Introduction to 3DLigandSite and CombFunc

www.sbg.bio.ic.ac.uk/~mwass

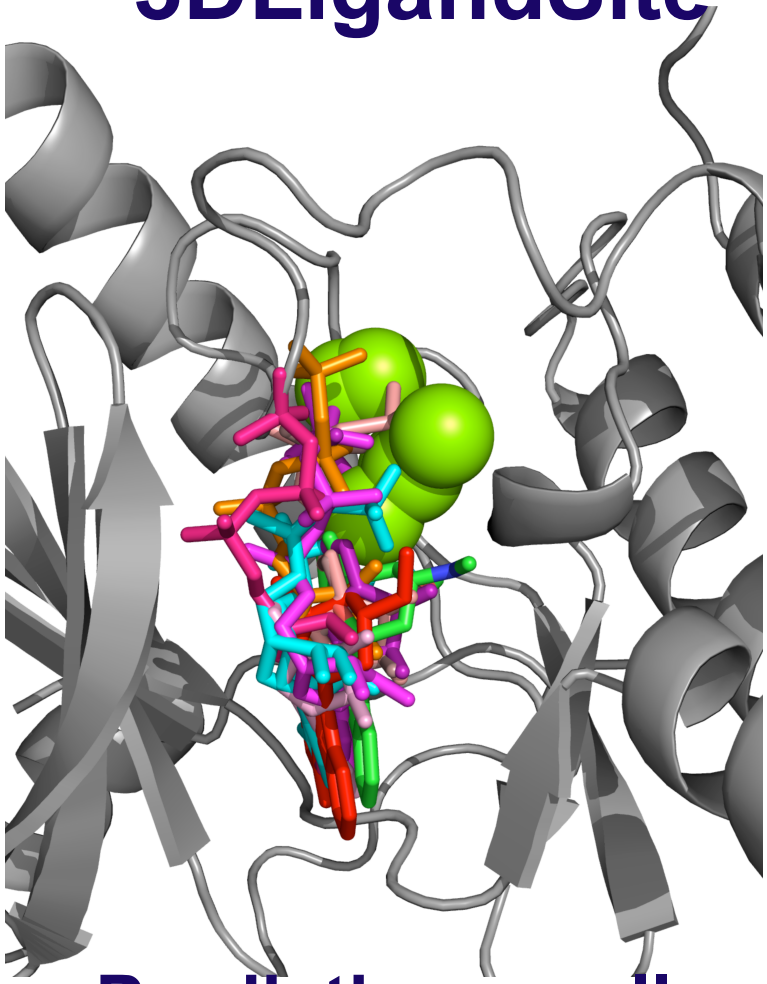
Links at Bottom of page

Mark Wass

m.n.wass@kent.ac.uk

3DLigandSite

3DLigandSite



Predicting small molecule binding sites

CombFunc

GO:0003924

GO:0004722

GO:0010564

GO:0005525

GO

Prediction

GO:0007067

GO:0008601

GO:0010458

Predicting protein function using Gene Ontology

CASP

MEEYKVVVCGSGPVALGCF Target sequence
(2 per day)

