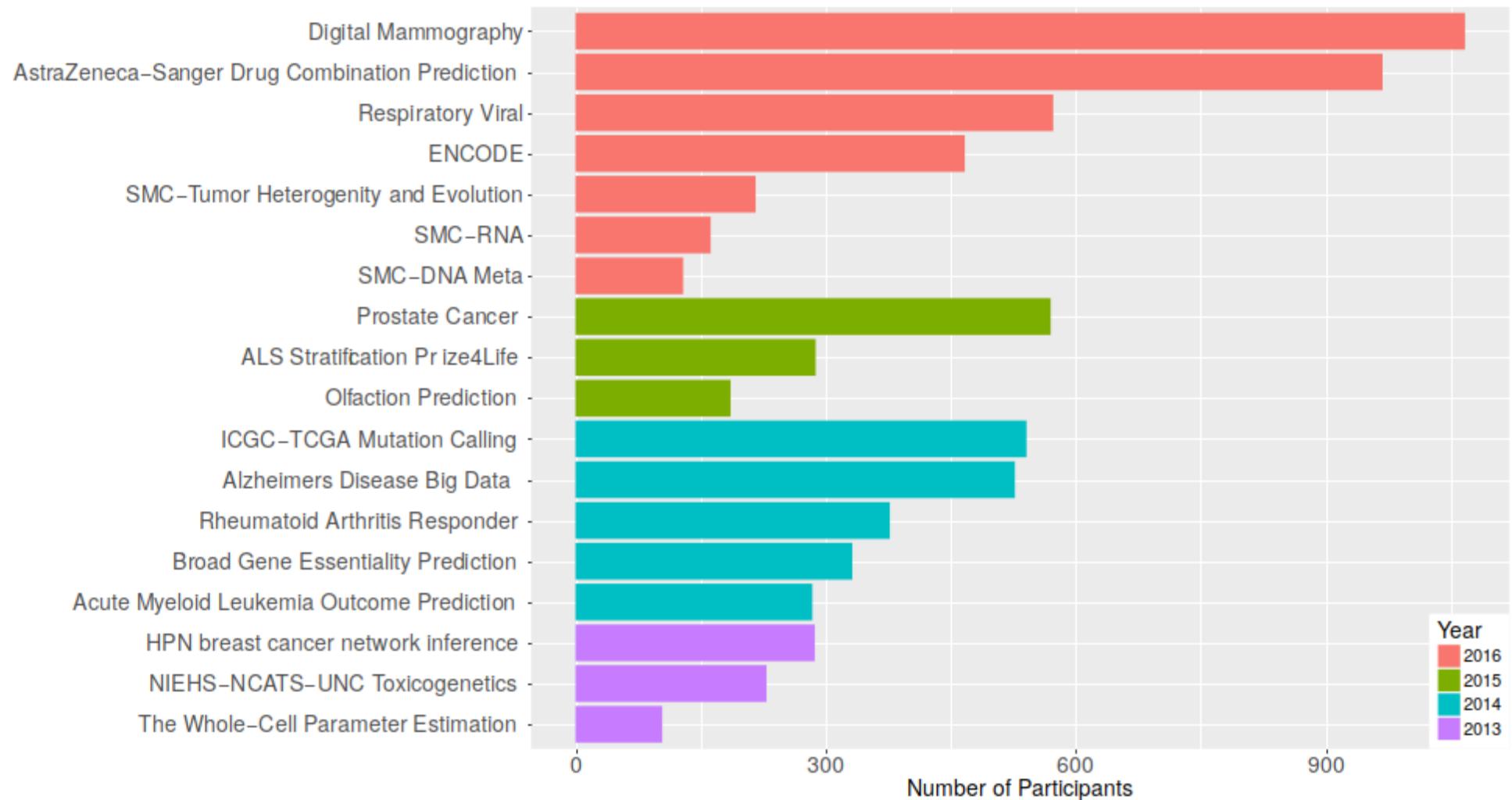
DIALOGUE FOR REVERSE ENGINEERING ASSESSMENT AND METHODS

- A crowdsourcing effort that poses quantitative, biomedical questions
- Mission:
  - to contribute to the solution of important biomedical problems
  - to foster collaboration between research groups
  - to democratize data
  - to accelerate research
  - to objectively assess algorithm performance

