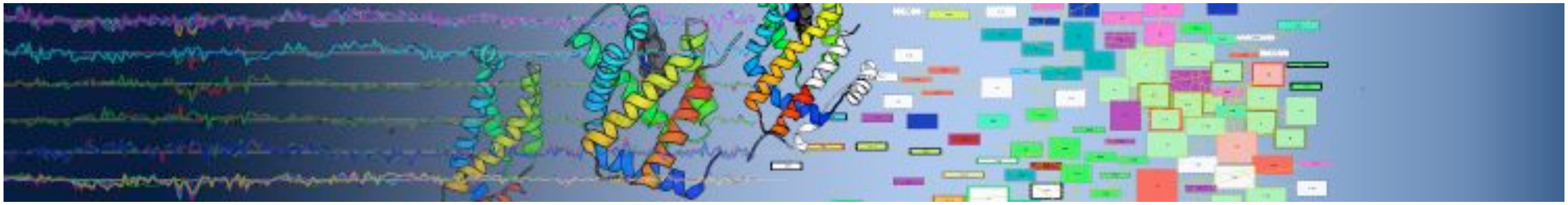
Protein Function Prediction Using Structural Homology



Kevin Drew,

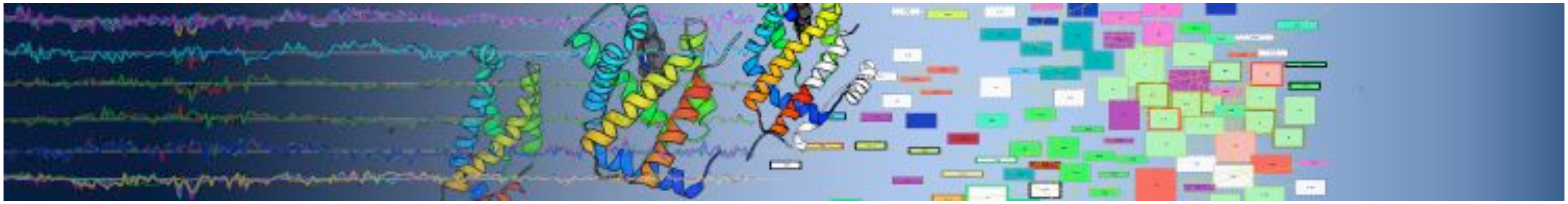
Lars Malmstroem,

Glenn Butterfoss,

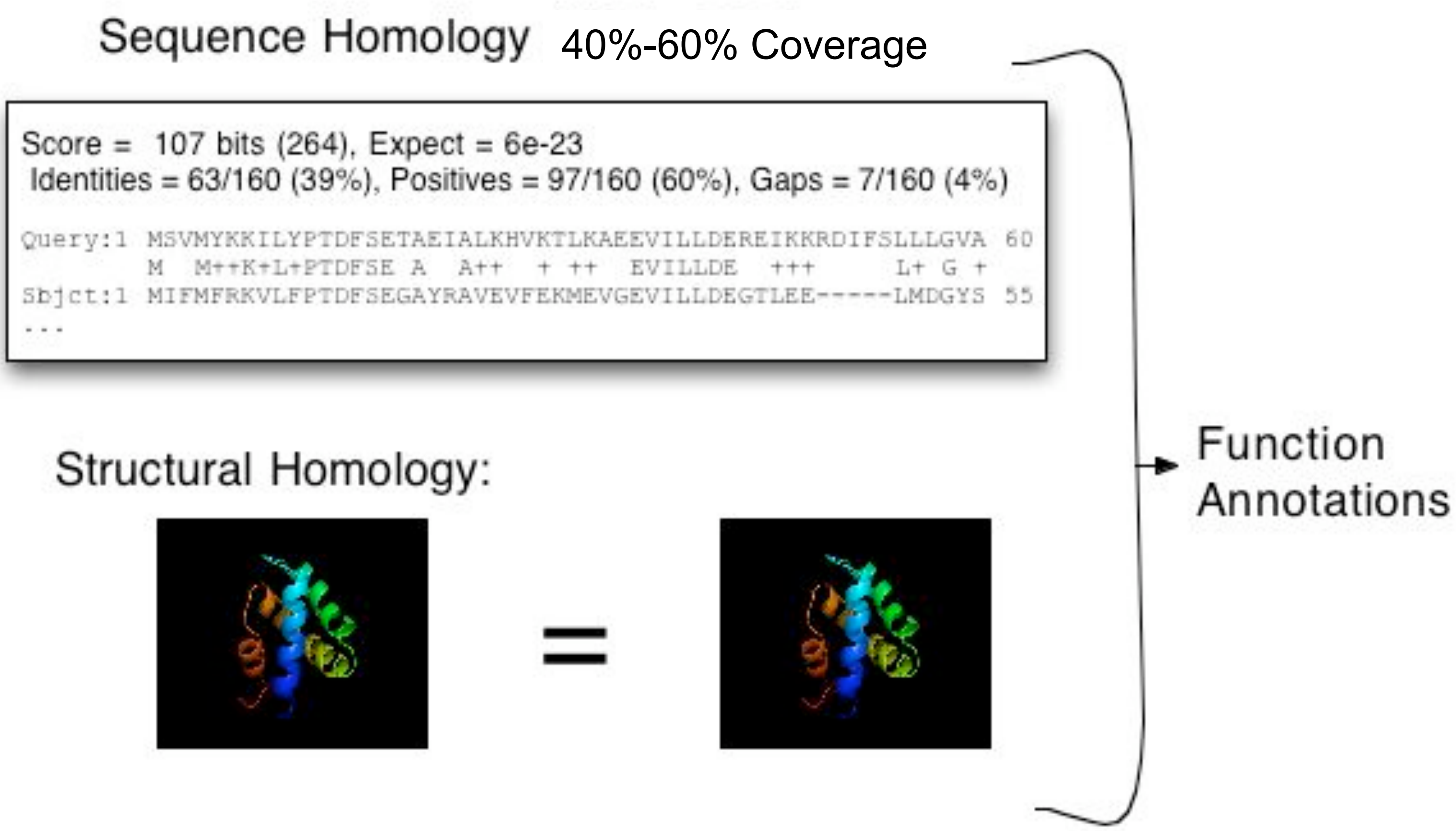
Richard Bonneau

CENTER FOR GENOMICS
AND SYSTEMS BIOLOGY
NEW YORK UNIVERSITY

COURANT
INSTITUTE



Genome Annotation using Homology



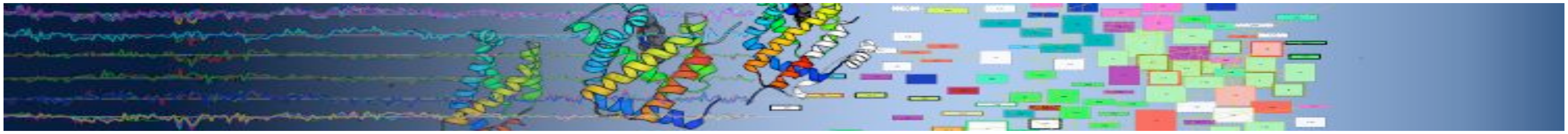
Sequence Homology 40%-60% Coverage

```
Score =  107 bits (264), Expect = 6e-23
 Identities = 63/160 (39%), Positives = 97/160 (60%), Gaps = 7/160 (4%)

Query:1 MSVMYKKILYPTDFSETAEIALKHVKTLKAEEVILLDEREIKKRDIFSLLLGVA 60
        M  M++K+L+PTDFSE A  A++  + ++  EVILLDE  +++     L+ G +
Sbjct:1 MIFMFRKVLFPTDFSEGAYRAVEVFEKMEVGEVILLDEGTLEE-----LMDGYS 55
...
```

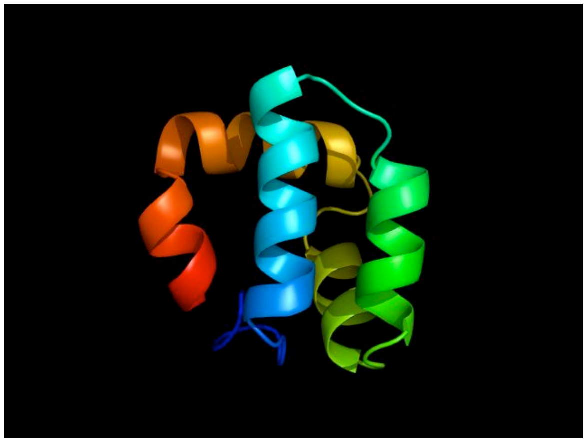
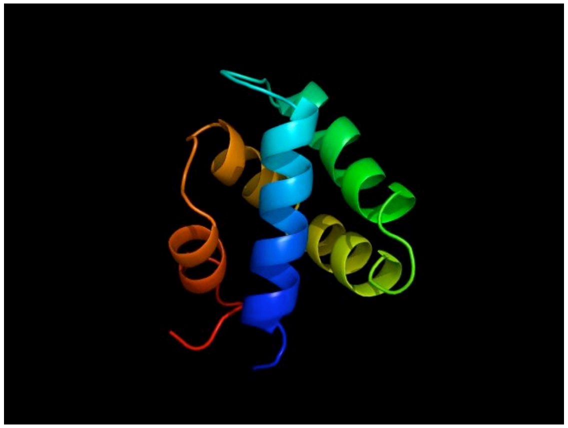
Structural Homology:

Function Annotations

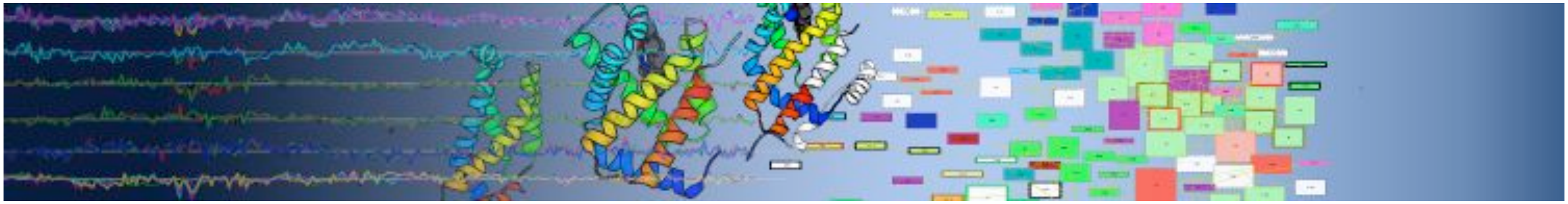


Structural Homology: Example

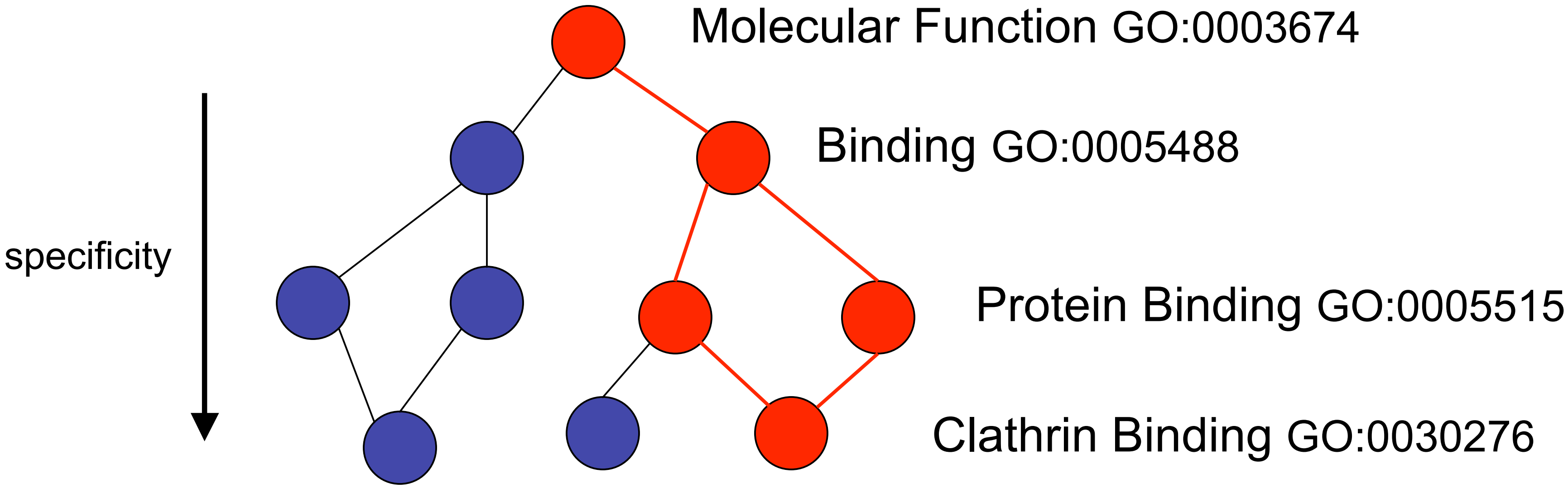
Bacteriocin AS-48, Casp 4

	1E68		1NKL
Sequence:	MAKEFGIPAAVAGTVLNVVEAGGW VTTIVSILTAVGSGGLSLLAAAGRES IKAYLKKEIKKKGKRAVIAW	4%=	GYFCESCRKIIQKLEDMVGPQPNEDTVTQAAS QVCDKLKILRGLCKKIMRSFLRRISWDILTGKKP QAICVDIKICKE
Structure:		=	
Function:	Cyclic Bacterial Lysin	=	NK Lysin