Human
predictors

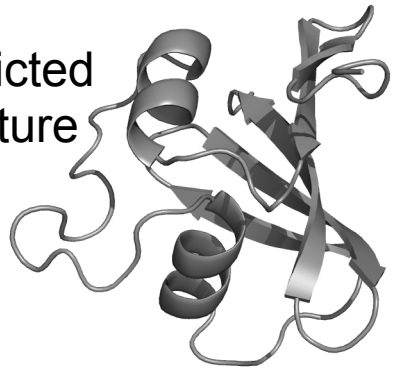
Server
predictors

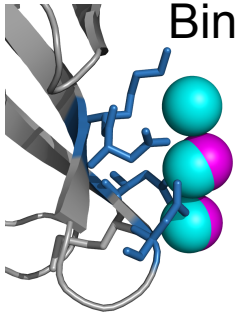
3 days

3 weeks

Predicted
Binding site

Predicted
structure



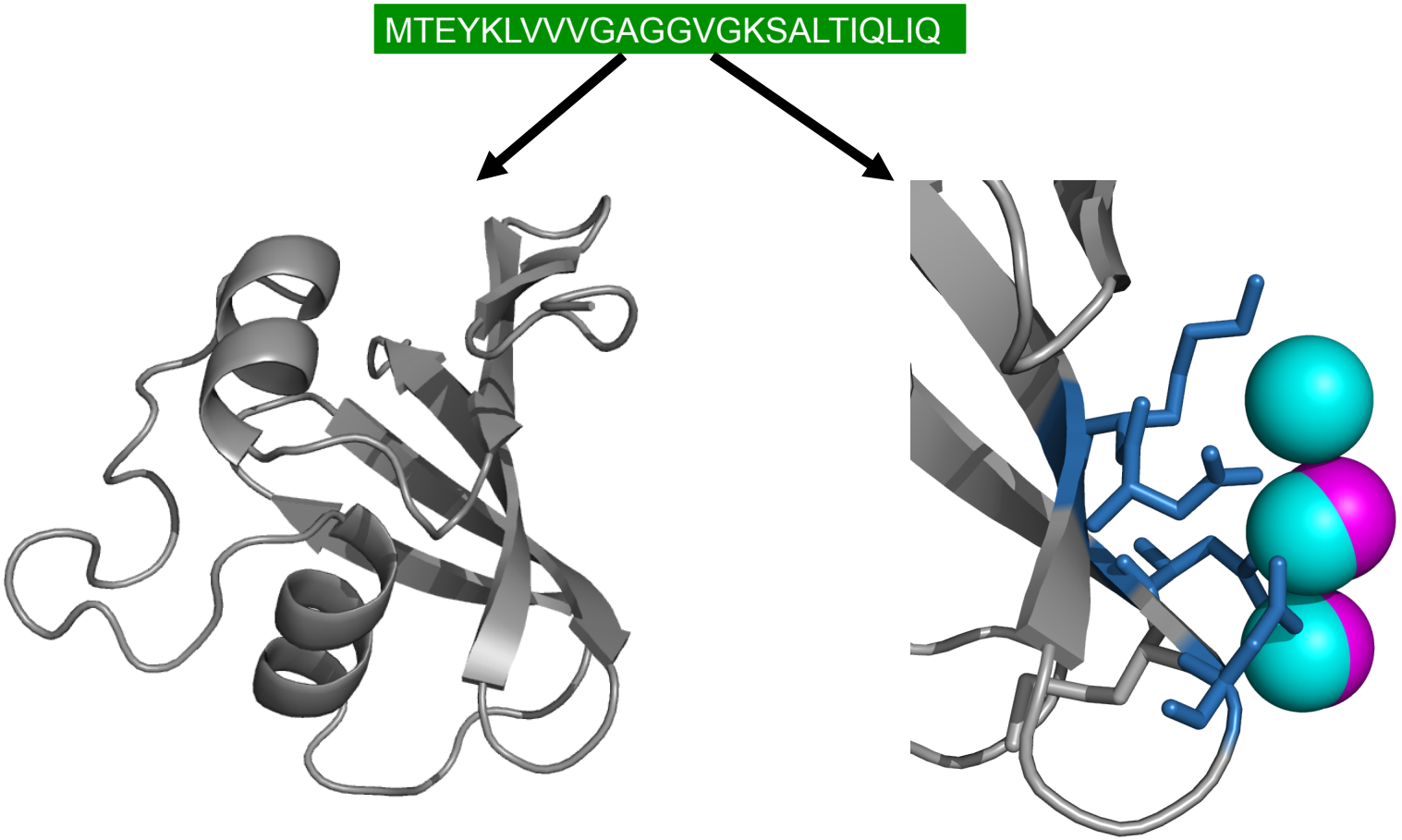


Assessment

Results
Performance
Compared to
other groups

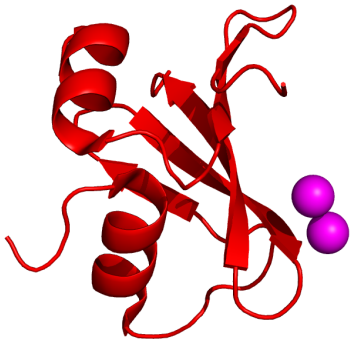
3DLigandSite

Developed as a result of participation in CASP

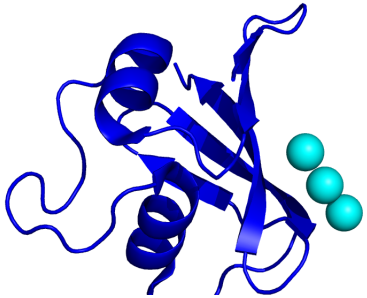


3DLigandSite

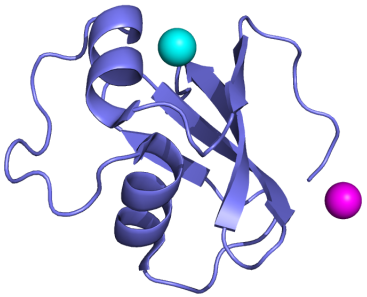
Homologous structures



2oai



2pls



2p4p

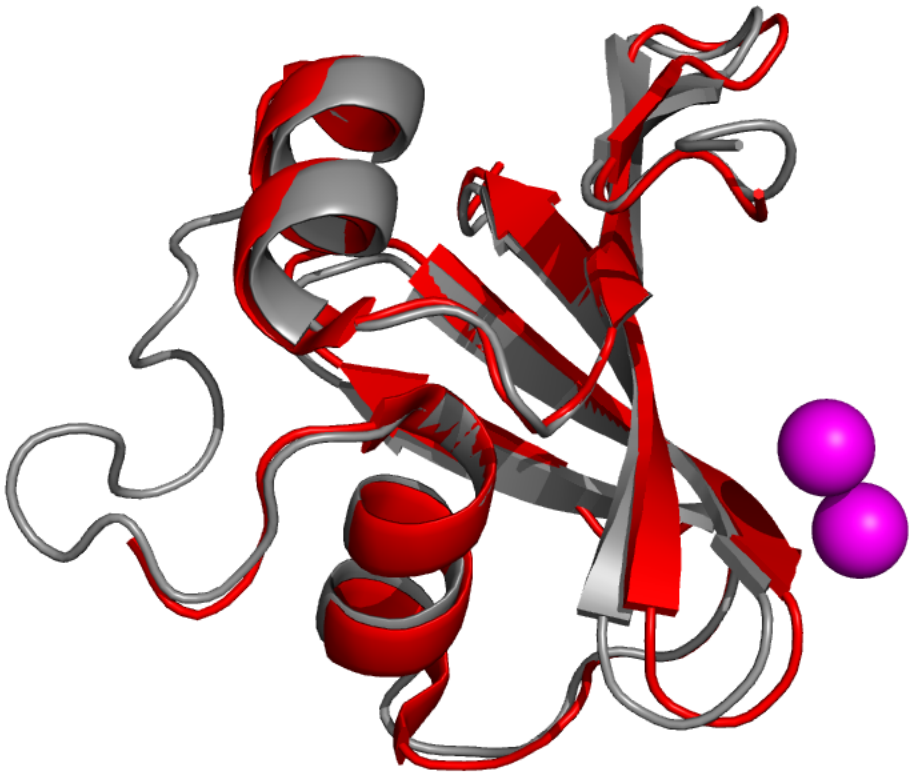
Magnesium

Calcium

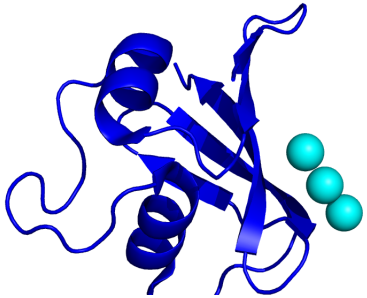
Wass & Sternberg *Proteins* 2009

3DLigandSite

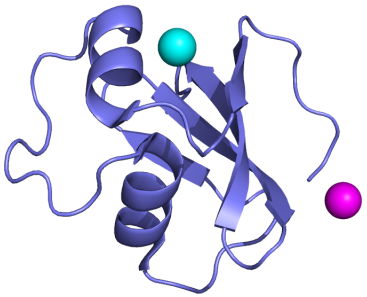
Homologous structures



2oai



2pls



2p4p

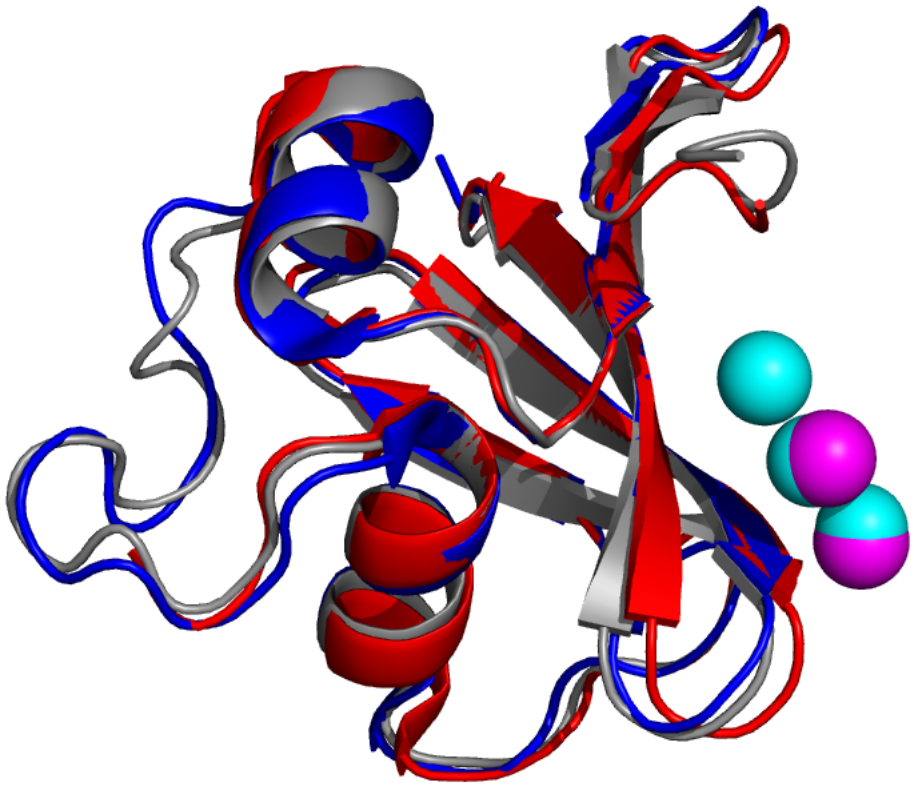
Magnesium

Calcium

Wass & Sternberg *Proteins* 2009

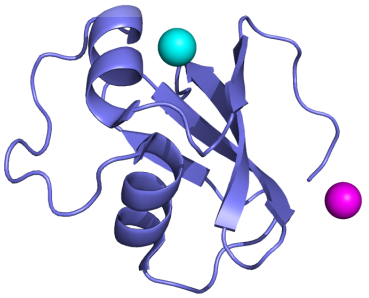
3DLigandSite

Homologous structures



2oai

2pls



2p4p

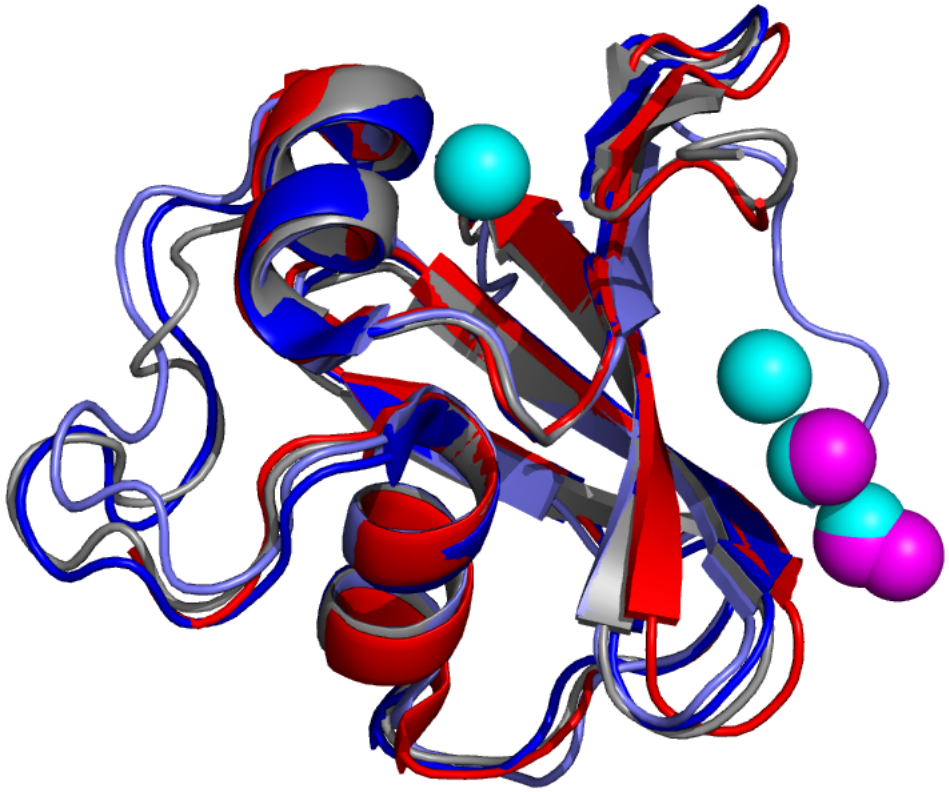
Magnesium

Calcium

Wass & Sternberg *Proteins* 2009

3DLigandSite

Homologous structures

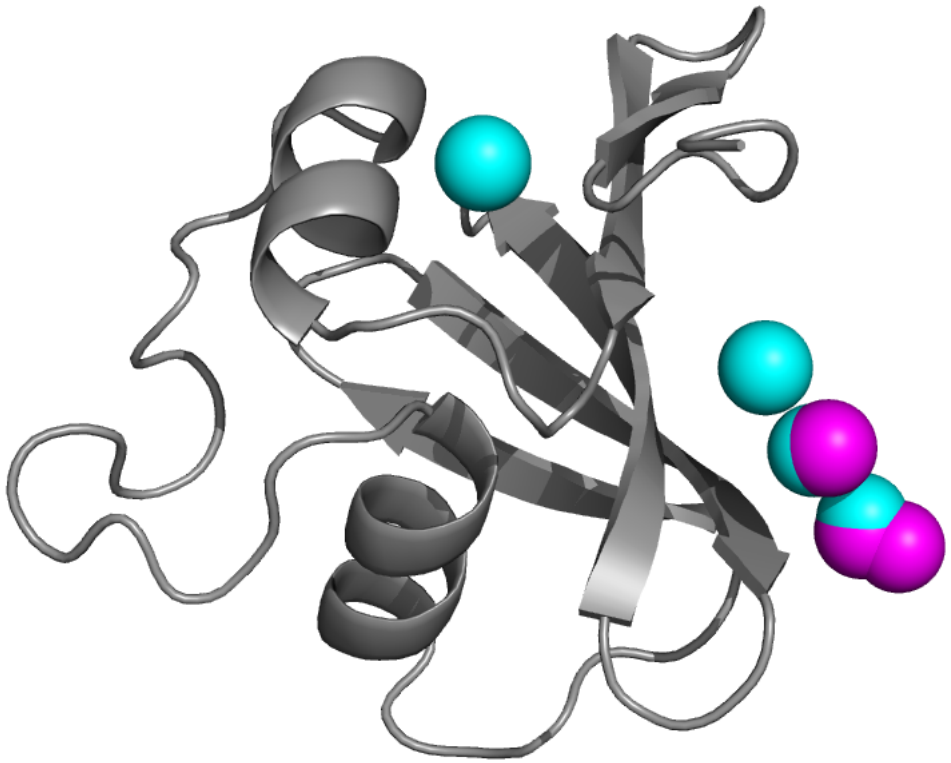


Magnesium Calcium

Wass & Sternberg *Proteins* 2009

3DLigandSite

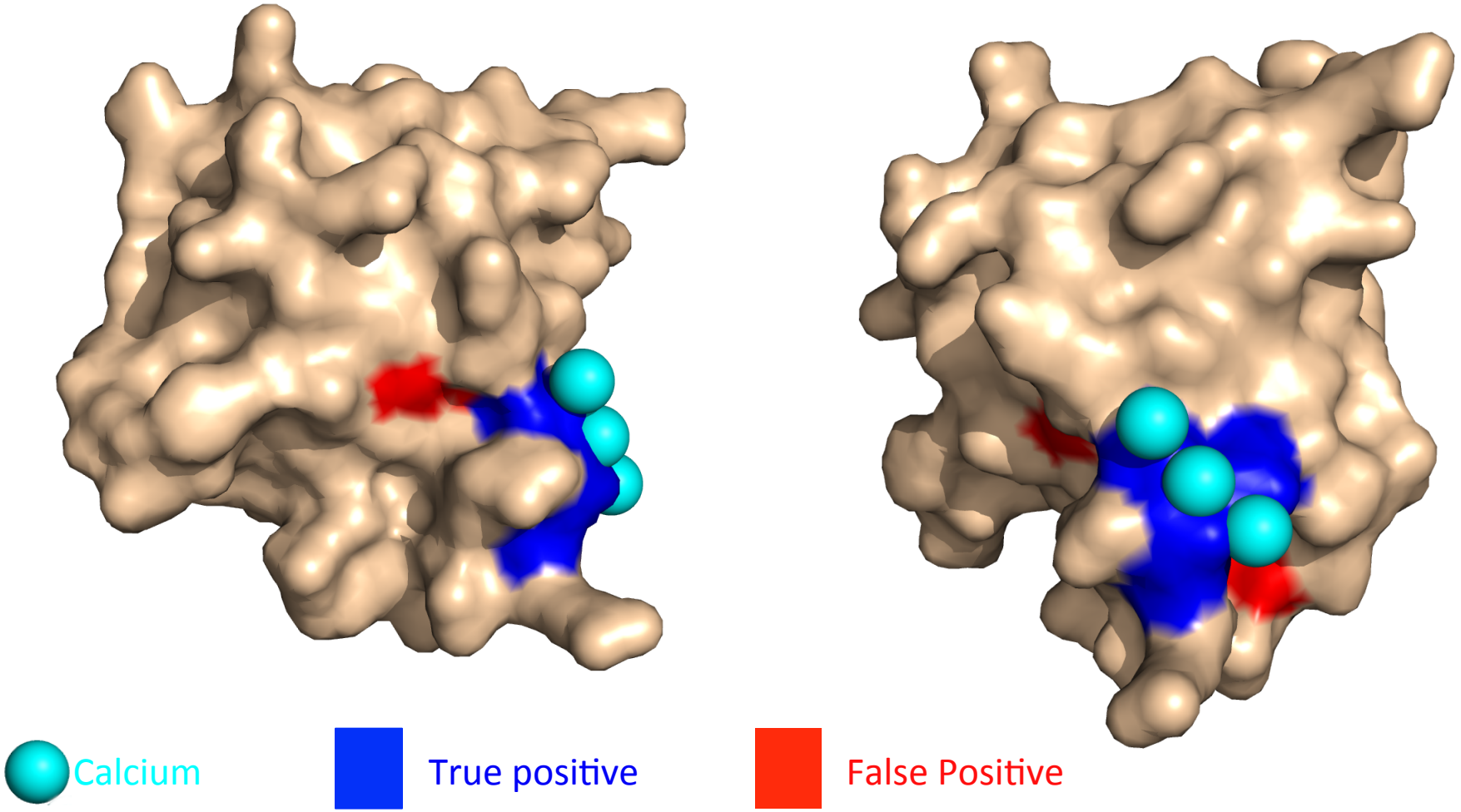
Homologous structures



Calcium Magnesium

