Leveraging Mathematical Models to Predict Allosteric Hotspots in the Age of Deep Sequencing

Declan Clarke

Dissertation Director: Mark Gerstein
Committee: Gary Brudvig, Patrick Loria

March 8 2016

Allosteric Hotspot Prediction Using Dynamics

applications to inter- and intra-species conservation

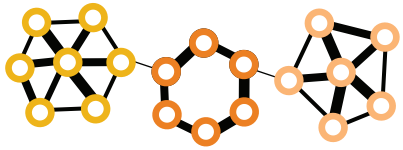
Networks

Bhardwaj et al, 2011 (*Protein Sci.*)

Clarke et al, 2012 (*J. Struct. Biol.*)

Gerstein et al, 2012 (*Nature*)

Sethi et al, 2015 (*COSB*)



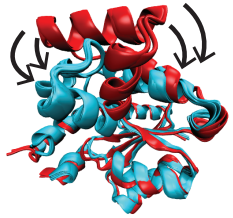
Macromolecular Motions

Bhardwaj et al, 2011 (*Protein Sci.*)

Clarke et al, 2012 (*J. Struct. Biol.*)

Sethi et al, 2015 (*COSB*)

MolMovDB items



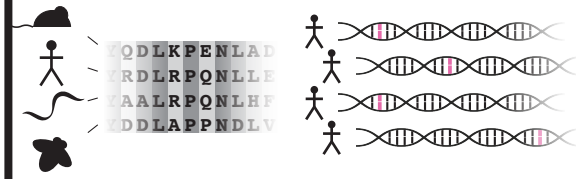
Next-Gen Sequencing & Variation

Habegger et al, 2012 (*Bioinformat.*)

Khurana et al, 2013 (*Science*)

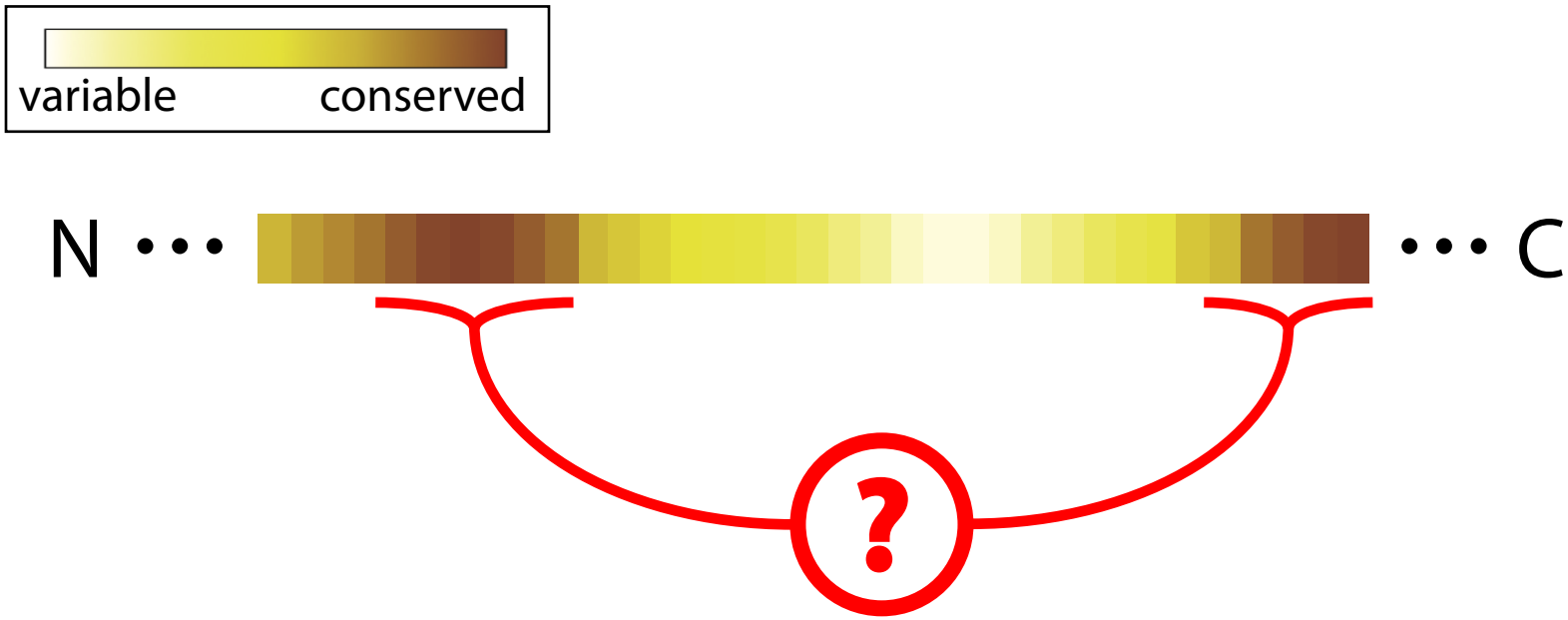
Sethi et al, 2015 (*COSB*)