Bonneau, R., Tsai, J., Ruczinski, I., Baker, D. Functional Inferences from Blind ab Initio Protein Structure Predictions. J. Structural Biology. (2001)



Gene Ontology (GO)



Molecular Function GO:0003674

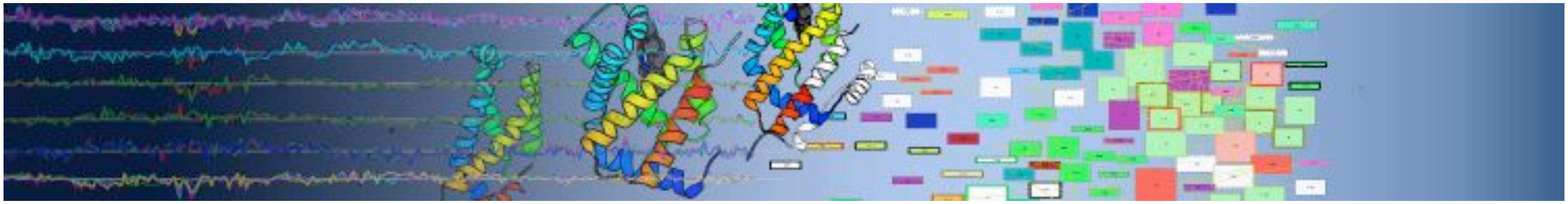
Binding GO:0005488

specificity

Protein Binding GO:0005515

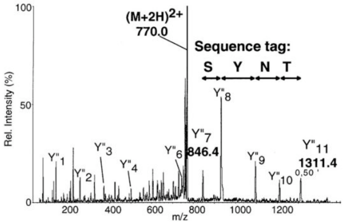
Clathrin Binding GO:0030276

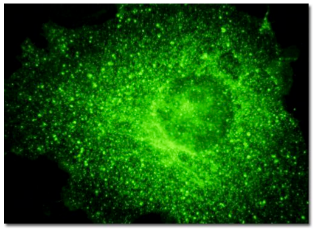
Molecular Function GO Graph



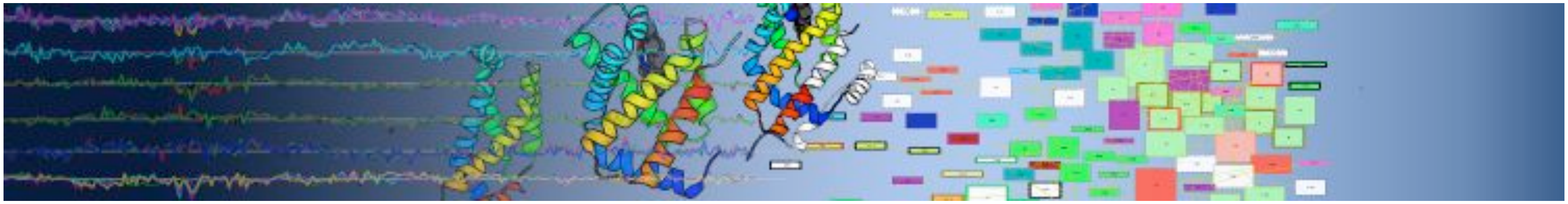
Additional Evidence of Function for Integration with Structure

- GO Biological Process
- GO Cellular Component
- Experimental Data
    - Mass Spec Pull Down
    - Fluorescent Localization