# DREAM over the years

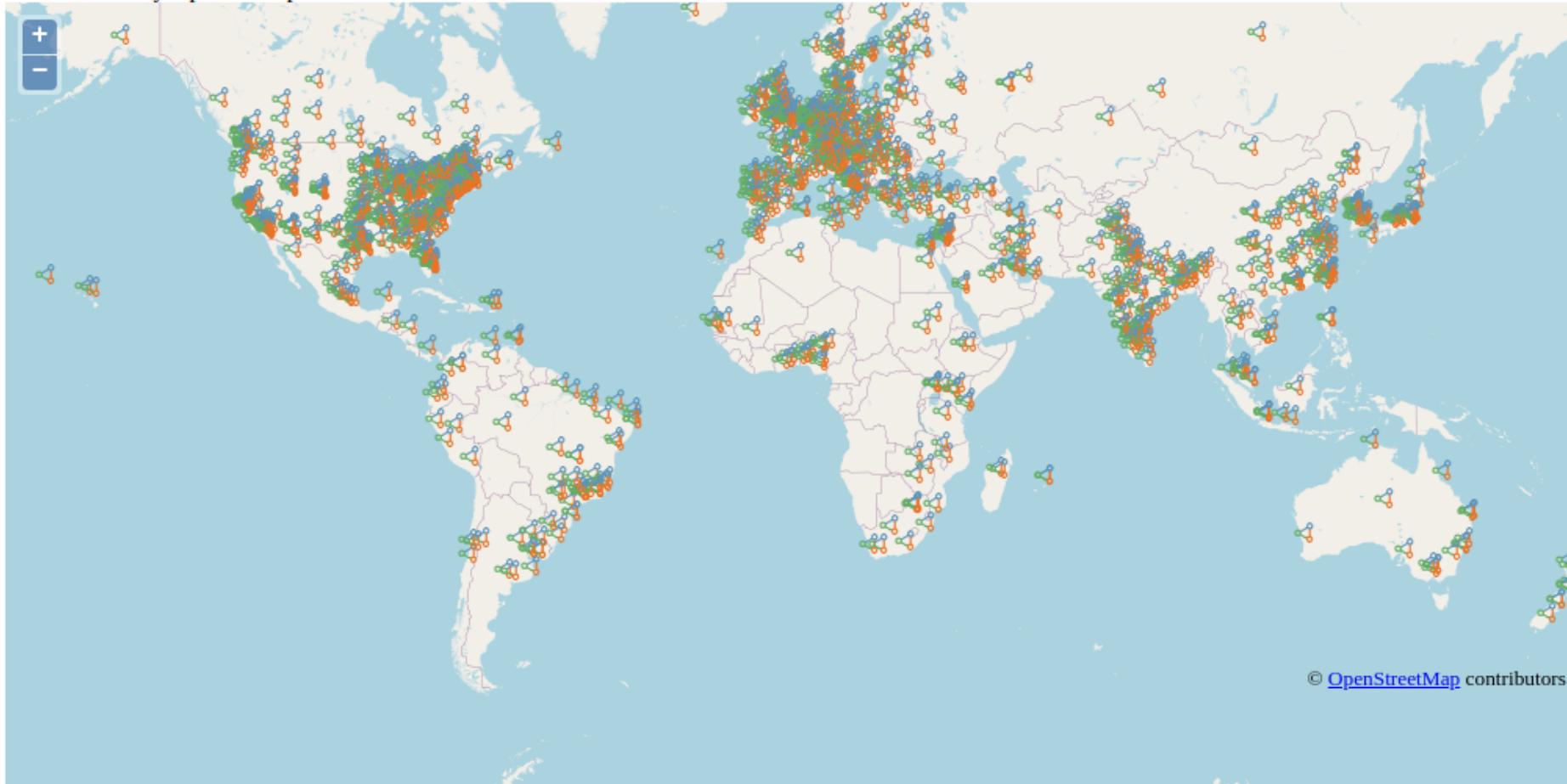


## Some challenge participation figures



## Where in the World are Synapse Users?

These 11383 Synapse users provide a 'location' in their User Profile. Scroll down to see locations of Teams.

