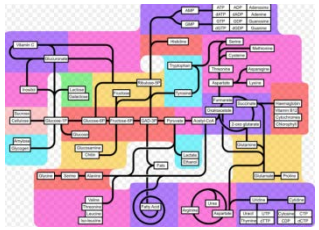
Perturbation by sequence variation: impacts of coding mutations on protein fitness

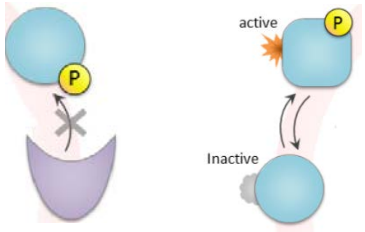
Anna Panchenko
National Center for Biotechnology Information, NIH

What do we study?

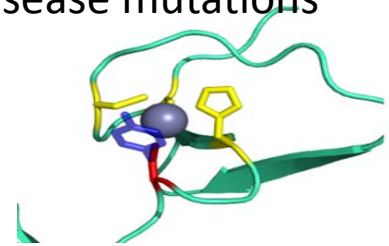
Biochemical pathways and interactomes



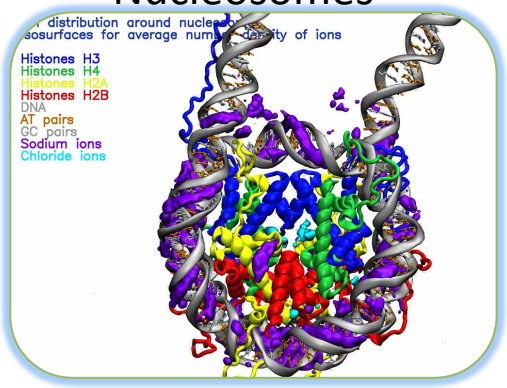
PTMs, phosphorylation



Genetic variants, disease mutations

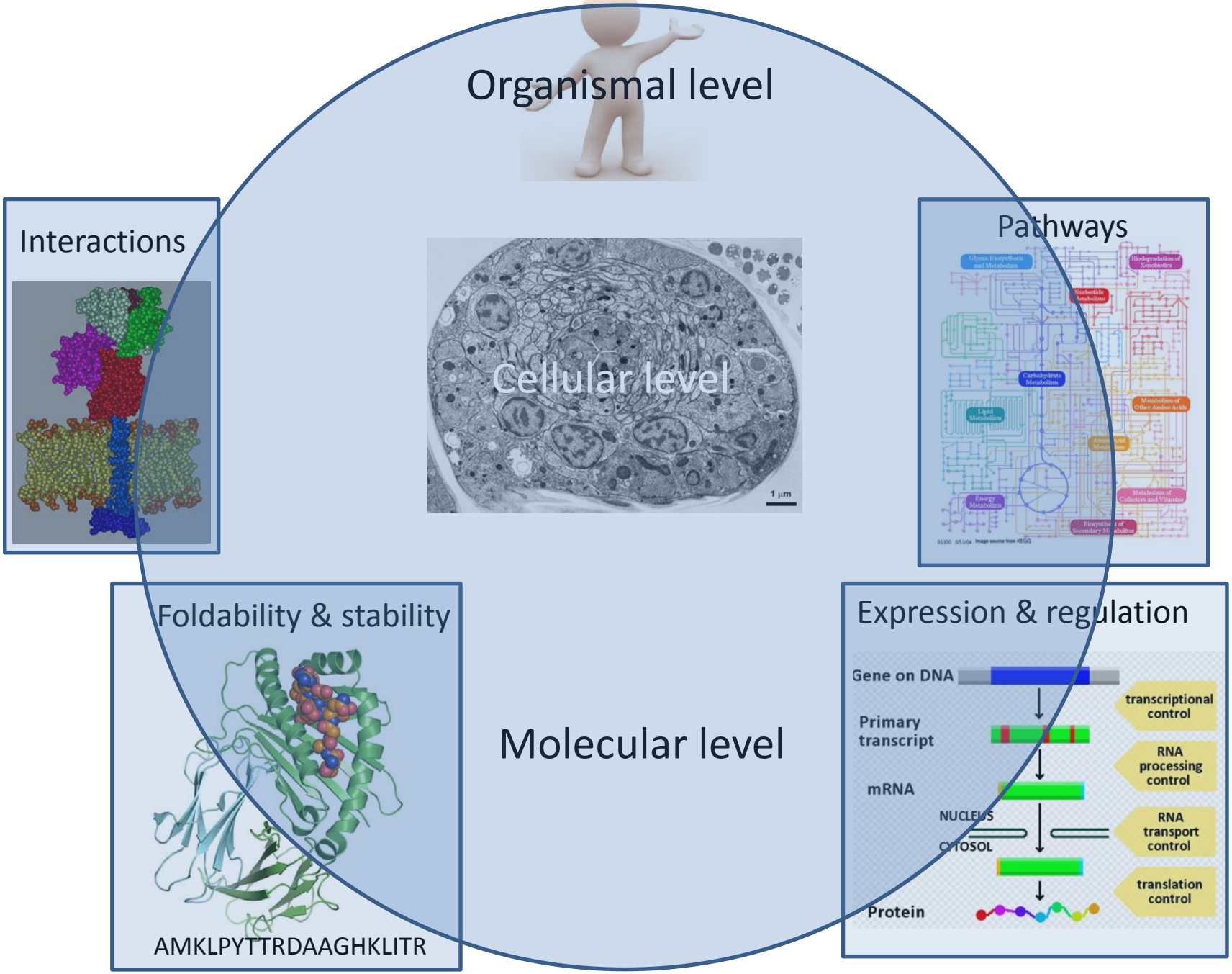


Nucleosomes



Methods

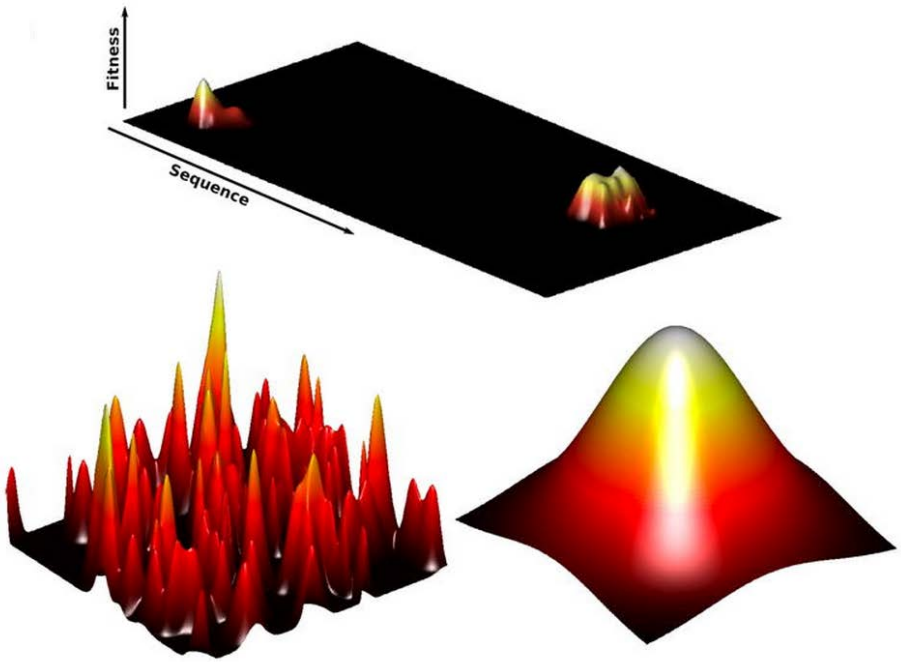
Bioinformatics	Sequence alignment, structure superposition, annotation of domains, intrinsically disordered regions, functional sites Programs: Blast, Vast, Muscle, IBIS, SPEER, CDD, …
Structural Modeling	Homology modeling of protein structures and structural complexes Programs: VMD, NAMD
Energy calculations	Empirical and statistical energy potentials, Molecular Mechanics Poisson-Boltzmann approach Force fields and programs: CHARMM27, CHARMM36, FoldX, BeatMusic and PopMusic,…
Dynamics	All-atom Molecular Dynamics simulations Programs: NAMD
Evolutionary analysis	Evolutionary conservation, phylogenetic analysis Programs: Mega, PAML, FastTree, …


Organismal level
Interactions
Cellular level
1 µm
Pathways
Foldability & stability
AMKLPYTTRDAAGHKLITR
Molecular level
Expression & regulation
Gene on DNA
Primary transcript
mRNA
NUCLEUS
CYTOSOL
Protein
transcriptional control
RNA processing control
RNA transport control
translation control

Variation may arise through genetic mutations and rearrangements

Types of variations:

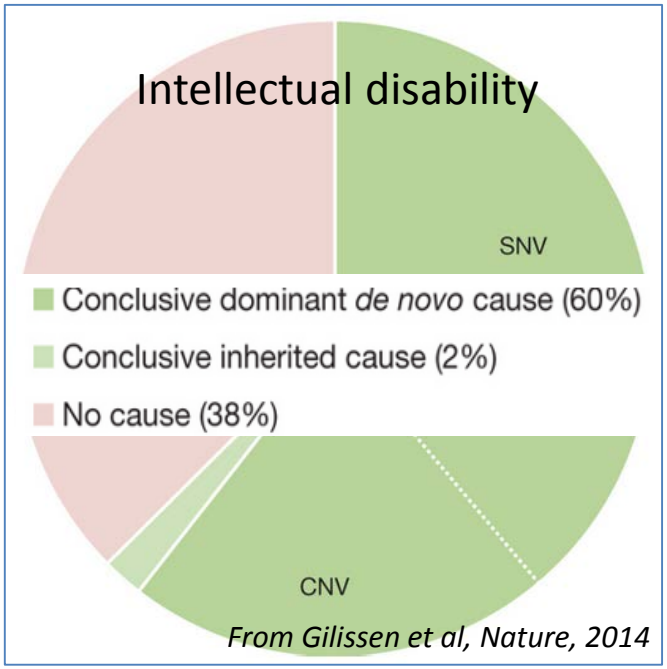
- missense mutations – change in amino acid type;
- Insertion and deletions;
- Polypeptide chain truncations;
- Domain shuffling;
- Post-translational modifications



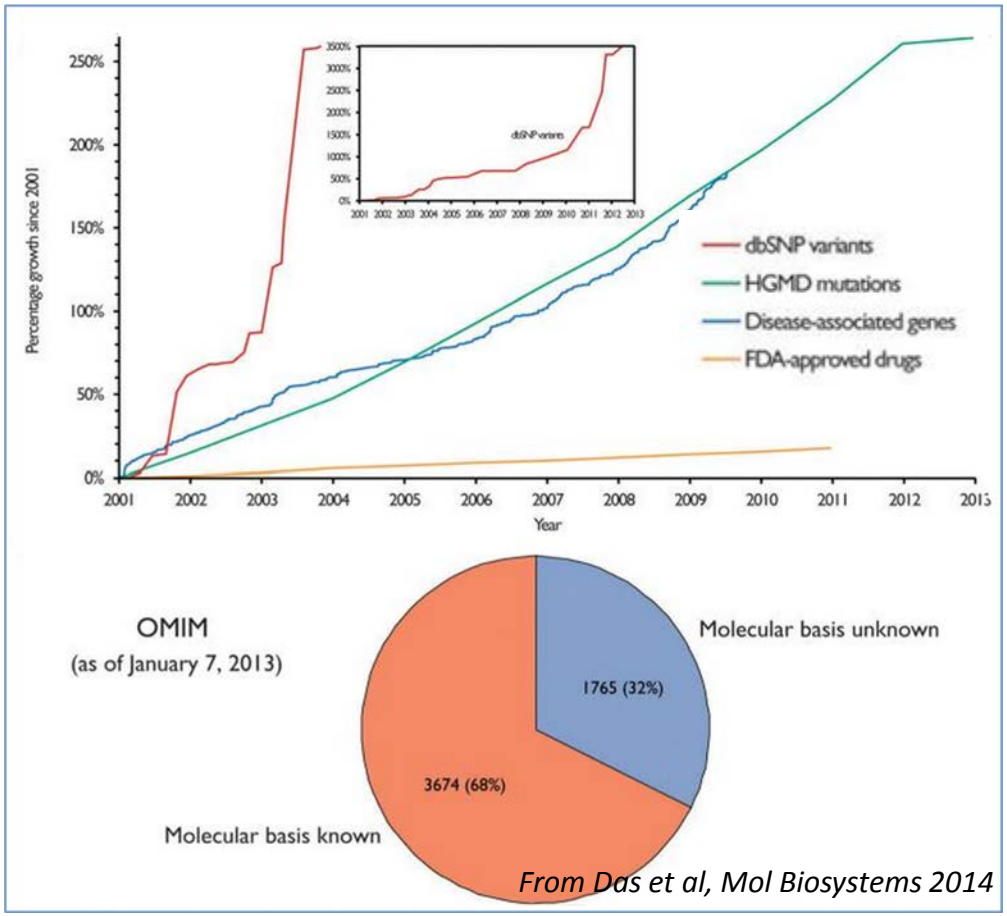
From Romero & Arnold, *Nat Rev Mol Cell Biol.* 2009

Hundreds of mutations with unknown molecular mechanisms

Type of variation	Number in human
De novo variants	50-100
SNV	3.5 million
nsSNV, sSNV	10,000-15,000
Protein loss-of-function nsSNV (from HGMD)	50-500 (50-100)

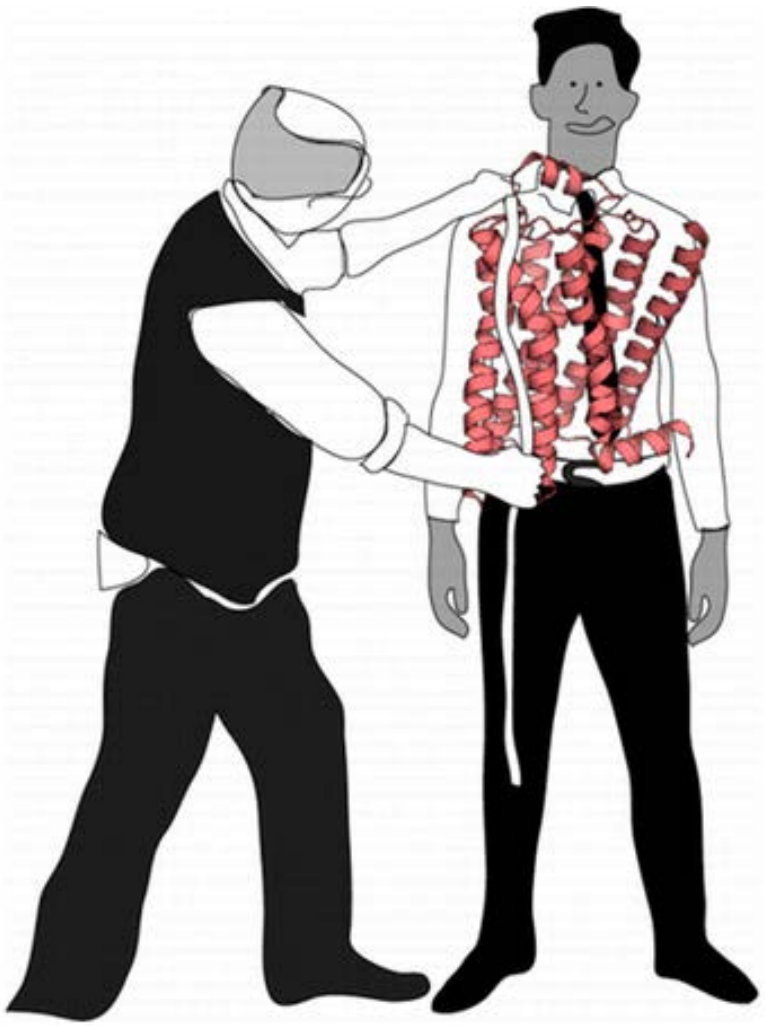


From Gilissen et al, Nature, 2014



From Das et al, Mol Biosystems 2014

Personal genomics meets biophysics



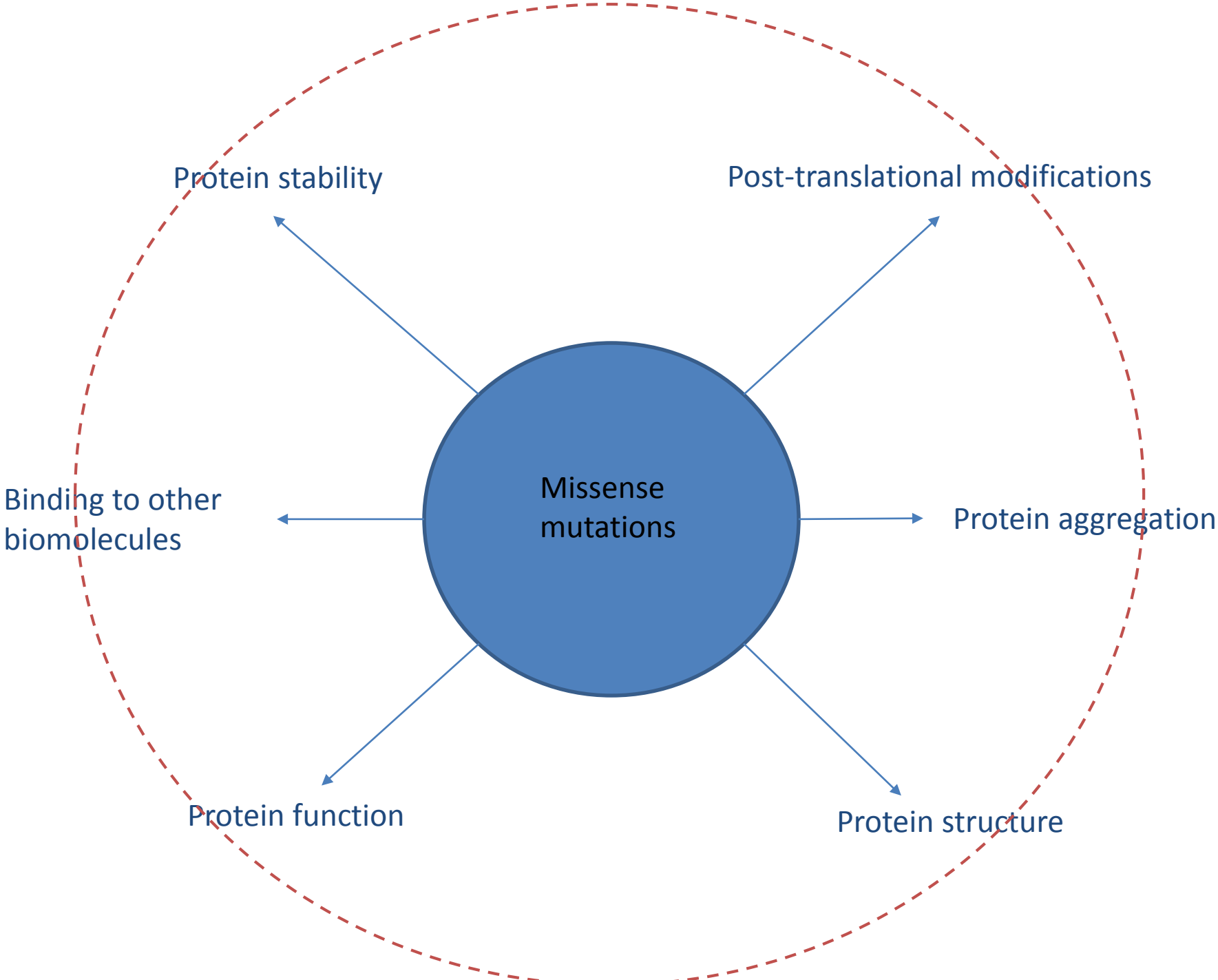
From Kroncke et al, Biochemistry 2015

Talk synopsis

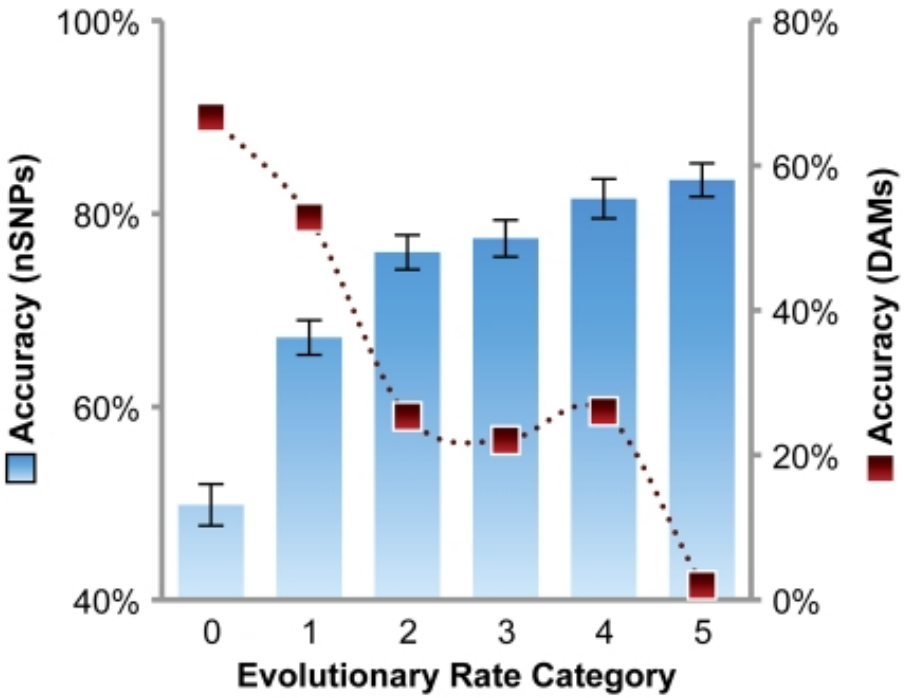
- Impact of missense mutations on proteins: stability, binding and activity
- Deciphering the human protein interactome with interactions with resolved binding interfaces

Why do we need to learn about the mechanisms of effect of mutations on proteins?