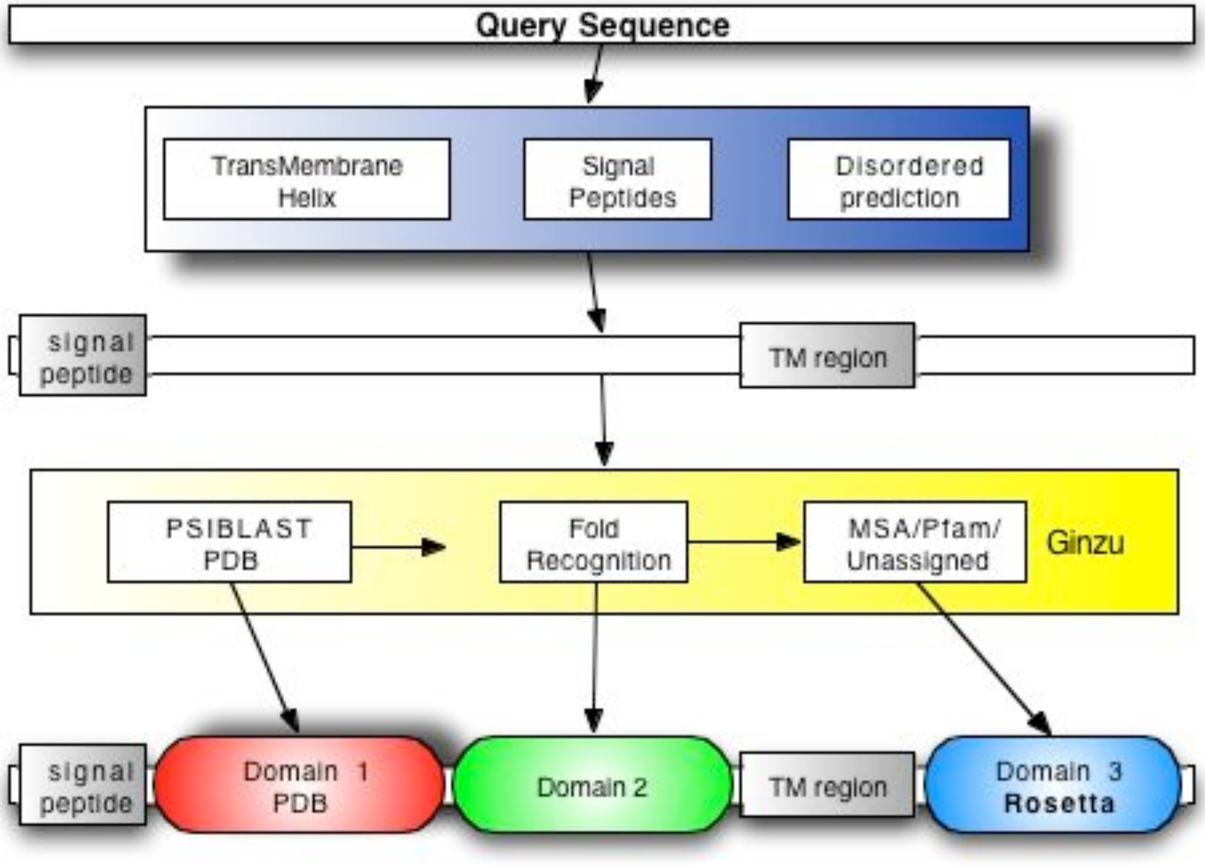


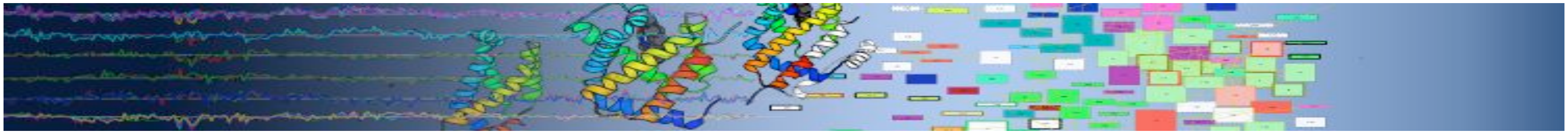


- Generally boosts confidence of predictions

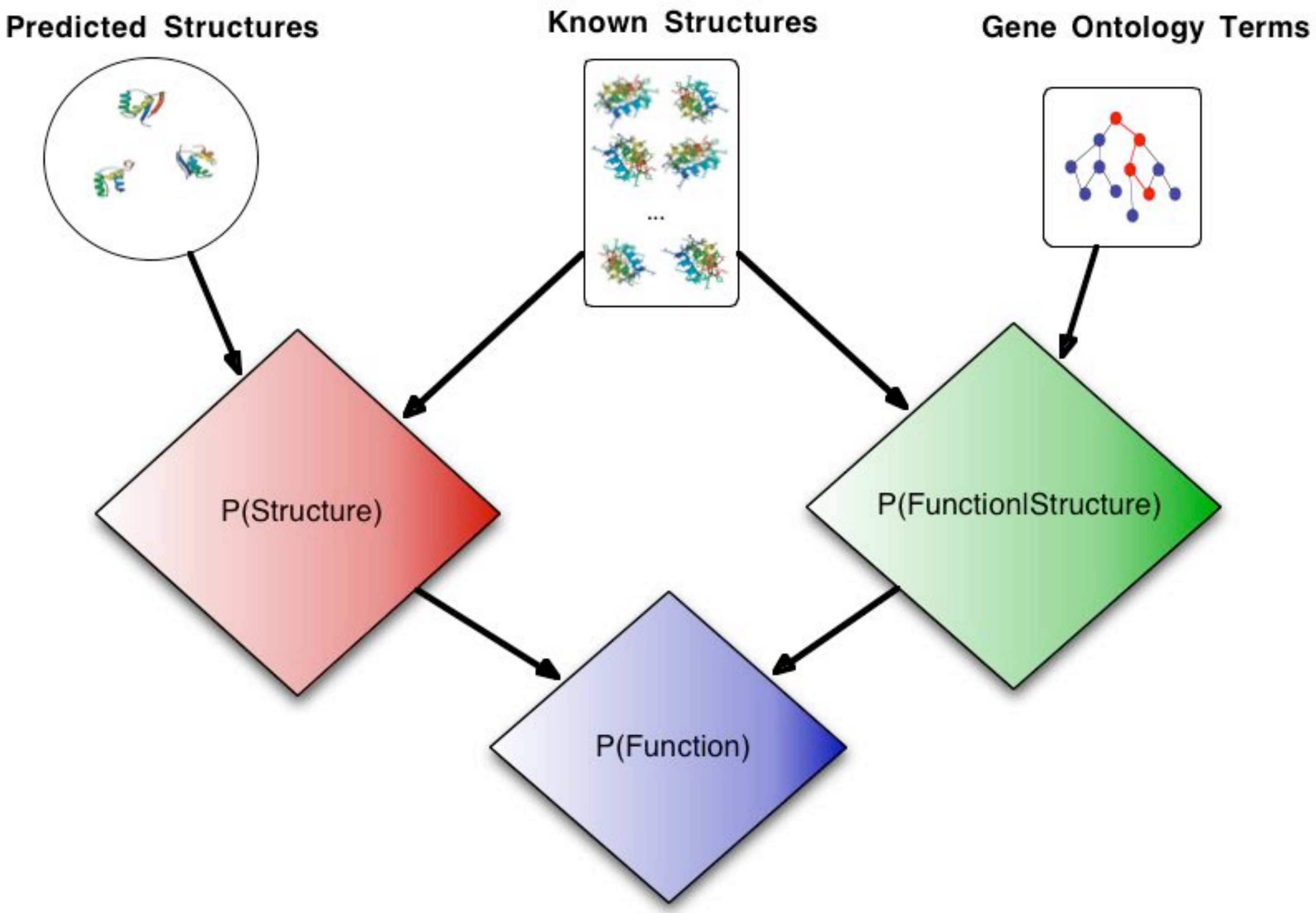


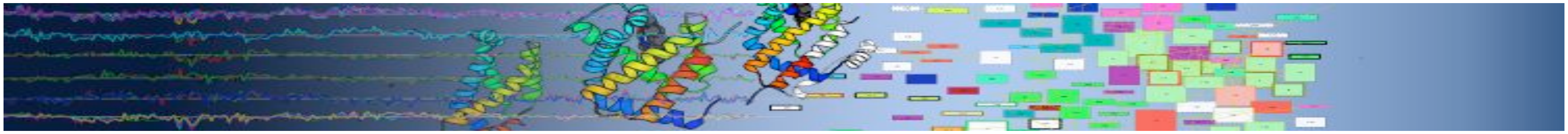
Protein Domain Prediction



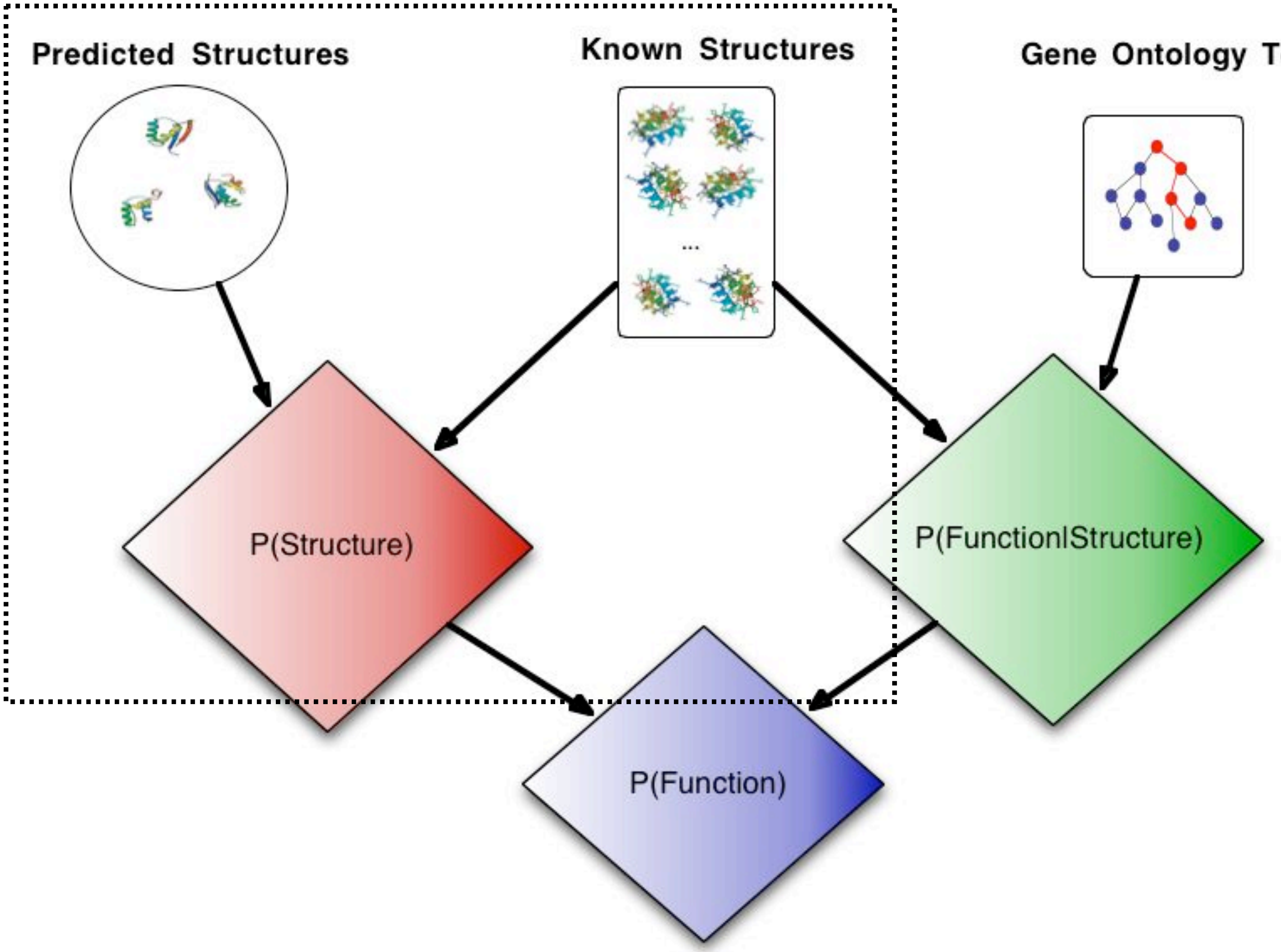


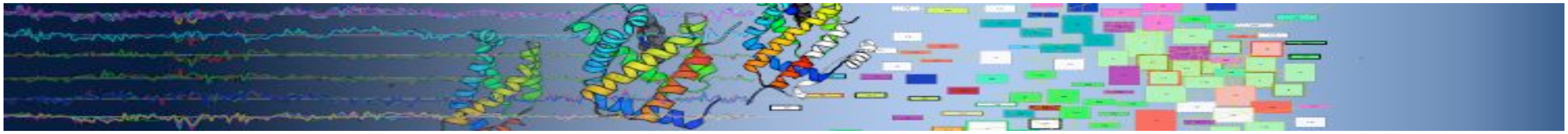
Function Prediction Overview



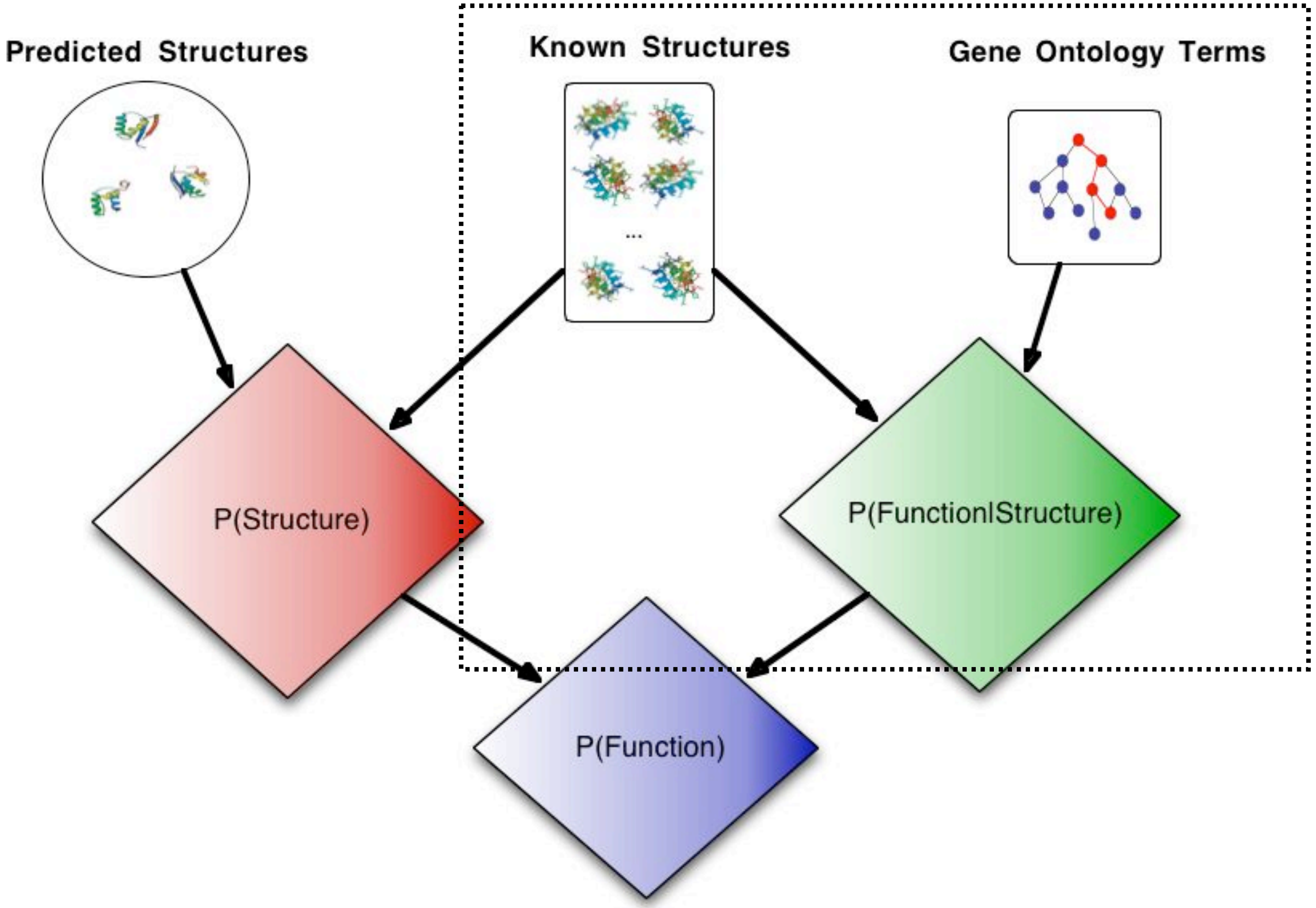


Function Prediction Overview





Function Prediction Overview



Predicted Structures

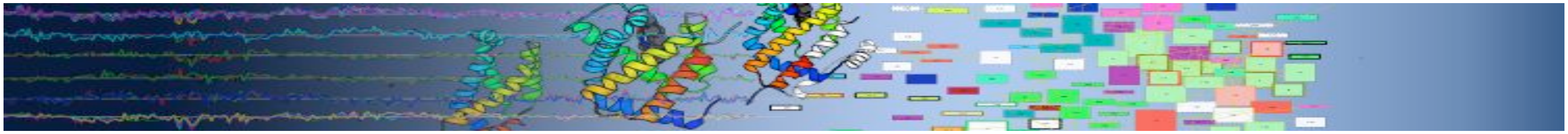
Known Structures

Gene Ontology Terms

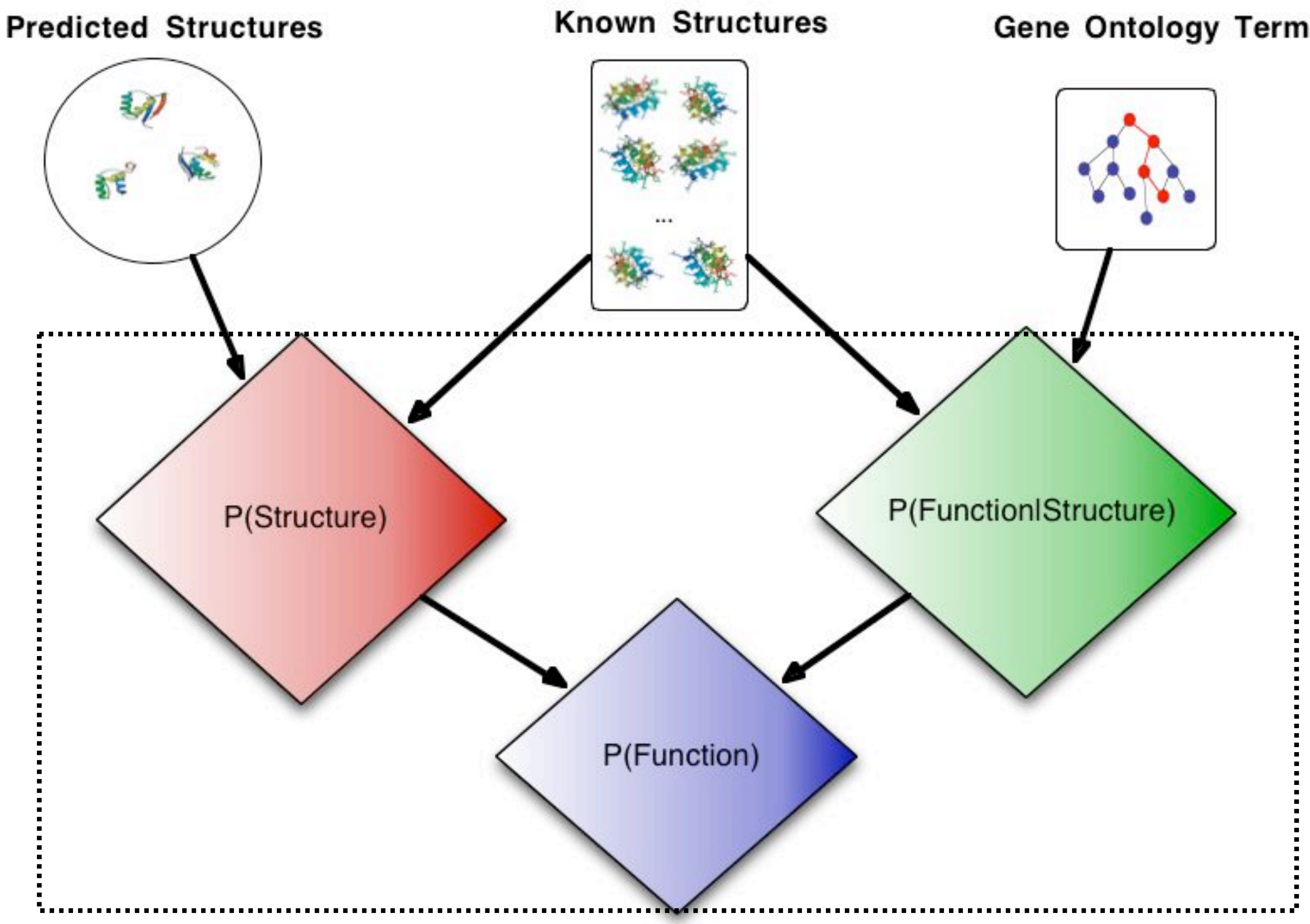
P(Structure)

P(Function|Structure)