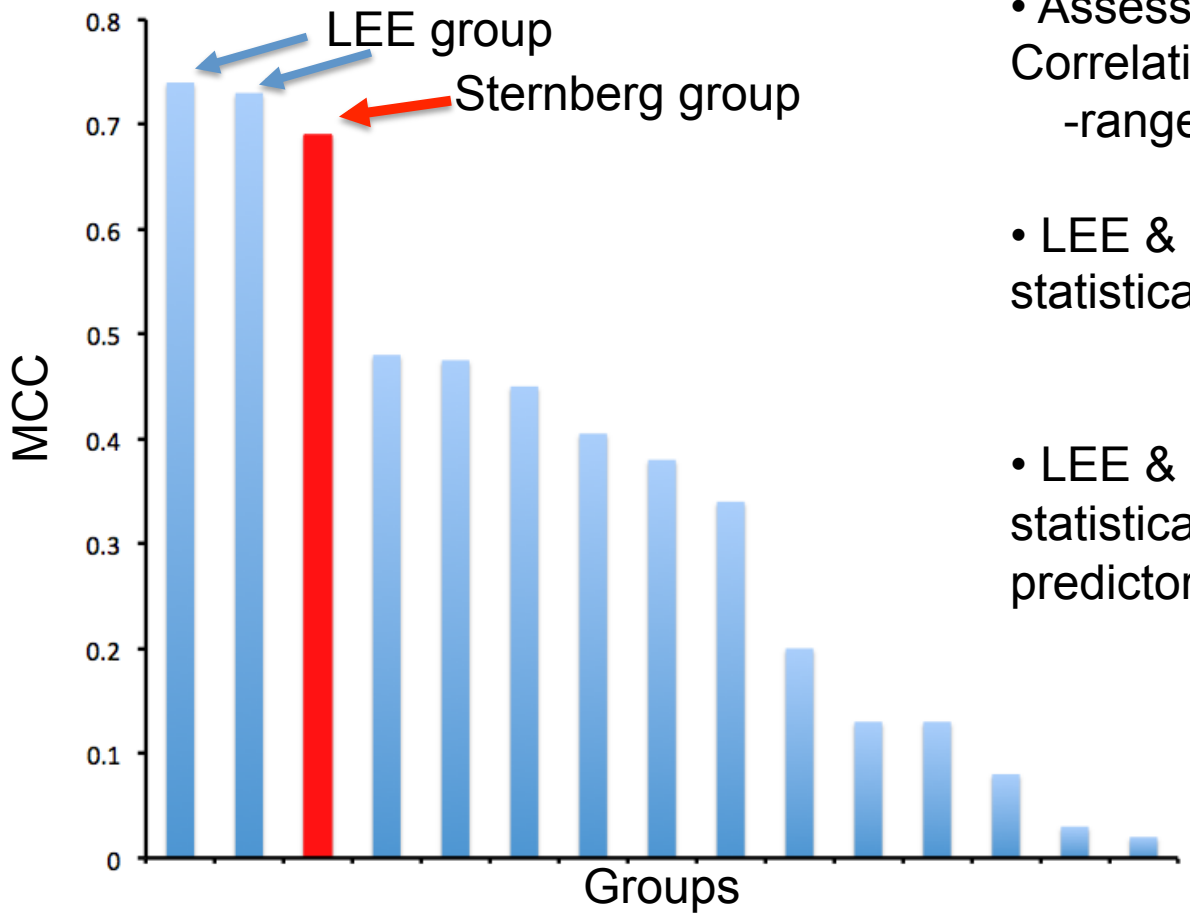
Wass & Sternberg *Proteins* 2009

3DLigandSite



Wass & Sternberg *Proteins* 2009

Performance at CASP8



- Assessed using Matthews' Correlation coefficient –
  -range -1 – +1

- LEE & Sternberg not statistically different.

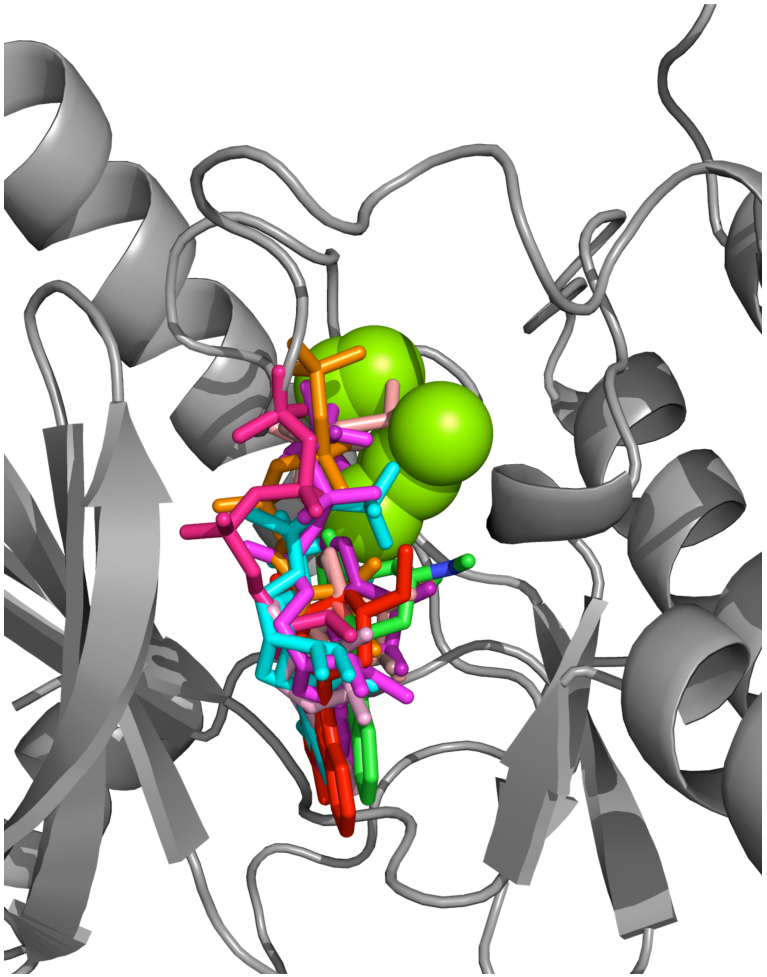
- LEE & Sternberg are statistically different to all other predictors ($p < 0.01$).

Adapted from Lopez et al., 2009

3DLigandSite

Automating our CASP8 approach

Structural View of Prediction



Jmol

Display Modification

Whole protein

colour by: ⦿ prediction ○ Jensen Shannon Divergence

spacefill: ⦿ off ○ 20% ○ 100%

wireframe: ⦿ off ○ wireframe ○ wireframe 50 ○ wireframe 100

☑ cartoon

Predicted residues

spacefill: ⦿ off ○ 20% ○ 100%

wireframe: ⦿ off ○ on ○ wireframe 50 ○ wireframe 100

☑ cartoon
☐ label

Heterogens

Display of Metalic heterogens
spacefill: ○ off ○ 20% ⦿ 100%

Display of Non Metalic heterogens
spacefill: ⦿ off ○ 20% ○ 100%

wireframe: ○ off ○ standard ○ wireframe 50 ⦿ wireframe 100

View

Reset to original orientation

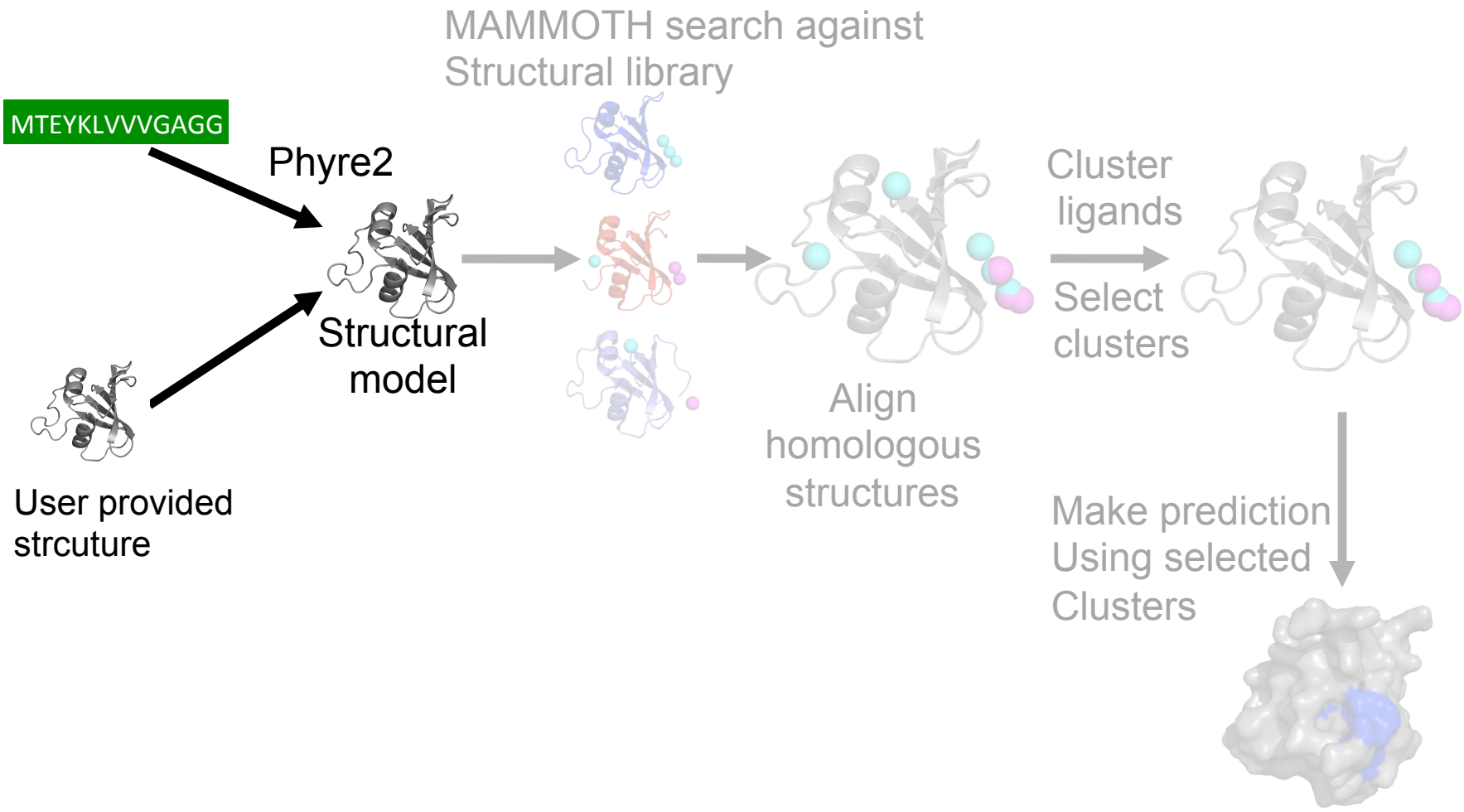
☐ spin

background black

Prediction colour legend:	Other residues	Predicted Binding Site			
Conservation Score Colour legend:	0-0.15	0.16-0.30	0.31-0.40	0.41-0.50	
	0.51-0.60	0.61-0.70	0.71-0.80	0.81-1.00	