- To decipher how proteins evolved.
- To predict which mutations are damaging.
- To distinguish functionally important mutations, distinguish driver from passenger mutations.
- Prioritize mutations for experimental research.
- Drug design.


Protein stability
Post-translational modifications
Binding to other biomolecules
Missense mutations
Protein aggregation
Protein function
Protein structure

Evolutionary conservation is related to functional importance

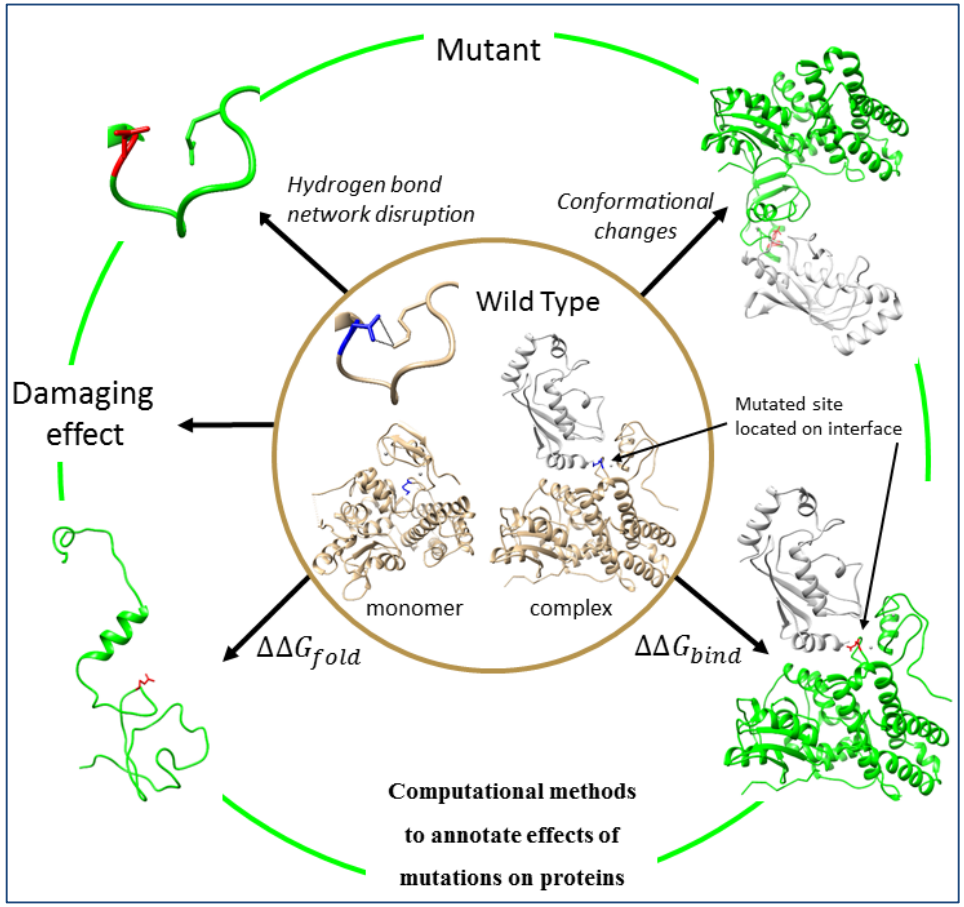


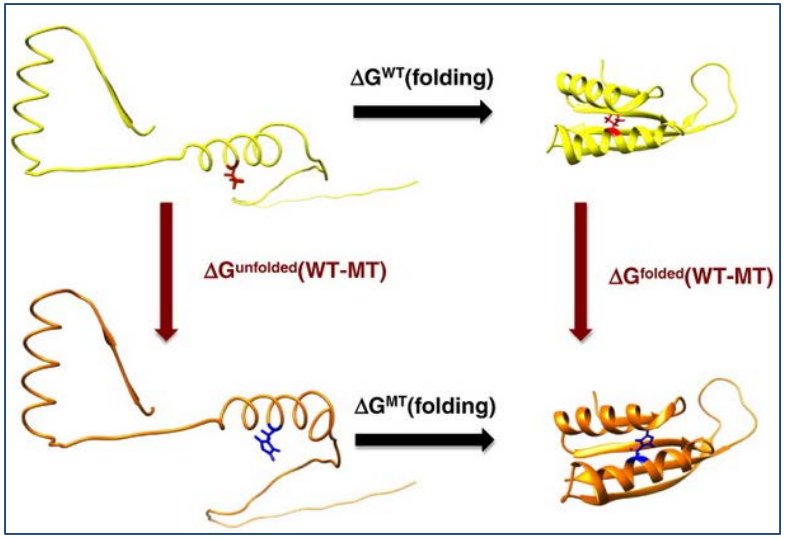
From Kumar et al, *Genome Research*, 2009

Mutations in functionally relevant sites might be damaging. Many methods exploit this observation.

Methods that predict functional effects of mutations on proteins use evolutionary conservation

Deciphering of molecular basis of mutational impacts

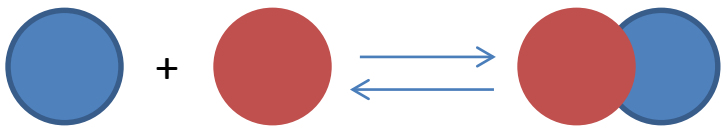


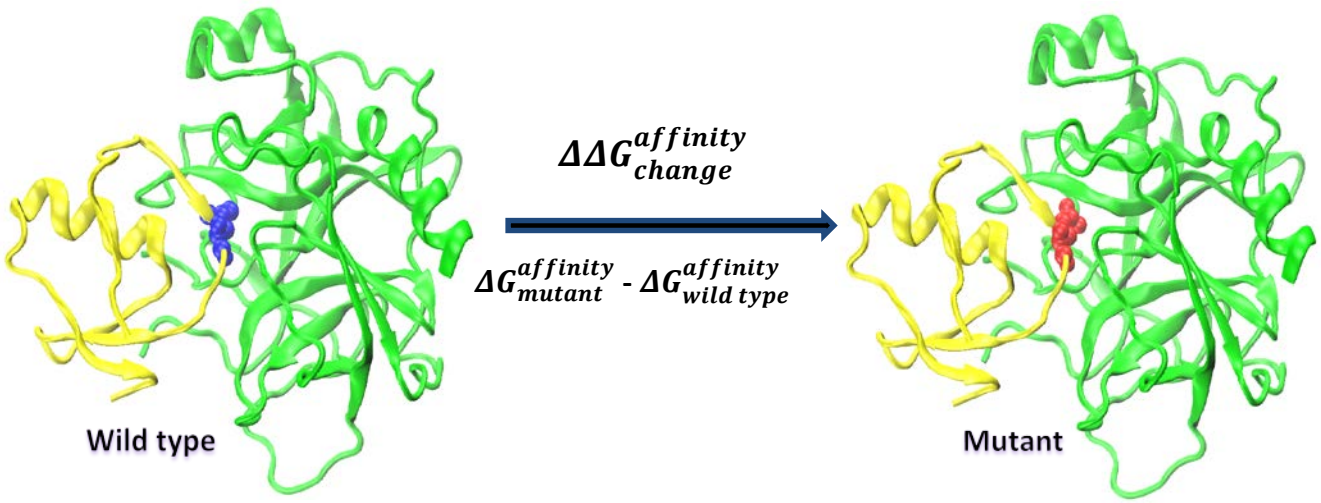


Stefl et al, *JMB* 2013
Li et al, *Methods Mol Biol* 2015

Effects of missense mutations on protein-protein binding affinity

$$\Delta\Delta G_{bind} = \Delta G_{wt} - \Delta G_{mut}$$





$$G = E^{MM}_{gas} + G^{p}_{solv} + G^{np}_{solv} - TS$$

Modified Molecular Mechanics Poisson-Boltzmann Surface Area approach

$$\Delta\Delta G^{bind}_{Pred1} = \alpha\Delta\Delta E_{vdw} + \beta\Delta\Delta G_{solv} + \gamma\Delta SA_{mut} + \delta$$

Li et al, *J Chem Theory & Comput*, 2014

Energy terms

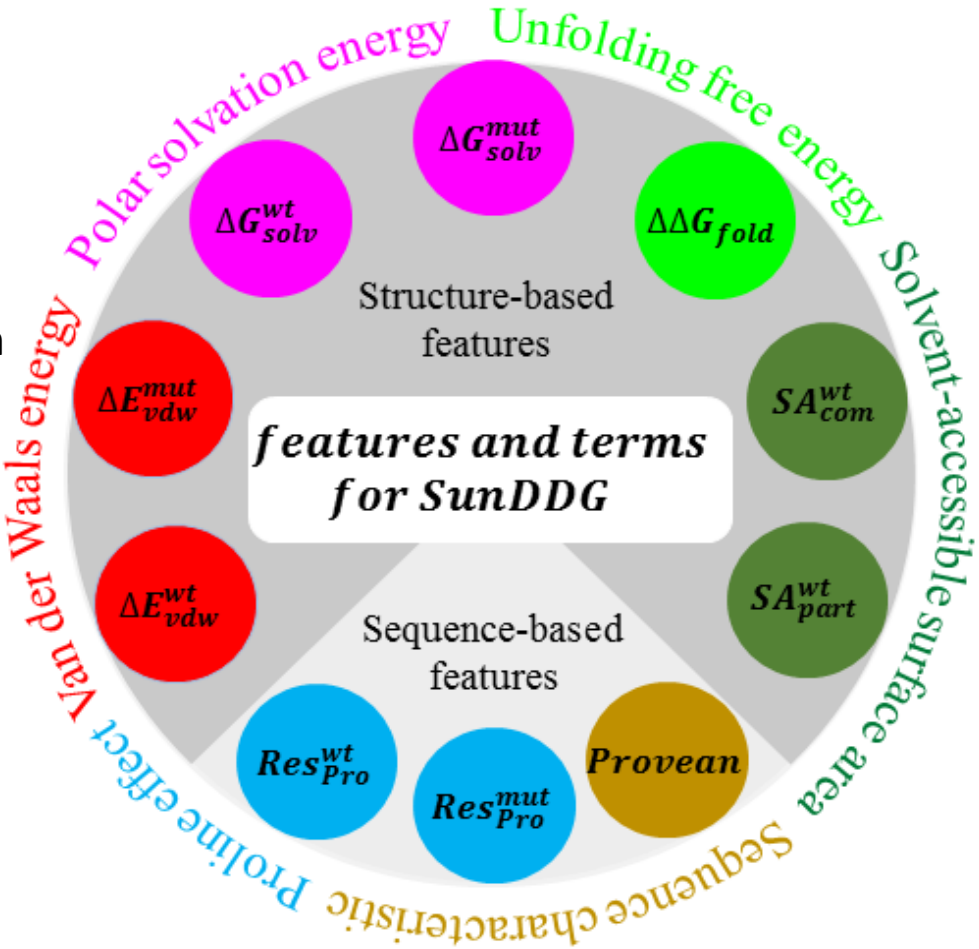
Polar solvation energies of a complex and each interacting partner in water, from Poisson-Boltzmann equation;

Van der Waals energies of complex and each interacting partner;

Unfolding free energy of mutant and wild type, from BuildModel module of FoldX.

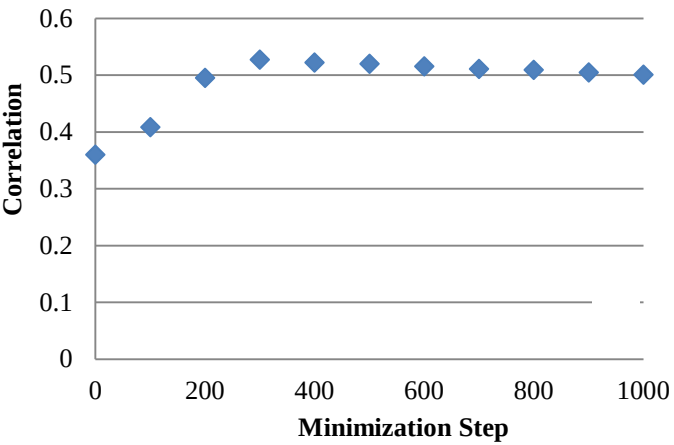
Solvent accessibility of wild type and mutant site

Evolutionary conservation and accounts for the sequence context around the mutated site.



Li et al, submitted

How to minimize mutant and wild type structures



Wild-type PDB structures

Solvent model or gas phase (VMD)

Minimization (NAMD)

Mutation structures (VMD)

Without restrains on the backbone atoms upon minimization (NAMD)

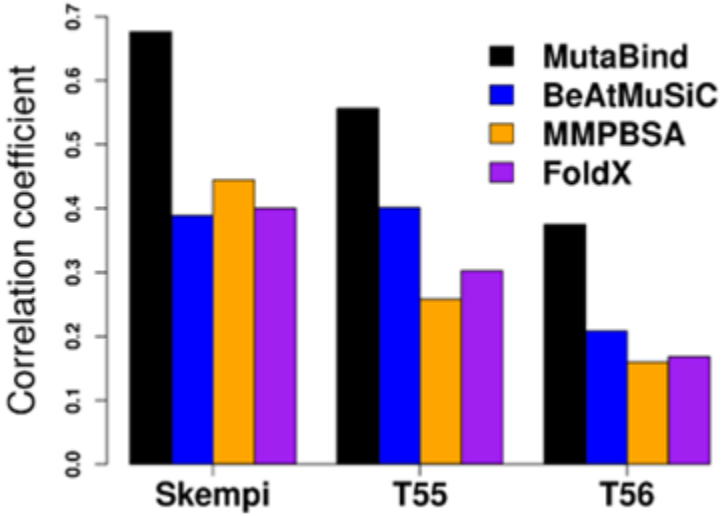
1ns unconstrained Molecular Dynamics (MD) simulations

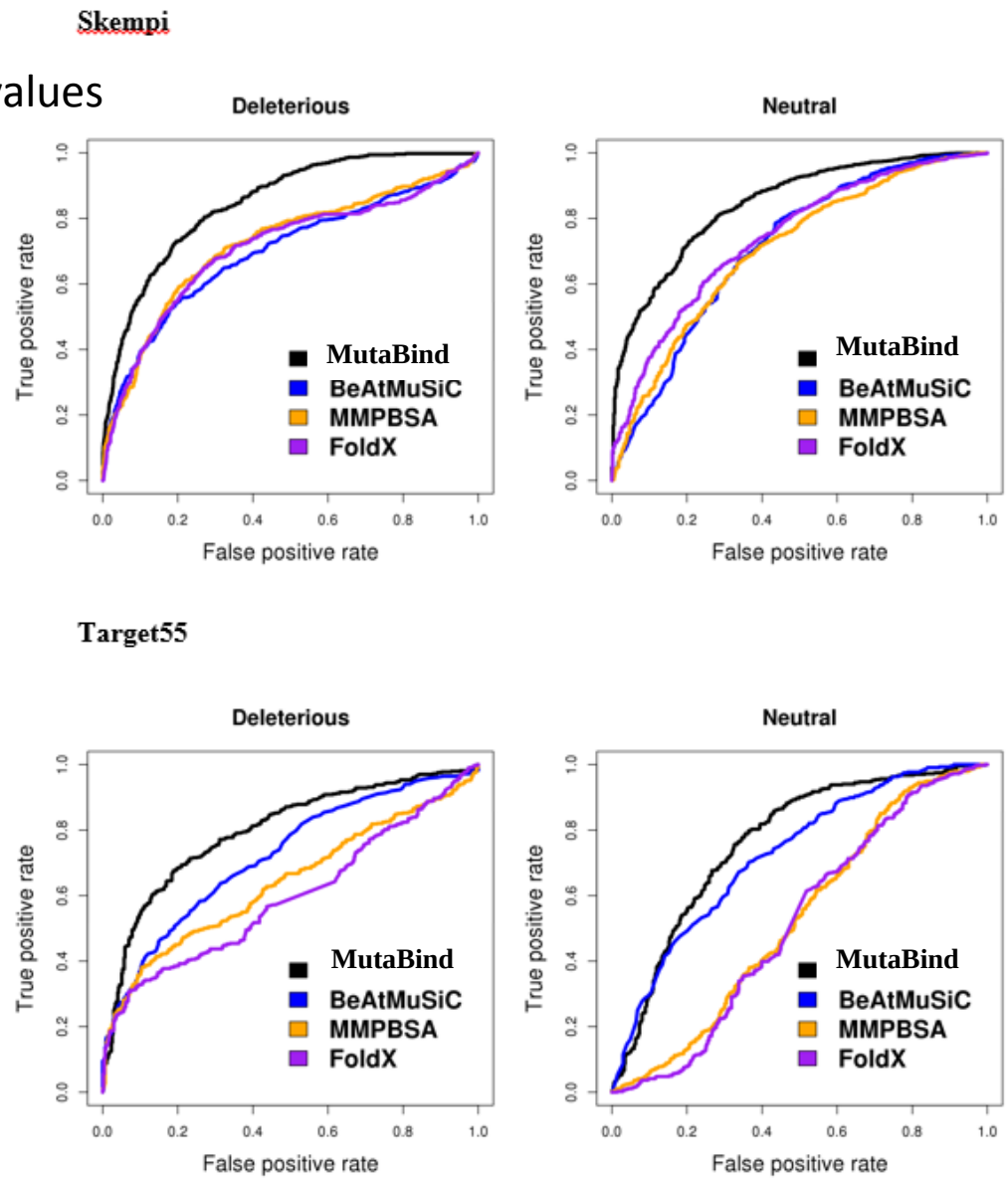
With restrains on the backbone atoms upon minimization