P(Function)



Function Prediction Overview



Predicted Structures

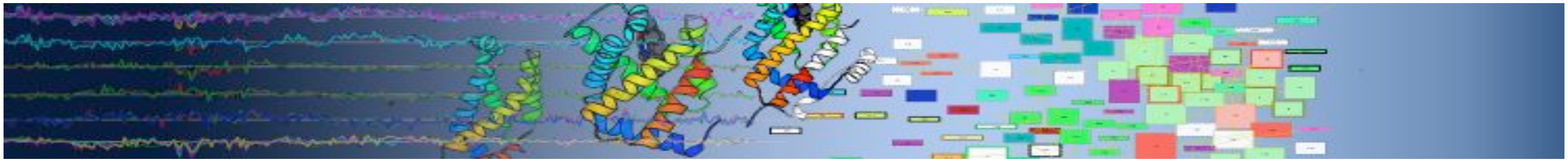
Known Structures

Gene Ontology Terms

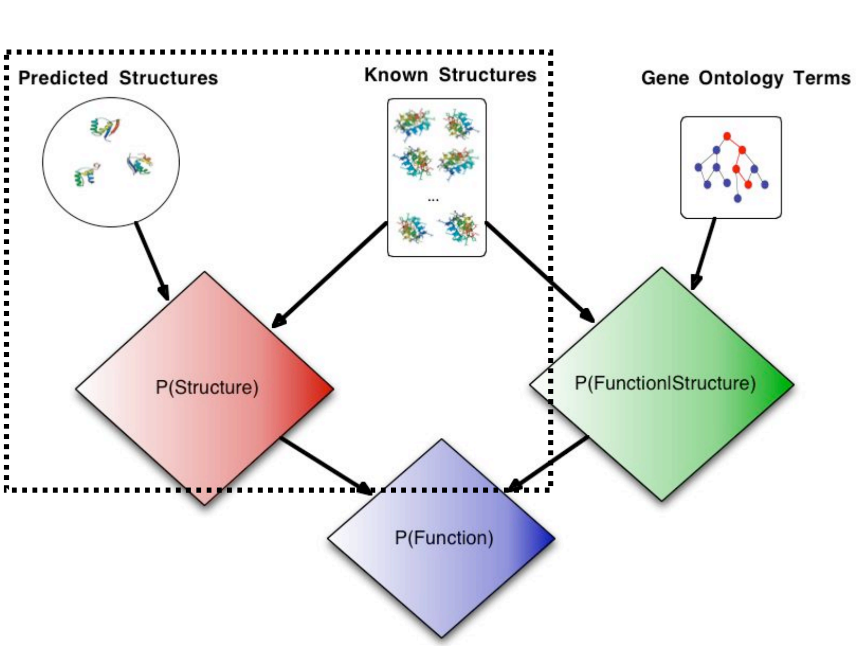
P(Structure)

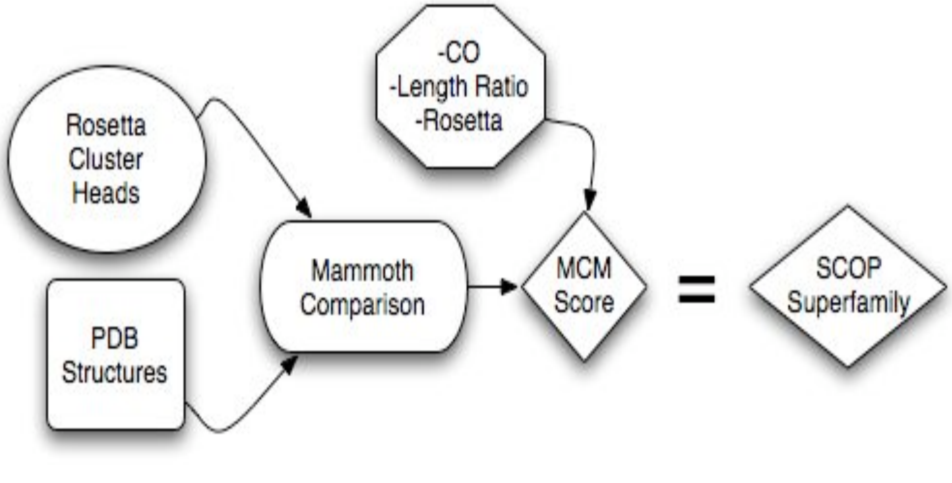
P(Function|Structure)

P(Function)



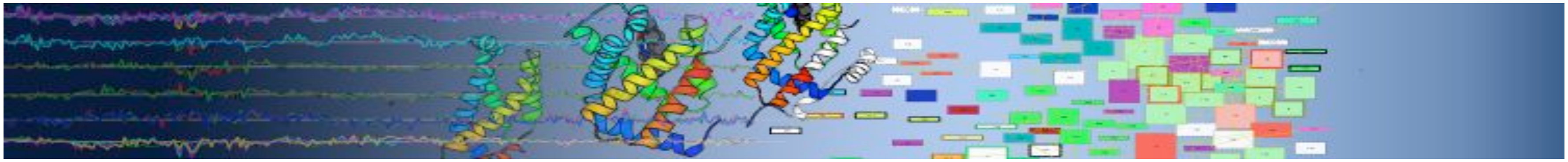
Matching Predicted Structures to Known Structures



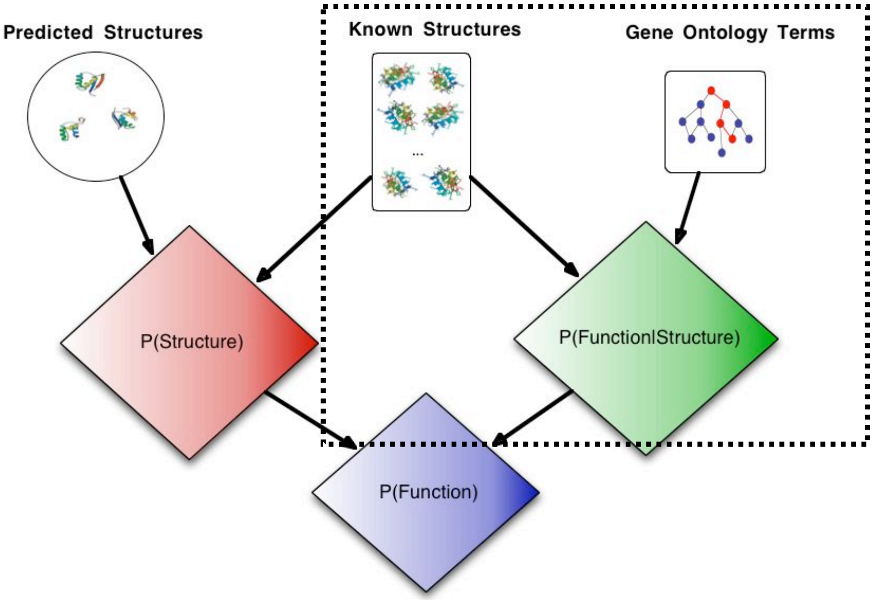


$$\log\left(\frac{P_{MCM}}{1-P_{MCM}}\right) = a\cdot zscore + b\cdot CO + c\cdot converg + d\cdot\left|\log\left(\frac{L_{Astral}}{L_{predicted}}\right)\right| + C$$

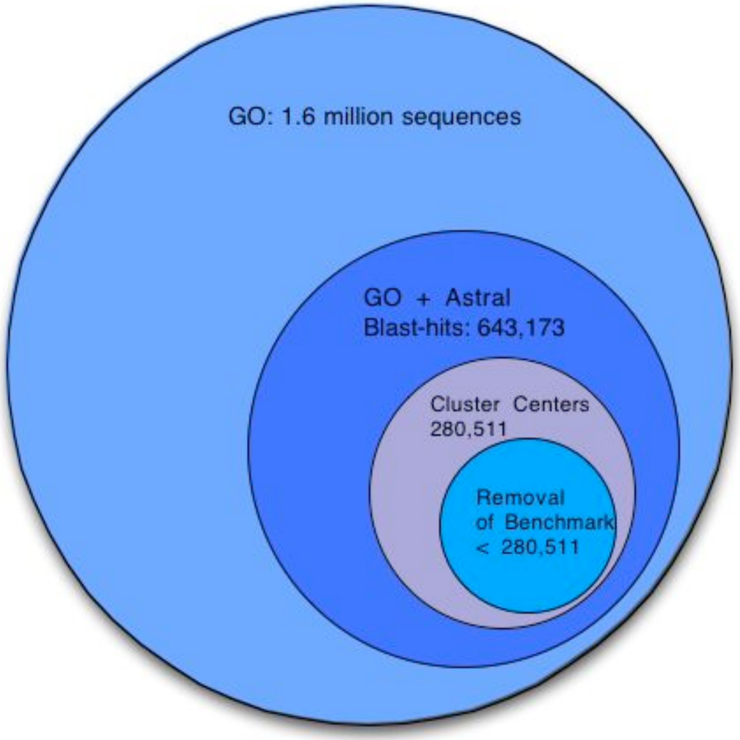
$$P(Structure) = 1 - \prod_{k=1}^{n}(1-p_k)$$

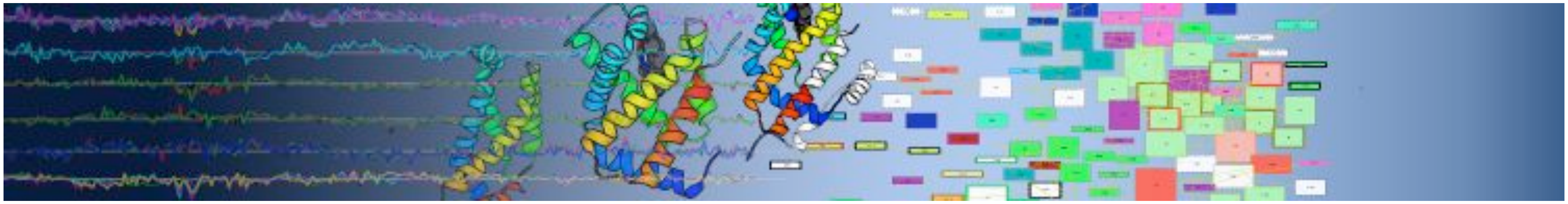


Training Data Derived from GO and Known Structures



Training Set Reduction



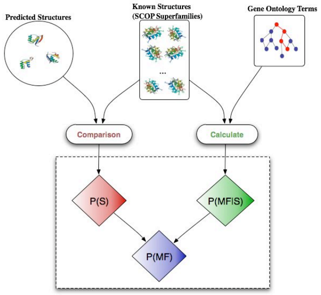


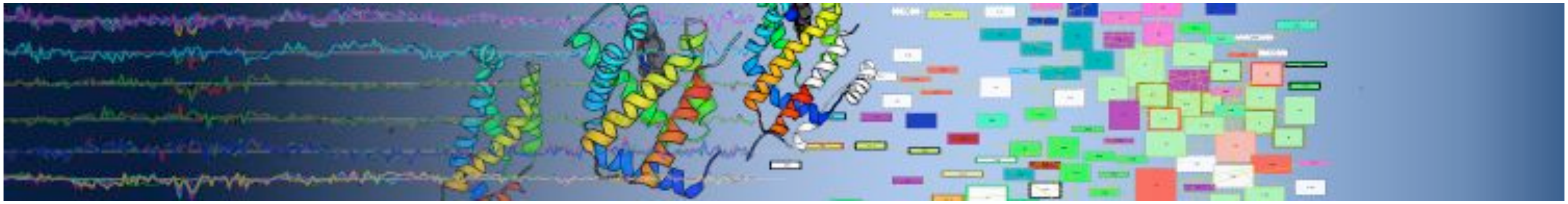
Naïve Bayes

- In words: what is the probability of a variable, y, is true given features, **x**, over the probability y is false given the features **x**.
  - Take the log and if its >0 its more likely to be true than false.