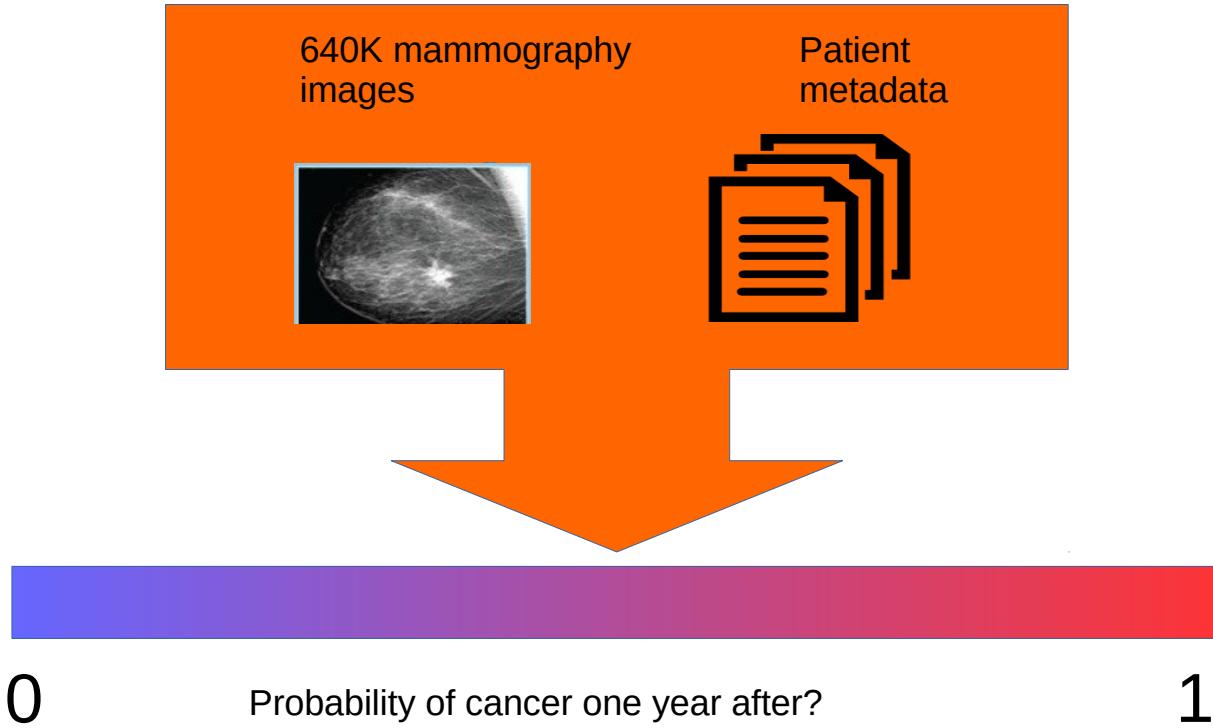


# Mammography DREAM Challenge

TP: patient sick, predicted sick  
FP: patient healthy, predicted sick  
TN: patient healthy, predicted healthy  
FN: patient sick, predicted healthy