Li et al, *J Chem Theory & Comput*, 2014

Performance of method

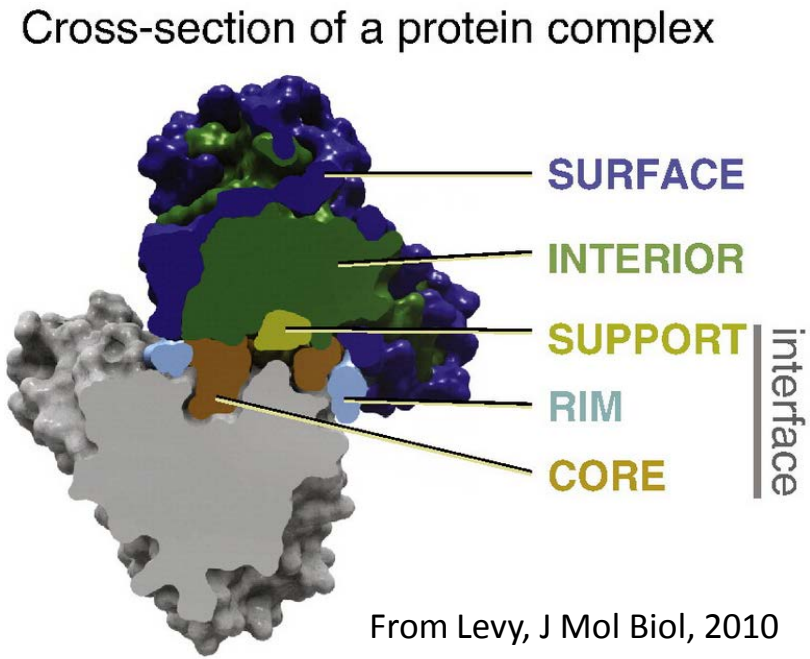
SKEMPI - database of experimental ddG values (*Moal et al, Bioinformatics 2012*) – 1844 mutations



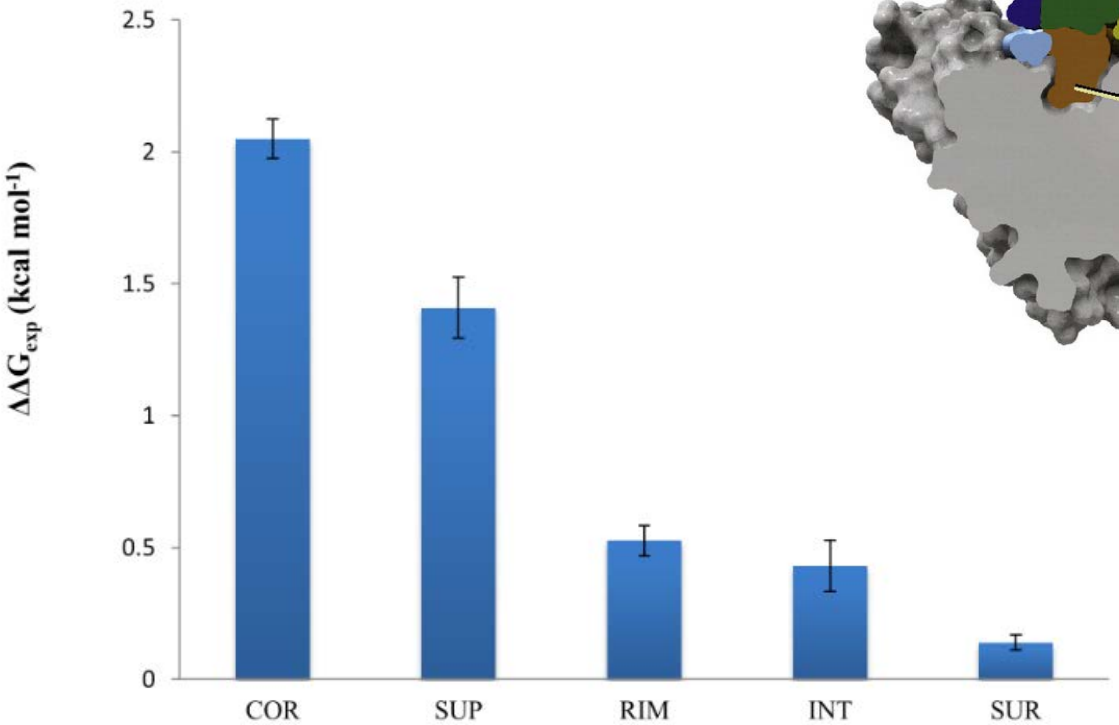


Factors influencing prediction quality

1. Locations of mutations



From Levy, J Mol Biol, 2010



Factors influencing prediction quality

2. Volume and charge of substituted amino acids

	Mutant		
	Small	**Medium**	**Large**
Wild-type	R/# mutations	R/# mutations	R/# mutations
Small	0.52/97	0.51/123	0.67/39
Medium	0.61/590	0.58/450	0.34/130
Large	0.63/210	0.64/142	0.58/63

	Mutant		
	Negative	Neutral	Positive
Wild-type	R/# mutations	R/# mutations	R/# mutations
Negative	-	0.33/232	-
Neutral	0.72/86	0.58/1042	0.48/89
Positive	0.81/33	0.67/300	-

Factors influencing prediction quality

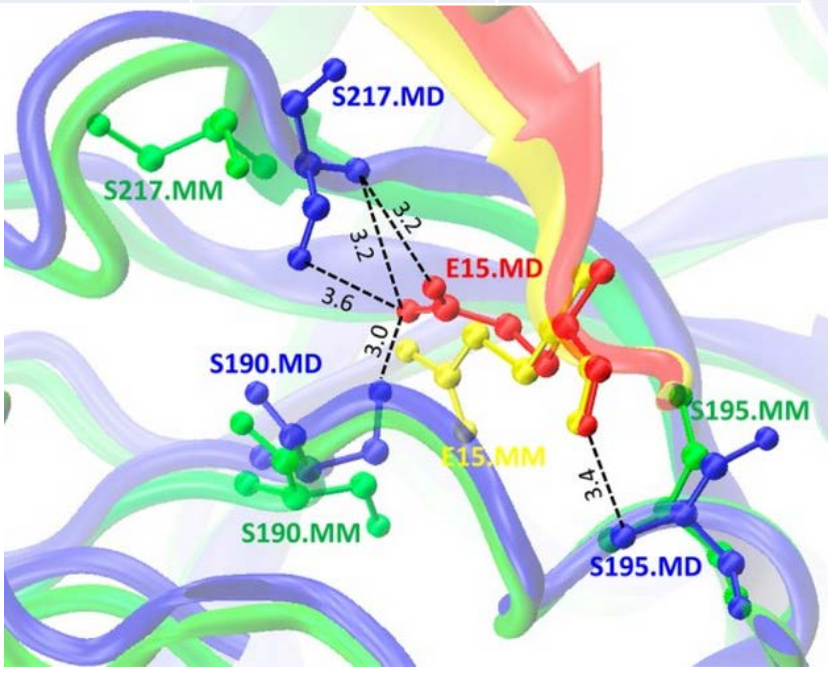
3. Low accuracy of predicting stabilizing mutations, probably due to lack of stabilizing mutants in experimental data sets

Method	Training/testing	Destabilizing	Stabilizing
Pred2	SKEMPI/NM	100%	7%
CC/PBSA	NM/NM	99%	32%
FoldX	test: NM	72%	48%
	test: SKEMPI	67%	41%
BeatMusic	test: NM	95%	23%
	Test: SKEMPI	90%	18%

Factors influencing prediction quality

4. Difficult to adequately account for the flexibility of proteins.

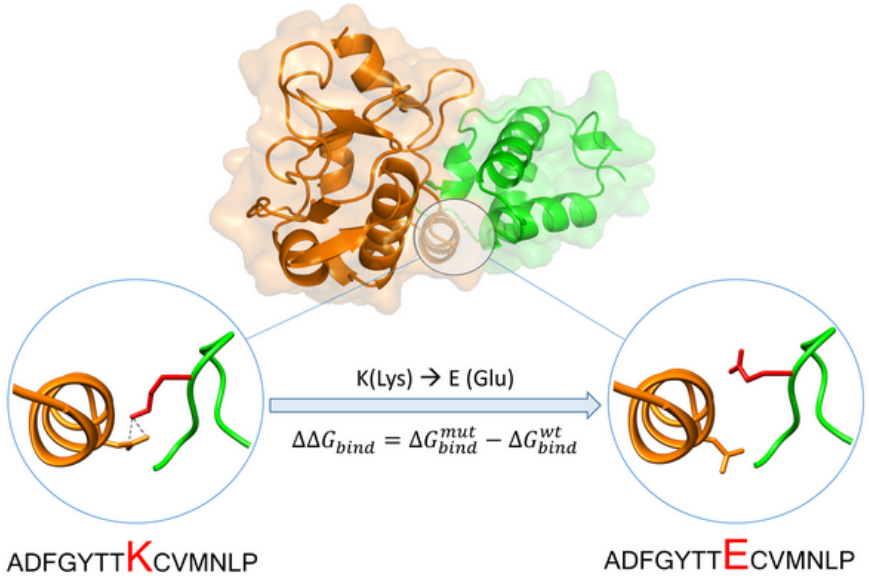
Simulation methiod	Flexibility	Cross-validated R	RMSE (kcal/mol)
Minimization	Flexible backbone	0.61	1.22
1ns MD simulations	Flexible backbone	0.26	1.48



MutaBind

Home | Results | Help | Method | References

MutaBind evaluates the effects of variations and disease mutations on protein-protein interactions. It predicts if a mutation may disrupt an interaction and calculates the changes in binding affinity. Structure of protein-protein complex is required for this method.



Step 1 - Select Protein Complex

Input PDB code: PDB id code [4-letter code]

Example: 1CSE

Bioassembly

Asymmetric Unit

Upload PDB file: Choose File No fil...hosen

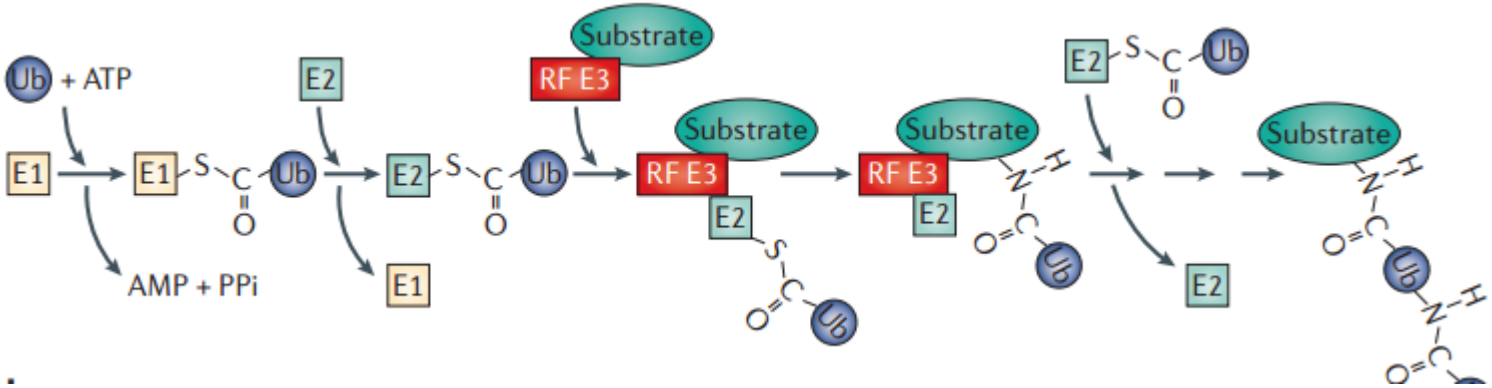
Additional Options

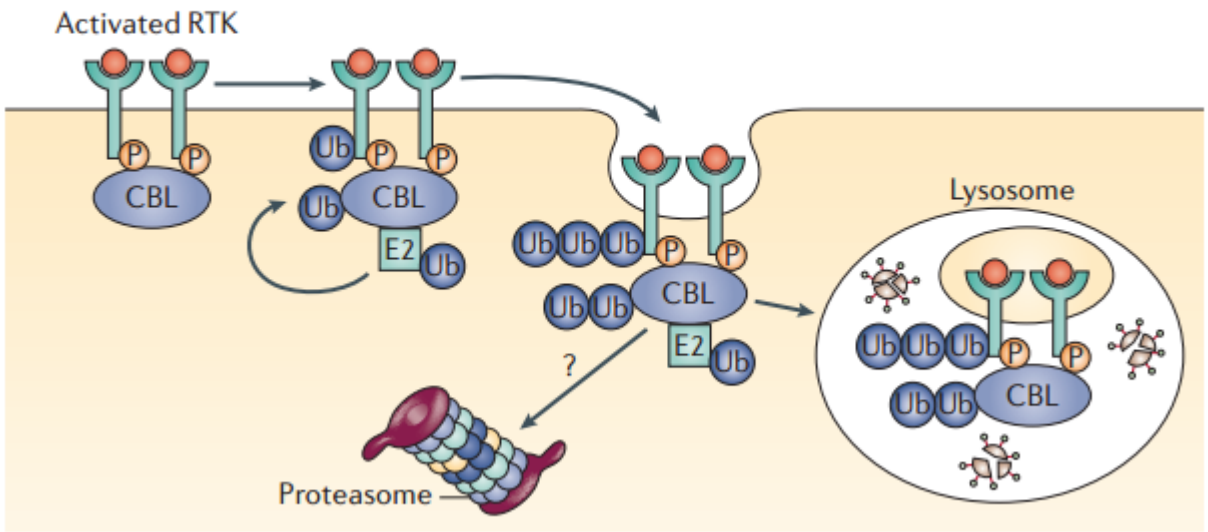
Is protease-inhibitor complex?

Next

Mechanisms of action of cancer mutations: CBL case

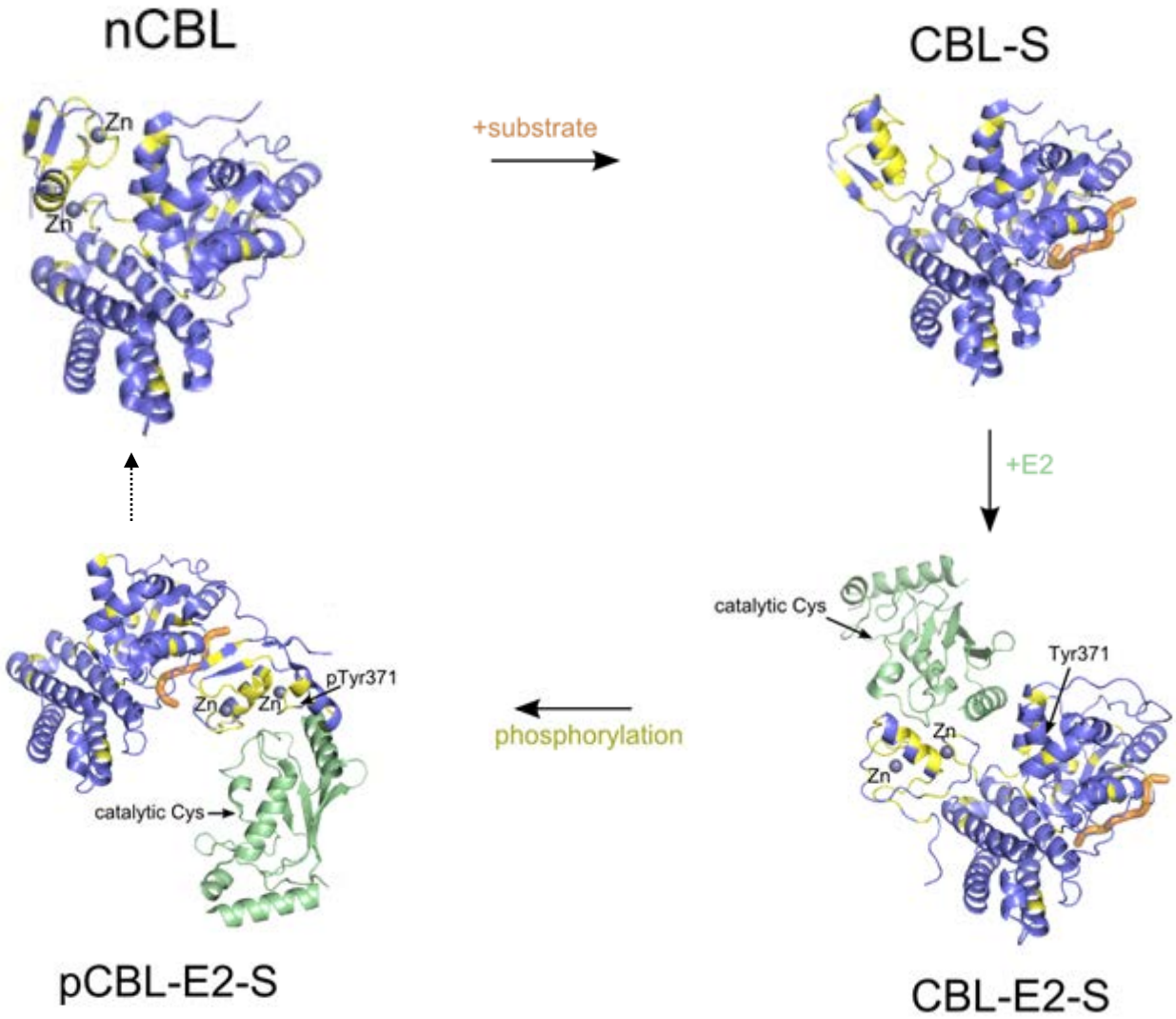
CBL ubiquitin ligase - CBL





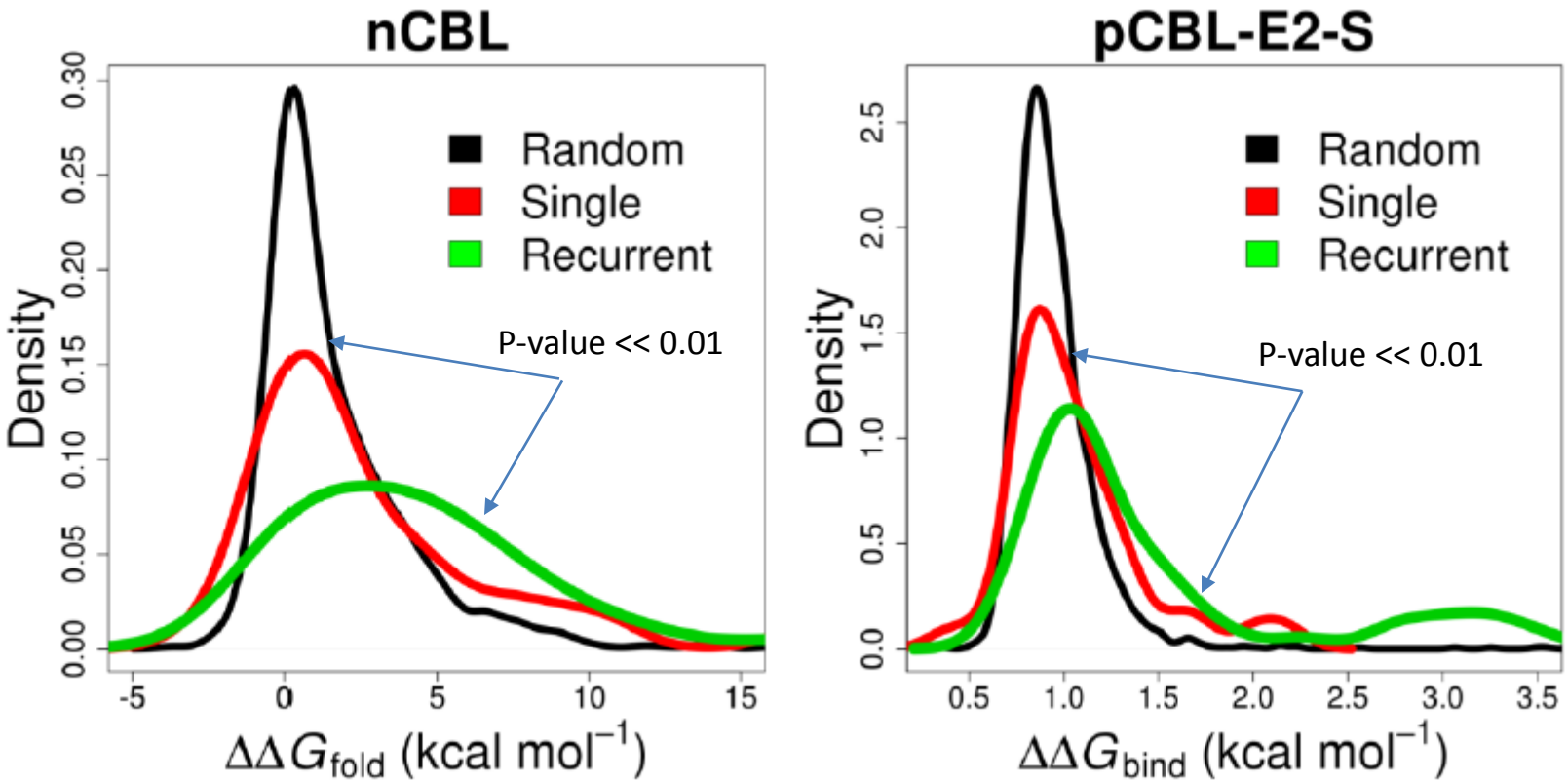
From Lipkowitz & Weissman, Nature Reviews Cancer, 2011