- y = molecular function and **x** = {sf, bp, cc}

$$LL_X = log\left(\frac{P(y = TRUE)}{P(y = FALSE)}\right) + \sum_{j=1}^{d} log\left(\frac{P(x_j|y = TRUE)}{P(x_j|y = FALSE)}\right)$$



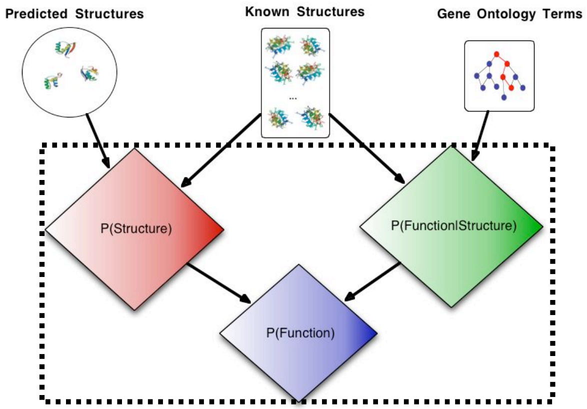


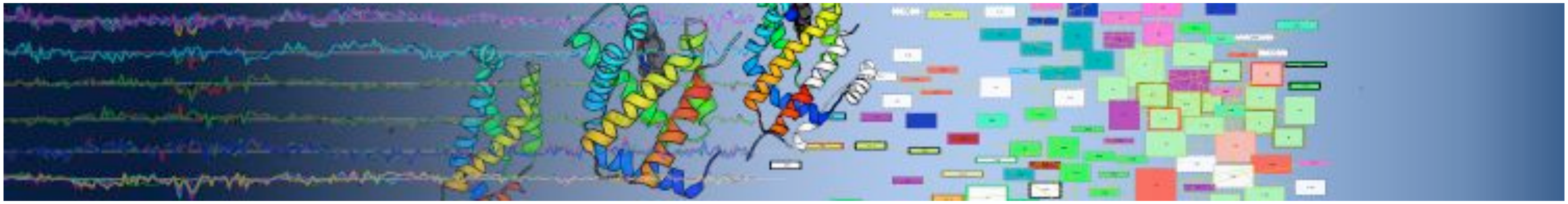
Full Function Prediction Formula

$$LL_X = log\Big(\frac{P(y = TRUE)}{P(y = FALSE)}\Big) + \sum_{j=1}^{d} log\Big(\frac{P(x_j|y = TRUE)}{P(x_j|y = FALSE)}\Big)$$

Naive Bayes

$$LL_{PLS} = log\Big(\frac{P(Function)}{P(\overline{Function})}\Big) + \sum_{i=1}^{N}\Big[P(Structure_i) * log\Big(\frac{P(Structure_i|Function)}{P(Structure_i|\overline{Function})}\Big)\Big] + \sum_{j=P,L} log\Big(\frac{P(x_j|Function)}{P(x_j|\overline{Function})}\Big)$$





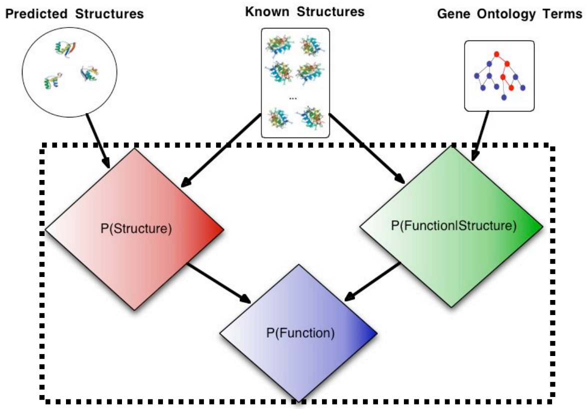
Full Function Prediction Formula

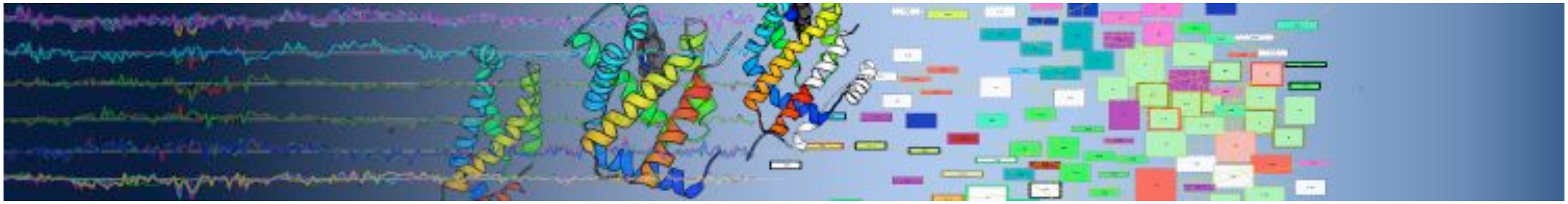
$$LL_X = log\Big(\frac{P(y = TRUE)}{P(y = FALSE)}\Big) + \sum_{j=1}^{d} log\Big(\frac{P(x_j|y = TRUE)}{P(x_j|y = FALSE)}\Big)$$

Naive Bayes

Structure Contribution

$$LL_{PLS} = log\Big(\frac{P(Function)}{P(Fun\bar{c}tion)}\Big) + \sum_{i=1}^{N} \Big[P(Structure_i) * log\Big(\frac{P(Structure_i|Function)}{P(Structure_i|Fun\bar{c}tion)}\Big)\Big] + \sum_{j=P,L} log\Big(\frac{P(x_j|Function)}{P(x_j|Fun\bar{c}tion)}\Big)$$



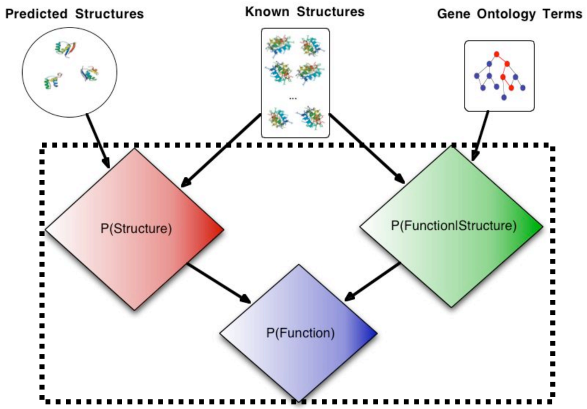


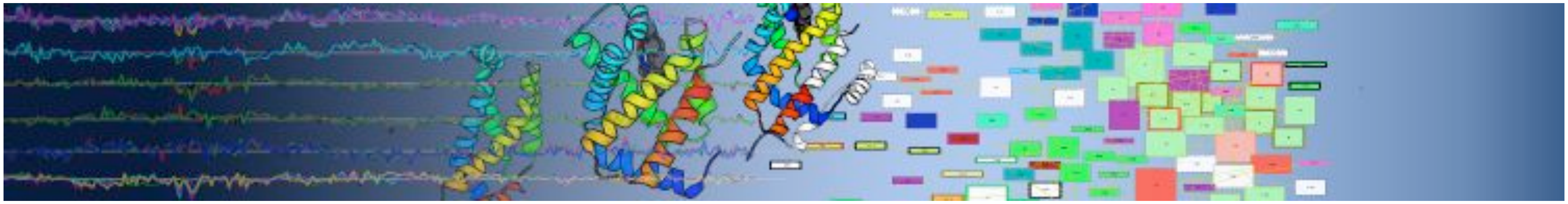
Full Function Prediction Formula

$$LL_X = log\left(\frac{P(y = TRUE)}{P(y = FALSE)}\right) + \sum_{j=1}^{d} log\left(\frac{P(x_j|y = TRUE)}{P(x_j|y = FALSE)}\right)$$

Naive Bayes

$$LL_{PLS} = log\left(\frac{P(Function)}{P(\overline{Function})}\right) + \sum_{i=1}^{N} \left[\overset{\text{P(Structure)}}{\boxed{P(Structure_i) *}}\ \overset{\text{P(Function|Structure)}}{\boxed{log\left(\frac{P(Structure_i|Function)}{P(Structure_i|\overline{Function})}\right)\right]}}} + \sum_{j=P,L} log\left(\frac{P(x_j|Function)}{P(x_j|\overline{Function})}\right)$$





Full Function Prediction Formula

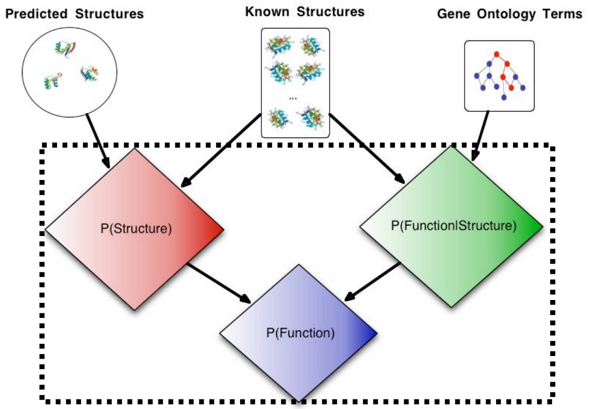
$$LL_X = log\Big(\frac{P(y = TRUE)}{P(y = FALSE)}\Big) + \sum_{j=1}^{d} log\Big(\frac{P(x_j|y = TRUE)}{P(x_j|y = FALSE)}\Big)$$

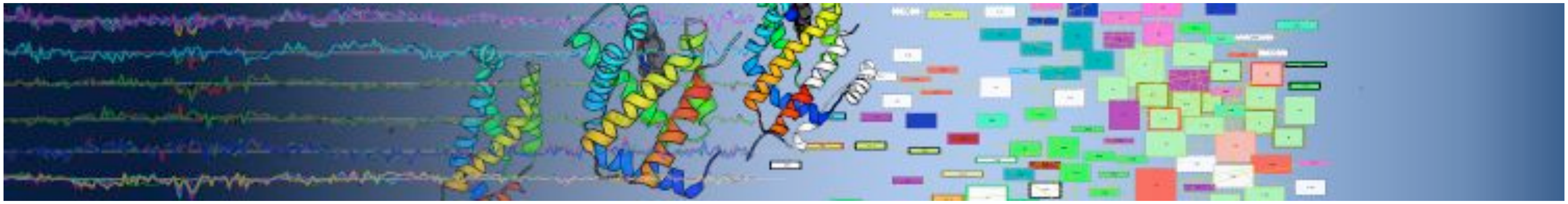
Naive Bayes

Prior

Additional Evidence

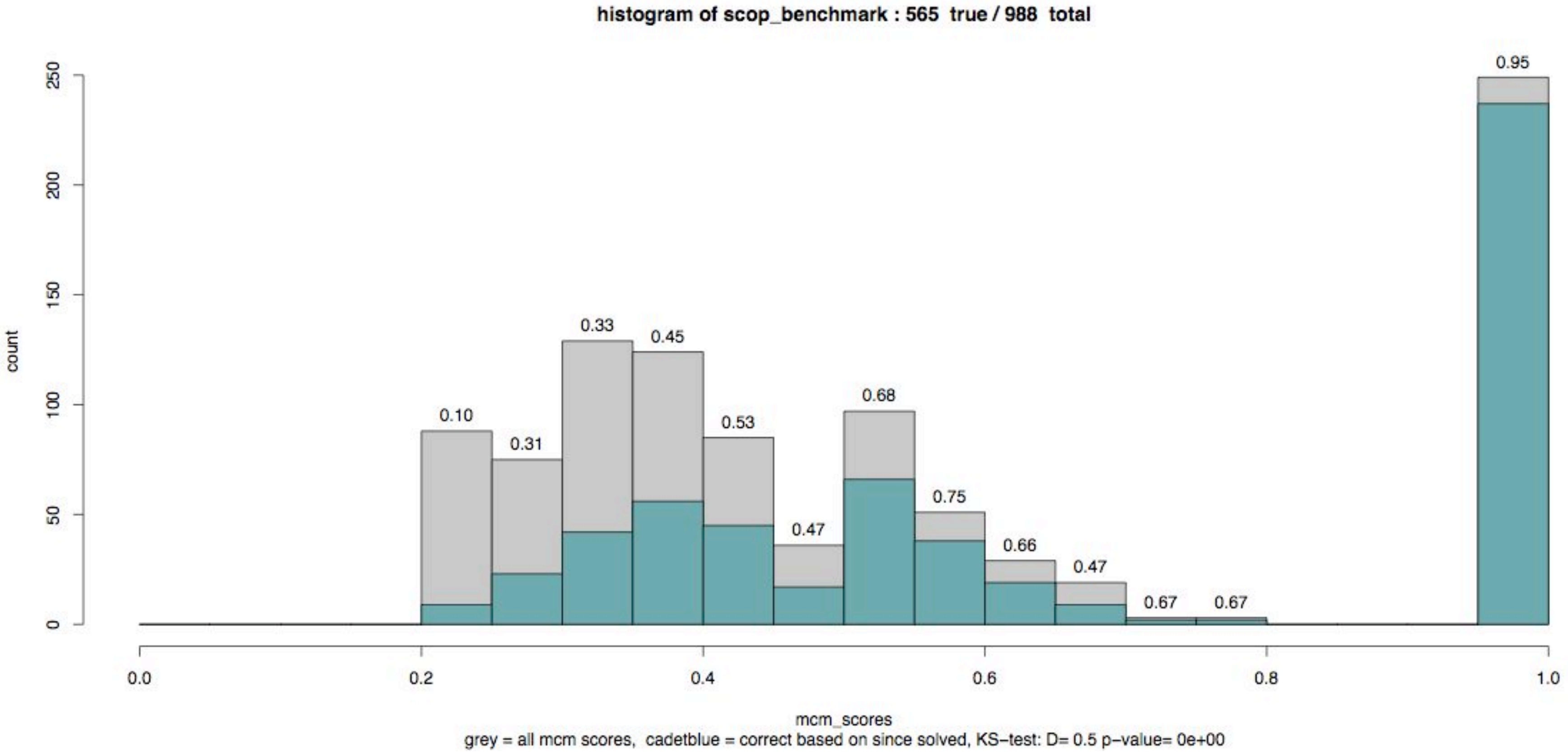
$$LL_{PLS} = \boxed{log\Big(\frac{P(Function)}{P(\bar{Function})}\Big)} + \sum_{i=1}^{N} \Big[P(Structure_i) * log\Big(\frac{P(Structure_i|Function)}{P(Structure_i|\bar{Function})}\Big)\Big] + \boxed{\sum_{j=P,L} log\Big(\frac{P(x_j|Function)}{P(x_j|\bar{Function})}\Big)}$$