Sensitivity:  $TP/(TP+FN)$

Specificity:  $TN/(TN+FP)$

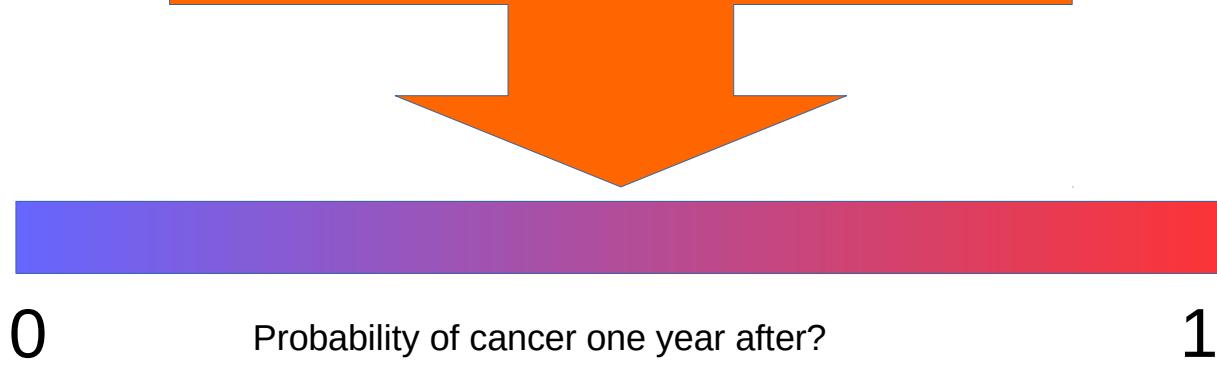
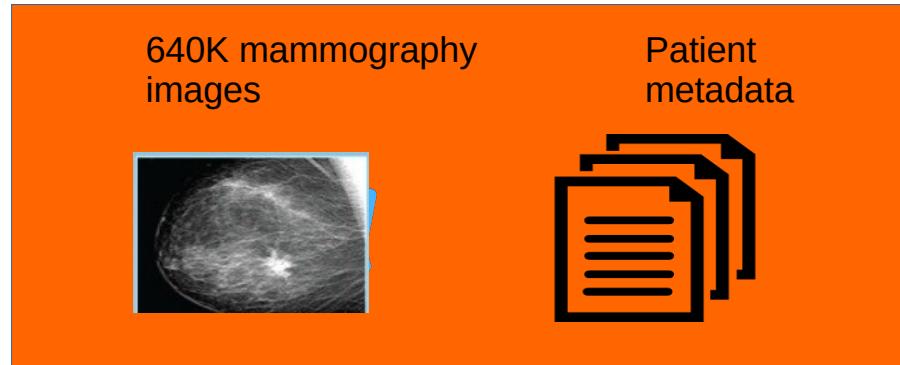
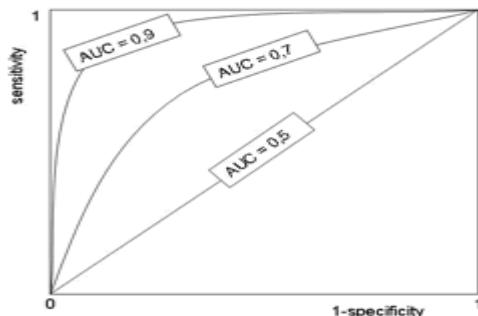


# Mammography DREAM Challenge

TP: patient sick, predicted sick  
FP: patient healthy, predicted sick  
TN: patient healthy, predicted healthy  
FN: patient sick, predicted healthy

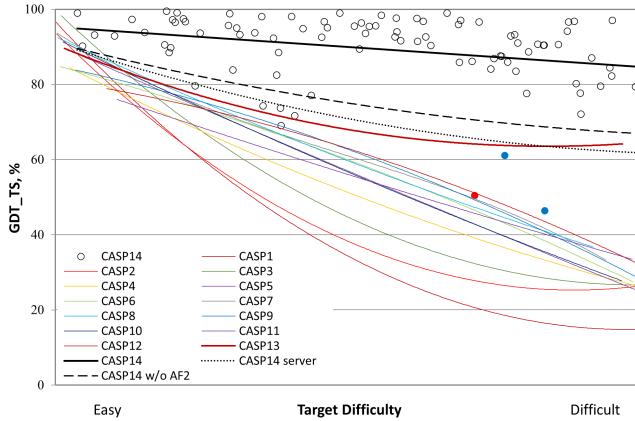
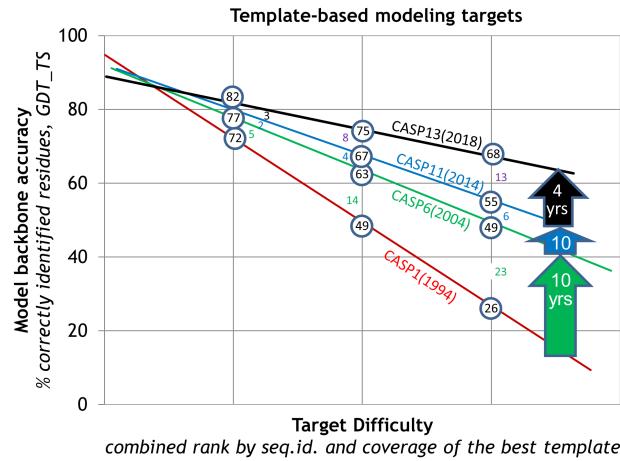
Sensitivity:  $TP/(TP+FN)$