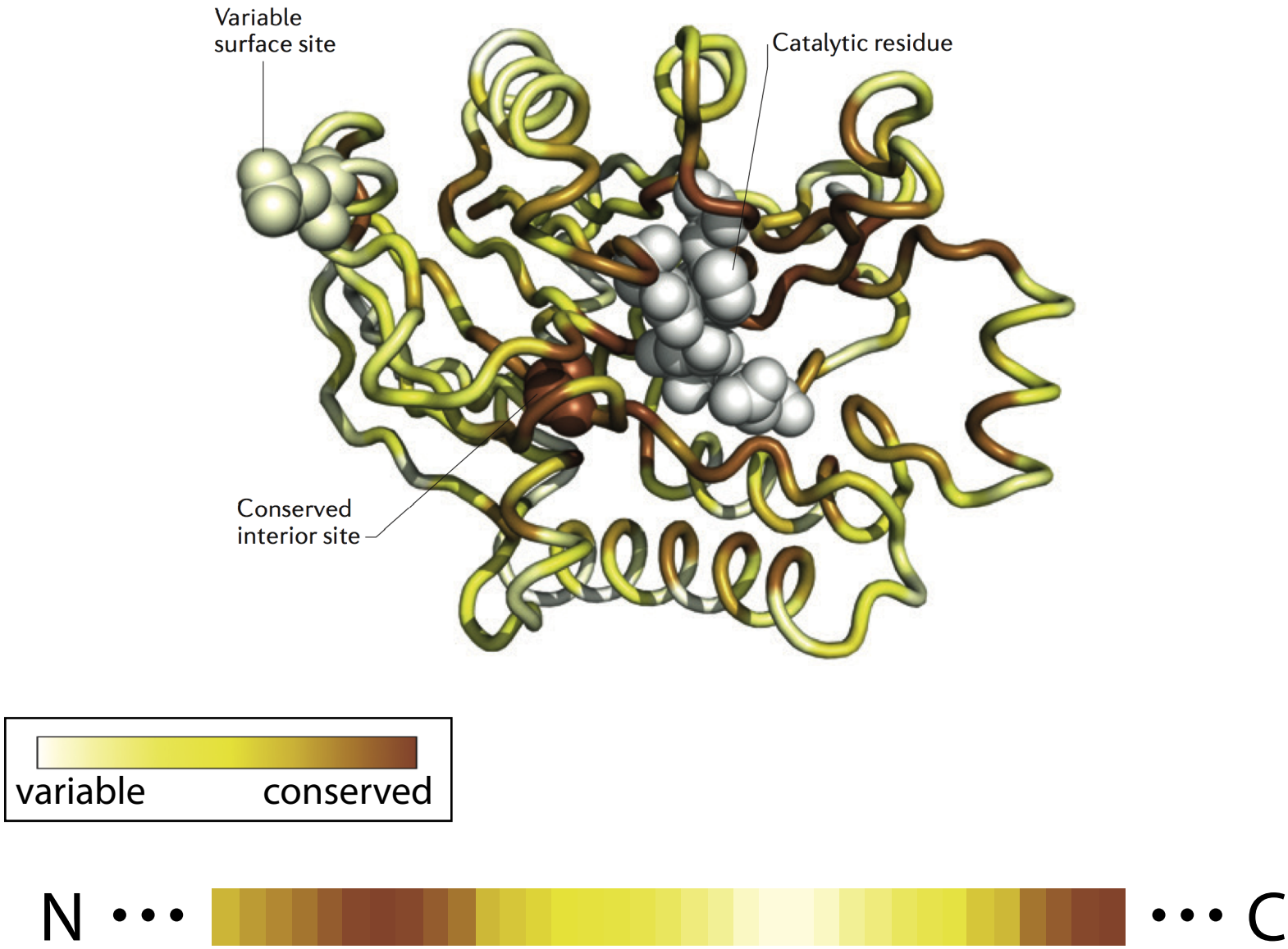
Kumar et al *(in prep)*