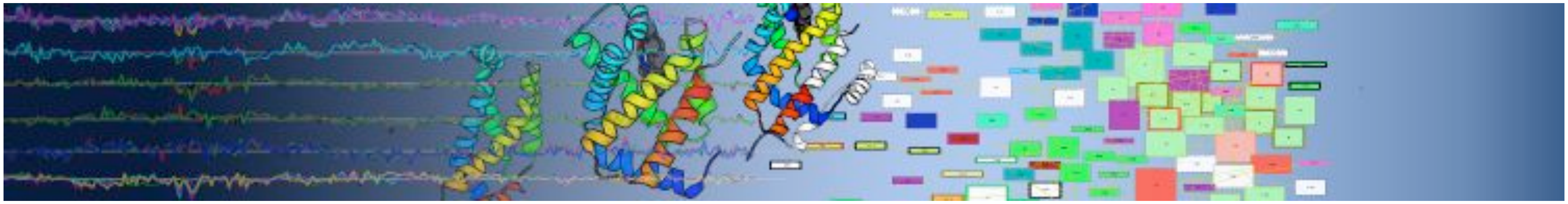




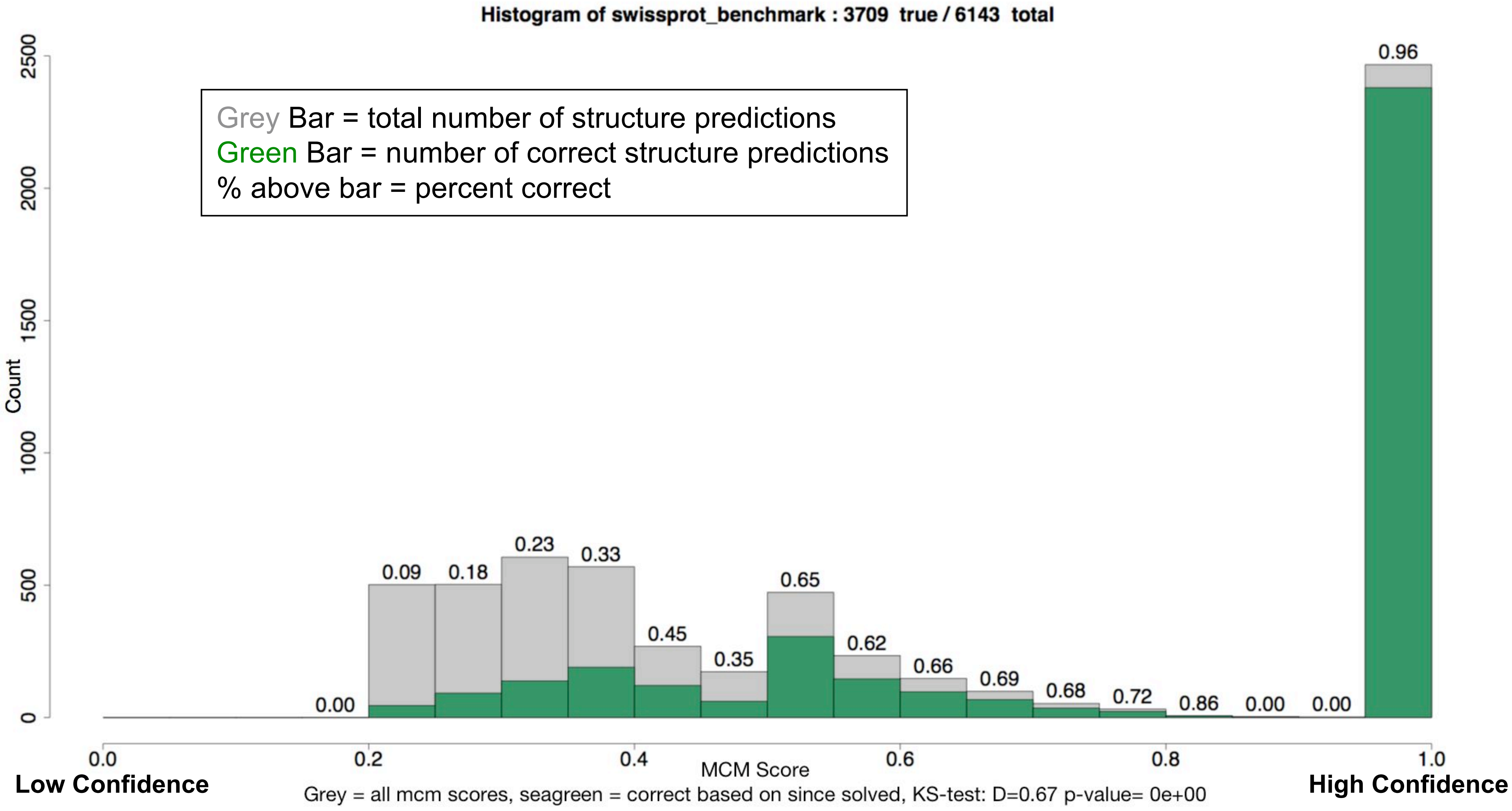
Results: Solved Structures

How accurate are we when we predict SCOP Superfamily for PDB Structures?



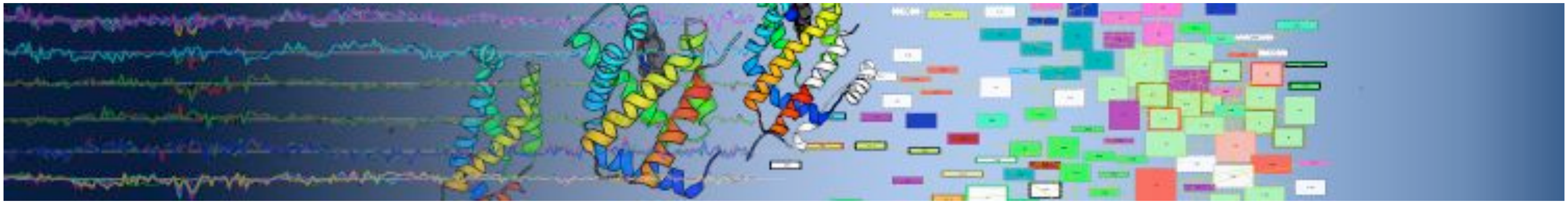


How accurate are we when we predict Structure for Swissprot Proteins?



Low Confidence

High Confidence

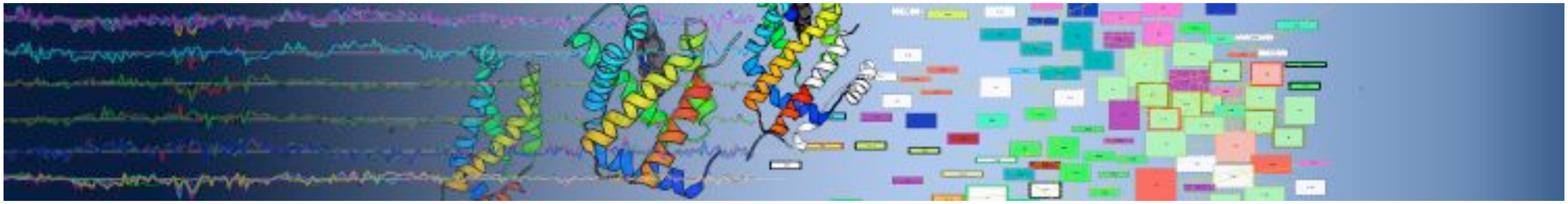


How do we measure preformance of function predictions?

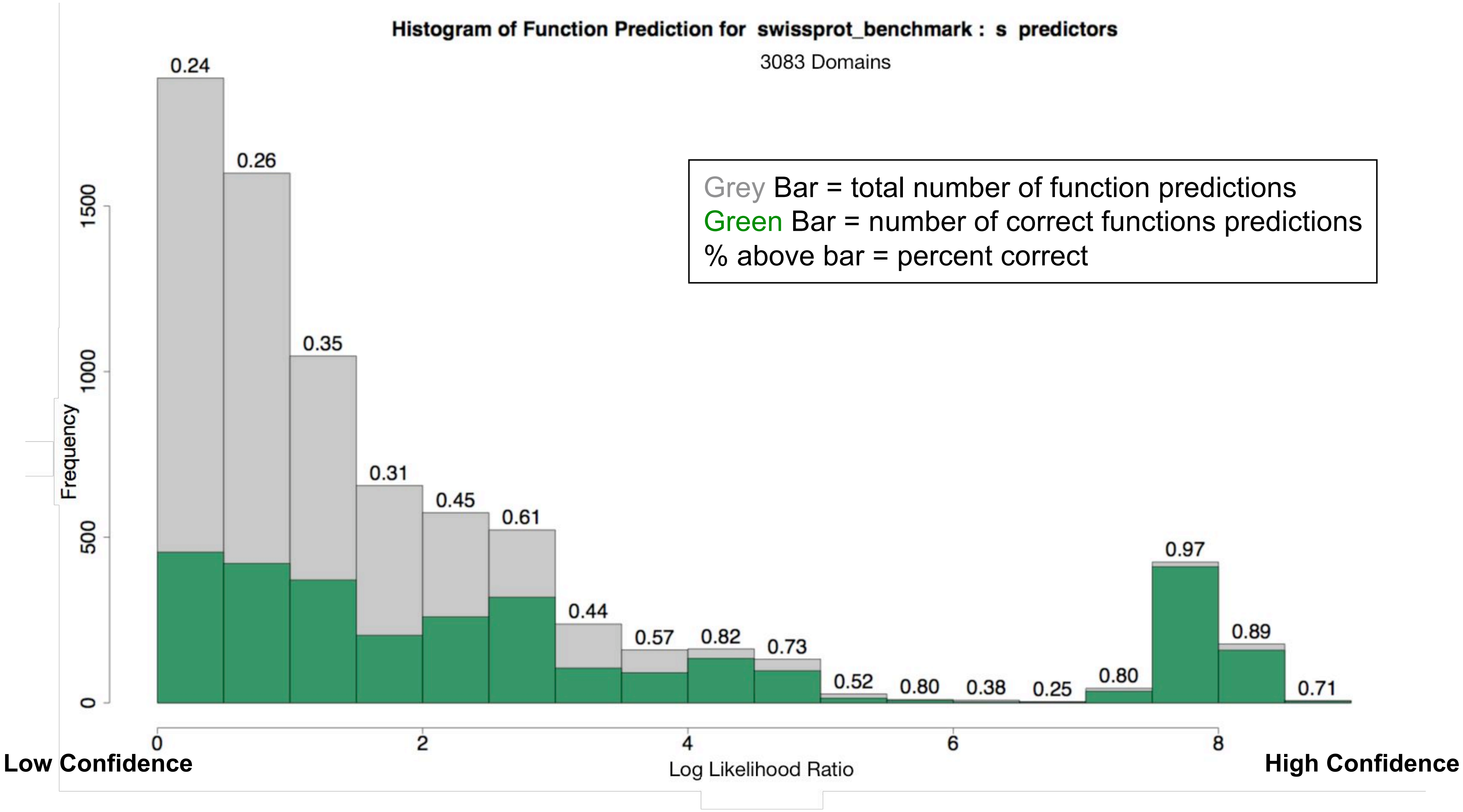
Accuracy - # of Correct / # of Total

Coverage - # of Proteins with High Confidence Predictions

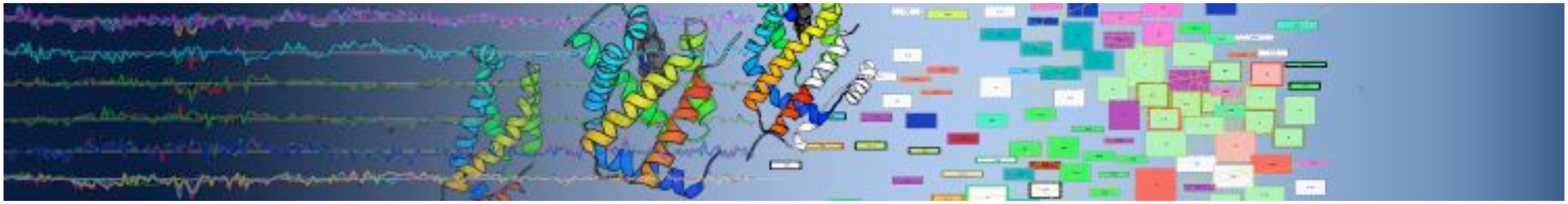
Specificity and Uniqueness - # of proteins with annotations (background probability)