CBL ubiquitin ligase activation cycle



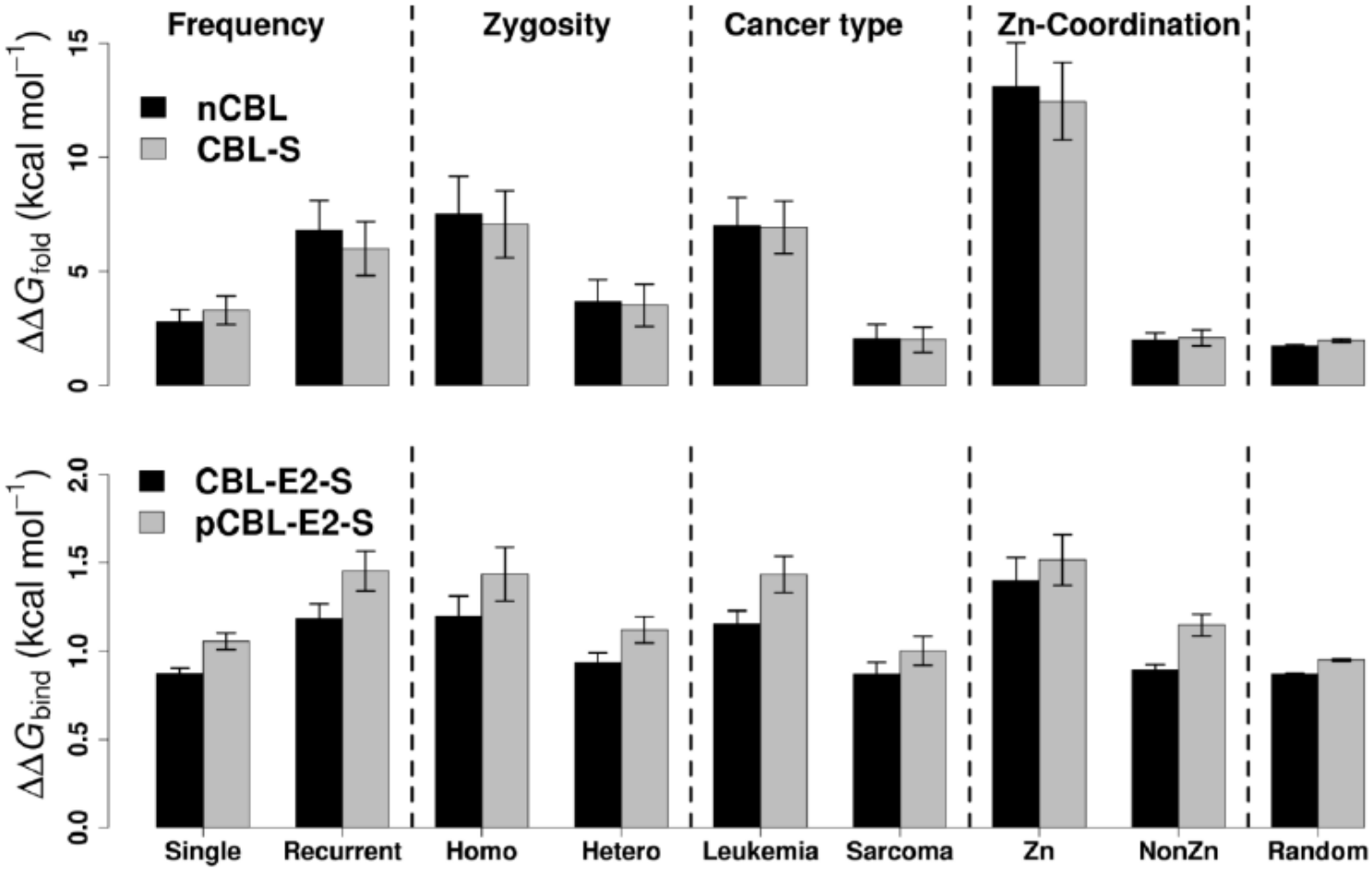
Cancer mutations impact CBL stability and CBL-E2 binding

~110 cancer and 2100 random missense mutations



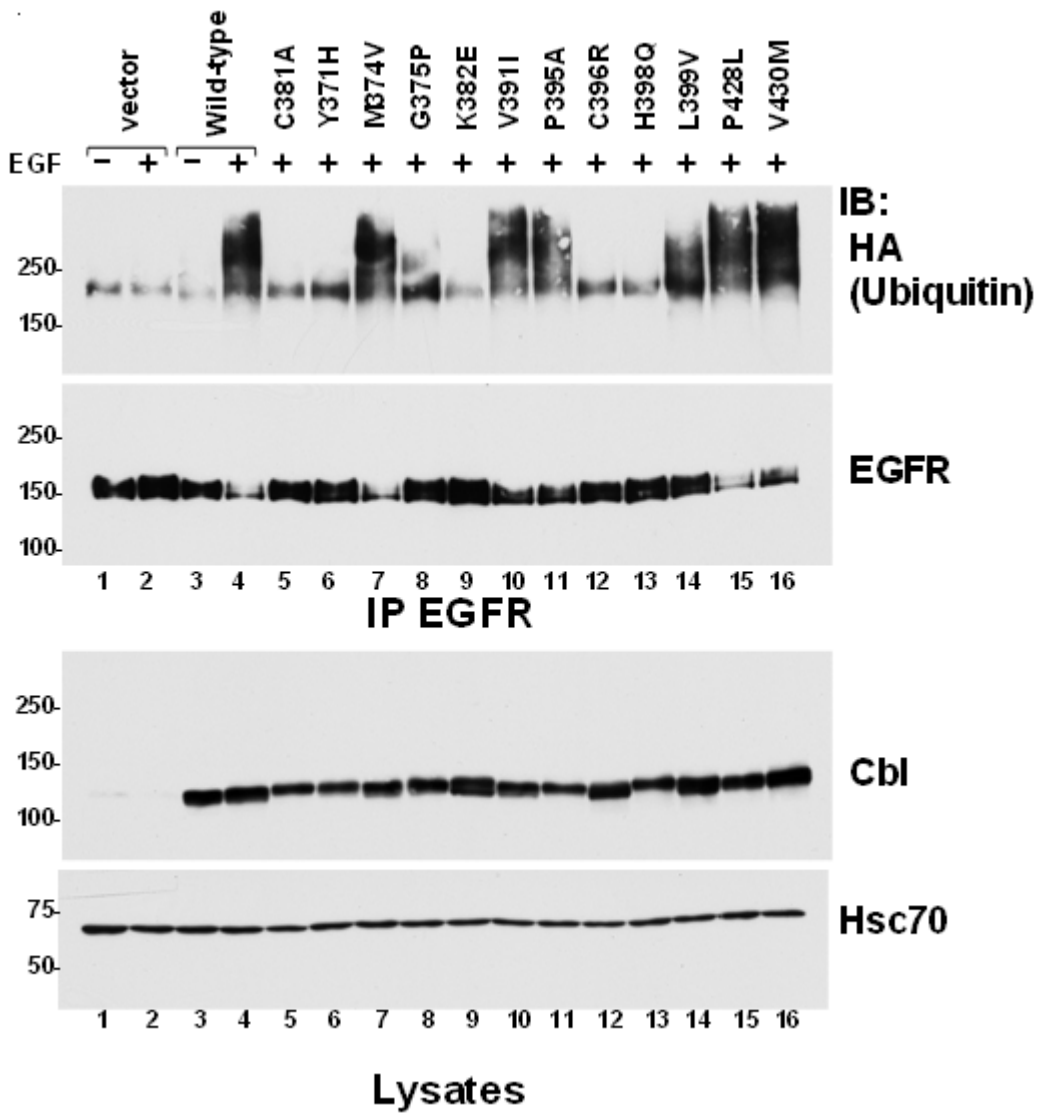
Li et al, *Cancer Research, 2016*

Homozygous mutations and mutations found in Zn-clusters and leukemia patients have largest effects

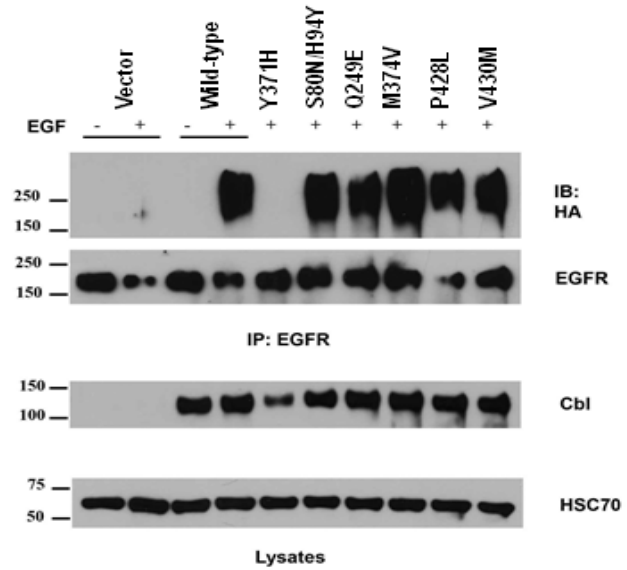


Experimental verification of activity of CBL mutants

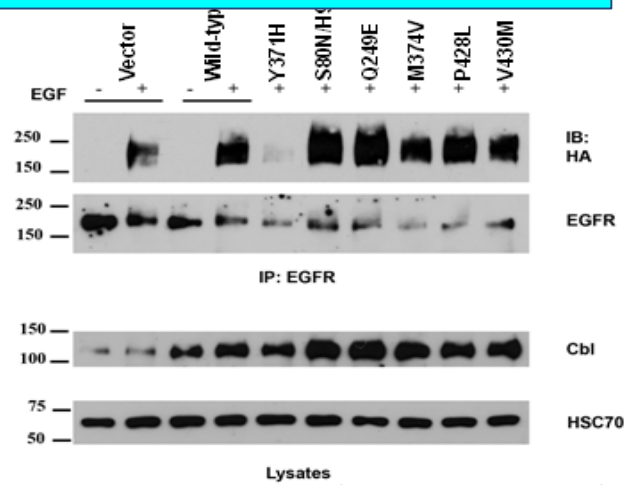
HEK293T - human embryonic kidney cells



A549 - human nonsmall cell lung cancer cells



HELA - cervical cancer cells



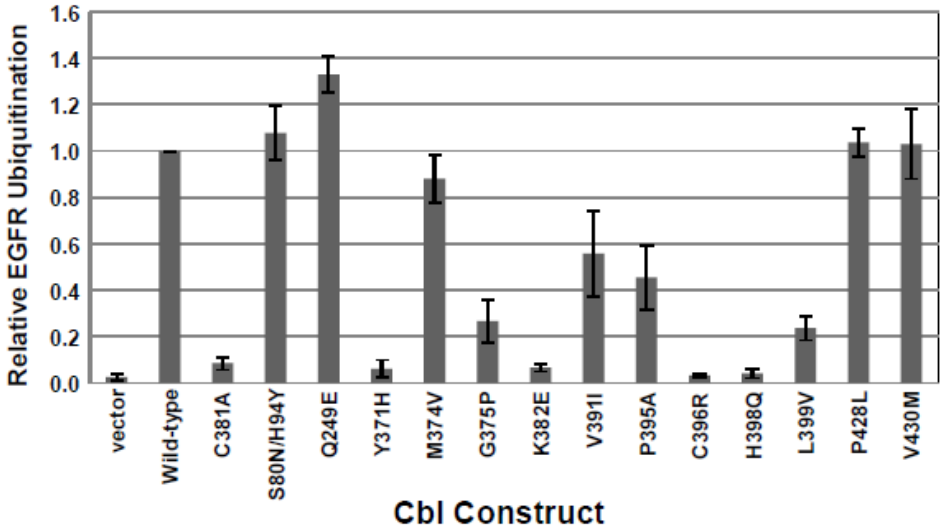
Li et al, *Cancer Research, 2016*

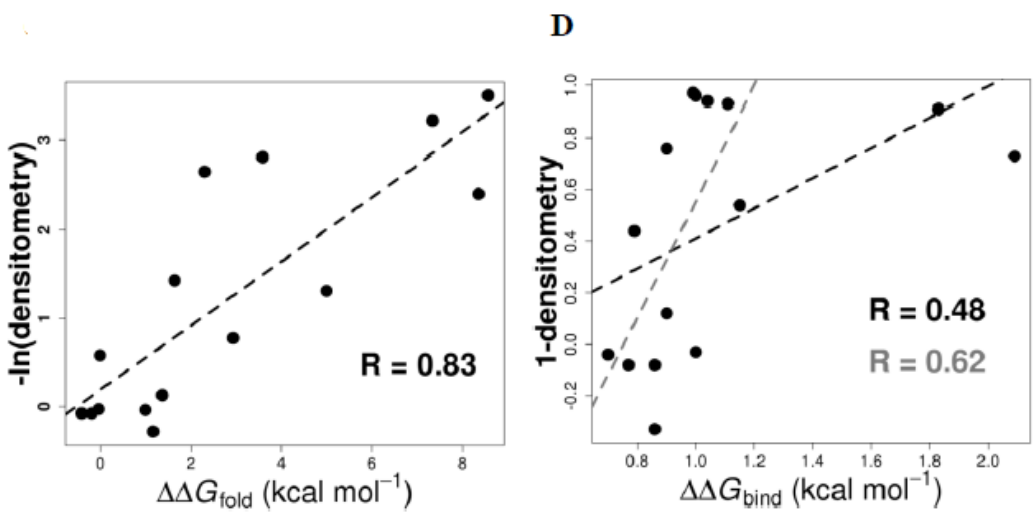
Comparing experiments and computational models

Mutations	Densitometry	Stability		Binding affinity		Methods to predict phetypic effects			
		nCBL	CBL-S	CBL-E2-S	pCBL-E2-S				
C396R	**0.03±0.007**	**4.65**	**8.57**	**0.98**	**0.99**	**-11.39**	**0.99**	**4.69**	**0.88**
H398Q	**0.04±0.019**	**7.28**	**7.34**	0.79	**1.00**	**-7.57**	**1**	**4.34**	**0.73**
Y371H	**0.06±0.038**	**3.58**	**3.43**	**0.98**	**1.04**	-4.70	**1**	**2.35**	**0.92**
K382E+	**0.07±0.016**	**2.30**	0.56	0.82	**1.11**	-3.74	**1**	**2.56**	**0.77**
C381A	**0.09±0.028**	**8.36**	**7.55**	**2.29**	**1.83**	**-8.48**	**1**	**4.72**	**0.59**
L399V	*0.24±0.051*	1.64	0.22	0.80	0.90	-2.83	**1**	1.70	0.34
G375P+	*0.27±0.094*	-0.19	**5.00**	0.67	**2.09**	**-7.56**	**1**	**2.09**	**0.94**
P395A	*0.46±0.138*	**1.97**	**2.93**	0.66	**1.15**	**-7.54**	**1**	**2.54**	0.34
V391I	*0.56±0.185*	-0.14	-0.01	0.81	0.79	-0.37	0.13	0.45	**0.78**
M374V+	0.88±0.104	1.36	0.87	0.81	0.90	-3.56	0.83	**2.28**	0.35
V430M	1.03±0.150	-0.04	-1.15	0.87	**1.00**	-2.19	**1**	**2.16**	0.35
P428L	1.04±0.059	0.80	0.99	0.85	0.70	-4.29	**0.98**	**2.13**	**0.71**
S80N	1.08±0.115	-0.42	-0.5	0.71	0.86	-2.69	**1**	**2.56**	**0.84**
H94Y	1.08±0.115	-1.00	-0.2	0.80	0.77	-4.26	**1**	**2.22**	**0.78**
Q249E	1.33±0.077	0.88	1.16	0.79	0.86	-2.81	**1**	**2.78**	**0.84**
Cutoff		1.80	2.04	0.87	0.95	-4.75	0.87	2.07	0.49