Wass et al., NAR 2010

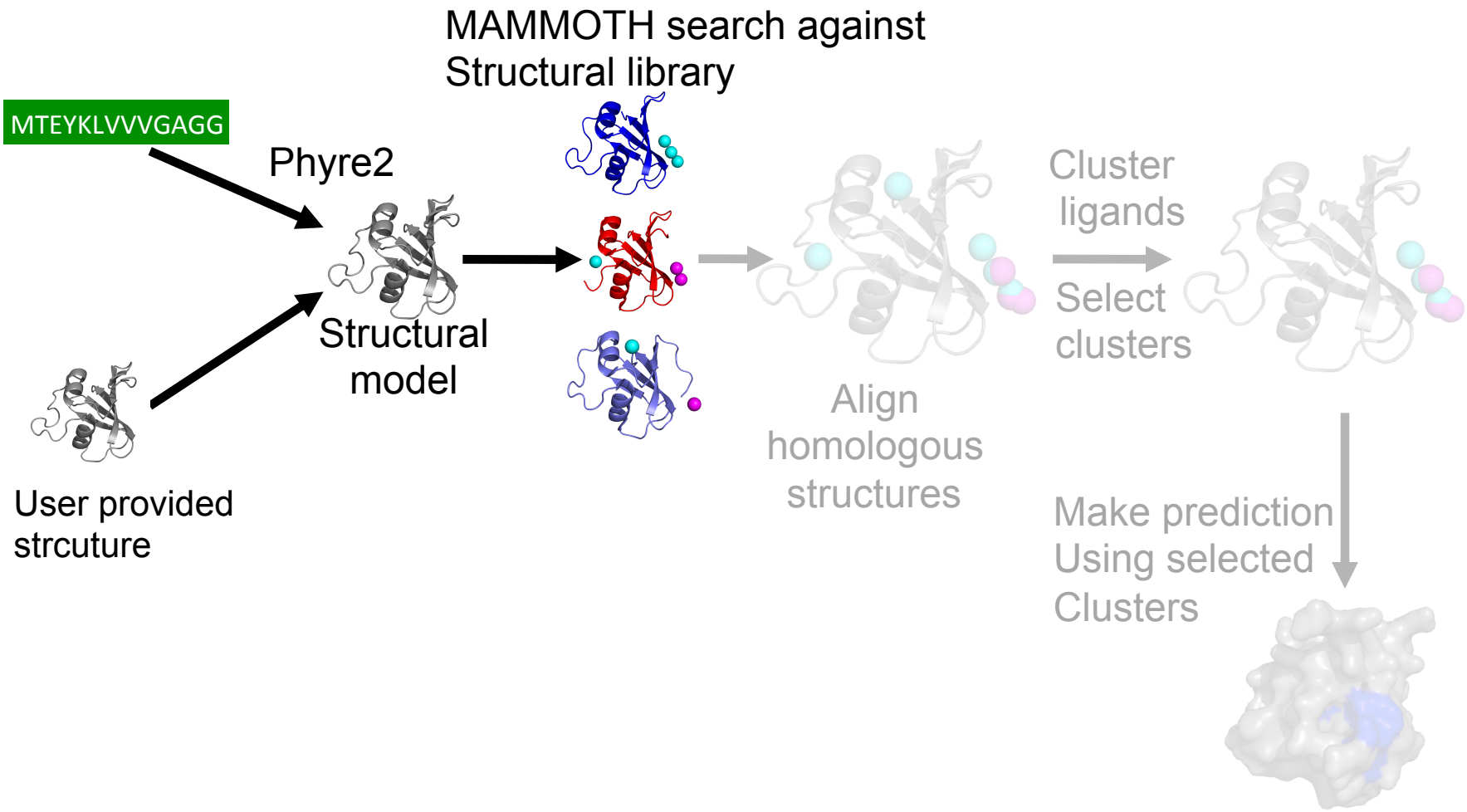
3DLigandSite



Wass *et al*., NAR 2010

Imperial College
London

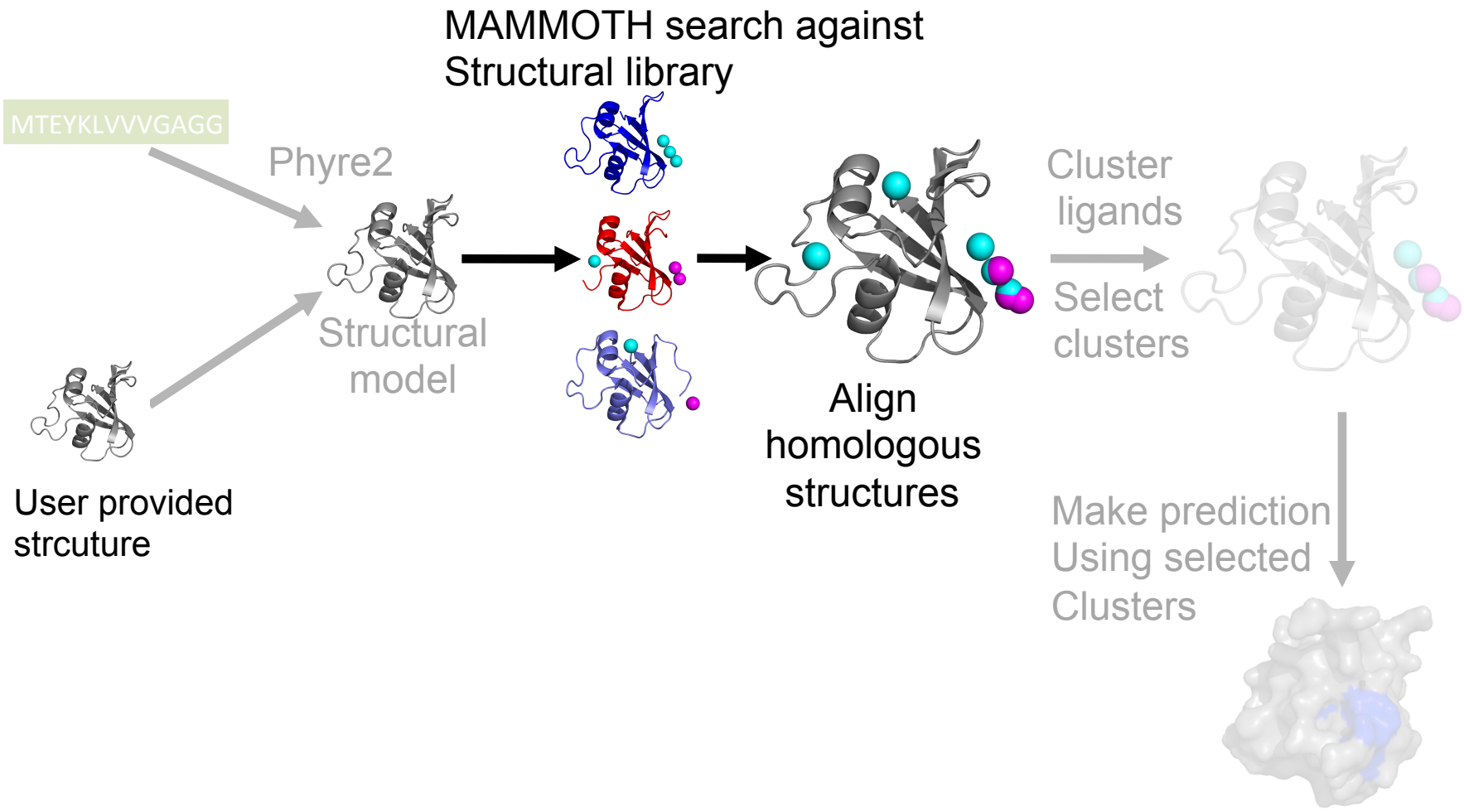
3DLigandSite



Wass *et al*., NAR 2010

Imperial College
London

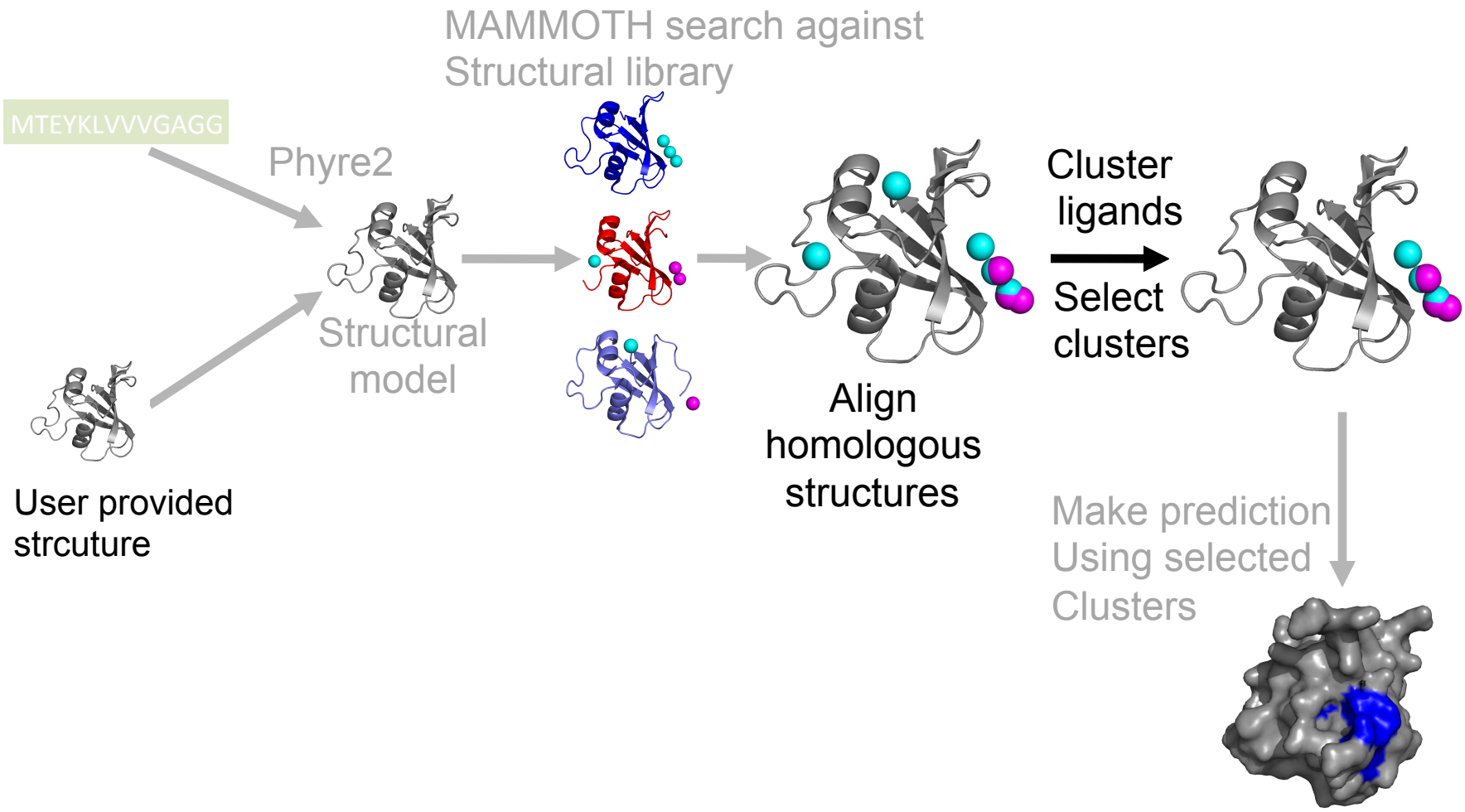
3DLigandSite



Wass *et al*., NAR 2010

Imperial College
London

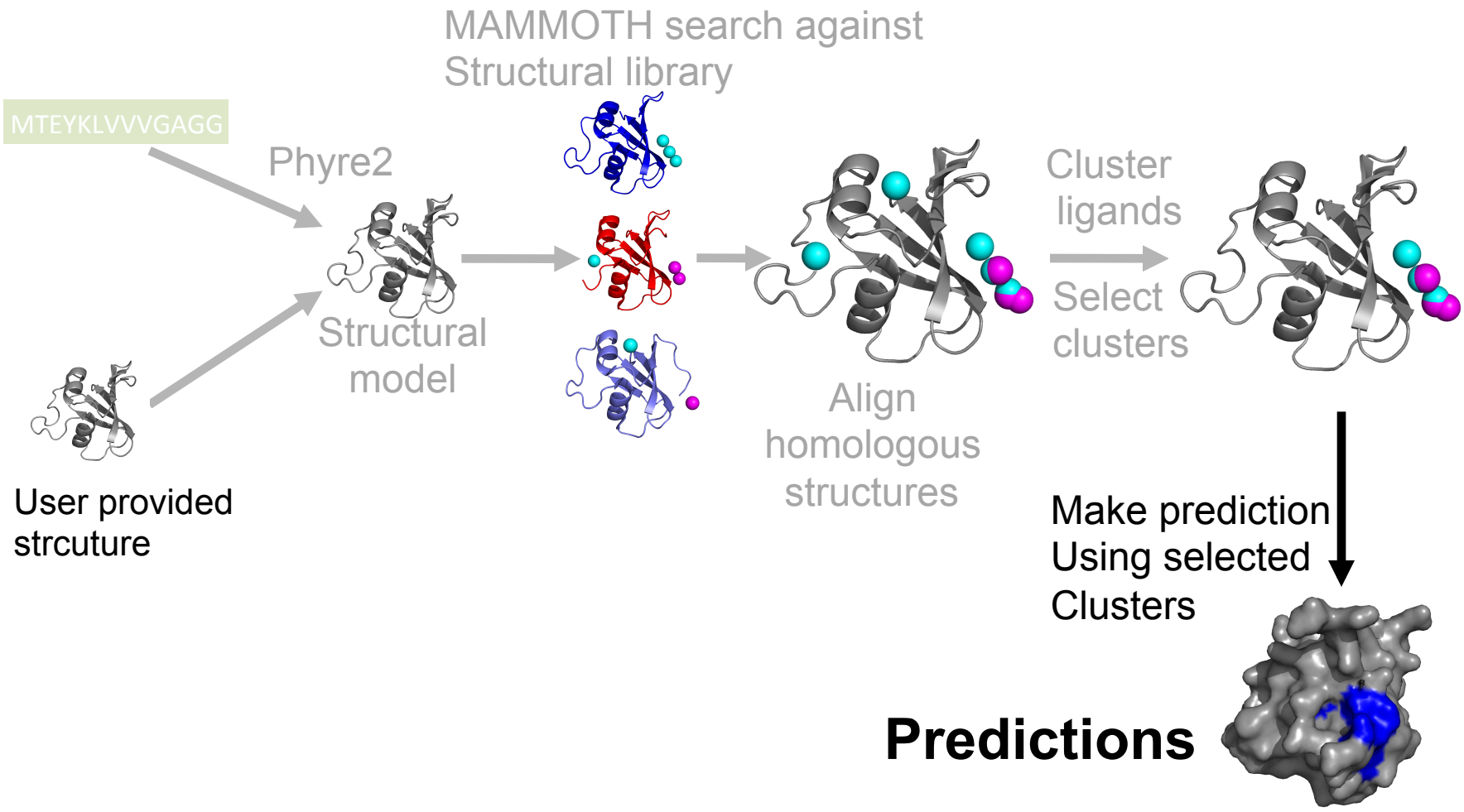
3DLigandSite



Wass *et al*., NAR 2010

Imperial College
London

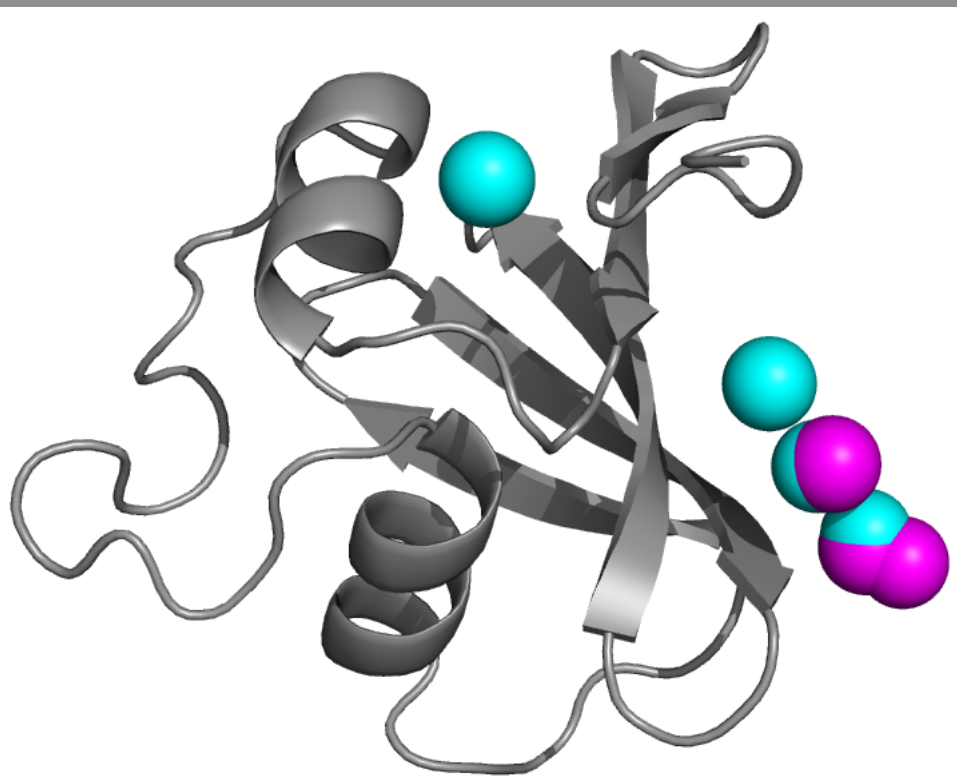
3DLigandSite

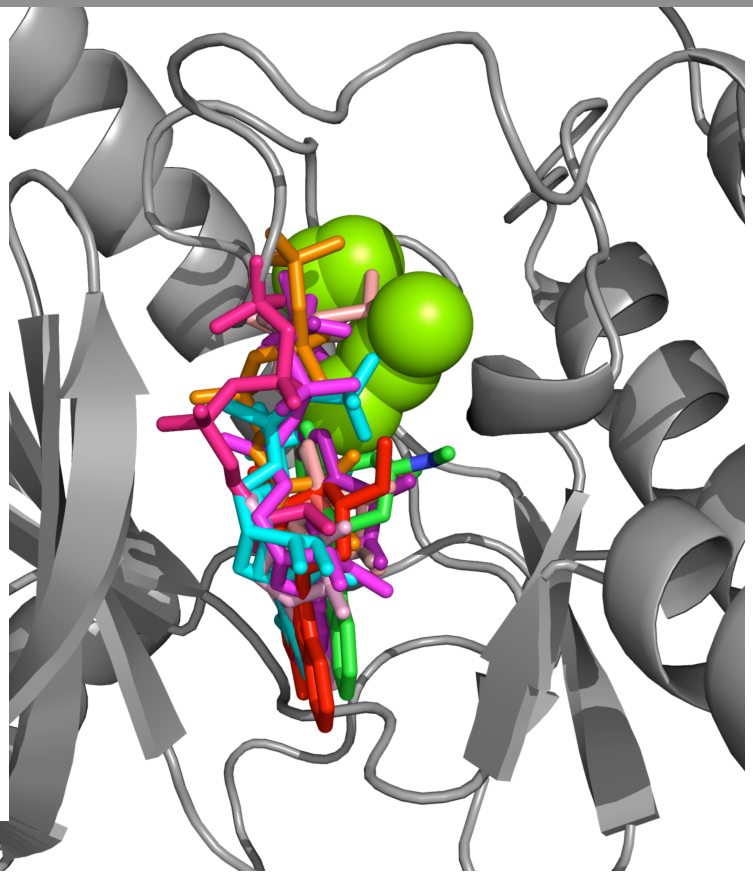


Wass *et al*., NAR 2010

Imperial College London

Predicting contacting residues





Multiple molecules in cluster but where is the actual binding site?

Threshold for prediction = Contact 25% ligands

3DLigandSite Benchmarking

FINDSITE set (617)

Measure	3DLigandSite
MCC	0.68
Recall	70%
Precision	70%

CASP8 targets (28)

Measure	3DLigandSite	Human CASP8
MCC	0.64	0.63
Recall	71%	83%
Precision	60%	56%

MCC – Matthews Correlation Coefficient
Recall– percentage of binding sites that are predicted (TP/(TP+FN))
Precision– percentage of predicted residues that are correct (TP/(TP+FP))

CombFunc

Protein Function Prediction

Why predict protein function?

- Many genomes sequenced but function unknown for many of the proteins they code
    - 80M protein sequences in UniProt!
- Experimental characterisation slow
- Direct experimental studies

Gene Ontology

- Graph Structure of functional terms from General to Specific Terms
- 3 main categories – Biological Process, Molecular Function, Cellular Component

What is Protein function?

- Many possible meanings
- *'Everything that happens to or through a protein'*

  Rost 2003
- *Can be described at different levels*
  - *Molecular function – (e.g. biochemical role e.g. enzyme function)*
  - *Cellular function – (larger functional process – e.g. system that that an enzyme is part of.*
  - *Phentypic function – (resulting phenotype e.g. related to disease)*

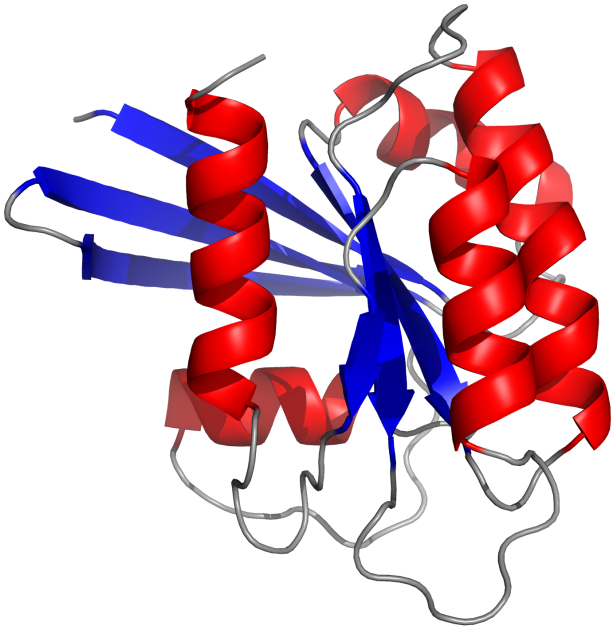
Gene Ontology

- ***Molecular Function*** = biochemical function
  - the tasks performed by individual gene products; examples are *carbohydrate binding* and *GTPase activity*
- ***Biological Process*** = biological goal or objective (higher level function)
  - broad biological goals, such as *mitosis* or *purine metabolism*, that are accomplished by combinations of individual molecular functions.
- ***Cellular Component*** = active location
  - subcellular structures, locations, and macromolecular complexes; examples include *nucleus*, *telomere*, and *RNA polymerase II holoenzyme*

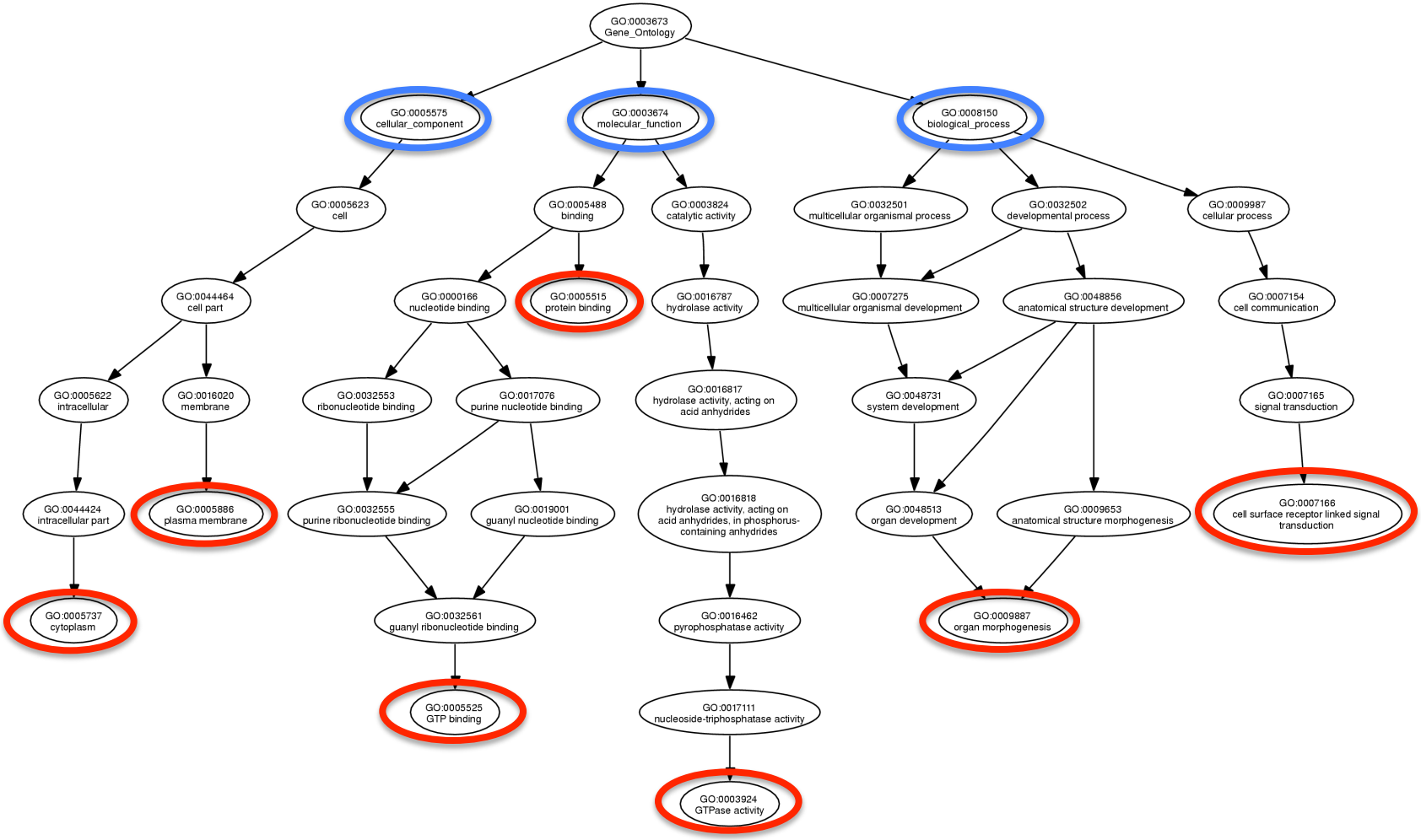
GO annotations of Ras

Ras is an oncogene – mutations present in many cancers

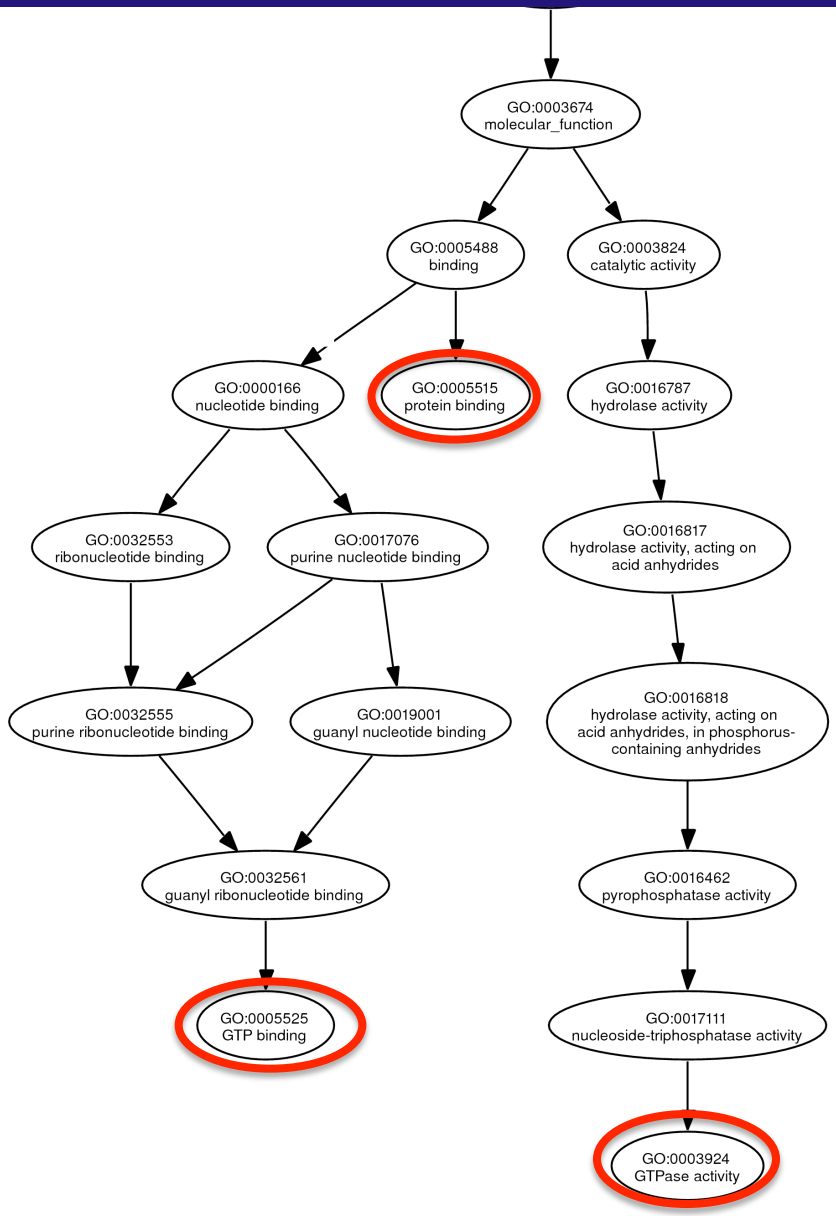
What are the GO annotations for Ras?