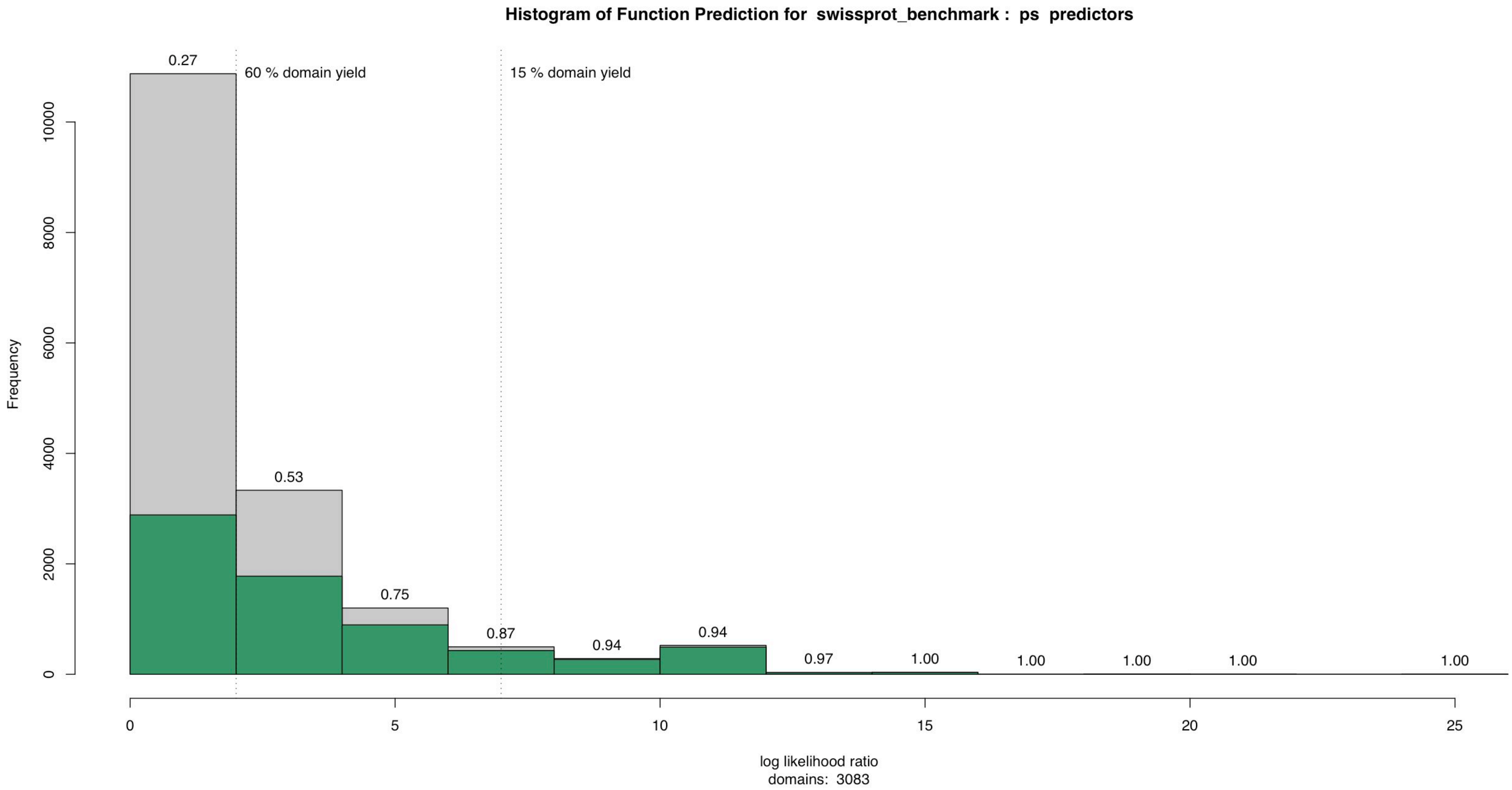
How accurate are our function predictions using structure only?

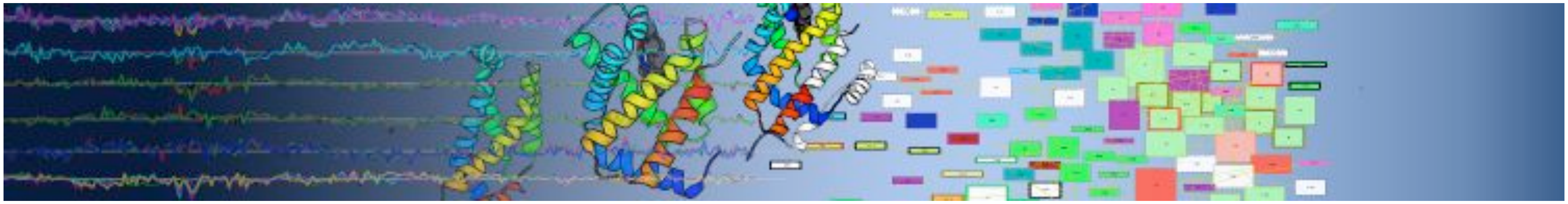


Grey Bar = total number of function predictions
Green Bar = number of correct functions predictions
% above bar = percent correct

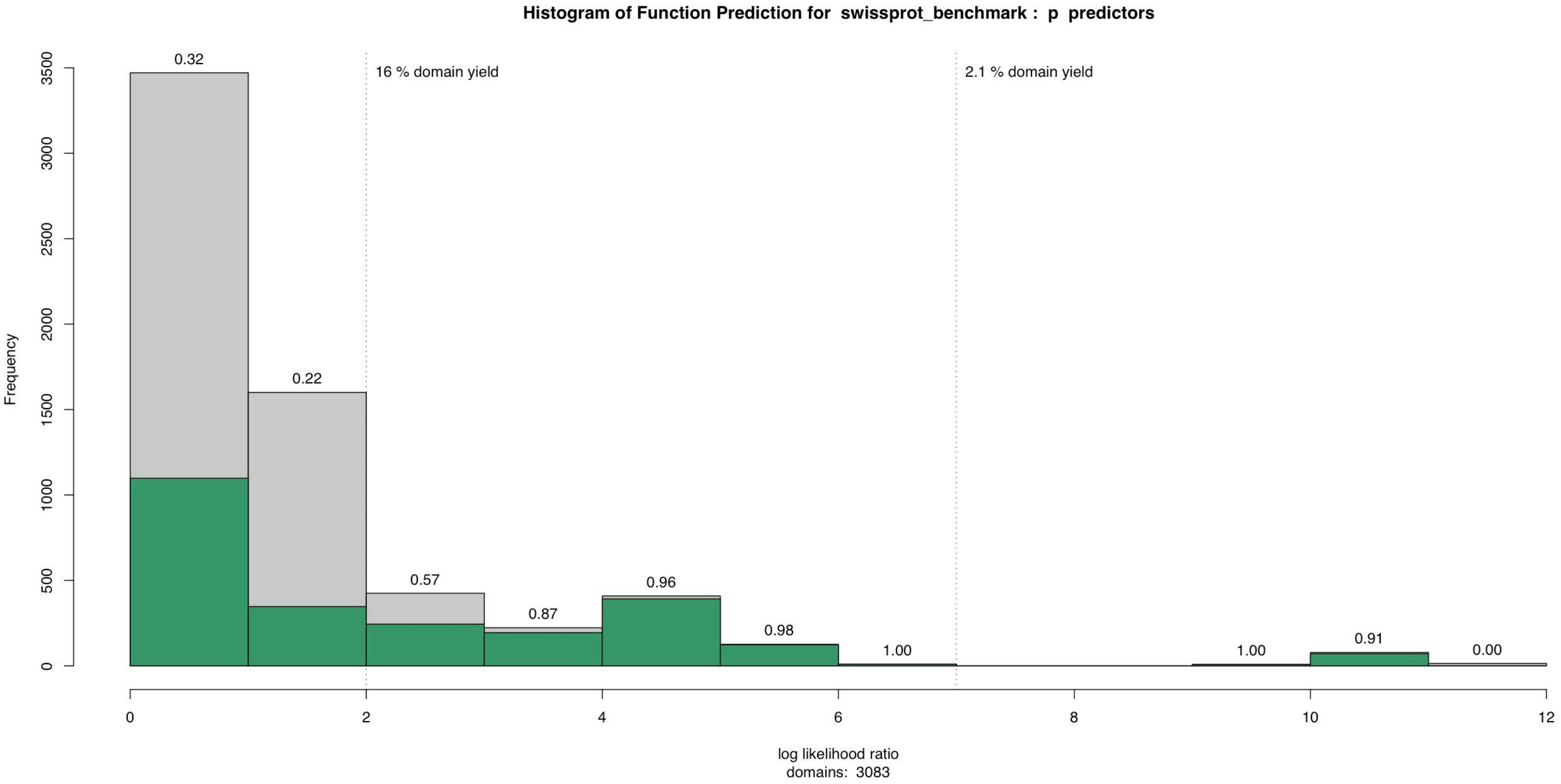


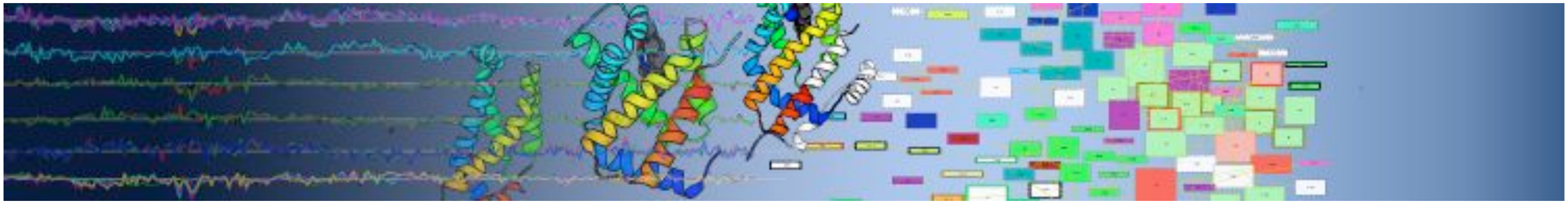
How accurate are our function predictions using GO process & structure?



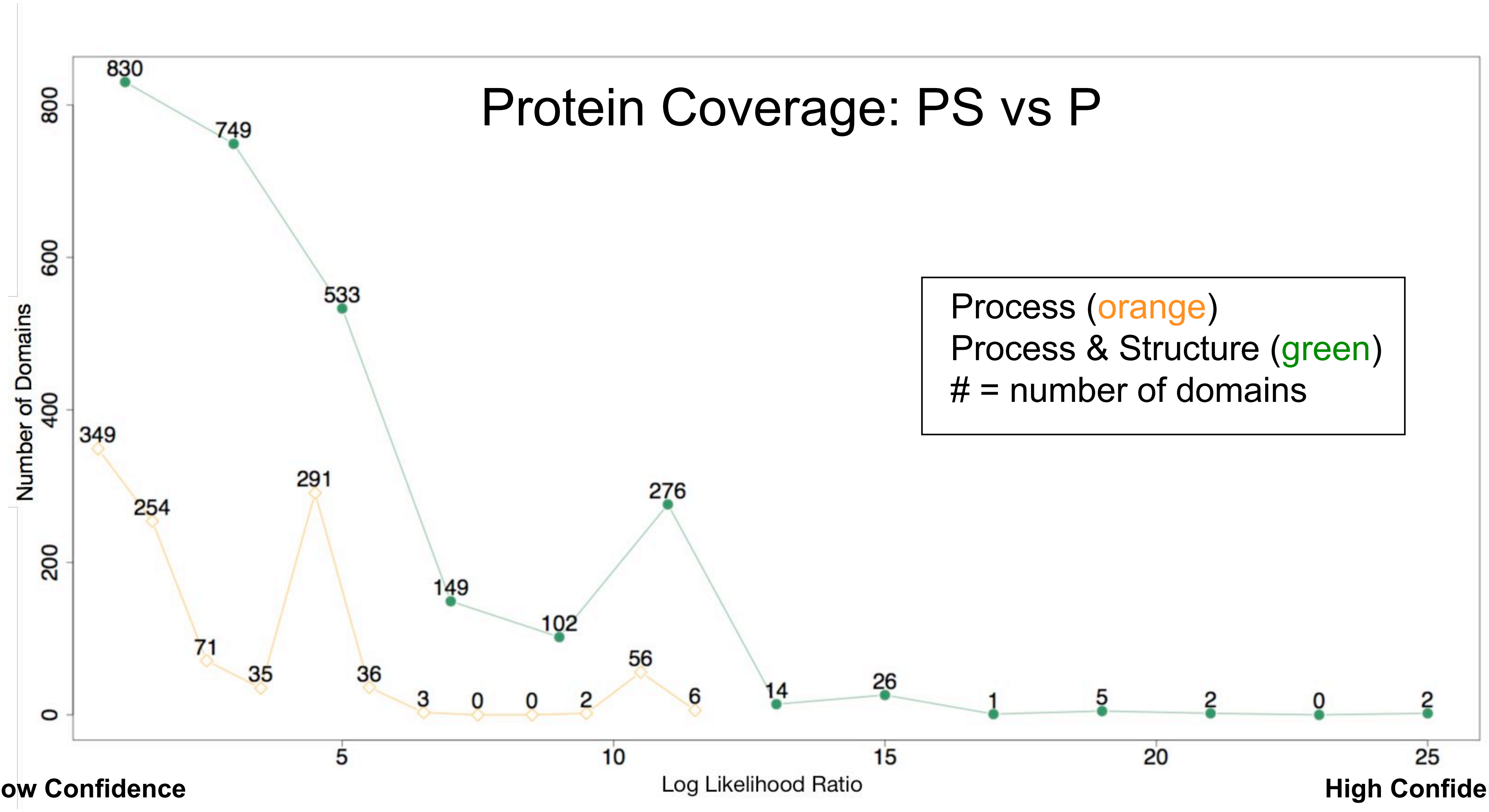


What does structure provide over GO process alone?