Li et al, *Cancer Research, 2016*

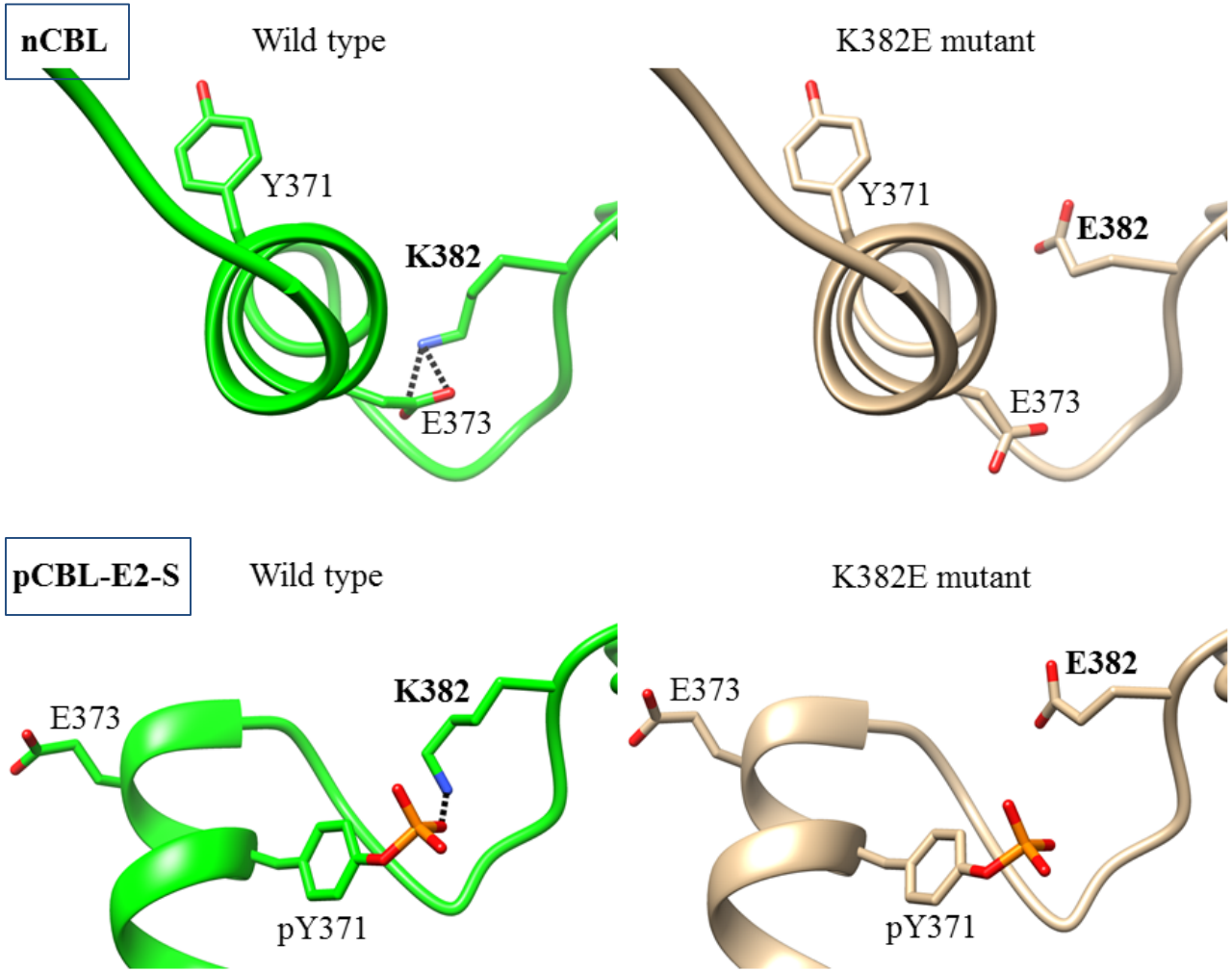
Stability-activity tradeoff





$$\frac{D_{mut}}{D_{WT}} \sim e^{-\Delta\Delta G_{fold}}$$

K382E mutations significantly destabilizes the closed and active CBL states

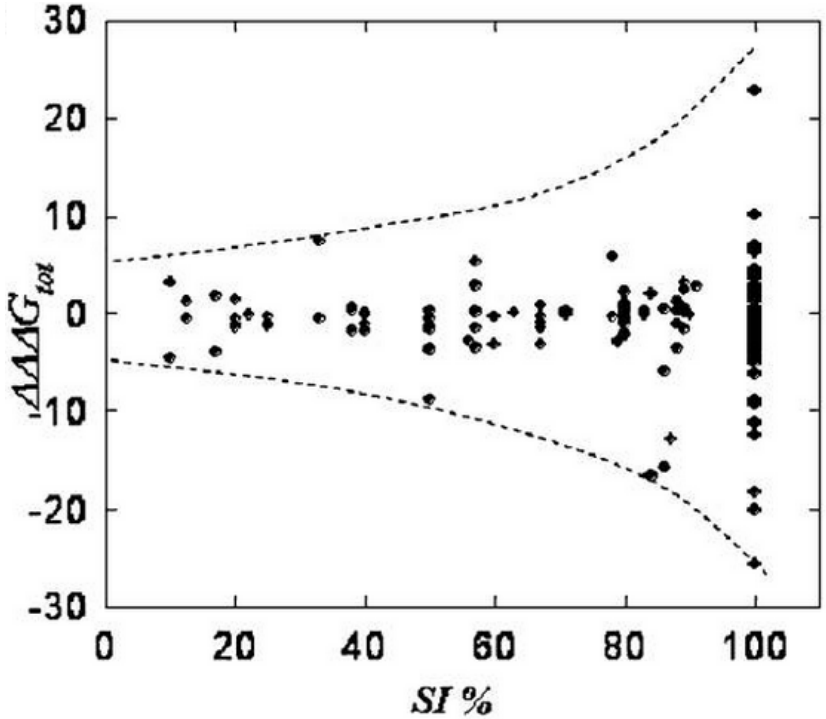


Effect of OMIM nsSNPs on protein complex stability

Mean values of ΔΔΔG distributions

	ΔΔΔG(total) kcal/mol	ΔΔΔG(van der Waals) kcal/mol	G(electrostatic) kcal/mol
OMIM	-1.65	-1.03	-2.35
Non-OMIM	-0.70	0.14	-0.45
	p-value > 0.01	**p-value > 0.01**	**p-value =0.006**

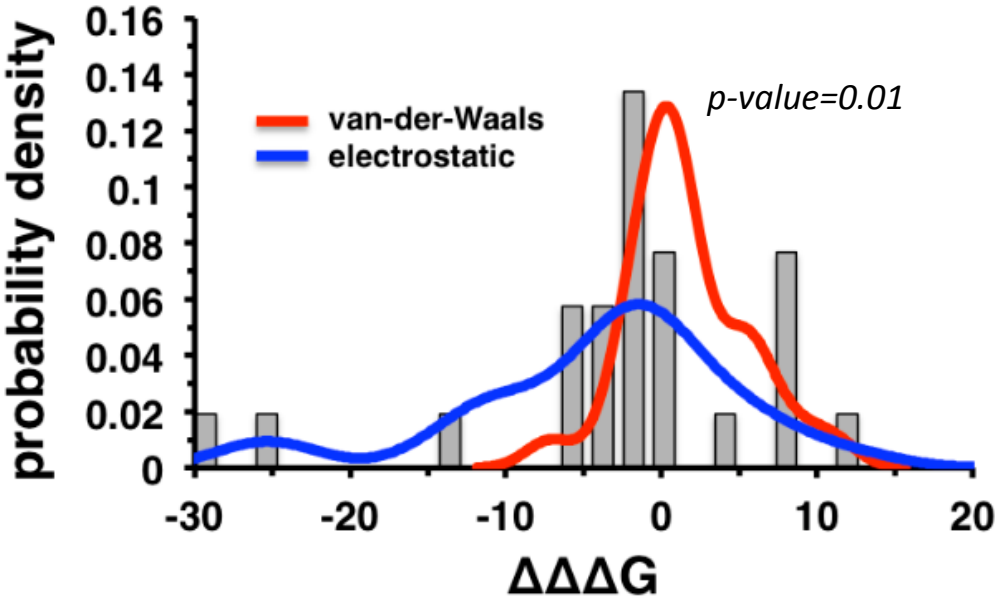
- OMIM mutations destabilize electrostatic components of binding energy;
- Largest effect of mutations is observed at evolutionary conserved sites.



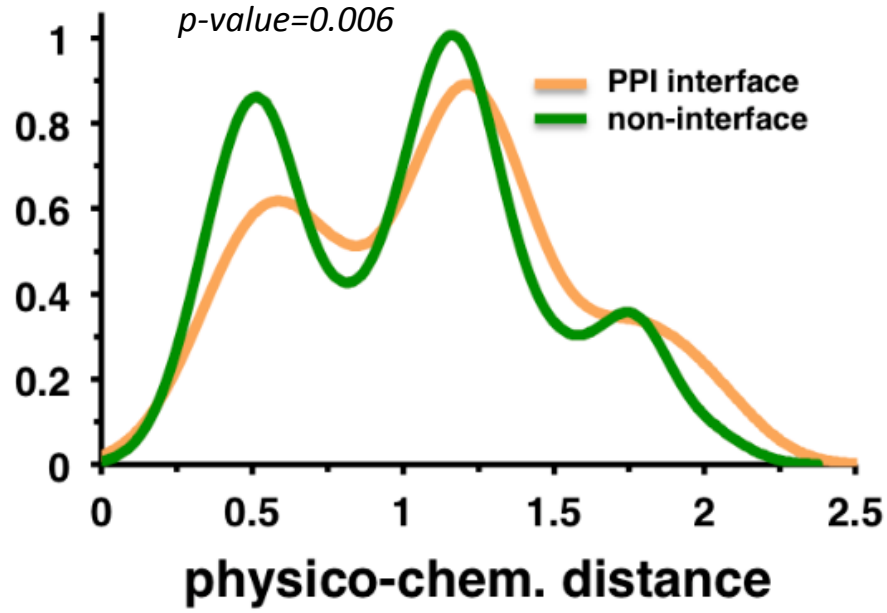
Teng et al, *Biophysical Journal*, 2009

Effect of glioblastoma mutations on protein binding

Binding energy difference upon mutation for electrostatic and van-der-Waals components

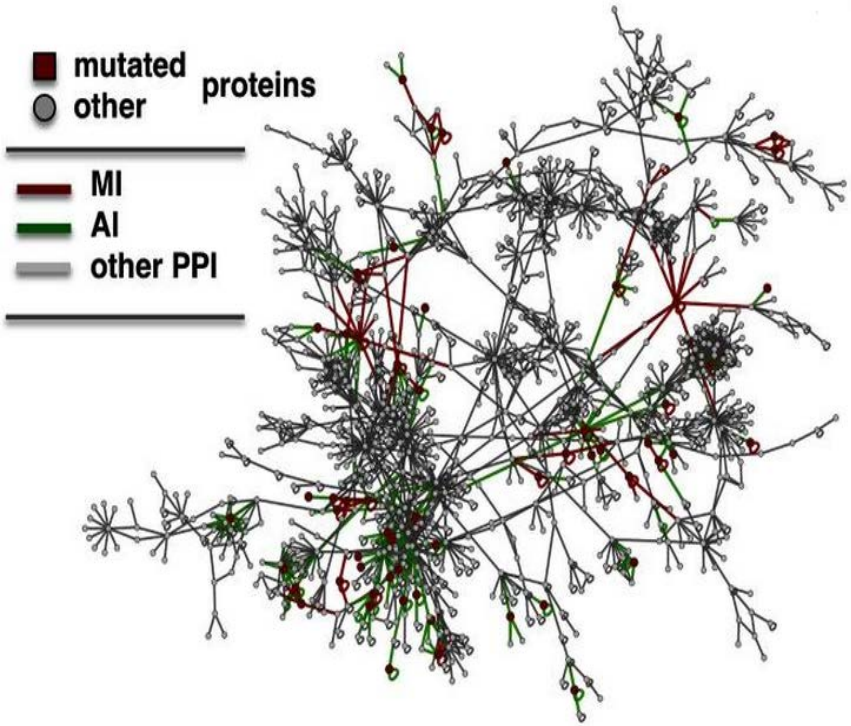


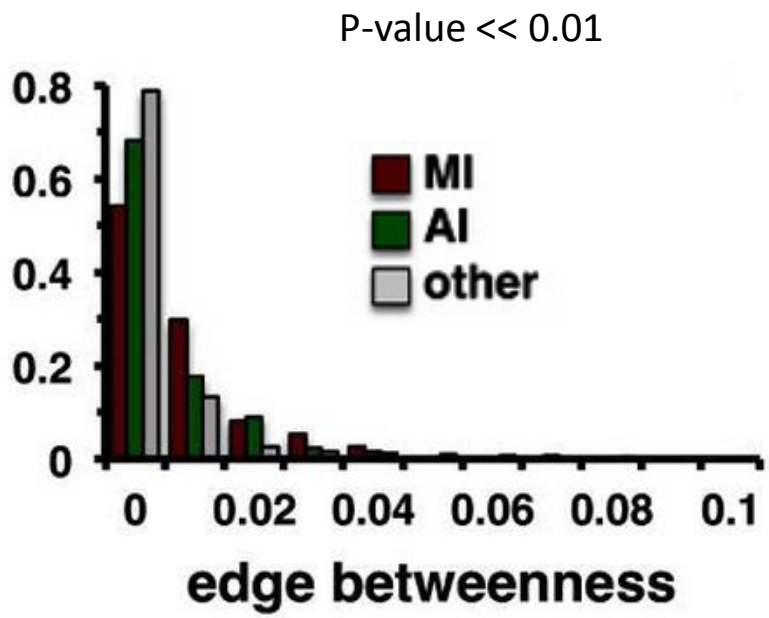
Physico-chemical distances between mutations on protein-protein interfaces and non-interface regions



Nishi et al, *Plos One*, 2013

Topological properties of mutated gene network





number of shortest paths going through a node

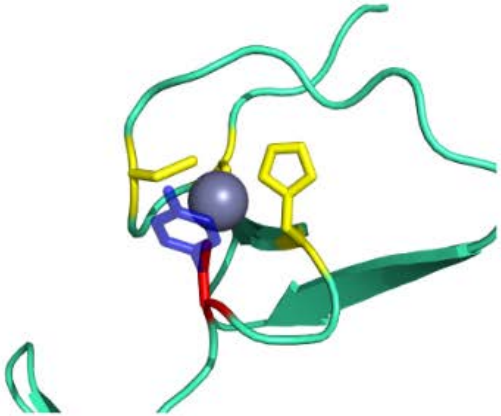
AI – 444 interactions between proteins with mutation anywhere in protein

MI – 160 interactions between proteins with mutation on interface

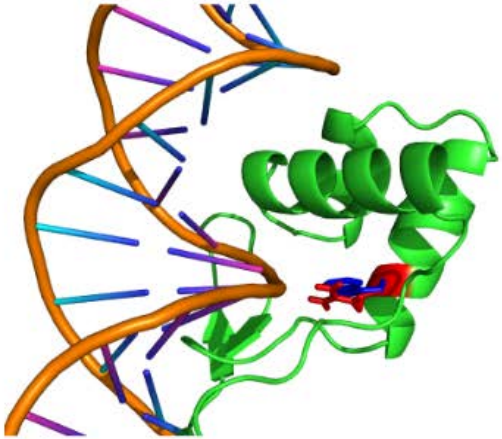
Interactions with mutations occur in central network positions!

Predicted driver mutations

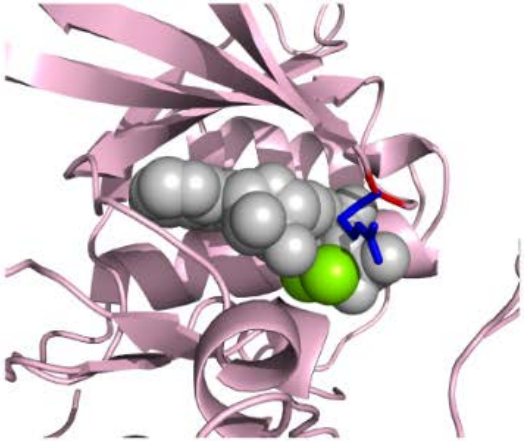
Protein-protein
ABL2
ARL1
EPHA2
IDH1
NLGN2
NRAS
RAB3C
RAC2
RAD52
TP53
Protein-DNA
BCL11A
PAX9
TP53
ZIK1
ZNF339
Protein-RNA
ELAVL2
KLK9
RBMS3
RPL11
Protein-ion
ADAMTS17
DSG4
GZMH
HPCAL4
LCT
LMX1A
MAPK9
NELL2
SGK2
TP53
ZIK1
ZNF497



Zinc binding motif of LMO-2 (homolog of LMX1A) , C→Y

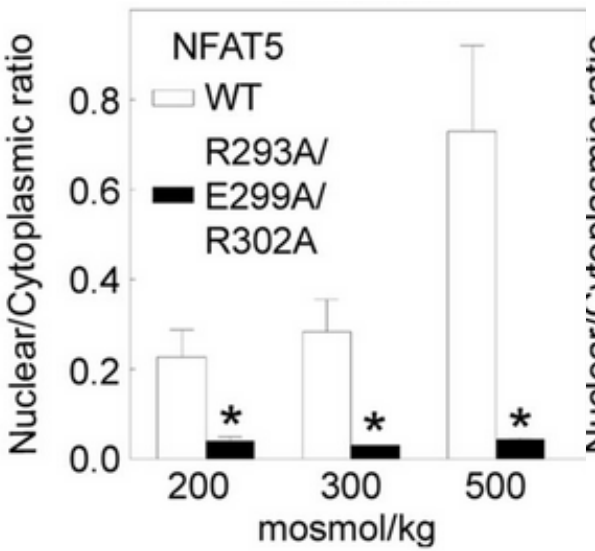


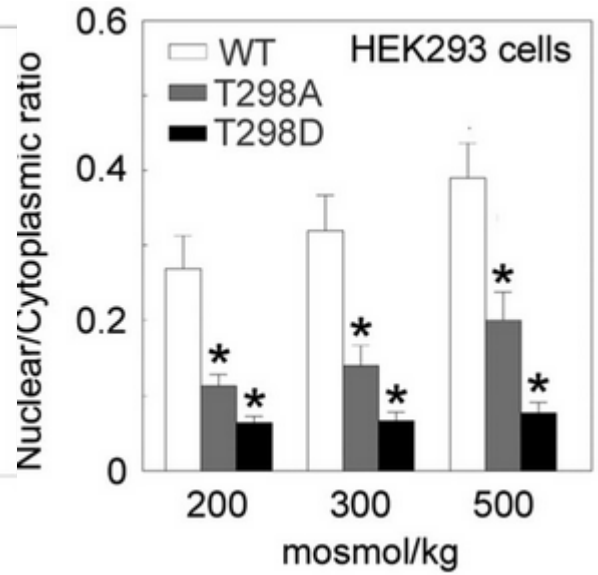
DNA binding site of Pax-6, R→W

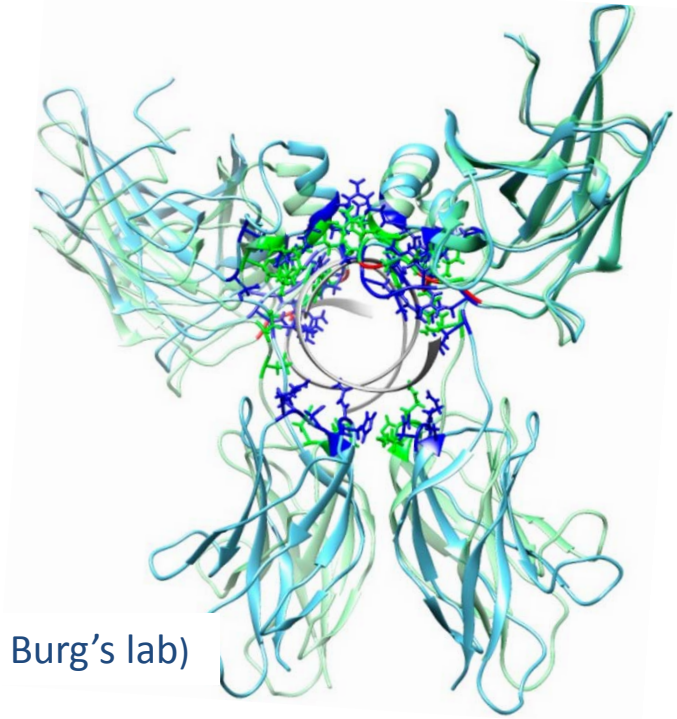


Protein-ion binding site of MAPK10, G→R

Mutations in DNA-binding loop of NFAT5 produce unique outcomes on binding and dynamics