Specificity:  $TN/(TN+FP)$



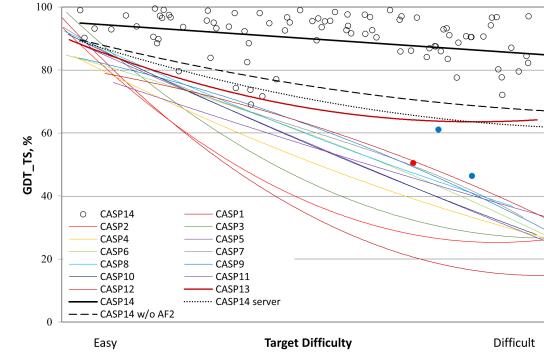
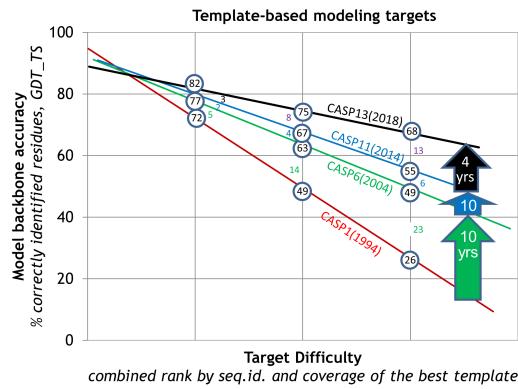
# Do Computational Challenges Improve Methods?

## Critical Assessment of Structure Prediction



# Do Computational Challenges Improve Performance?

## Critical Assessment of Structure Prediction



## Critical Assessment of Function Annotation



**A**

CAFA2 top models →

CAFA1 top models ↓	99.81	99.76	99.50	98.14	86.19
	99.97	100	99.94	99.95	94.87
	100	99.99	99.93	98.72	95.39
	100	100	100	100	99.70
	100	100	100	99.60	98.92

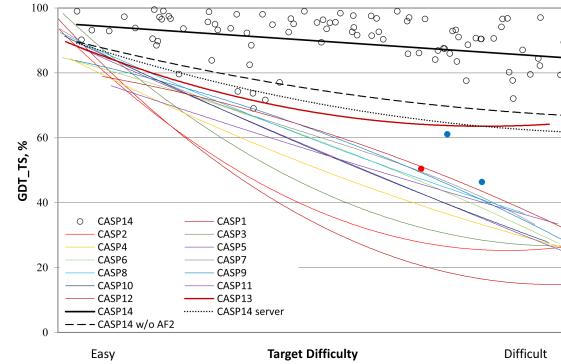
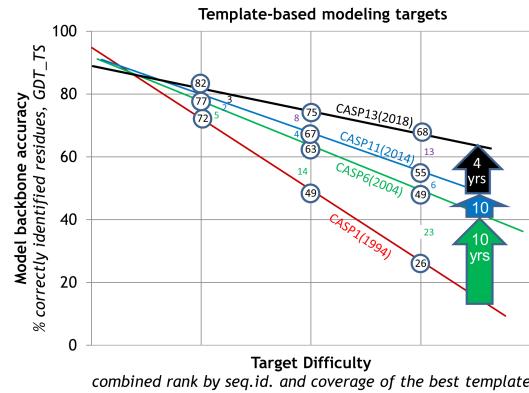
**A.**

CAFA4 top models →

CAFA3 top models ↓	99.90	99.13	89.58	88.97	88.49
	99.72	98.92	91.17	90.39	90.03
	99.83	99.21	93.67	93.56	93.21
	99.88	99.32	94.09	93.77	93.55
	99.95	99.65	95.62	95.37	95.10

# Do Computational Challenges Improve Performance?

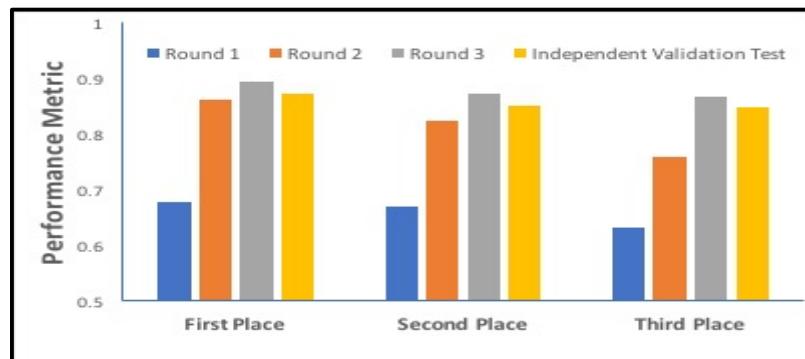
## Critical Assessment of Structure Prediction