GO annotations of Ras



GO annotations of Ras

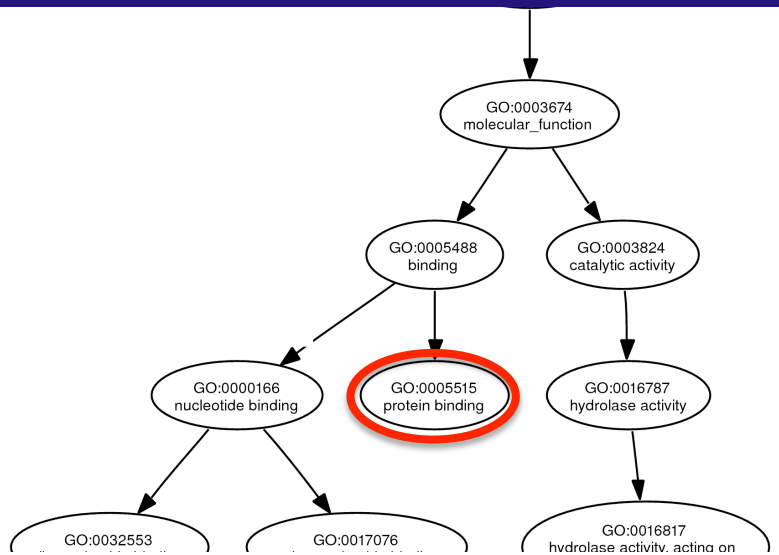


Molecular Function

$$GTP + H_2O \longrightarrow GDP + Pi$$

Ras is a GTPase

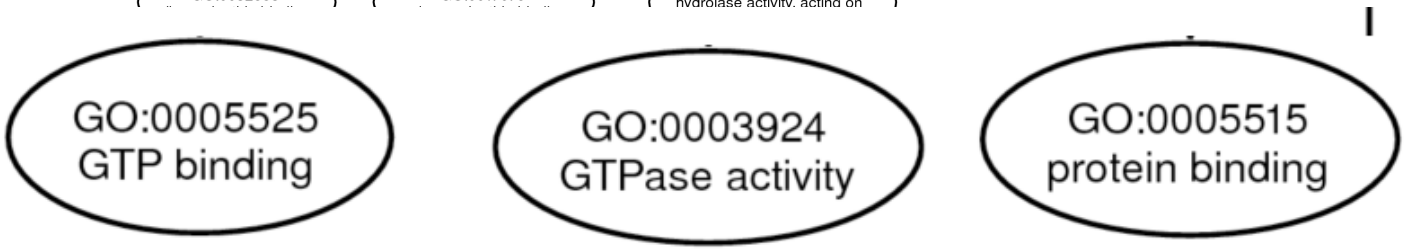
GO annotations of Ras

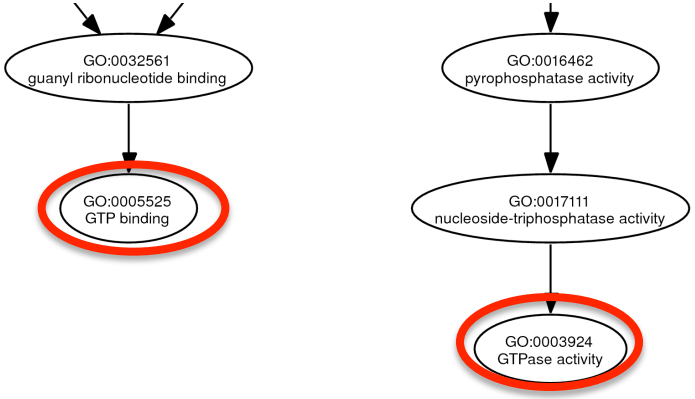


Molecular Function

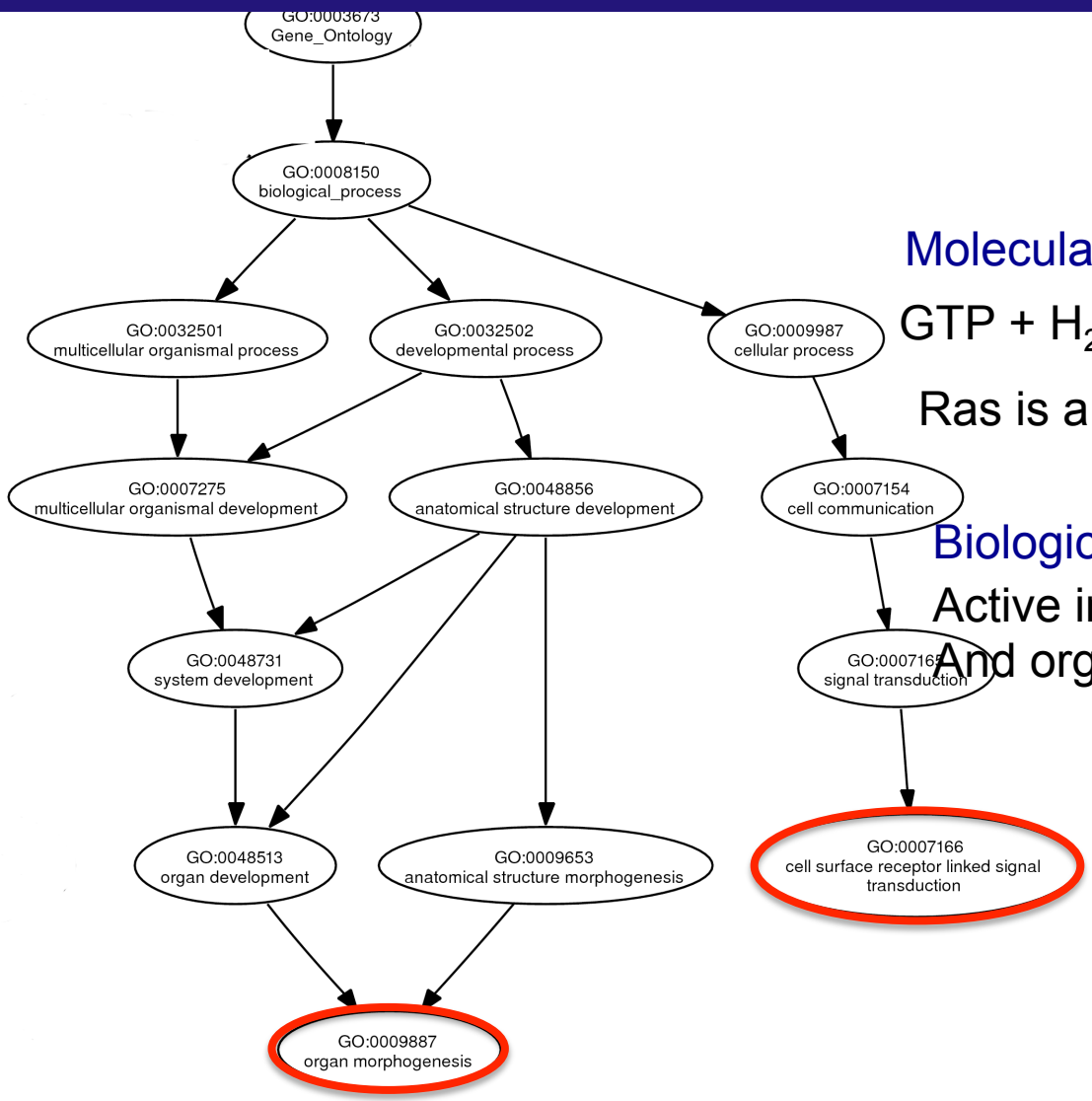
$$GTP + H_2O \longrightarrow GDP + Pi$$

Ras is a GTPase





GO annotations of Ras



Molecular Function

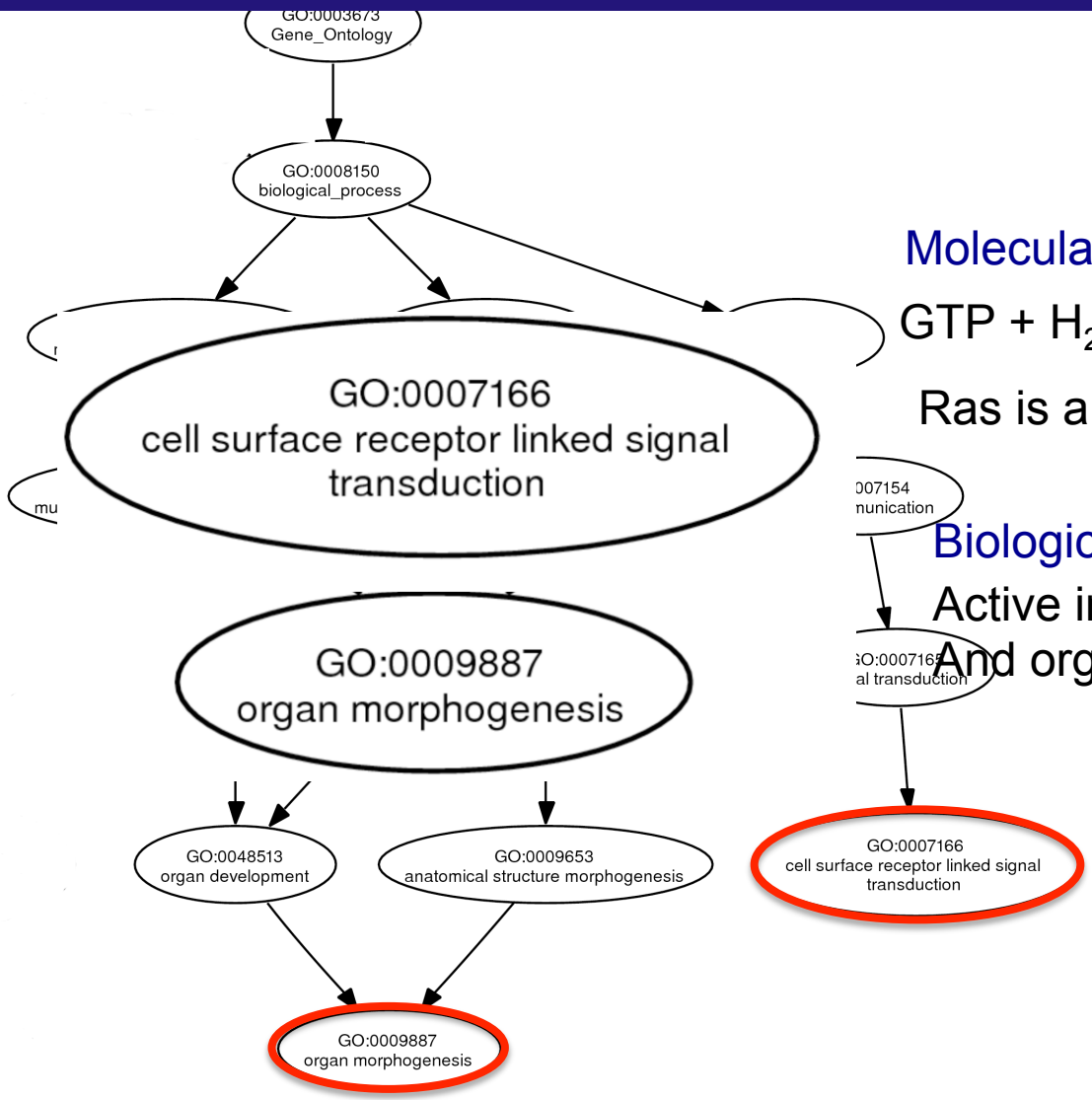
$GTP + H_2O \longrightarrow GDP + Pi$

Ras is a GTPase

Biological Process

Active in cell surface signal transduction
And organ morphogenesis

GO annotations of Ras



Molecular Function

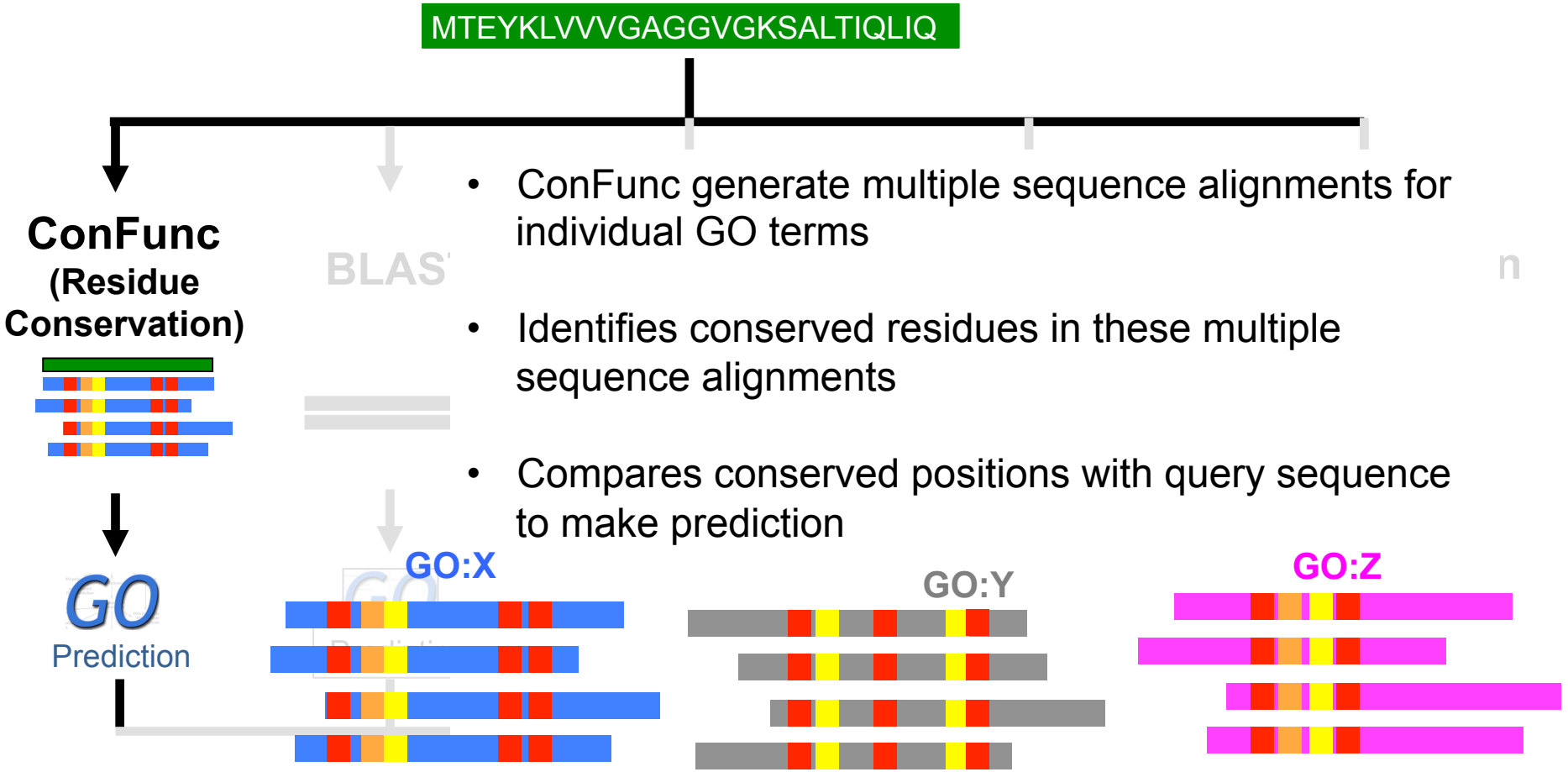
$GTP + H_2O \longrightarrow GDP + Pi$

Ras is a GTPase

Biological Process

Active in cell surface signal transduction
And organ morphogenesis

ConFunc/CombFunc

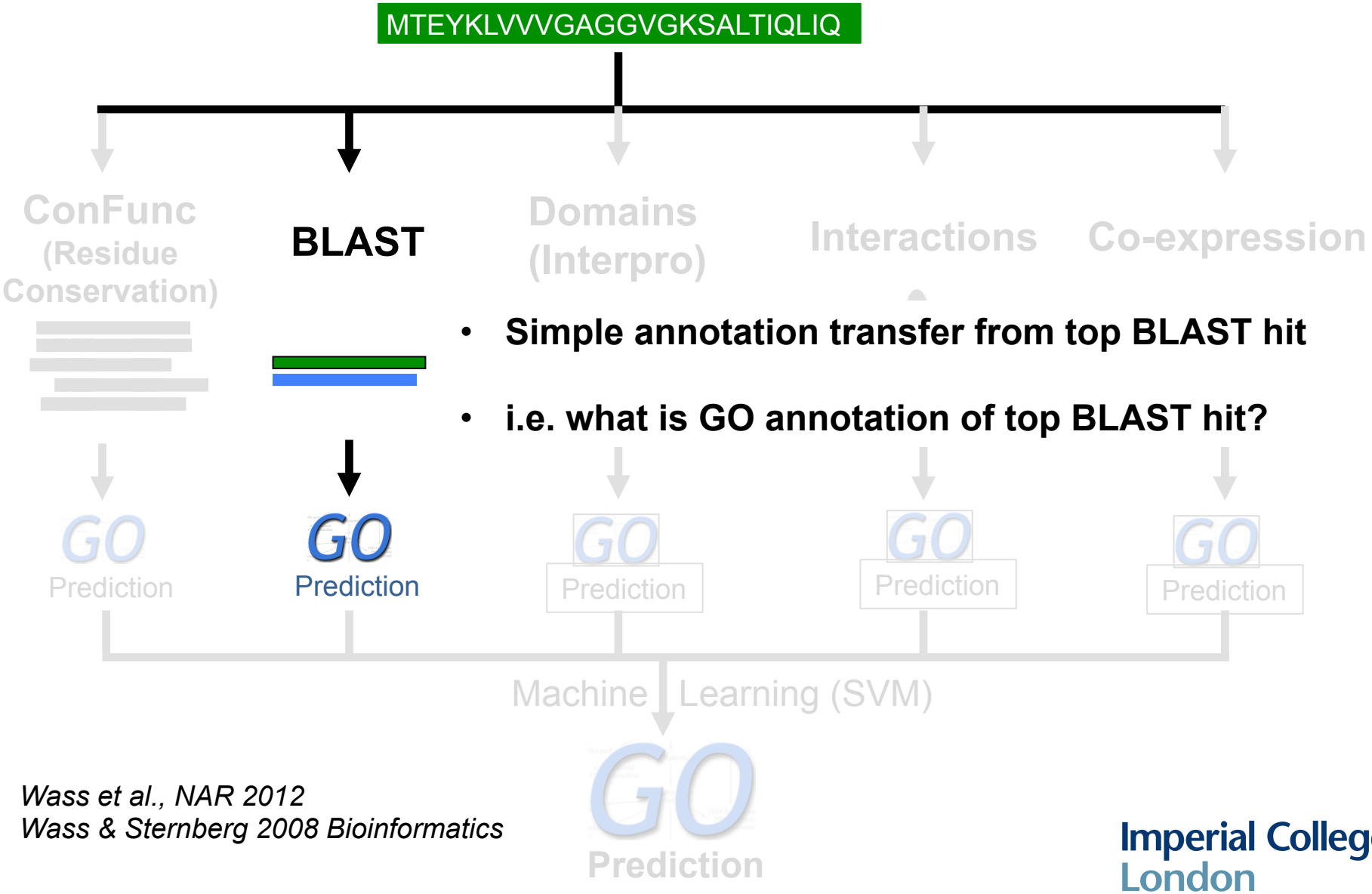


- ConFunc generate multiple sequence alignments for individual GO terms
- Identifies conserved residues in these multiple sequence alignments
- Compares conserved positions with query sequence to make prediction

Wass et al., NAR 2012
Wass & Sternberg 2008 Bioinformatics

Imperial College London

ConFunc/CombFunc

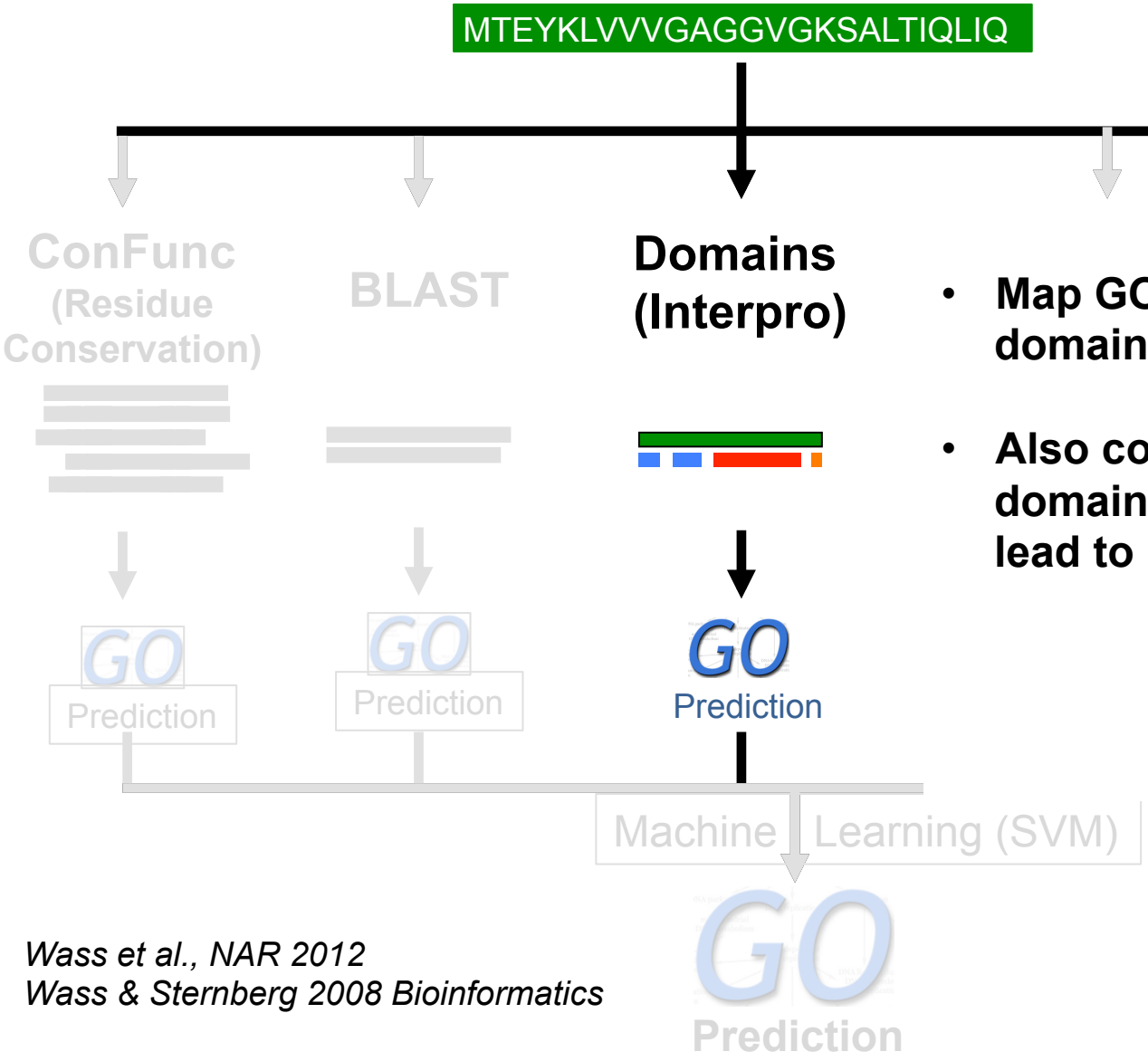


- **Simple annotation transfer from top BLAST hit**
- **i.e. what is GO annotation of top BLAST hit?**