Izumi et al, *Am J Physiol Cell Physiol.* 2012 (experimental data from M. Burg's lab)

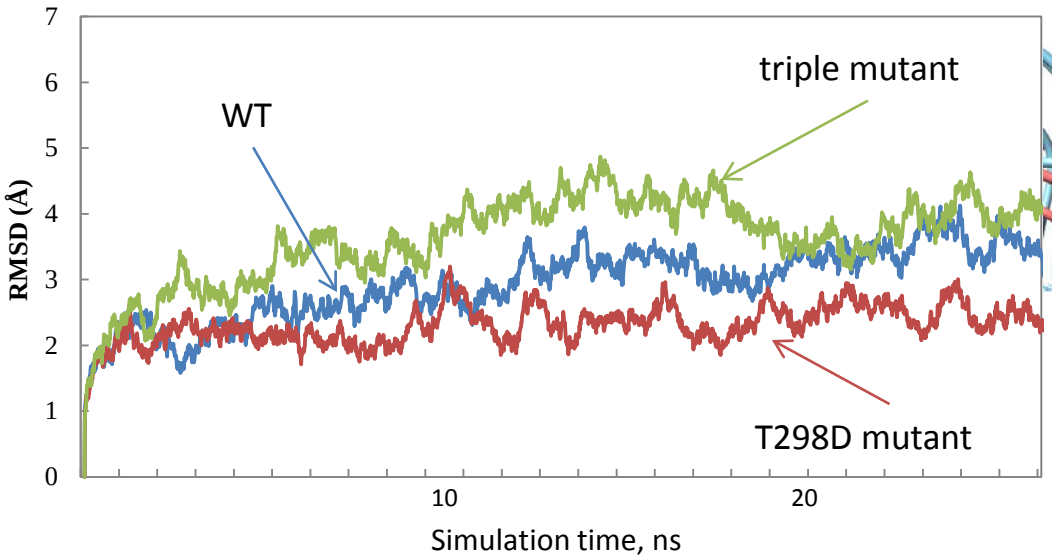
Binding energy, kcal/mol

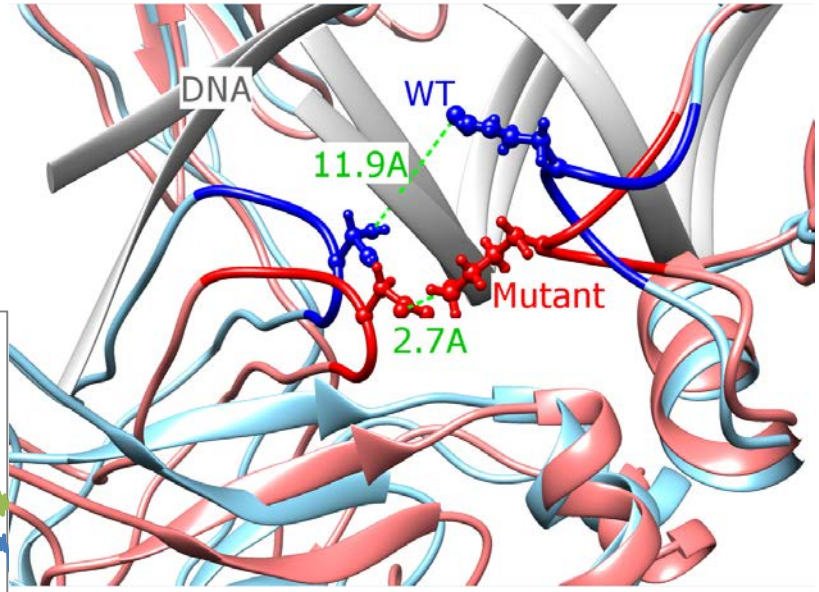
	Complex - DNA	Chain C - DNA	Chain D - DNA	Chain C – Chain D
Native	10.00 (0.62)	3.61 (0.39)	2.89 (0.39)	2.73 (0.23)
T298D	9.24 (0.54)	3.79 (0.23)	2.76 (0.23)	2.92 (0.23)
R293, E299, R302	6.47 (0.62)	3.13 (0.23)	2.36 (0.23)	2.78 (0.31)

Li et al, *J of Phys Chem*, 2013

Phosphomimetic mutation T298D constraints movements of two chains

Changes in conformation , Molecular Dynamics simulations (NAMD program, explicit solvent)

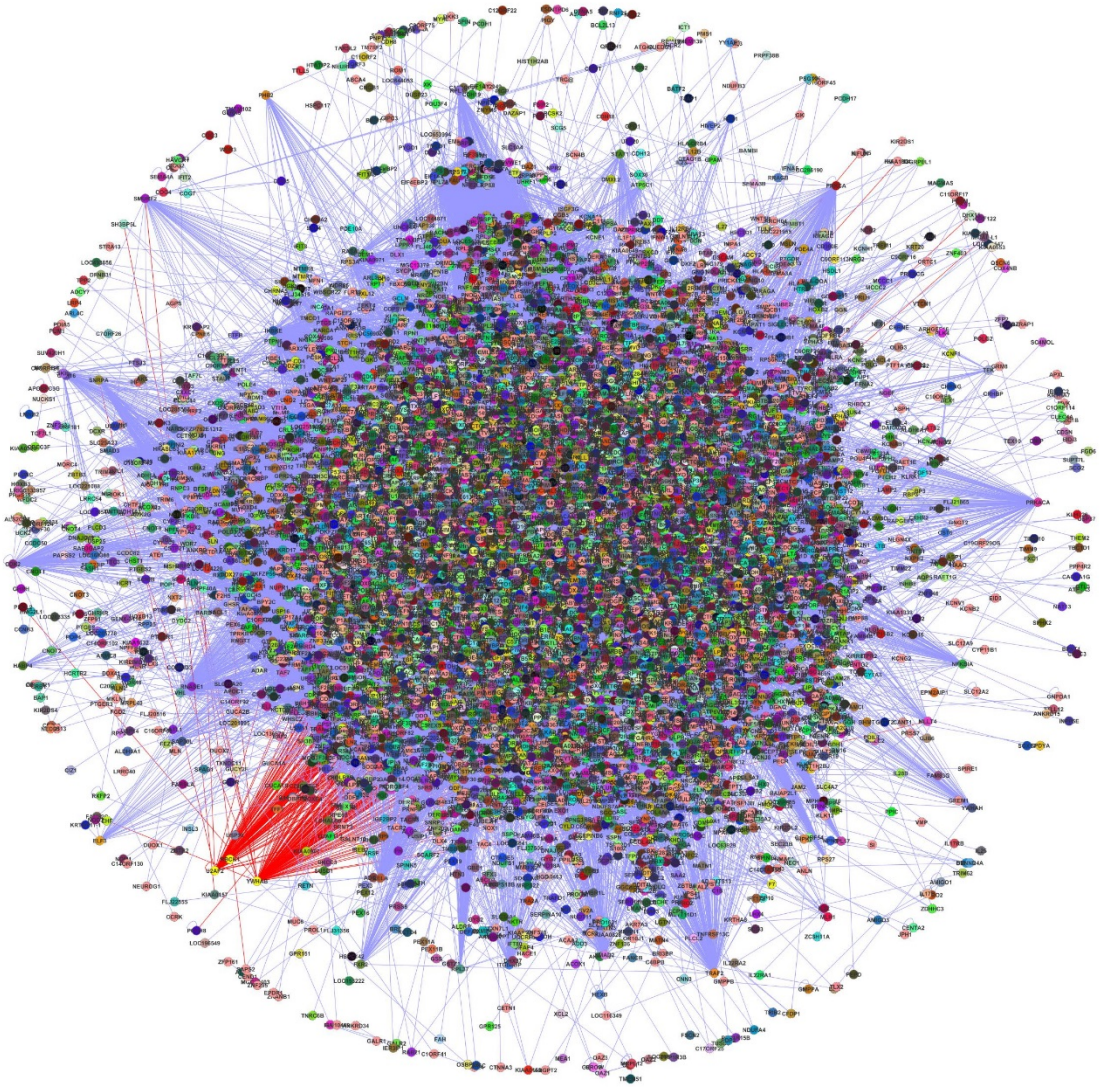




Effect of T298D on structure, formation of an extra salt bridge between two chains in a dimer

Li et al, *J of Phys Chem*, 2013

Reality: PPI Interactomes



All hairball graphs look alike

Protein interactions in the cell

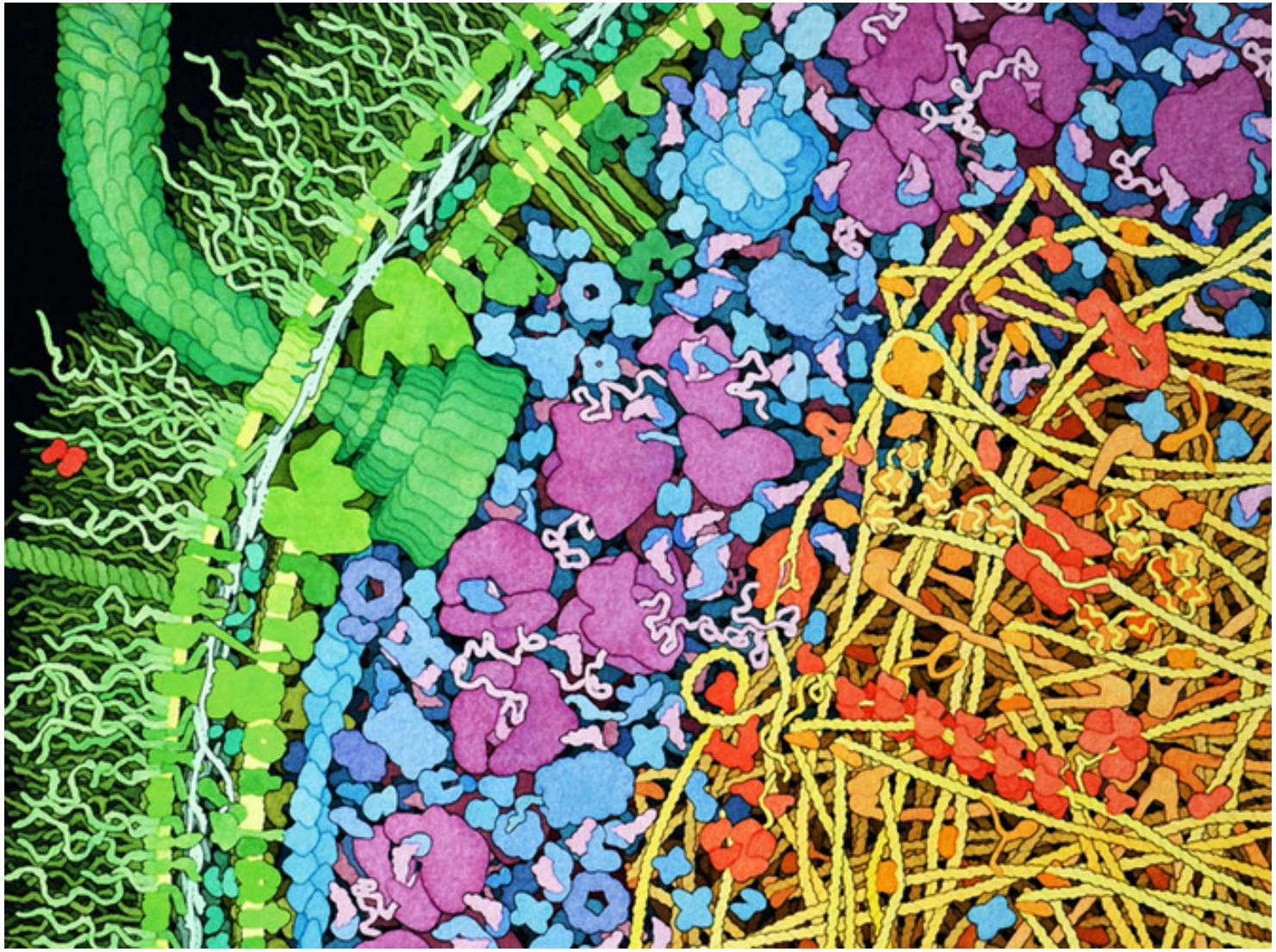


Illustration by David S Goodsell

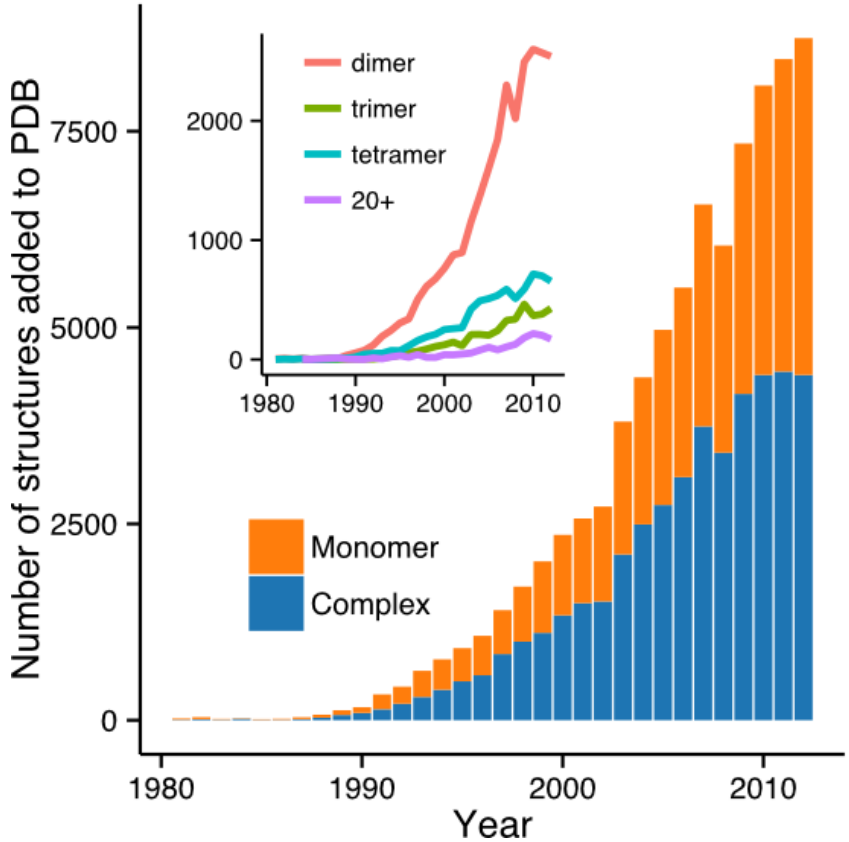
Structural interactomes are informative and useful

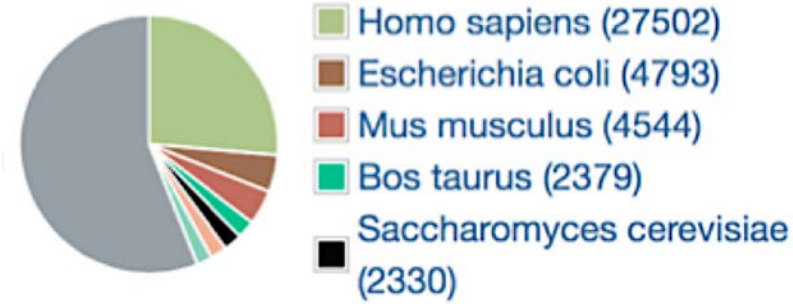
- Interactome with structural details:
  - Which proteins interact?
  - How they interact:
    - Which domains interact?
    - Which residues form binding sites?

- Atomic-resolution interfaces are needed to study:
  - The mechanisms of interactions.
  - The effects of mutations on stability of proteins and their complexes.
  - To modulate interactions (drugs)

- Strategies:
  - Use available structures of protein complexes.
  - Dock structural monomers if structural complex is not available.
  - Template-based modeling of protein complexes (or interfaces).

What do we have so far? Growth of structural and PPI data

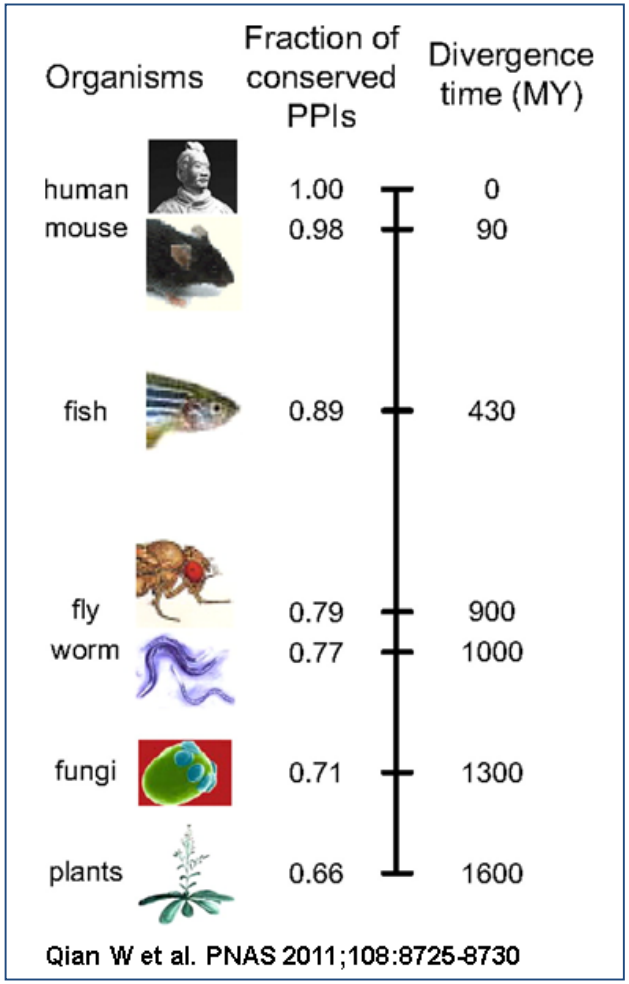




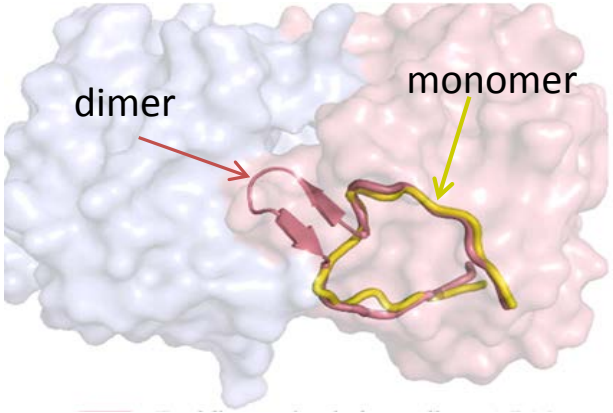
Conservation of protein interactions and oligomeric states