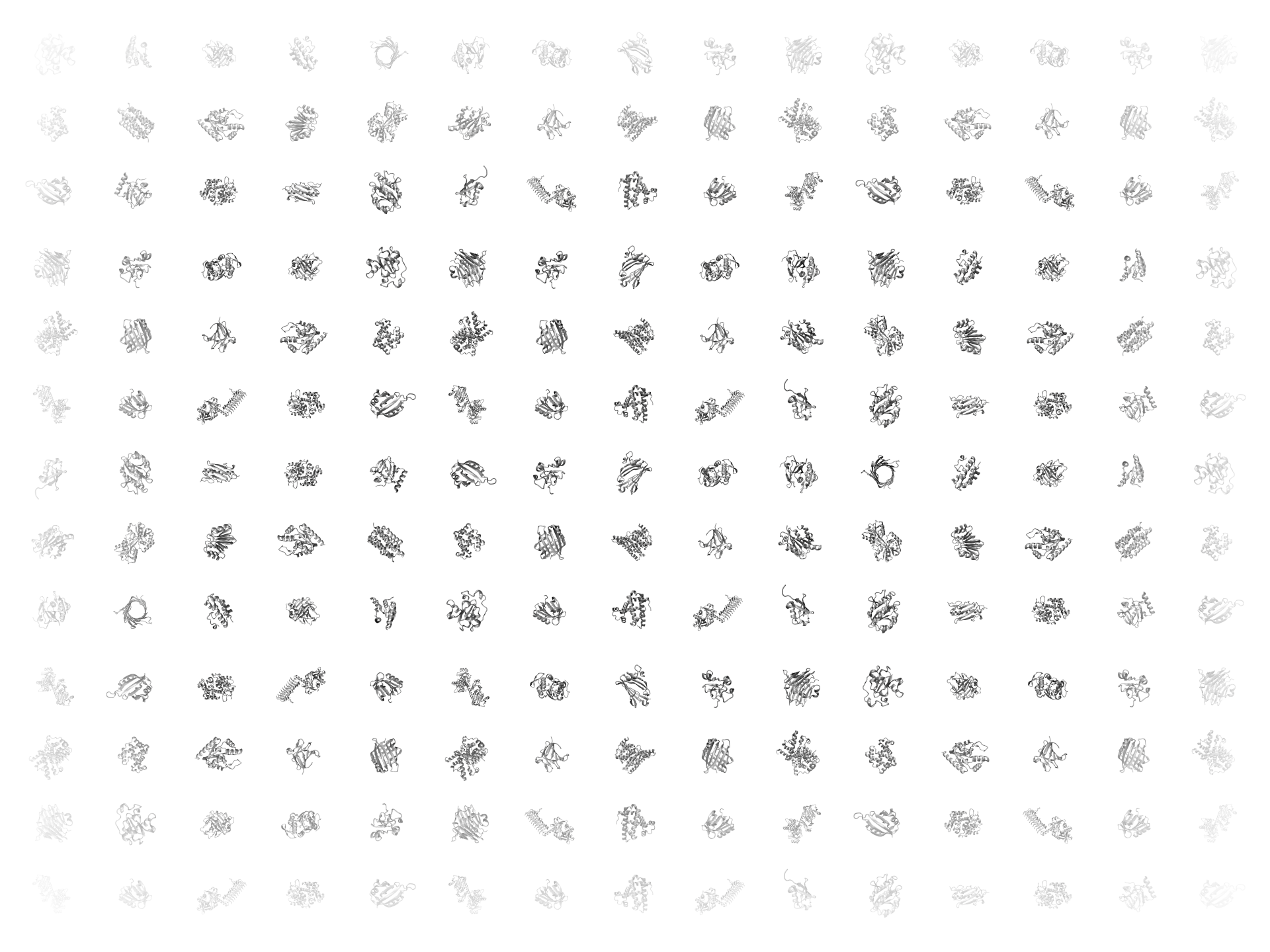
Protein Structure