Wass et al., NAR 2012
Wass & Sternberg 2008 Bioinformatics

Imperial College London

ConFunc/CombFunc



- **Map GO function to protein domains**
- **Also consider combinations of domains and the function they lead to**

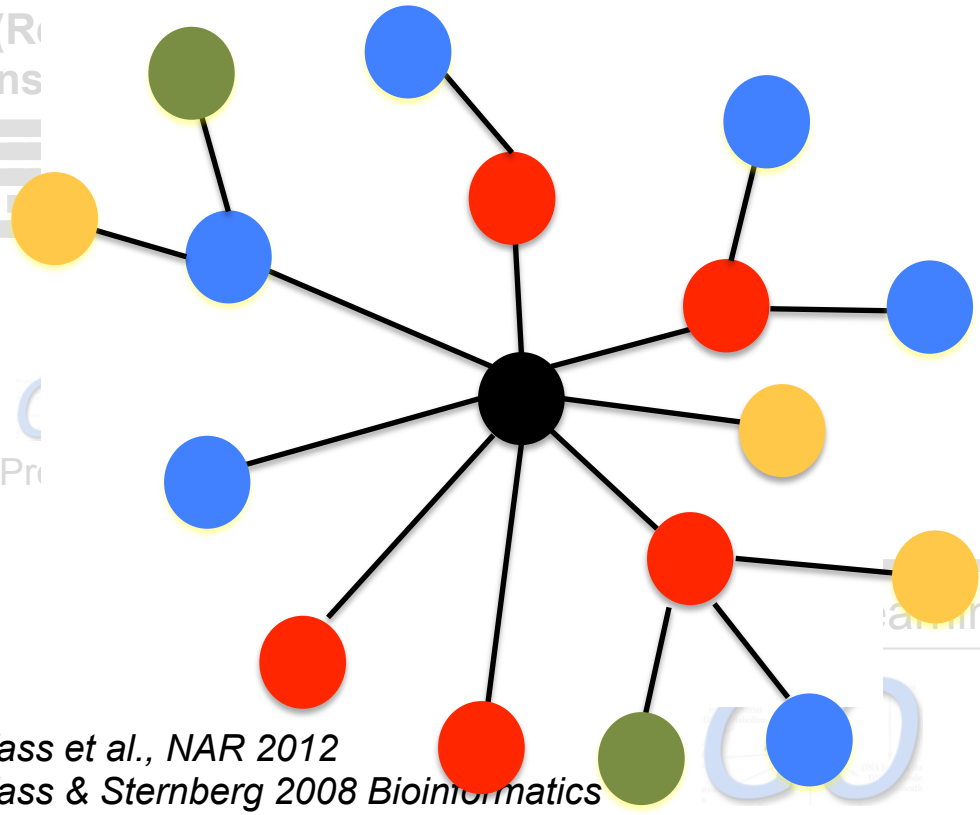
Wass et al., NAR 2012
Wass & Sternberg 2008 Bioinformatics

Imperial College London

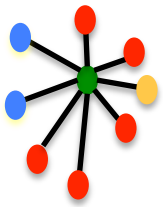
ConFunc/CombFunc

MTEYKLVVVGAGGVGKSALTIQLIQ

- **Similar function to interaction partners?**



Interactions



GO
Prediction

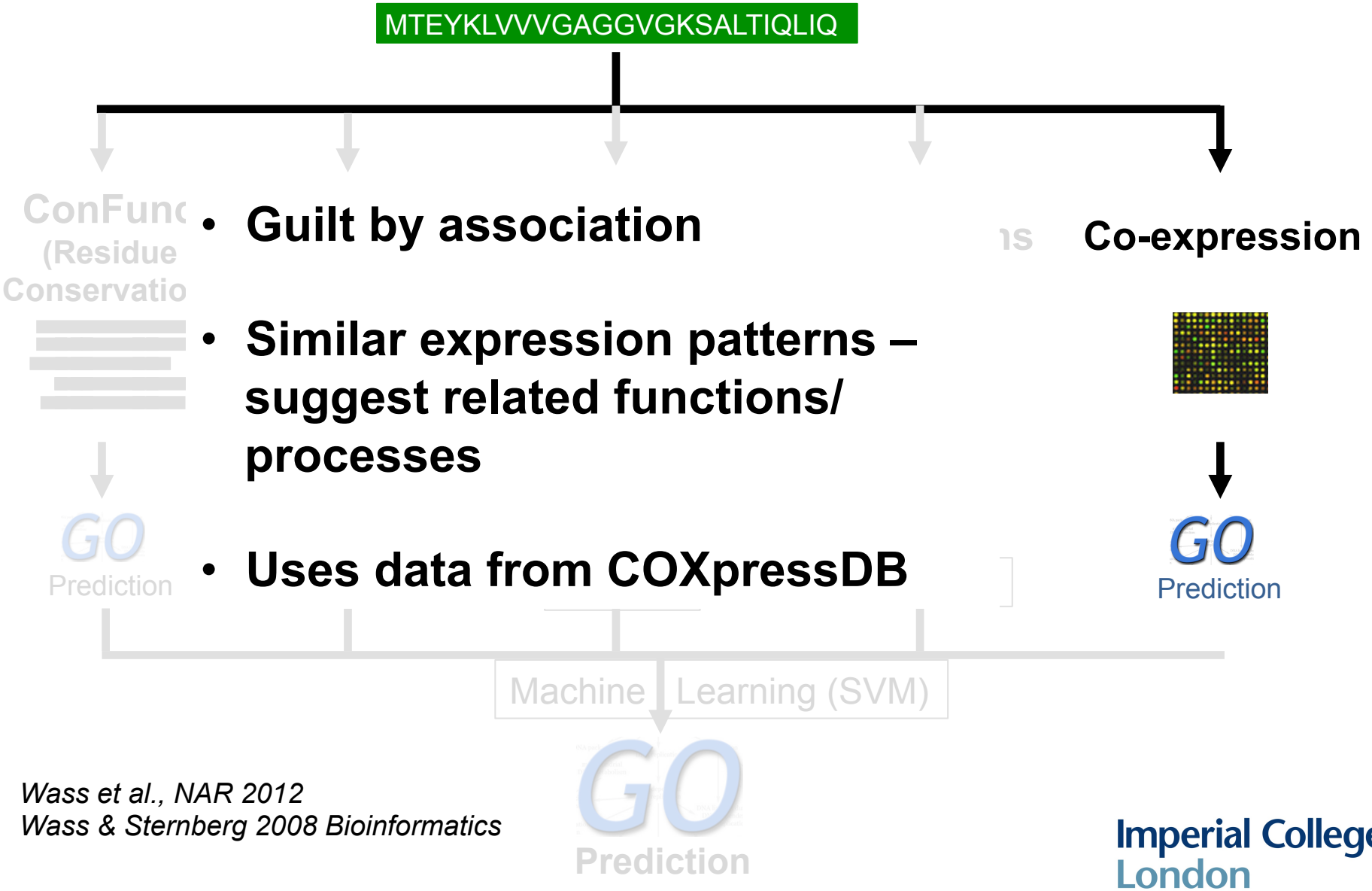
Co-expression

GO
Prediction

Wass et al., NAR 2012
Wass & Sternberg 2008 Bioinformatics

Imperial College London

ConFunc/CombFunc

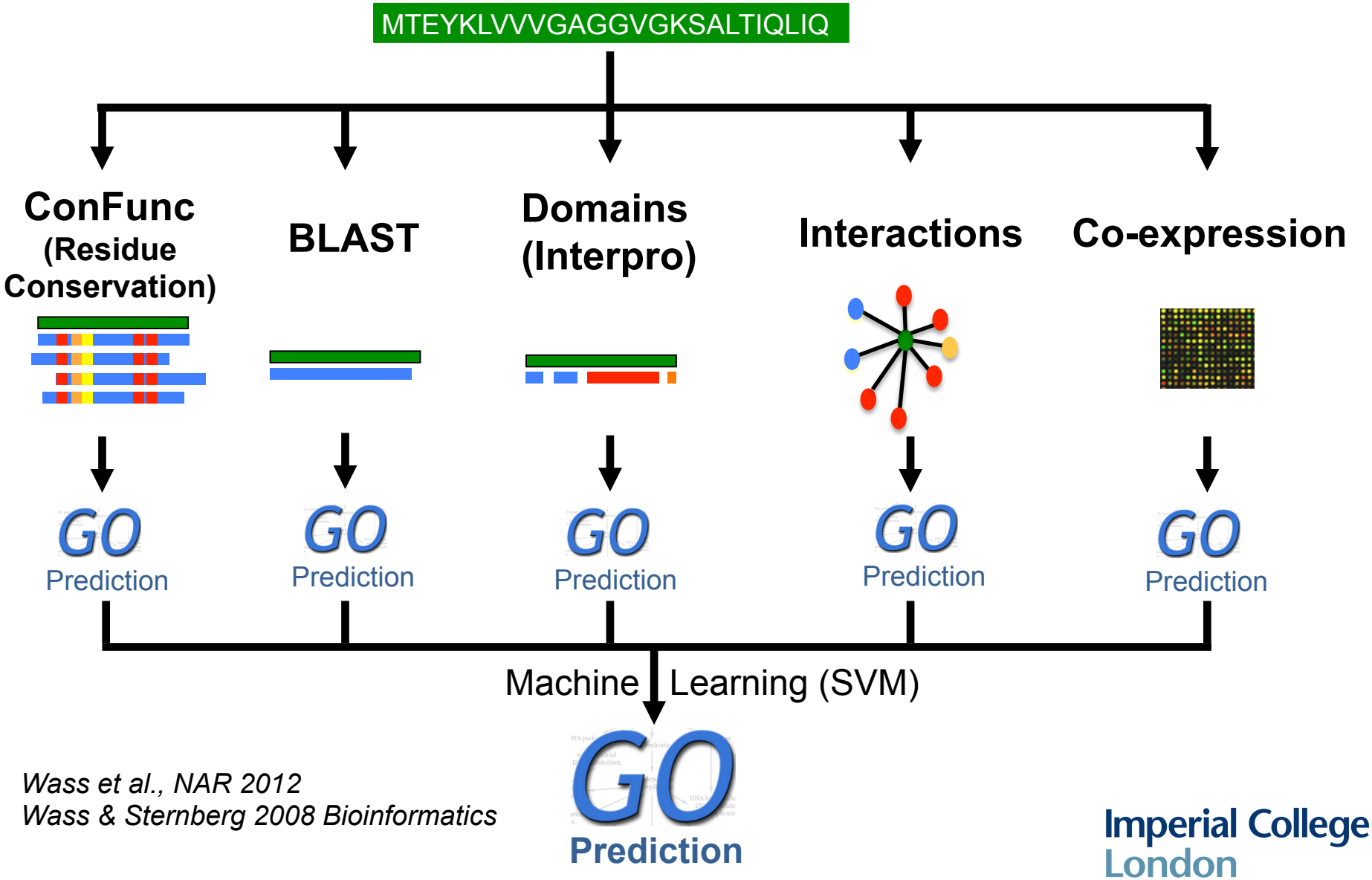


- **Guilt by association**
- **Similar expression patterns – suggest related functions/ processes**
- **Uses data from COXpressDB**

Wass et al., NAR 2012
Wass & Sternberg 2008 Bioinformatics

Imperial College London

ConFunc/CombFunc



Wass et al., NAR 2012
Wass & Sternberg 2008 Bioinformatics

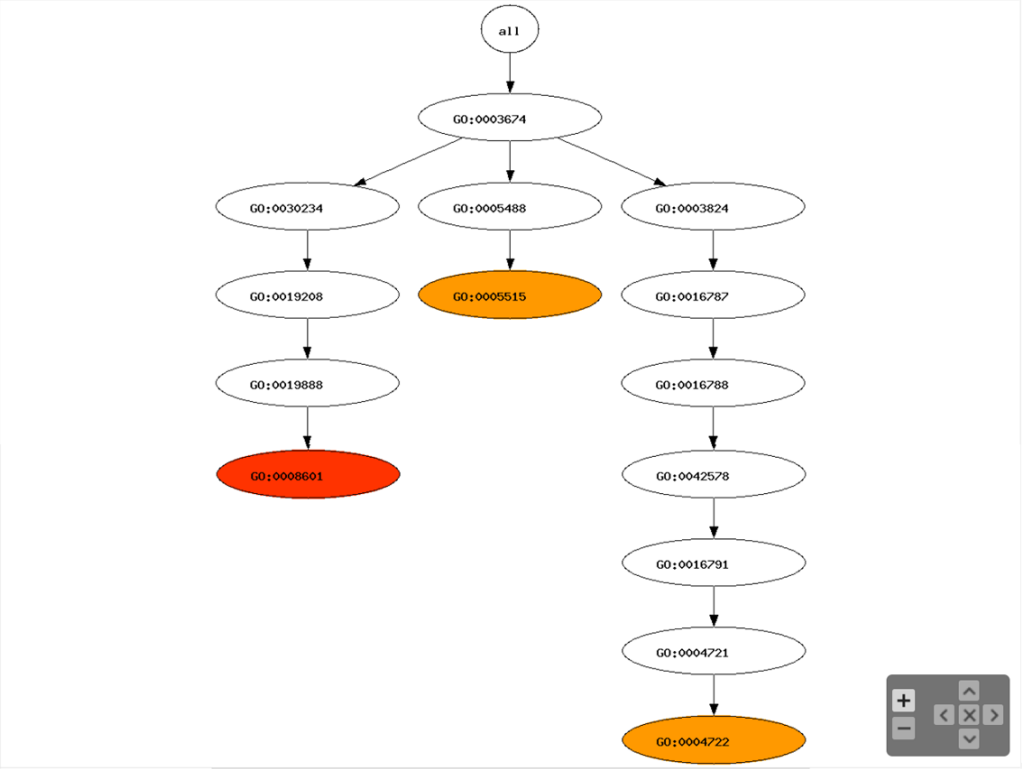
Imperial College London

CombFunc – web server

http://www.sbg.bio.ic.ac.uk/combfunc

Molecular Function Predictions

GO Term	Description	SVM Score	SVM Probability
GO:0008601 GO	protein phosphatase type 2A regulator activity	3.284	0.881
GO:0004722 GO	protein serine/threonine phosphatase activity	1.004	0.495
GO:0005515 GO	protein binding	1.002	0.477



The lists below can also be used to explore the predictions in terms of the GO graph