Conservation of interactions partners

Rate of PPI evolution = $(2.6 \pm 1.6) \times 10^{-10}$ per PPI per year



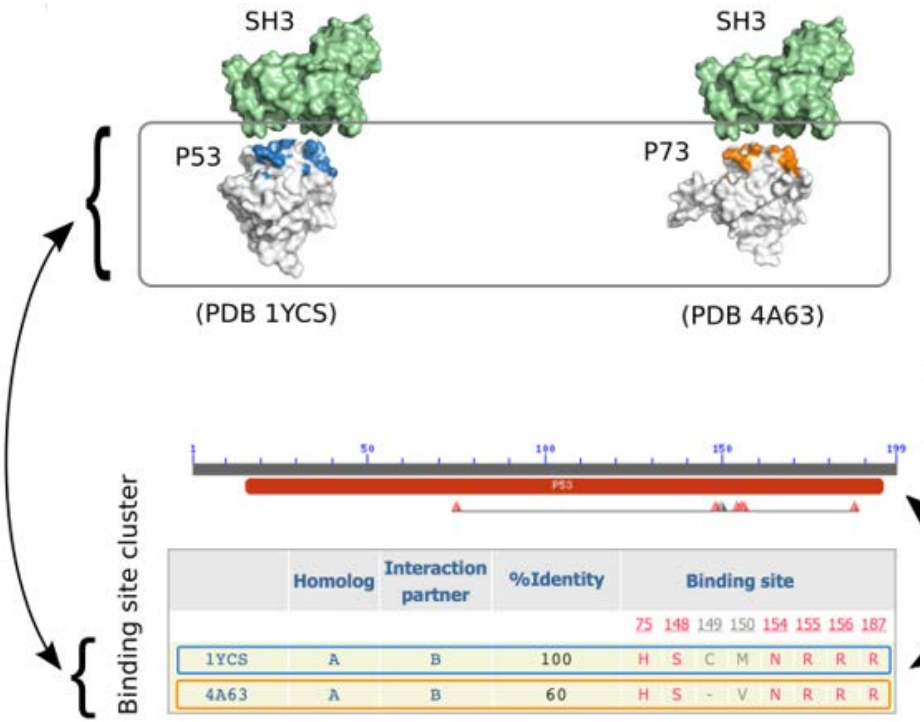
Conservation of oligomeric states

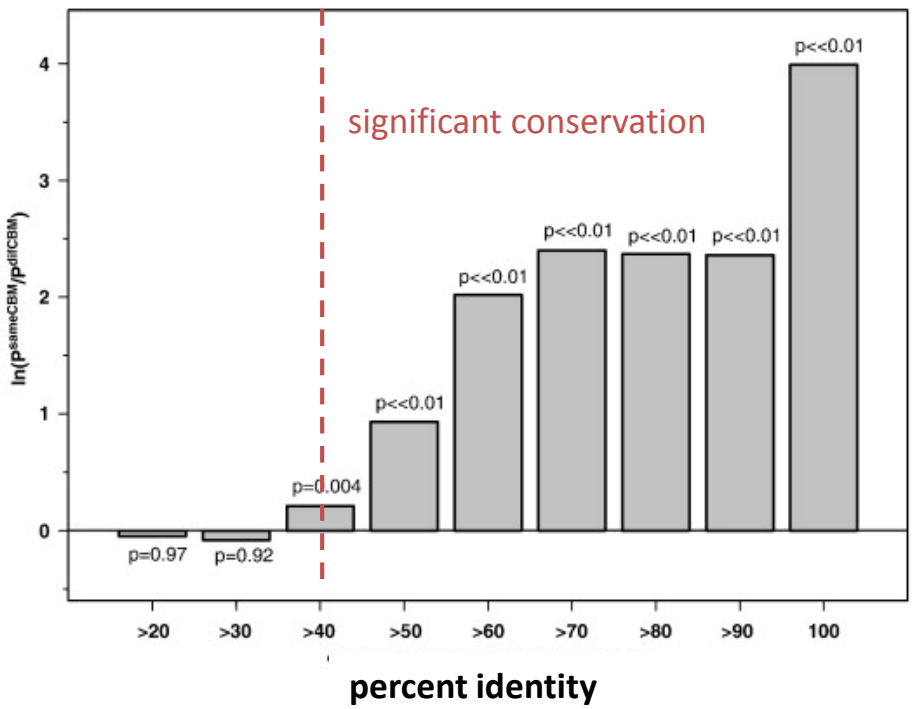


	Sensitivity TP/(TP + FN)	Specificity TN/(FP + TN)
Presence/absence of enabling and disabling features	0.70	0.74
Percent identity	0.71	0.62
RMSD	0.72	0.60

Hashimoto et al, *PNAS*, 2010

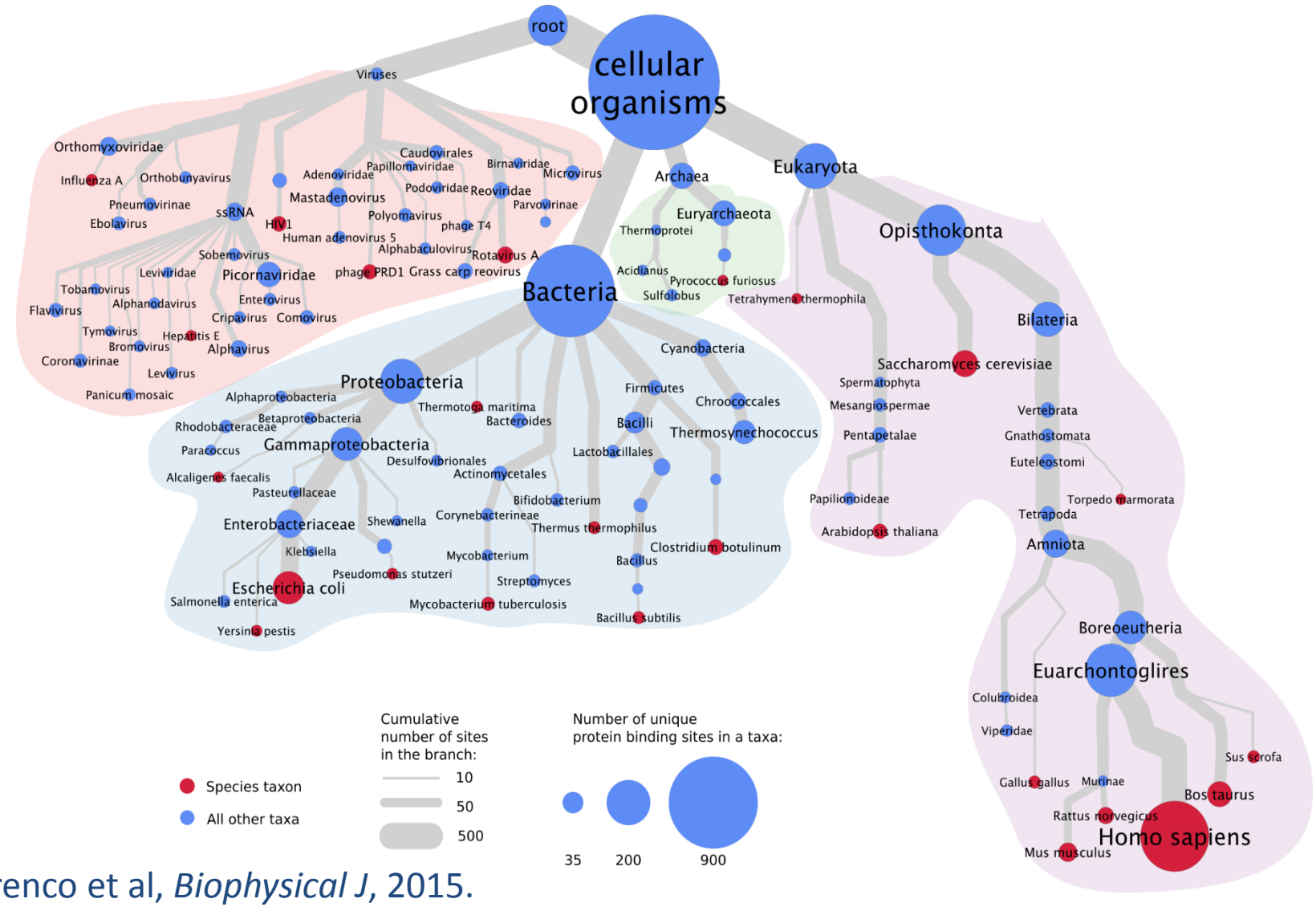
Different protein complexes might have similar binding interfaces





Dayhoff et al, *J Mol Biol*, 2010
Goncearenco et al, *Biophysical J*, 2015

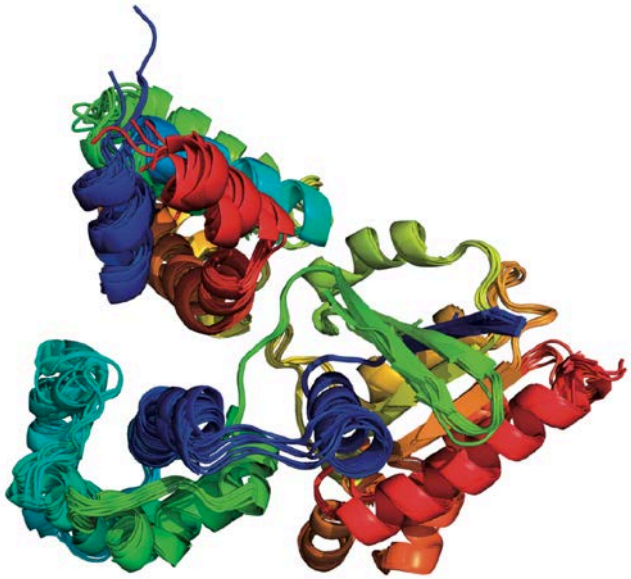
Tracing back evolution of protein binding sites to the root of all organisms

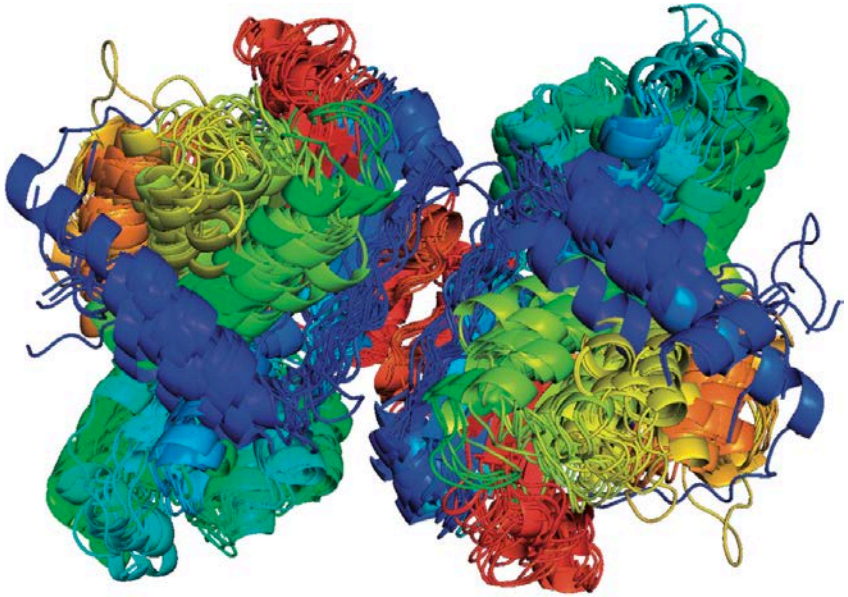


Goncearenco et al, *Biophysical J*, 2015.

The underlying interolog hypothesis

- If proteins are similar they may interact in a similar way
- Homologs may have similar interfaces





IBIS – NCBI server to analyze interactions and binding sites http://www.ncbi.nlm.nih.gov/Structure/ibis/ibis.cgi

Observed interactions – from structural complexes

Inferred interactions – from homologous structures with observed interactions

Types of interactions: protein-protein, protein-nucleic acids, protein-small molecule, protein-peptide, protein-ion

Biological relevance of binding sites:

- occurs in several non-redundant homologs;
- structurally and sequence conserved;
- validated biounit

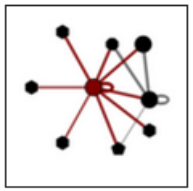
Shoemaker et al, *Nucleic Acids Res*, 2010
Shoemaker et al, *Nucleic Acids Res*, 2012

IBIS
Inferred Biomolecular Interactions Server

Protein PDB ID, Accession or GI | New Search

Query 1YCR_A

Mdm2



All interactions for query sequence

Download data

EXCEL ◉ XML ○

Search 1YCR A interactions

Similarity to query

- Sequence Identity: %Id list
- Structure RMSD: RMSD list

Interaction partner type

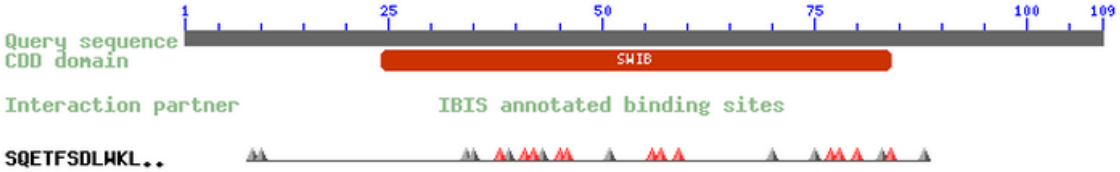
- PDB Code: PDB code list
- Taxonomy: Taxonomies list

Reset

Protein-Protein (8) | Protein-Chemical (4) | Protein-Peptide (1) | Protein-Ion (1)

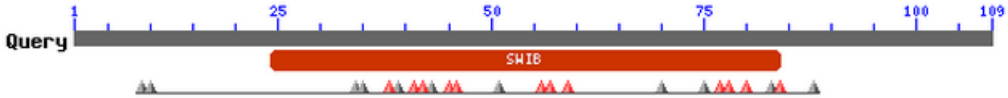
Mdm2 (sequence A)

1YCR sequence B



List of peptide interaction partners and binding sites. Similar binding sites of homologs of the query are grouped into clusters. To view the cluster members click on the plus sign. "**o**" denoted observed interactios.

	Interaction Partner	Ranking Score	Number of Cluster Members	Average %Identity to Query	Number of Binding Site Residues	Taxonomic Diversity
⊟ o	SQETFSDLWKLLPEN	n/a	24	83	23	Euteleostomi



View Binding Sites

Download Cn3D

tion er	%Identity to query	Binding site																						
	-	9 E	10 T	34 M	35 K	38 L	39 F	41 L	42 G	43 Q	45 I	46 M	51 Y	56 Q	57 H	59 V	70 F	75 F	77 V	78 K	80 H	83 I	84 Y	88 Y
	100	E	T	M	K	L	-	L	G	-	I	M	-	Q	H	V	-	-	V	K	H	-	Y	Y
	73	-	-	-	-	I	-	L	G	-	I	M	Y	Q	-	-	-	-	V	K	P	L	Y	-
	58	-	-	V	K	M	-	L	G	-	I	M	Y	Q	H	V	-	-	V	-	-	L	Y	-
	58	-	-	-	-	M	-	L	G	-	I	M	Y	Q	H	V	-	-	V	-	-	-	Y	-

See all members

* click structure accession or sequence letter to explore structure and sequence information.

NCBI

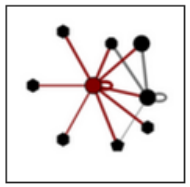
HOME | SEARCH | SITE MAP

IBIS

Inferred Biomolecular Interactions Server

Query 1YCR_A

Mdm2



All interactions for query sequenc

Download data

EXCEL ◉ XML ○

Search 1YCR A interactions

Similarity to query

☐ Sequence Identity: %Id list

☐ Structure RMSD: RMSD list

Interaction partner type

☐ Chemical: Compounds list

☐ PDB Code:

☐ Taxonomy

Compounds list
1-{[(5r,6s)-5,6-Bi
28W
(2\'s,3r,4\'s,5\'r)
3-{(1s)-2-(Tert-Bu
3-{(1s)-2-(Tert-Bu
4-({(4s,5r)-4,5-Bi
[(4s,5r)-2-(4-Tert-
(4s,5r)-2-(4-Tert-
[(4s,5r)-4,5-Bis(4-
(4s,5r)-4,5-Bis(4-
(5r,6s)-2-[((2s,5r)
(5z)-5-[(6-Chloro-
AC1L9LXV
AC1L9MCG
CHEMBL2024323
CHEMBL2059435
CHEMBL2177187

Clusters with

Show: ○

Referenc

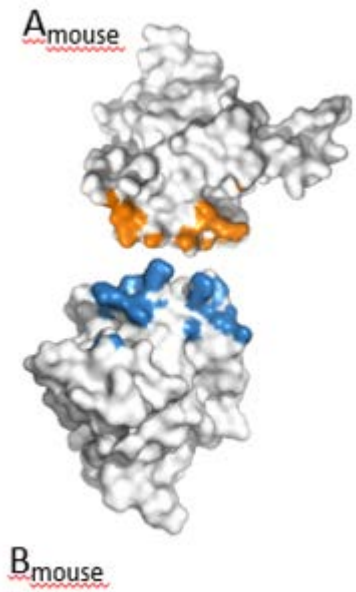
Shoem integrates

Shoem predict pro

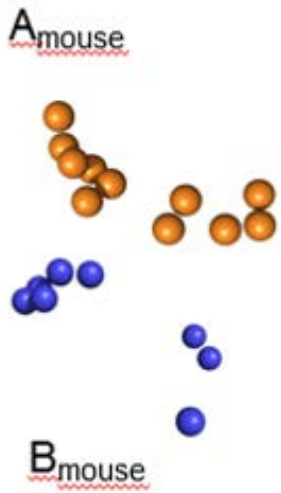
☐	Chemical	Homologous complex	Chain	%Identity to query	Number of binding site residues	Binding site	Taxonomic Diversity
Query		-	-	-	-	34 38 39 41 42 43 45 46 51 55 56 57 59 66 70 75 77 78 80 83 84 M L F L G Q I M Y Q Q H V L F F V K H I Y	-
☑	DIZ	1T4E	A	100	12	- L - L G - I M Y - Q - - - F F V - H I Y	Homo sapiens
☐	DIZ	1T4E	B	100	13	- L - L G - I M Y - Q - - - F F V - H I Y	Homo sapiens
☐	MI6	3LBL	A	100	11	- L - L G - I M - - - - - - F F V - H I Y	Homo sapiens
☐	CHEMBL2177187	4ERE	A	100	11	- L F - G - I - Y - - - - - F - V K H I Y	Homo sapiens
☐	CHEMBL2059435	4ERF	A	100	11	- L - L G - I M Y - - - - - - - V K H I Y	Homo sapiens
☐	SureCN9993627	4OAS	A	100	11	- L F - G Q I - - - - - - - F - V K H I Y	Homo sapiens
☐	CHEMBL2347399	4JV7	A	100	12	- L F L G Q I M Y - Q - - - - F V - - I -	Homo sapiens
☐	CHEMBL2347401	4JV9	A	100	11	- L F L G Q I M Y - Q - - - - - V - - I -	Homo sapiens
☐	CHEMBL2347383	4JVE	A	100	12	- L F L G Q I M - Q - - - - - - V - H I Y	Homo sapiens
☐	(2's,3r,4's,5'r)-N-(2-Aminoethyl)-6-Chloro-4'-(3-	4JVR	A	100	12	- L - L G - I M Y - - - - - F F V - H I Y	Homo sapiens
☐	CHEMBL2347393	4JWR	A	100	10	- L - L G - I M - - - - - - F - V - H I Y	Homo sapiens
☐	3-{(1s)-2-(Tert-Butylamino)-1-[(4-Chlorobenzyl)(F	3TJ2	A	100	13	- L F - G - I M Y - - - V - F F V - H I Y	Homo sapiens
☐	SureCN9993362	4HBM	A	100	10	- L - L G - I - - - - - - - F - V K H I Y	Homo sapiens
☐	3-{(1s)-2-(Tert-Butylamino)-1-[{4-[(4-Chlorobenzy	4MDN	A	100	12	M L F L G - I M - - - - - - F - V - H I Y	Homo sapiens
☐	28W	4MDQ	A	100	10	- L F - G - I - Y - Q - - - - F V - H I -	Homo sapiens
☐	1-{[(5r,6s)-5,6-Bis(4-Chlorophenyl)-6-Methyl-3-(P	3VZV	A	99	11	- L - L G - I M Y - Q - - - - - V - H I Y	Homo sapiens
☐	(5r,6s)-2-[((2s,5r)-2-{[(3r)-4-Acetyl-3-Methylpip	3W69	A	99	12	- L - L G - I M Y - Q - V - - - V - H I Y	Homo sapiens
☐	(5z)-5-[(6-Chloro-7-Methyl-1h-Indol-3-	3VBG	A	99	6	- L - L G - I - - - - - - - - - V - - I -	Homo sapiens

gi M, Fong JH, Marchler-Bauer A, Bryant SH, Madej T, Panchenko AR. (2010), *Inferred Biomolecular Interaction Server--a web server to analyze and sites.* **Nucleic Acids Res. 38(Database issue): D518-24.**