## Critical Assessment of Function Annotation



## Mammography DREAM challenge



A

CAFA2 top models →

CAFA1 top models ↓	99.81	99.76	99.50	98.14	86.19
99.97	100	99.94	99.95	94.87	
100	99.99	99.93	98.72	95.39	
100	100	100	100	99.70	
100	100	100	99.60	98.92	

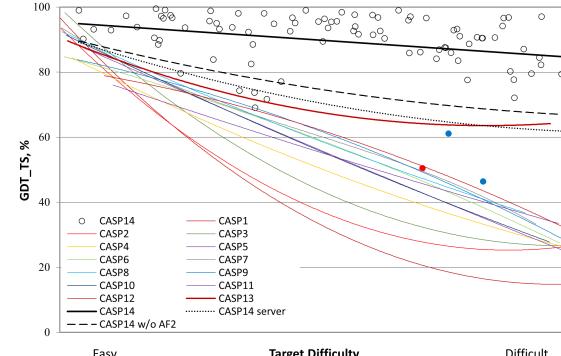
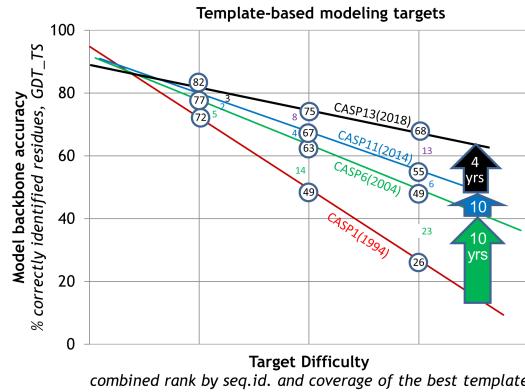
A.

CAFA4 top models →

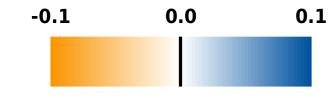
CAFA3 top models ↓	99.90	99.13	89.58	88.97	88.49
99.72	98.92	91.17	90.39	90.03	
99.83	99.21	93.67	93.56	93.21	
99.88	99.32	94.09	93.77	93.55	
99.95	99.65	95.62	95.37	95.10	

# Do Computational Challenges Improve Performance?

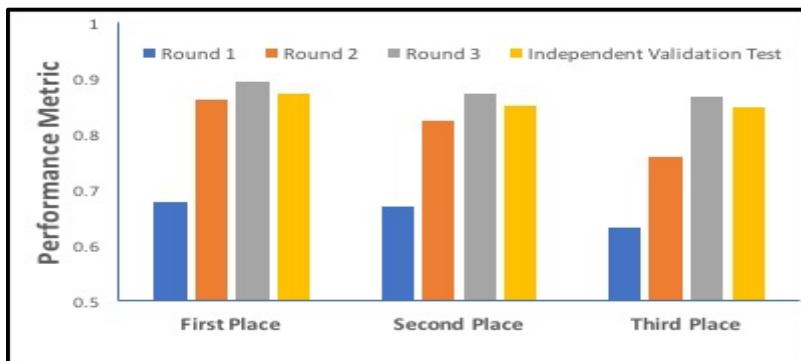
## Critical Assessment of Structure Prediction



YES



## Mammography DREAM challenge



A

CAFA2 top models →

99.81	99.76	99.50	98.14	86.19
99.97	100	99.94	99.95	94.87
100	99.99	99.93	98.72	95.39
100	100	100	100	99.70
100	100	100	99.60	98.92

A.

CAFA4 top models →

99.90	99.13	89.58	88.97	88.49
99.72	98.92	91.17	90.39	90.03
99.83	99.21	93.67	93.56	93.21
99.88	99.32	94.09	93.77	93.55
99.95	99.65	95.62	95.37	95.10

# More Information

- Kaggle (general AI challenges)
- BioFunctionPrediction.org

**Mark Your Calendars! JULY 10-14, 2022**



JOIN US AT THE PREMIER  
COMPUTATIONAL  
BIOLOGY MEETING OF THE  
YEAR!