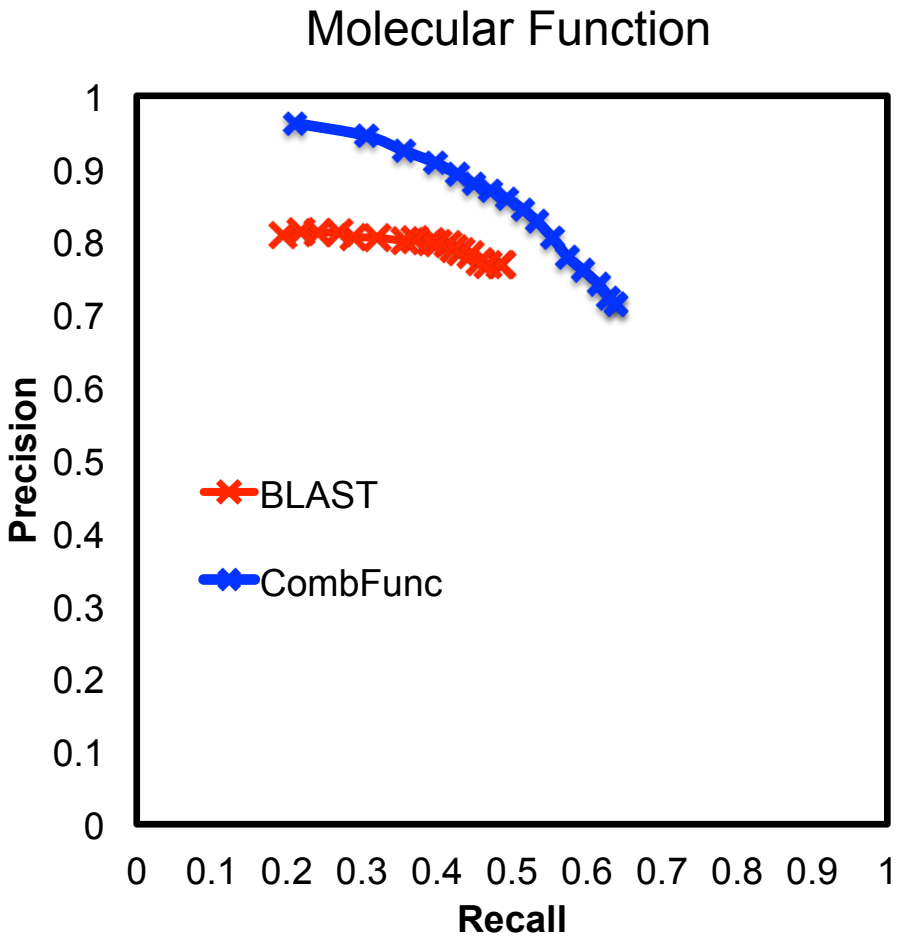
Molecular Function Predictions

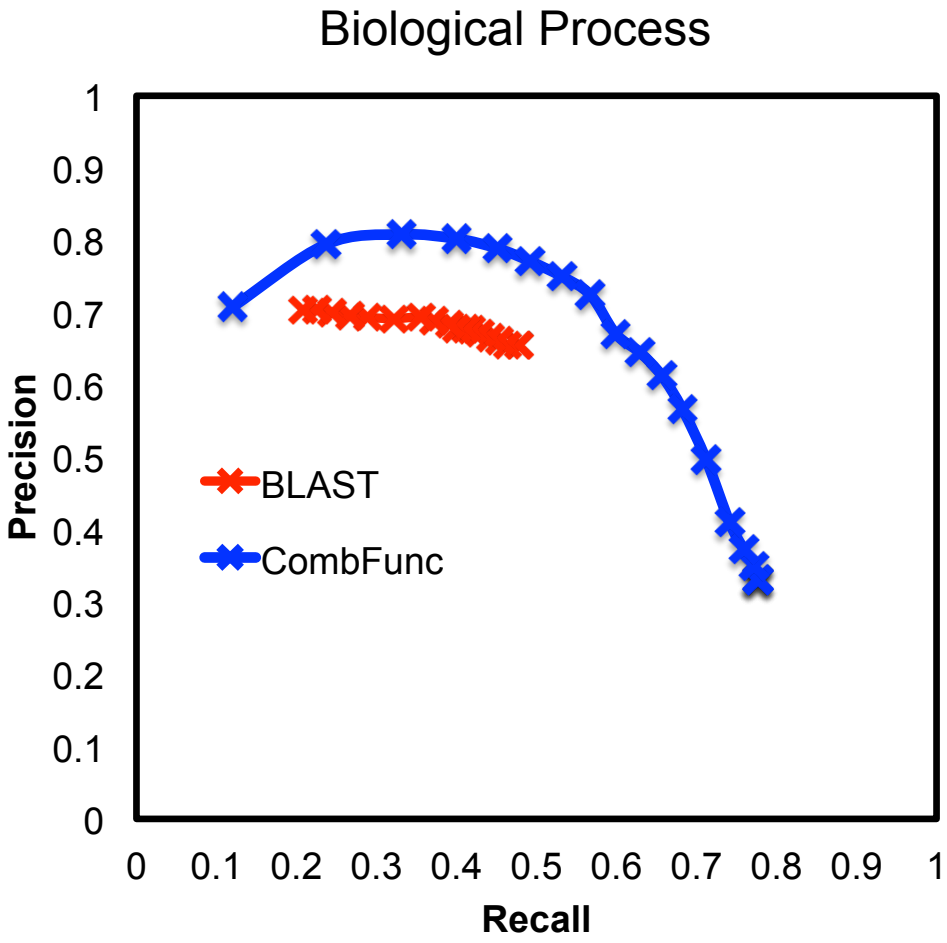
Expand All | Collapse All

- ⊟ GO:0008601 - protein phosphatase type 2A regulator activity GO
  - ⊞ GO:0019888 - protein phosphatase regulator activity GO
- ⊟ GO:0004722 - protein serine/threonine phosphatase activity GO
  - ⊟ GO:0004721 - phosphoprotein phosphatase activity GO
    - ⊟ GO:0016791 - phosphatase activity GO
      - ⊟ GO:0042578 - phosphoric ester hydrolase activity GO
        - ⊟ GO:0016788 - hydrolase activity, acting on ester bonds GO
          - ⊟ GO:0016787 - hydrolase activity GO
            - ⊟ GO:0003824 - catalytic activity GO
              - ▪ GO:0003674 - molecular_function GO
- ⊟ GO:0005515 - protein binding GO
  - ⊞ GO:0005488 - binding GO

Imperial College London

CombFunc - Benchmarking





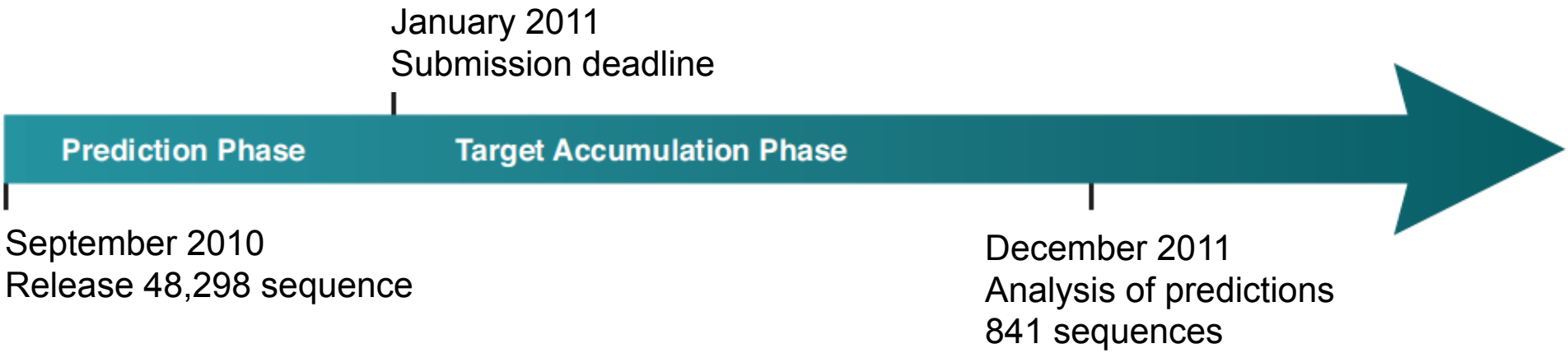
http://www.sbg.bio.ic.ac.uk/combfunc

1686 proteins

Imperial College London

CAFA - Critical Assessment of Functional Annotation

1st International Assessment of Protein Function Prediction



Sequences assessed

737 Eurkaryote

104 Prokaryote

841 Total

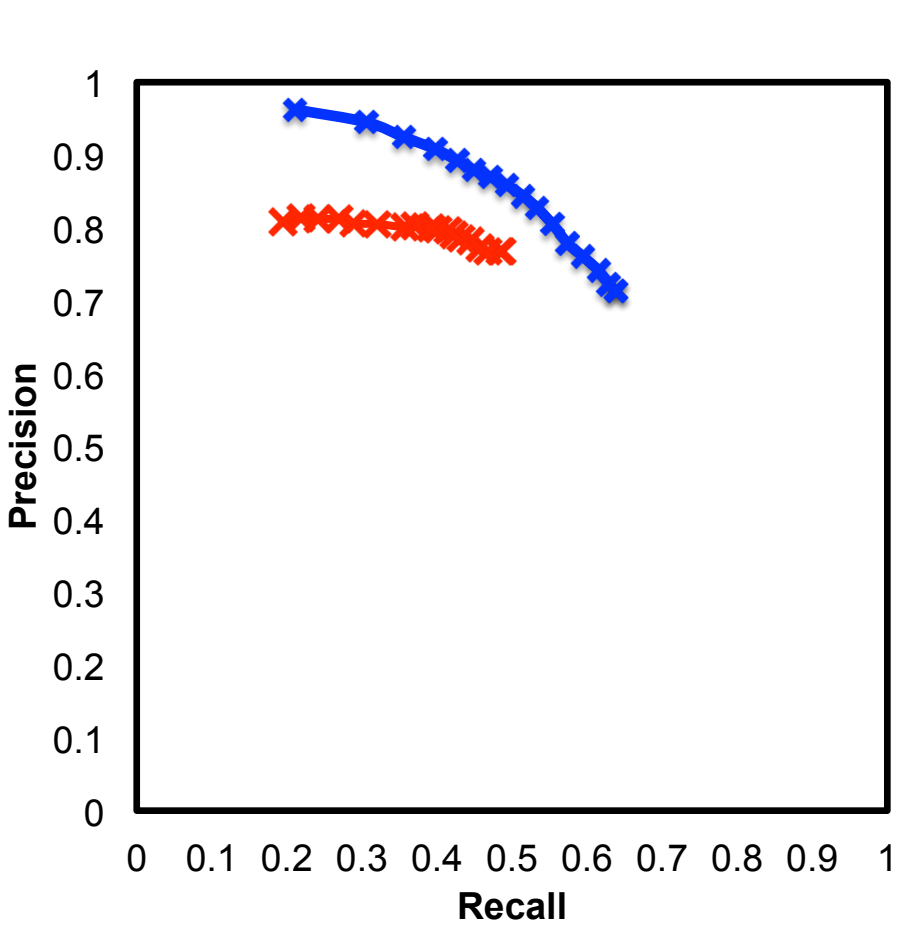
Radivojac et al., *Nature Methods* 2013

Imperial College London

CAFA – Assessing predictions

Precision Recall Graphs

P = TP/(TP+FP) – Fraction of prediction correct
R = TP/(TP+FN) – Fraction of annotations identified



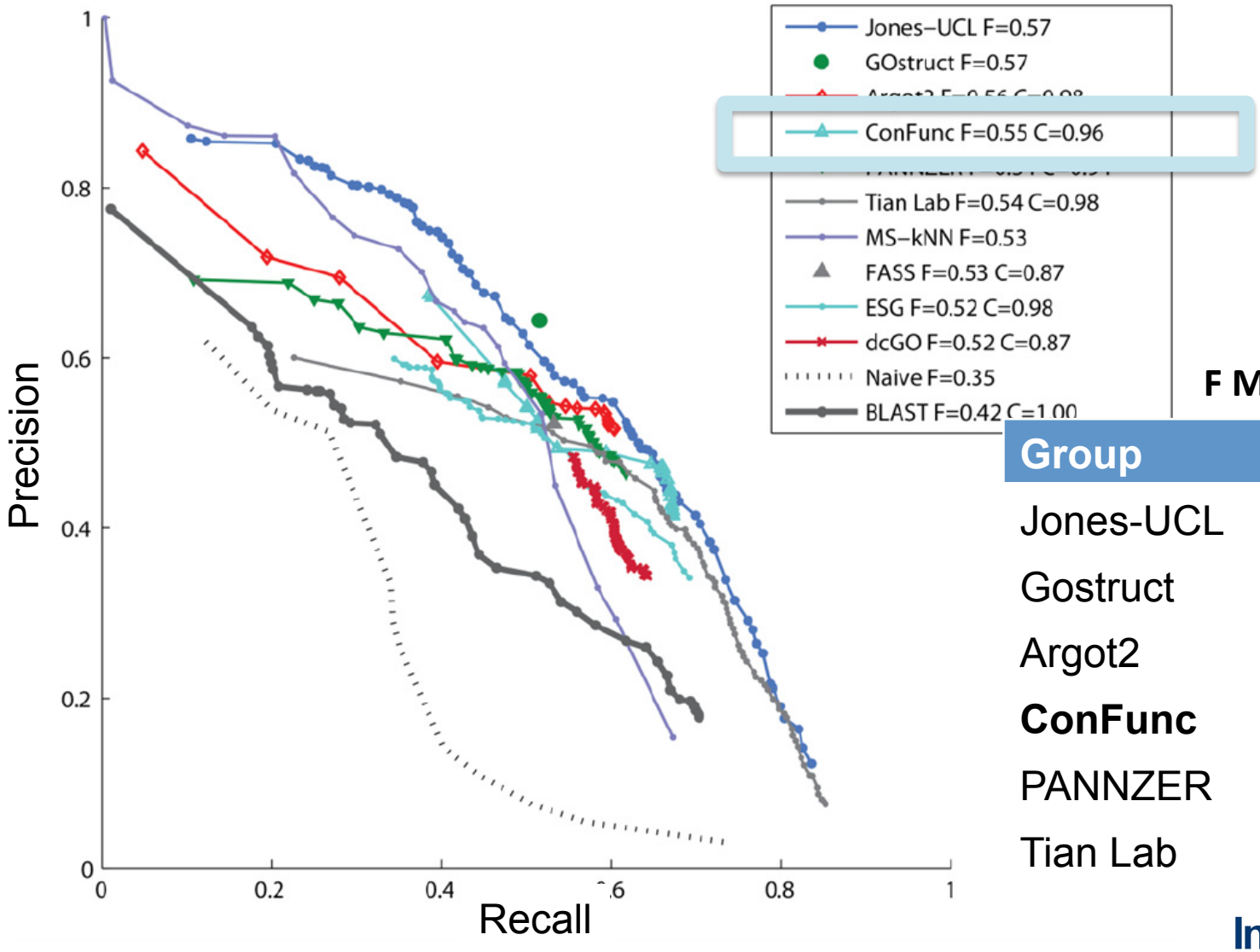
F Measure

$$F = 2 \times \frac{(P + R)}{(P \times R)}$$

They calculate the maximum F score obtained for each method

Imperial College London

CAFA – Eukaryote prediction



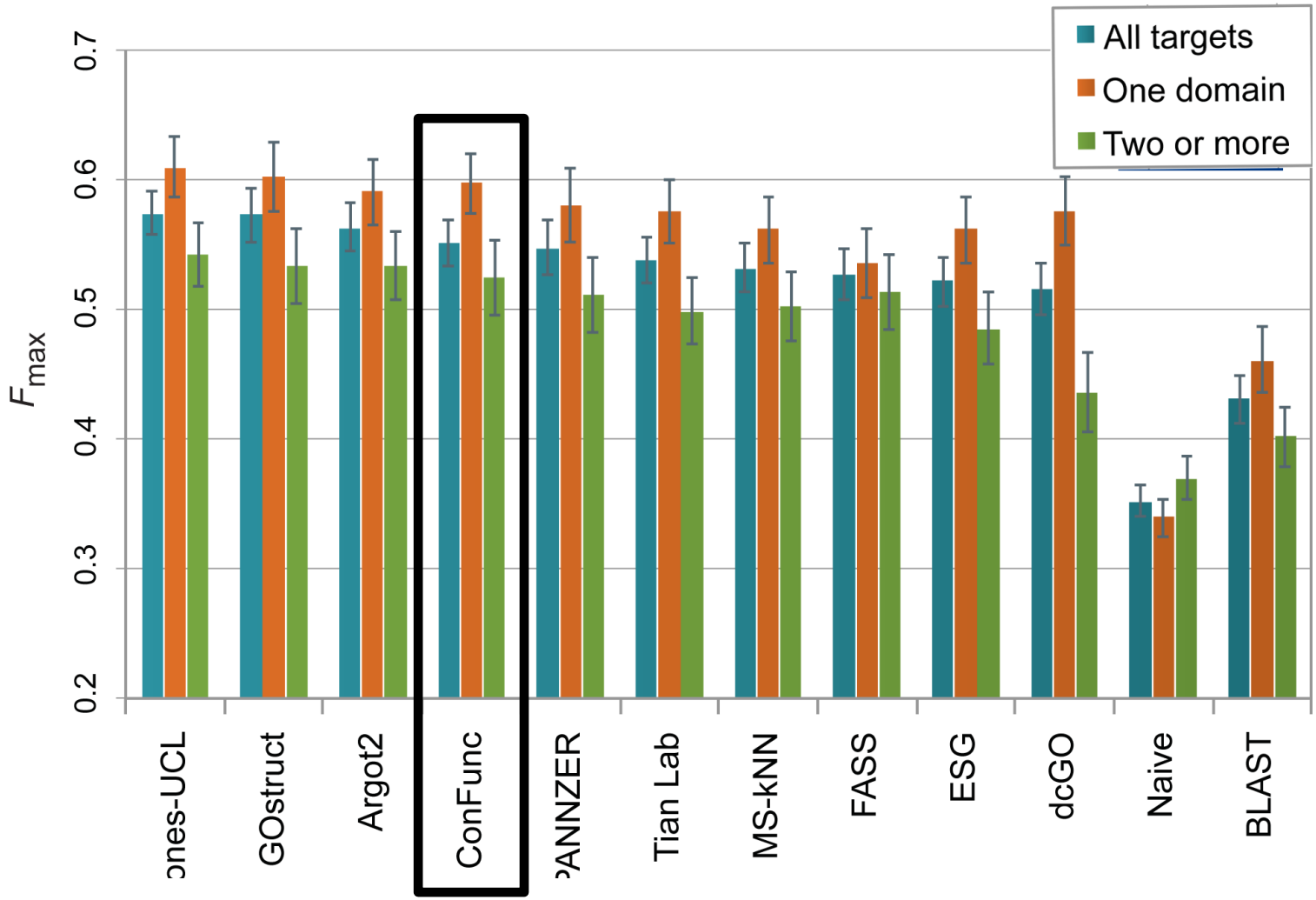
F MEASURE

Group	F Measure
Jones-UCL	0.57
Gostruct	0.57
Argot2	0.56
ConFunc	**0.55**
PANNZER	0.54
Tian Lab	0.54