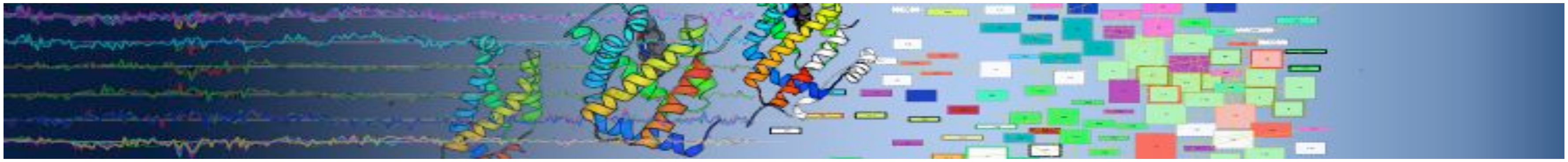




What does structure provide over GO process alone?

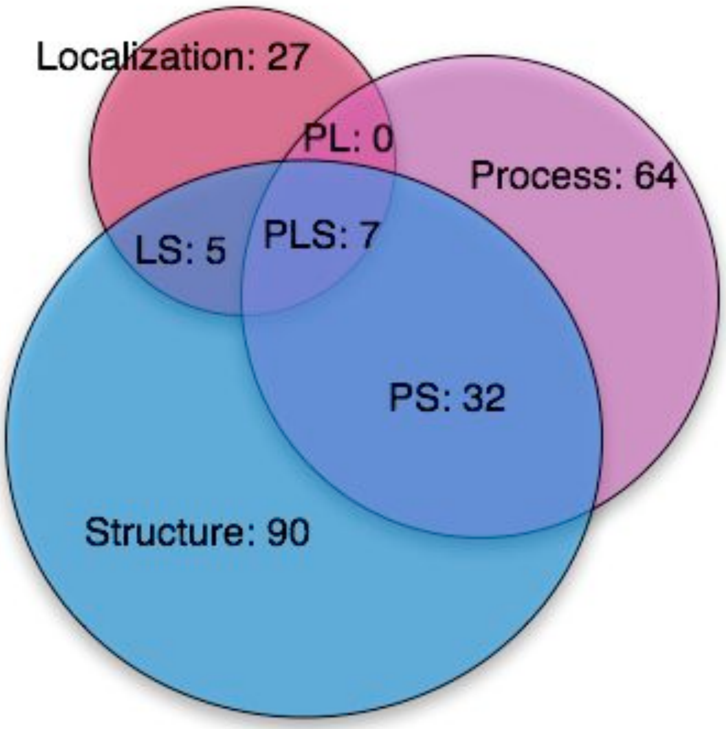


Low Confidence **High Confidence**



Uniqueness and Specificity of GO Functions

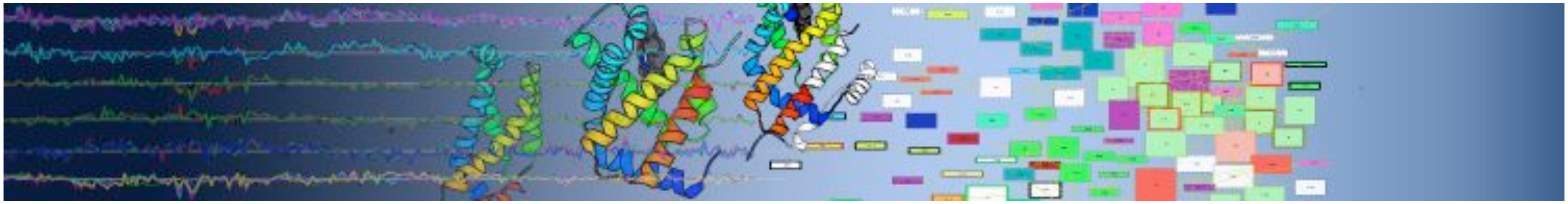
Unique Functions by Evidence



Swissprot LLR >= 2

Sampling of Predicted Terms

GO ID	GO Name	Percent of Genes with Terms
GO:0005198	structural molecule activity	0.03
GO:0003735	structural constituent of ribosome	0.02
GO:0003676	nucleic acid binding	0.17
GO:0003723	RNA binding	0.04
GO:0016491	oxidoreductase activity	0.16
GO:0046872	metal ion binding	0.11
GO:0016787	hydrolase activity	0.24
GO:0043167	ion binding	0.12
GO:0043169	cation binding	0.11
GO:0005509	calcium ion binding	0.01
	. . .	
GO:0004550	nucleoside diphosphate kinase activity	0.0009
GO:0005496	steroid binding	0.001
GO:0042379	chemokine receptor binding	0.0006
GO:0030234	enzyme regulator activity	0.01
GO:0016788	hydrolase activity, acting on ester bonds	0.04
GO:0008289	lipid binding	0.005
GO:0004812	aminoacyl-tRNA ligase activity	0.01
GO:0005506	iron ion binding	0.03
GO:0005216	ion channel activity	0.003



Conclusions

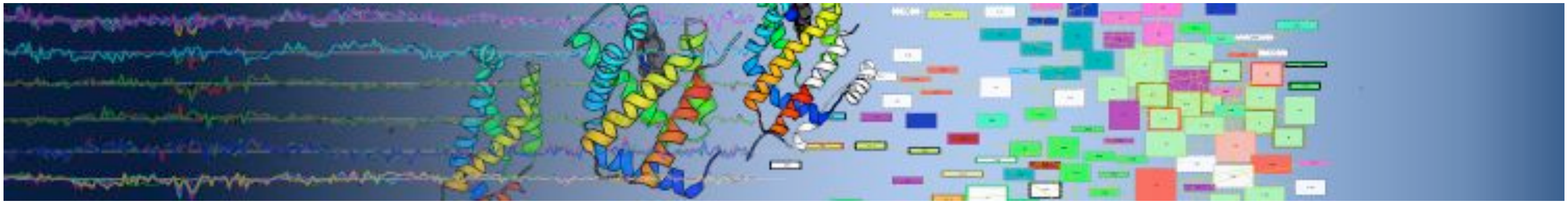
- Method for predicting function using Rosetta protein predictions

- Accurately match protein predictions to known structures

- Accurately prediction functions

- Integrate multiple pieces of evidence to increase coverage

- Predict specific and unique functions



Acknowledgements

Bonneau Lab

Glenn Butterfoss
Thadeous Kacmarczyk
Peter Waltman
Aviv Madar
Kevin Belasco
Alex Pine
Richard Bonneau

University of Washington

Lars Malmstroem
David Baker
Trisha N. Davis
Michael Riffle
Yeast Resource Center

NYU

Sasha Levy
Peter McKenney
Jane Carlton
Dennis Shasha
Kris Gunsalus
Fabio Piano
Patrick Eichenberger
Biology Department

IBM

Viktors Berstis
Keith J Uplinger
Bill Boverman

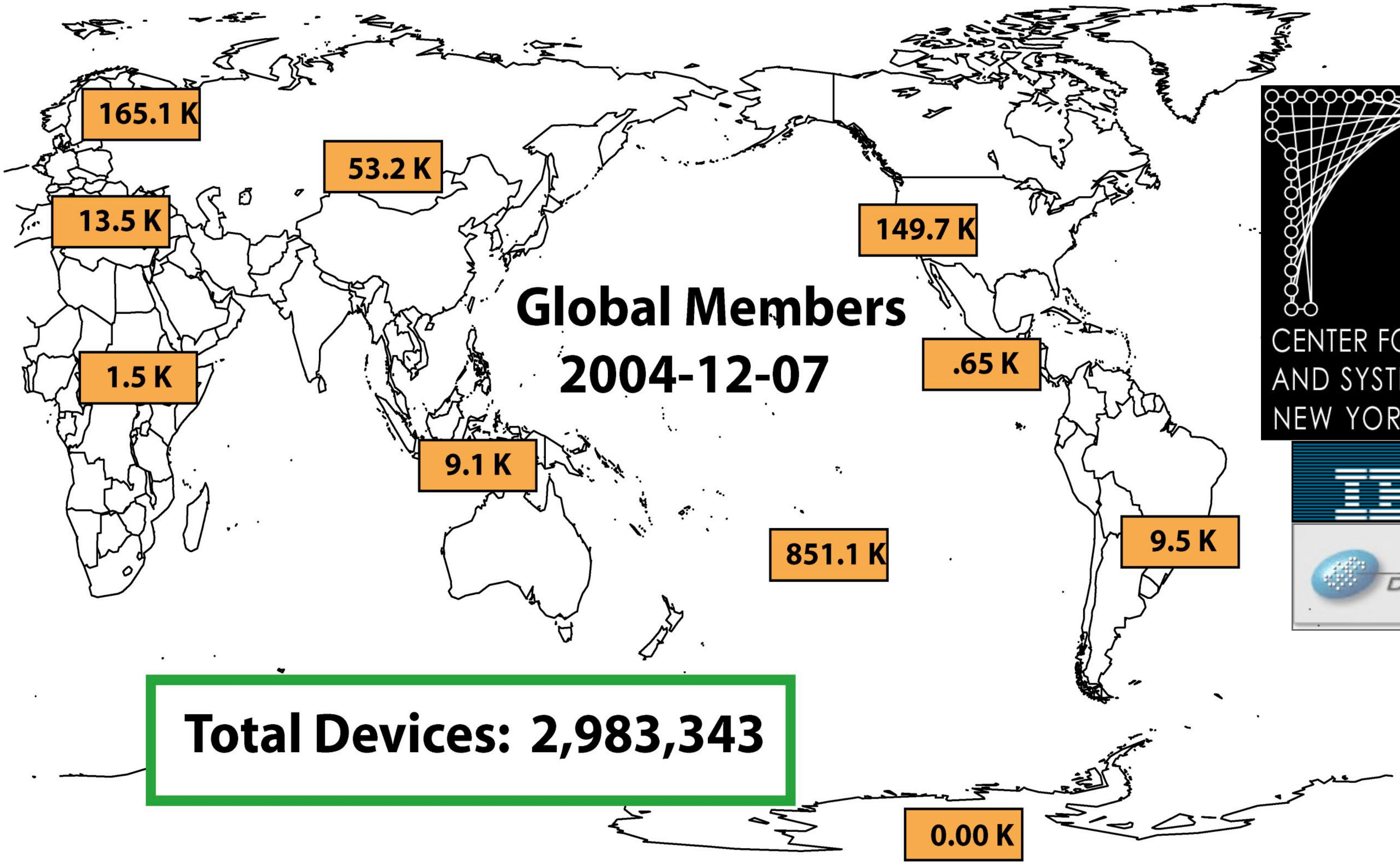
Funding

DOD
DOE
NSF

Rosetta-Commons

Data & Results: http://www.yeastrc.org/pdr/

WORLDCOMMUNITYGRID.ORG & GRID.ORG



Global Members
2004-12-07

Total Devices: 2,983,343

CENTER FOR GENOMICS AND SYSTEMS BIOLOGY
NEW YORK UNIVERSITY

IBM

UNITED DEVICES™

COLLABORATORS: LARS MALMSTROEM, VIKTORS BERSTIS, MIKE RIFFLE, LEROY HOOD, DAVID BAKER