<https://www.iscb.org/ismb2022>

# Acknowledgments



Casey Greene

Predrag Radivojac

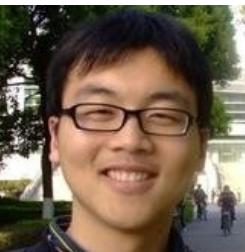
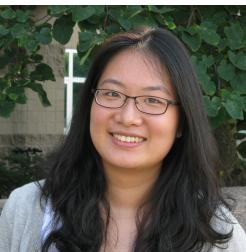
Sean Mooney

Deb Hogan

Gio Bosco

Mark Wass

Kimberly Reynolds



Wyatt Clark

Naihui Zhou

Yuxiang Jiang

Yisu Peng

Tim Bergquist

Balint Kacsoh

Scott Zarecor



Burkhard Rost

Christine  
Orengo

Marc RR

Dannie  
Durand

Steven  
Brenner

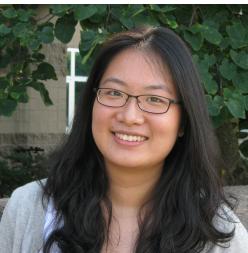


## Principal Investigators

# Acknowledgments



Casey Greene Predrag Radivojac Sean Mooney Deb Hogan Gio Bosco Mark Wass Kimberly Reynolds



Wyatt Clark

Naihui Zhou

Yuxiang Jiang

Yisu Peng

Tim Bergquist

Balint Kacsoh

Scott Zarecor



Burkhard Rost

Christine  
Orengo

Marc RR

Dannie  
Durand

Steven  
Brenner



# Acknowledgments

Students & Staff



Casey Greene

Predrag Radivojac

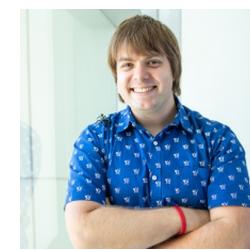
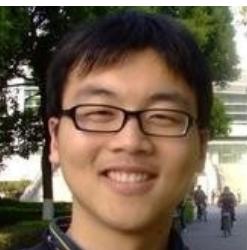
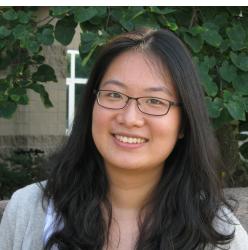
Sean Mooney

Deb Hogan

Gio Bosco

Mark Wass

Kimberly Reynolds



Wyatt Clark

Naihui Zhou

Yuxiang Jiang

Yisu Peng

Tim Bergquist

Balint Kacsoh

Scott Zarecor



Burkhard Rost

Christine  
Orengo

Marc RR

Dannie  
Durand

Steven  
Brenner



# Acknowledgments

cosI Leaders



Casey Greene

Predrag Radivojac

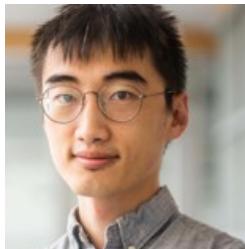
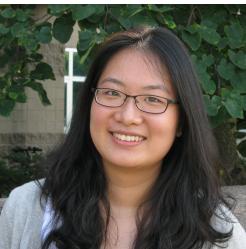
Sean Mooney

Deb Hogan

Gio Bosco

Mark Wass

Kimberly Reynolds



Wyatt Clark

Naihui Zhou

Yuxiang Jiang

Yisu Peng

Tim Bergquist

Balint Kacsoh

Scott Zarecor



Burkhard Rost

Christine  
Orengo

Marc RR

Dannie  
Durand

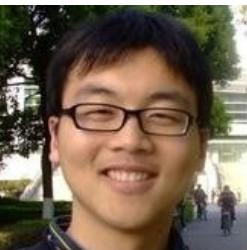
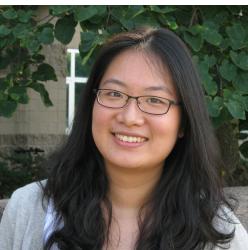
Steven  
Brenner



## Steering Committee



Casey Greene Predrag Radivojac Sean Mooney Deb Hogan Gio Bosco Mark Wass Kimberly Reynolds