Radivojac et al., *Nature Methods* 2013

Imperial College London

CAFA – Eukaryote prediction



Radivojac et al., *Nature Methods In press*

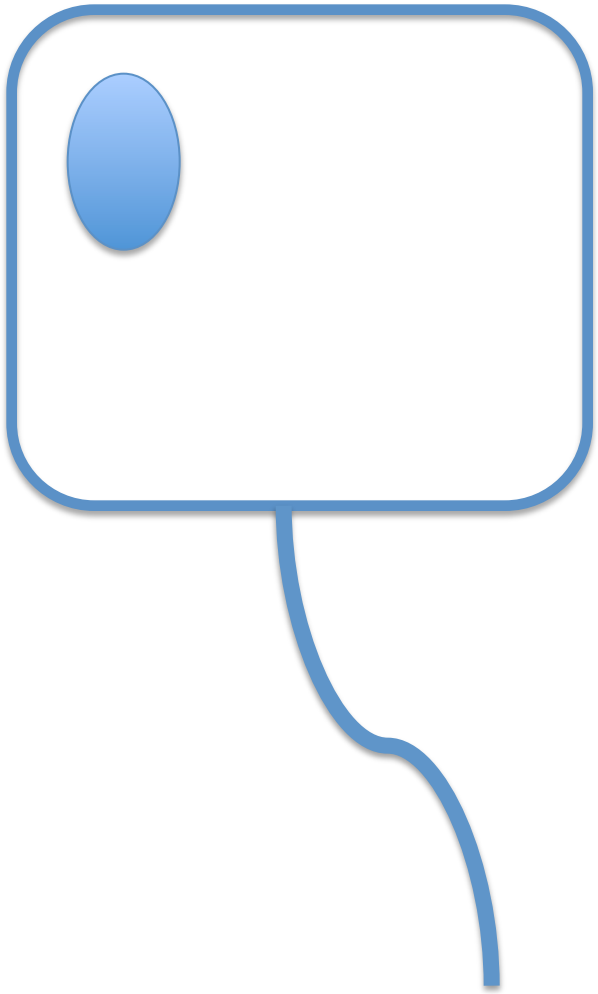
Imperial College London

Application to Plasmodium

Collaboration with: Bob Sinden, Andrew Blagborough,
Sara Marques & Arthur Talman

Imperial College
London

Plasmodium Male Gamete



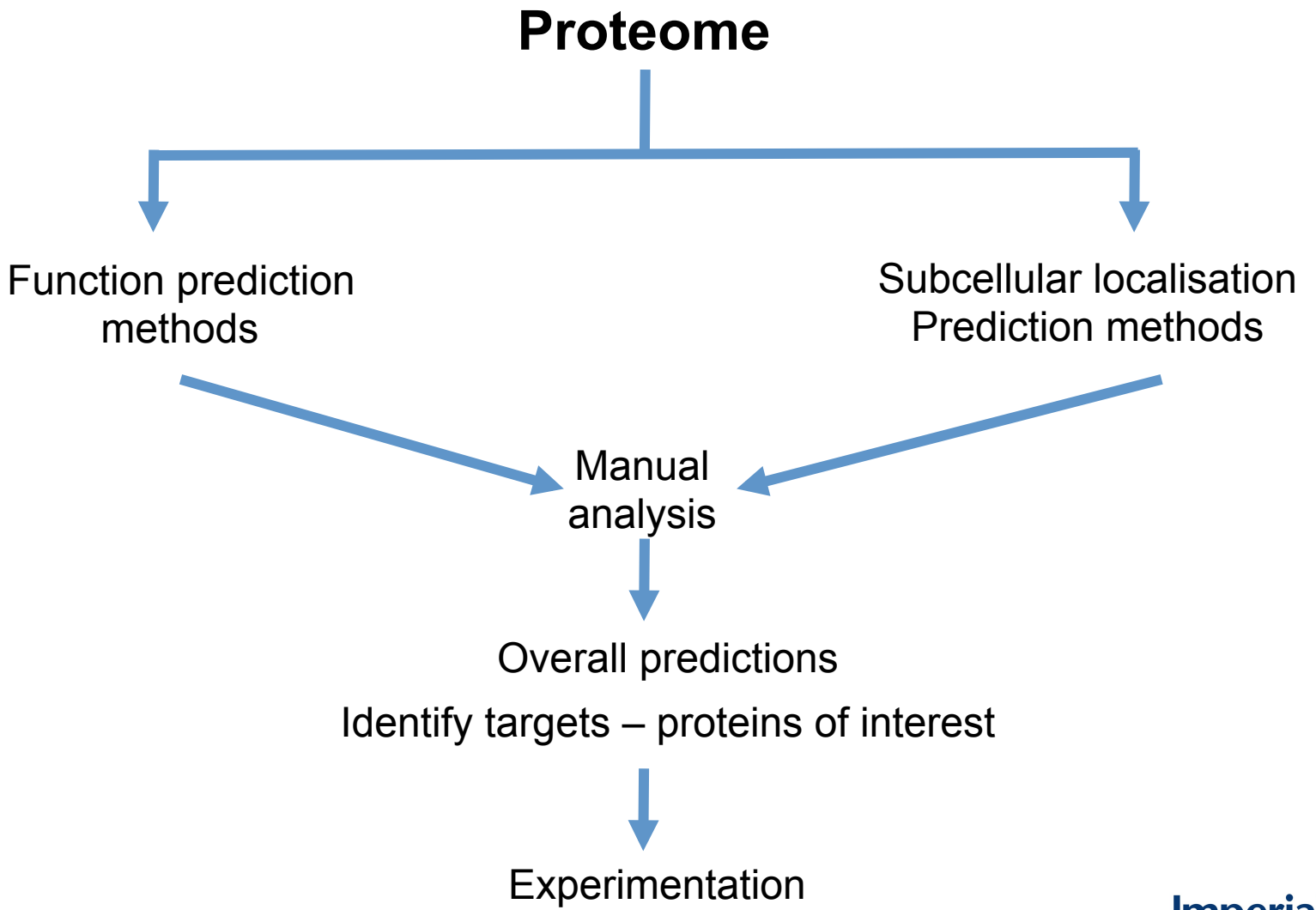
- Very simple cell – 424 proteins
- ~ 50% have unknown function
- Only 4 cellular compartments

2 Aims

- **Can we identify function/location of all proteins in gamete?**
- **Identify targets of interest for experimentation**

Imperial College London

Prediction Approach



Wass et al., 2012 Parasitology

Imperial College London

Overall Results

Focus on 182 Hypothetical Proteins

Function

Predictions for 98/182

Main Functional Processes

Location	Number
Transport	12
Signalling	11
DNA/RNA binding	11
Metabolism	4
Transcription	2

Cellular Localisation

Predictions for 152/182

Main Locations predicted

Location	Number
Extracellular	40
Membrane	34
Nucleus	27
Flagellum	18
Cytoplasm	8

Wass et al., 2012 Parasitology

Imperial College London

Identifying Targets

Transmission Blocking

Identify targets likely to be at surface or Have a direct role in fertilisation

Investigate - function/location

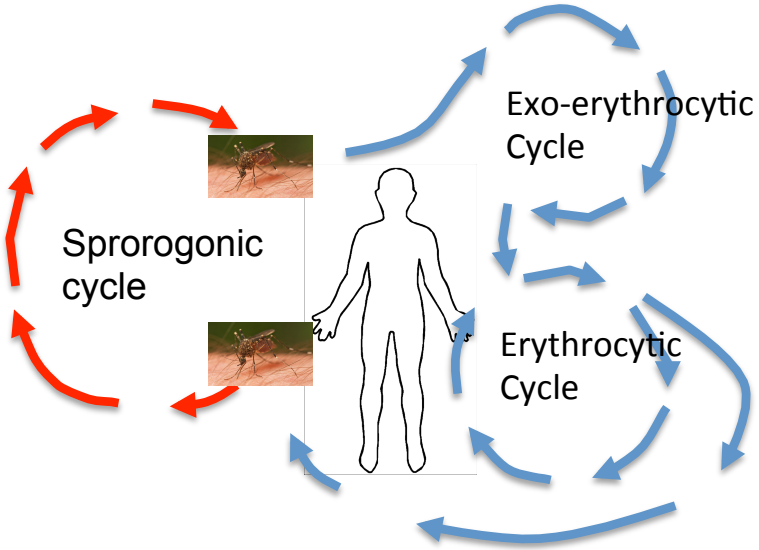
Surface Proteins

Produce protein

Immunise

Produce Antibody

Test for Transmission blocking



Imperial College London

Identifying Targets

Transmission Blocking

Identify targets likely to be at surface or Have a direct role in fertilisation

Investigate - function/location

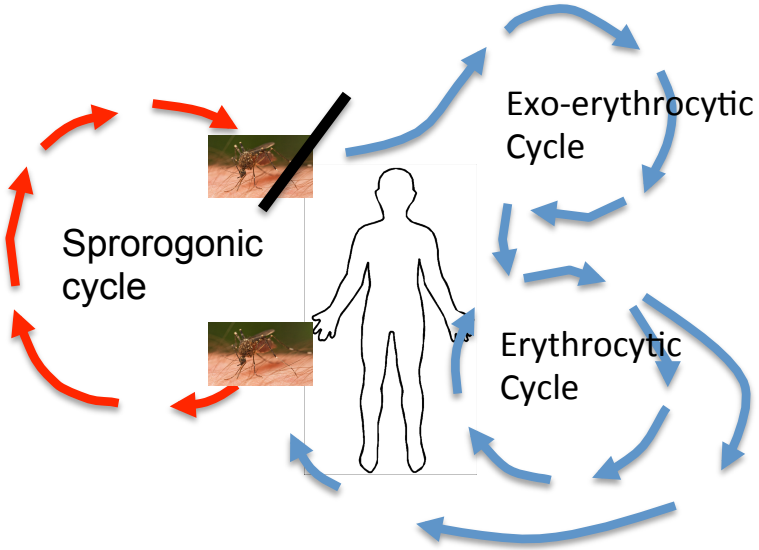
Surface Proteins

Produce protein

Immunise

Produce Antibody

Test for Transmission blocking



Imperial College London

Identifying Targets

Transmission Blocking

Identify targets likely to be at surface or Have a direct role in fertilisation

Investigate - function/location

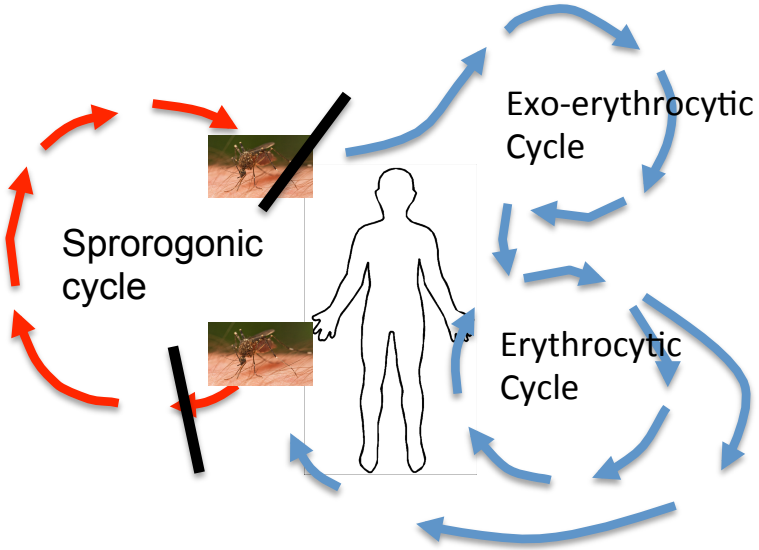
Surface Proteins

Produce protein

Immunise

Produce Antibody

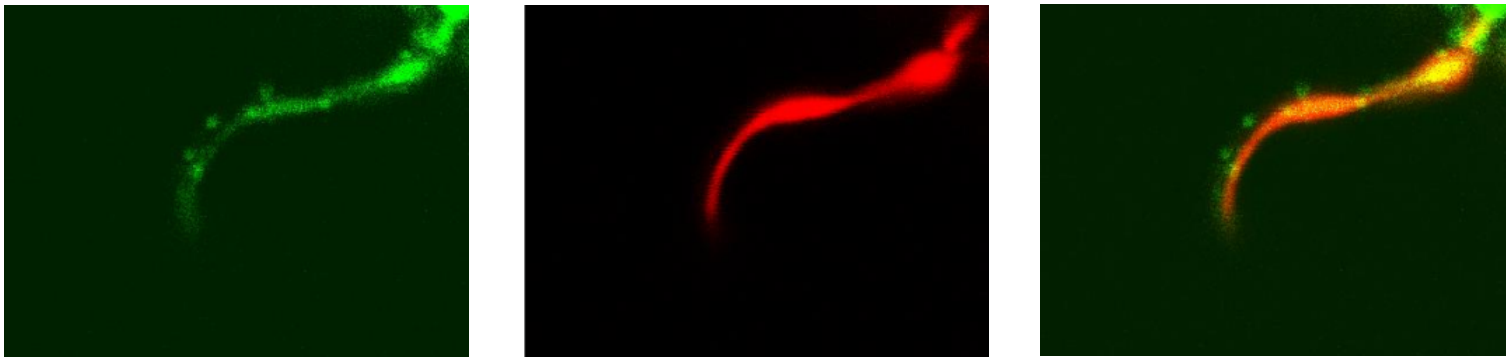
Test for Transmission blocking



Imperial College London

Identifying Targets

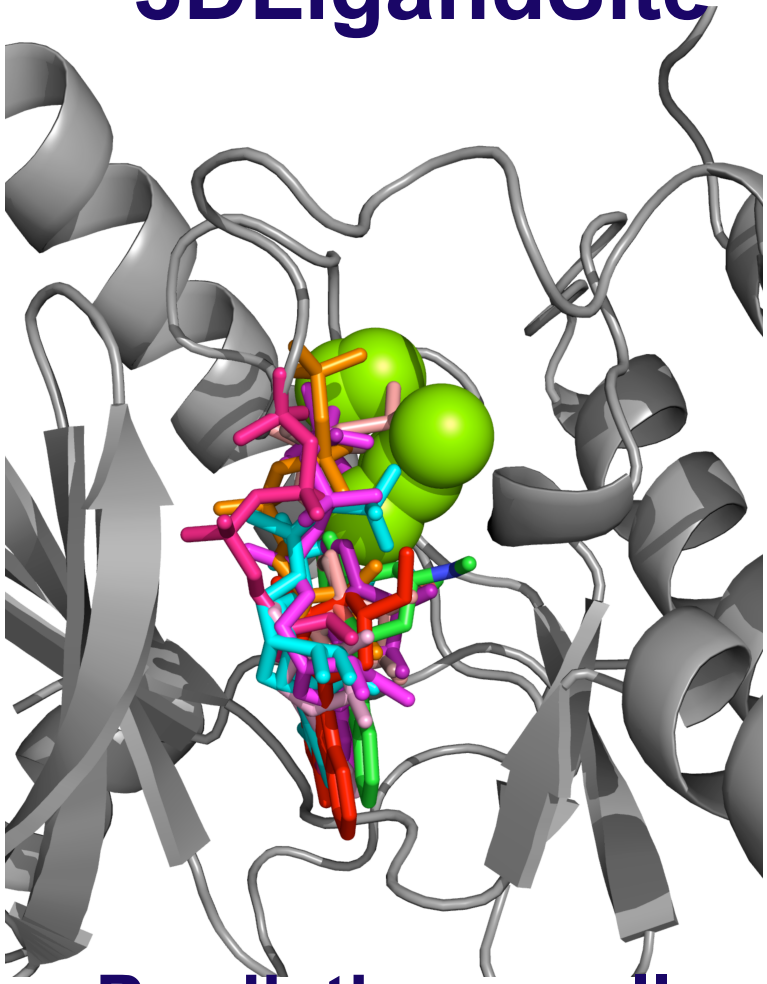
MoI of Interest **Tubulin** **Merge**



Currently examining 12 surface male gamete for transmission blocking potential.

Imperial College London

3DLigandSite



Predicting small molecule binding sites

CombFunc

GO:0003924

GO:0004722

GO:0010564

GO:0005525

GO

Prediction

GO:0007067

GO:0008601

GO:0010458

Predicting protein function using Gene Ontology

Acknowledgements

Structural bioinformatics
Michael Sternberg
Lawrence Kelley
Suhail Islam

Imperial College London

Funding:

bbsrc
biotechnology and biological sciences research council

FEBS