Modeling of interactions and interfaces



Structural PPI complex



Library of interfaces

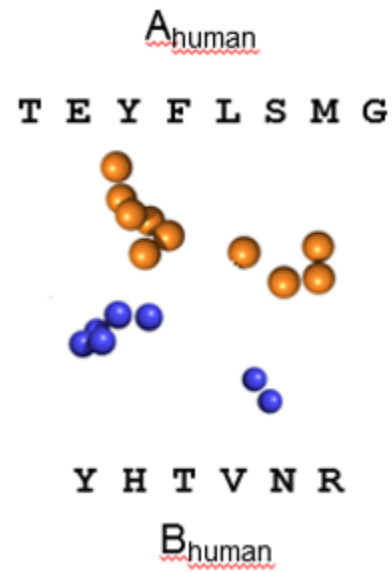
A_h: T E Y F L S M – G

A_h: T E Y F L S M – G
A_m: T D Y F R T M L K

Alignment of binding sites

B_m: Y H – S L V R
B_h: Y H L T V N R

B_h: Y H L T V N R

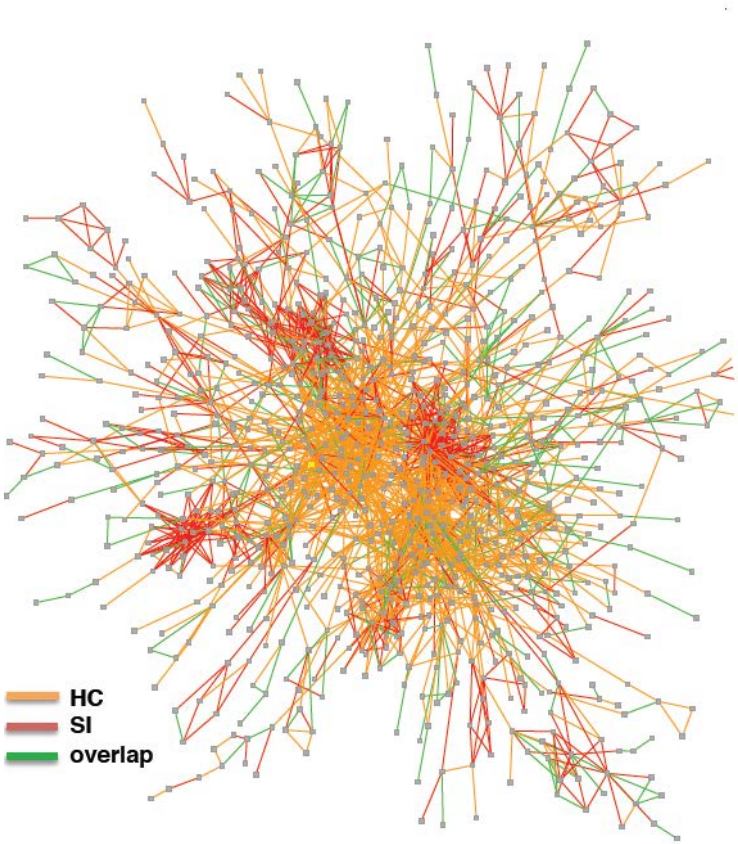


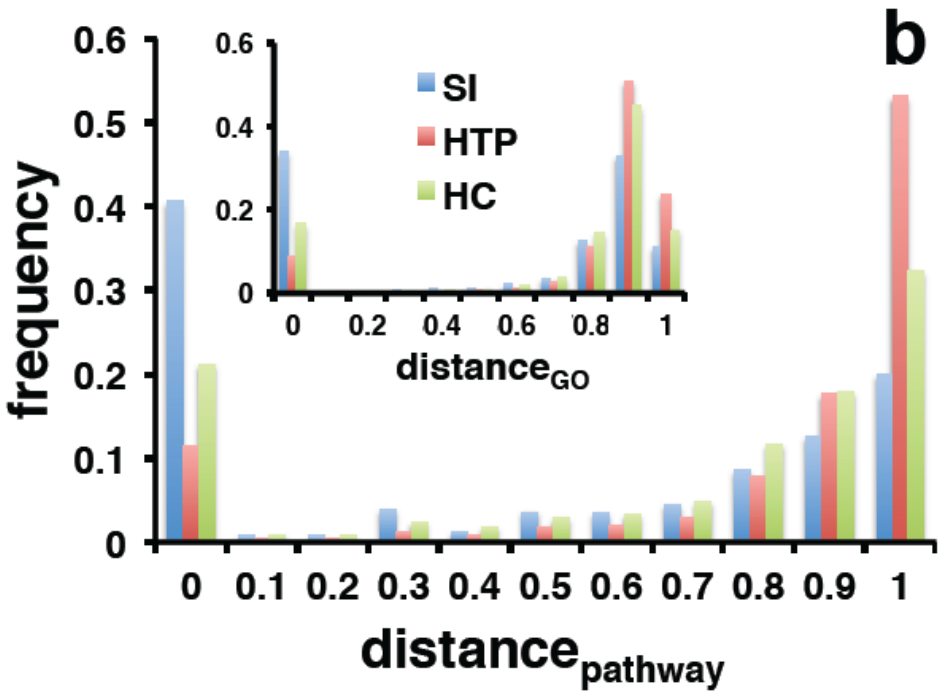
Model of human interface/complex

Verification:
- minimization of complex;
- interface complementarity;
- interface conservation;
- co-localization;
- co-expression.

Mapping of human interactome using structural complexes

4400 human genes
20000 interactions





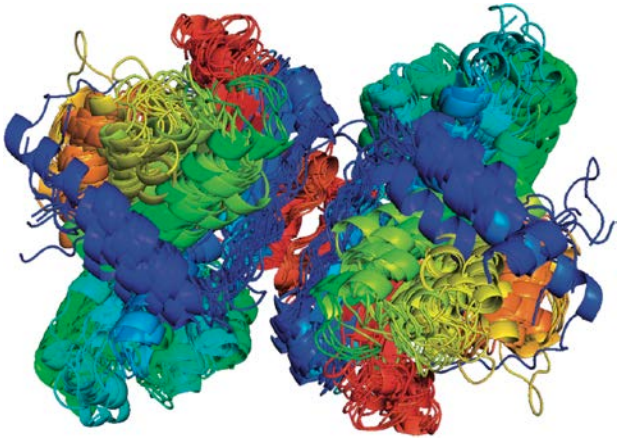
$$\Delta_{ij} = 1 - \frac{|\Gamma_i \cap \Gamma_j|}{|\Gamma_i \cup \Gamma_j|}$$

Γ_i - list of attributes for protein *i* (GO or pathway annotations),

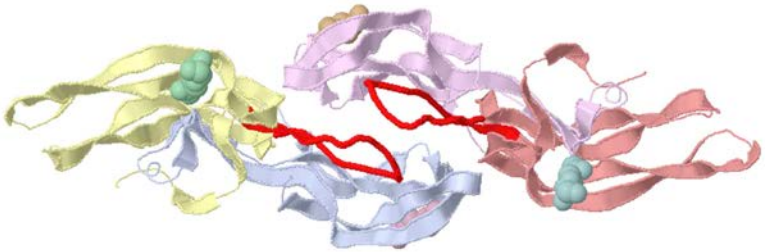
Structurally inferred networks (SI) are more functionally coherent than high-throughput networks (HTP, HC)

Tyagi et al, *EMBO Rep*, 2012

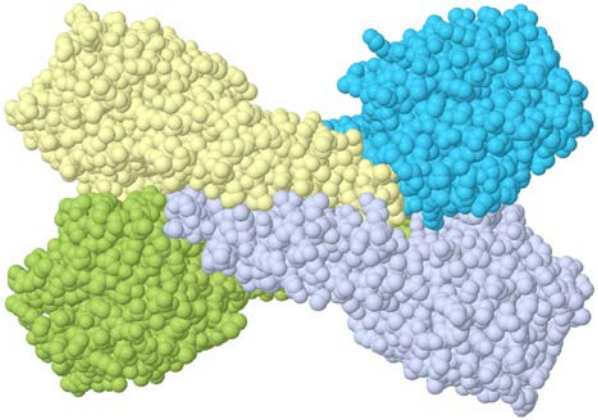
Challenges in computational analysis and prediction of PPI



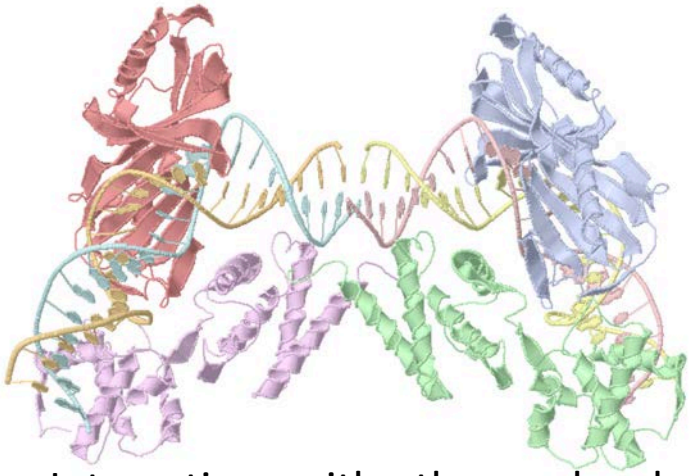
Substantial conformational changes



Highly variable regions



Higher oligomers



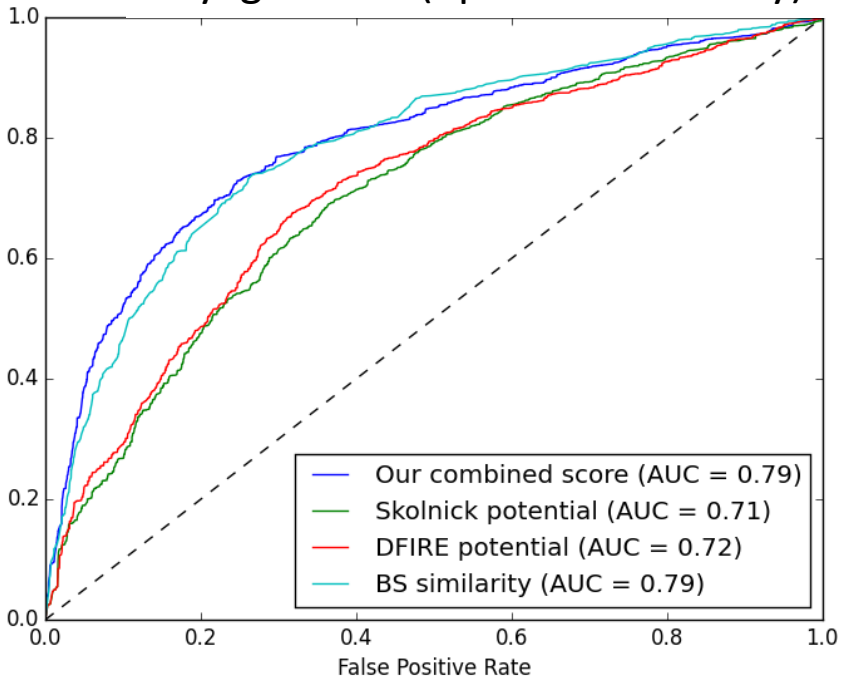
Interactions with other molecules

Benchmarking predicted PPIs

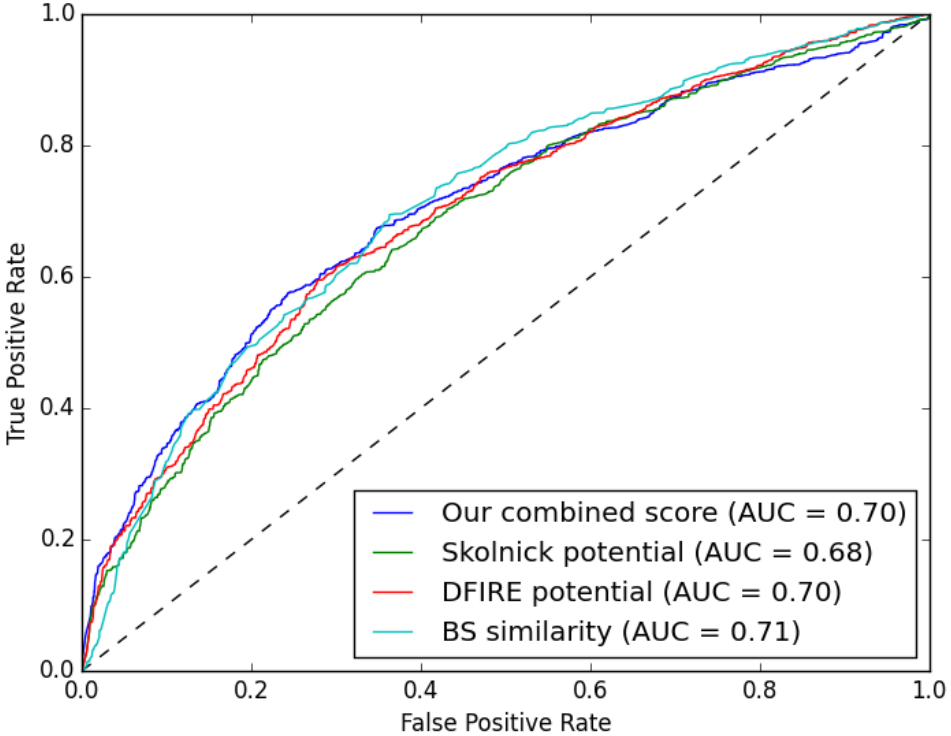
- Residue level:
  - Positive set: interface residues in PDB structures
  - Negative set: non-interface residues
- Protein level:
  - Positive set: pairs of protein with experimentally identified interactions (typically Y2H)
  - Negative set: ?

Residue level benchmark on structural complexes from yeast

Daylight zone (up to 98% identity)

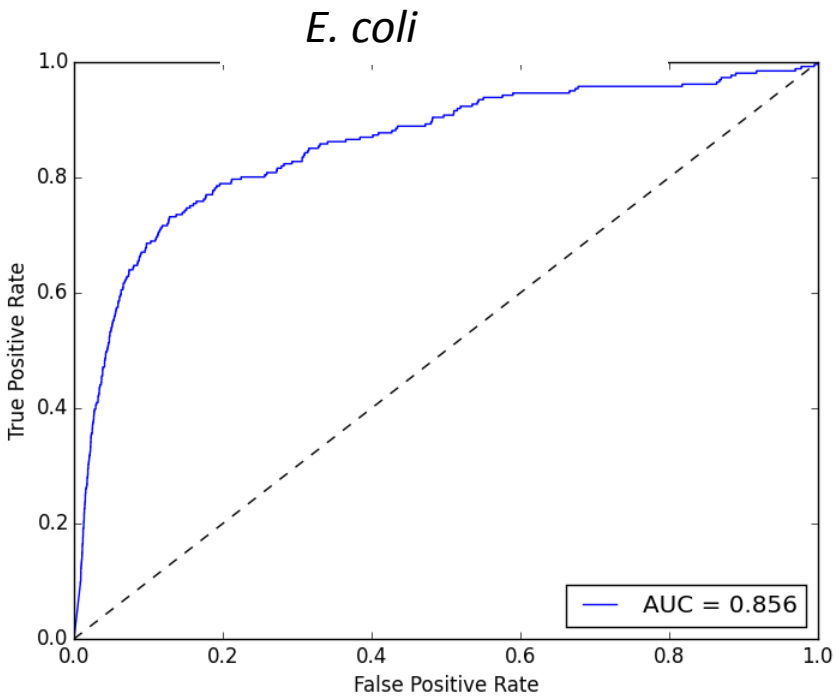


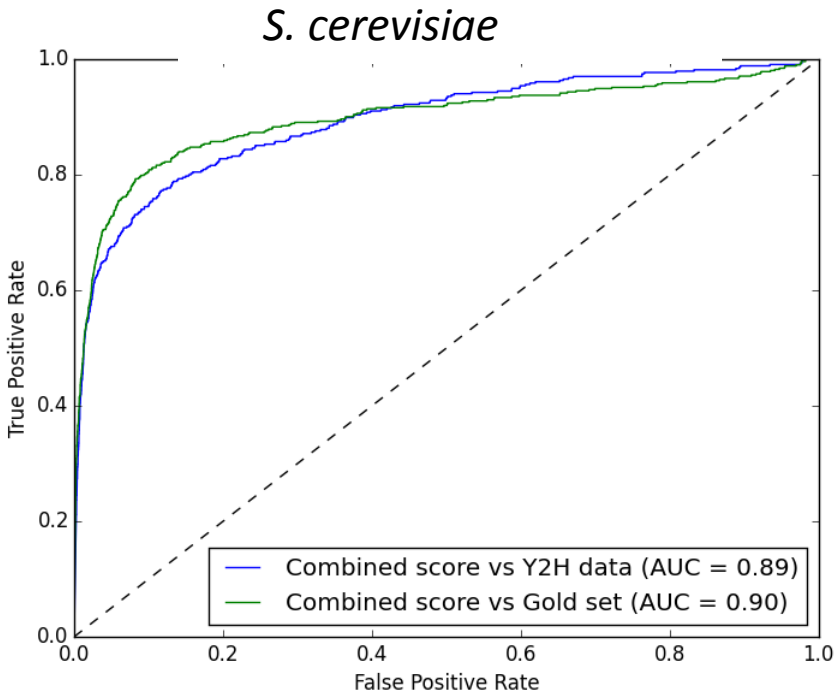
Twilight zone (up to 25% identity)



Goncearenco et al, to be submitted

Protein-level benchmarks with *in vitro* Y2H PPI data





Experimental data from:
Rajagopala et al. Nat Biotechnol 2014
Yu et al. Science 2008

Goncearenco et al, to be submitted

Acknowledgements

NCBI:

Minghui Li

Alexandr Goncearenco

Franco Simonetti

Benjamin Shoemaker

Alumni:

Manoj Tyagi (*Washington University School of Medicine*)

Hafumi Nishi (*Tohoku University, Japan*)

Kosuke Hashimoto (RIKEN, Japan)

Collaborators:

Emil Alexov (*Clemson University*)

Stefan Wuchty (*University of Miami, USA*)

Stanley Lipkowitz (*NCI, NIH*)