Wyatt Clark

Naihui Zhou

Yuxiang Jiang

Yisu Peng

Tim Bergquist

Balint Kacsoh

Scott Zarecor



Burkhard Rost

Christine  
Orengo

Marc RR

Dannie  
Durand

Steven  
Brenner



# Acknowledgments

# Acknowledgments

Funding



Casey Greene

Predrag Radivojac

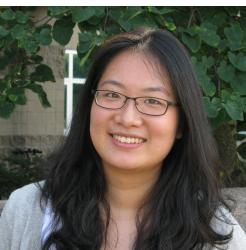
Sean Mooney

Deb Hogan

Gio Bosco

Mark Wass

Kimberly Reynolds



Wyatt Clark

Naihui Zhou

Yuxiang Jiang

Yisu Peng

Tim Bergquist

Balint Kacsoh

Scott Zarecor



Burkhard Rost

Christine  
Orengo

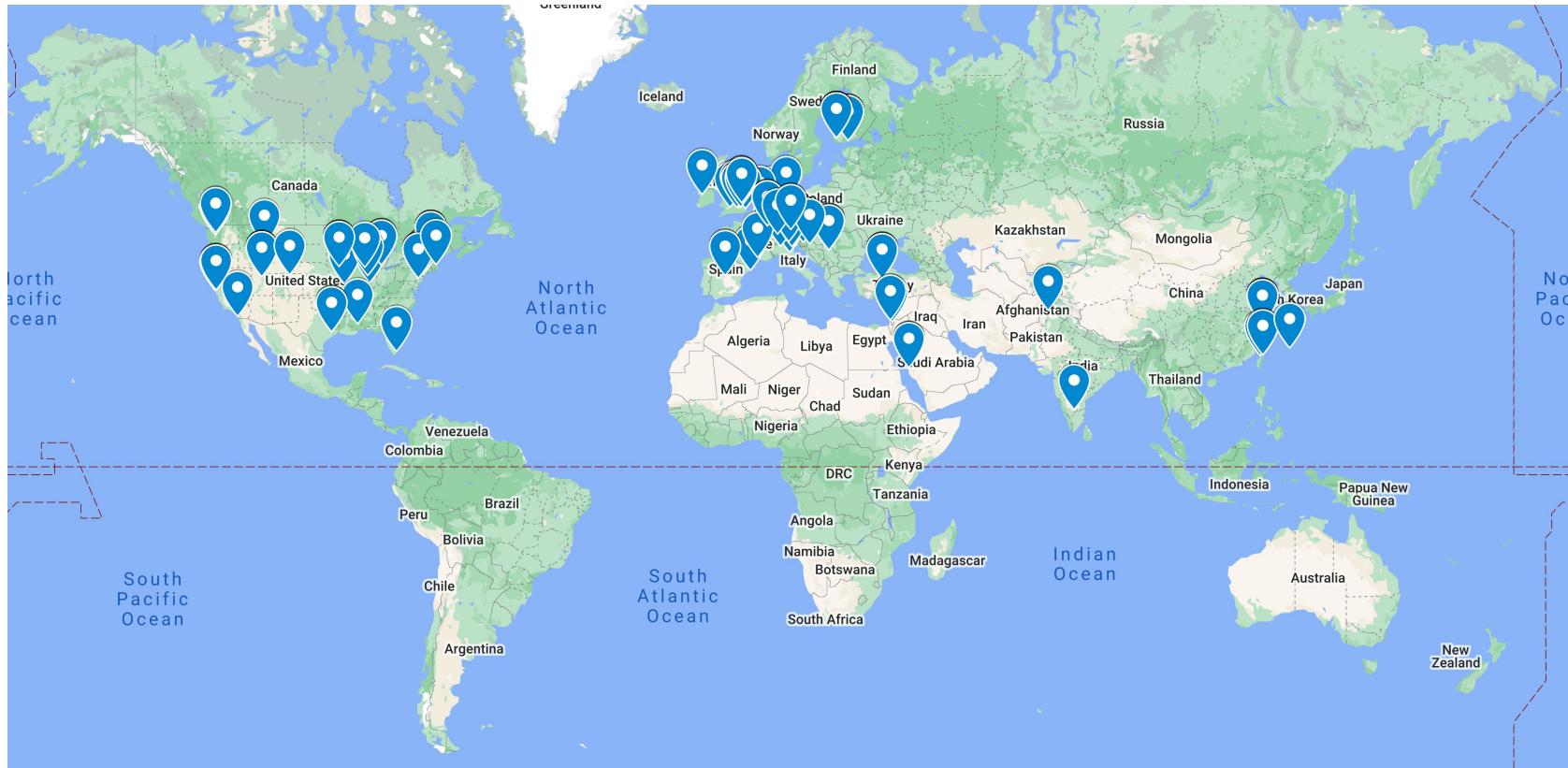
Marc RR

Dannie  
Durand

Steven  
Brenner



# Acknowledgments



>60 participating CAFA groups  
from 25 countries