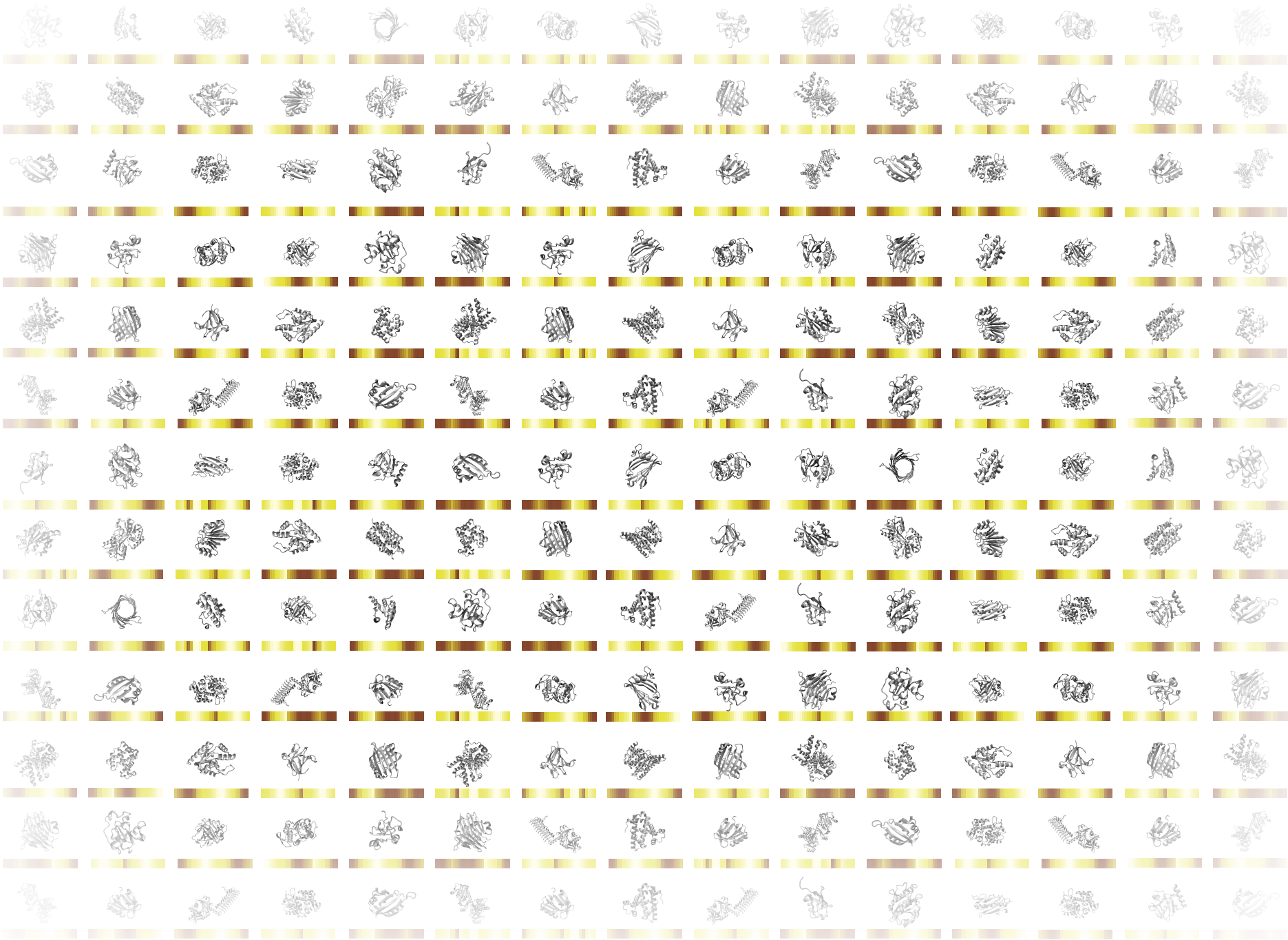

variable
conserved
N
C

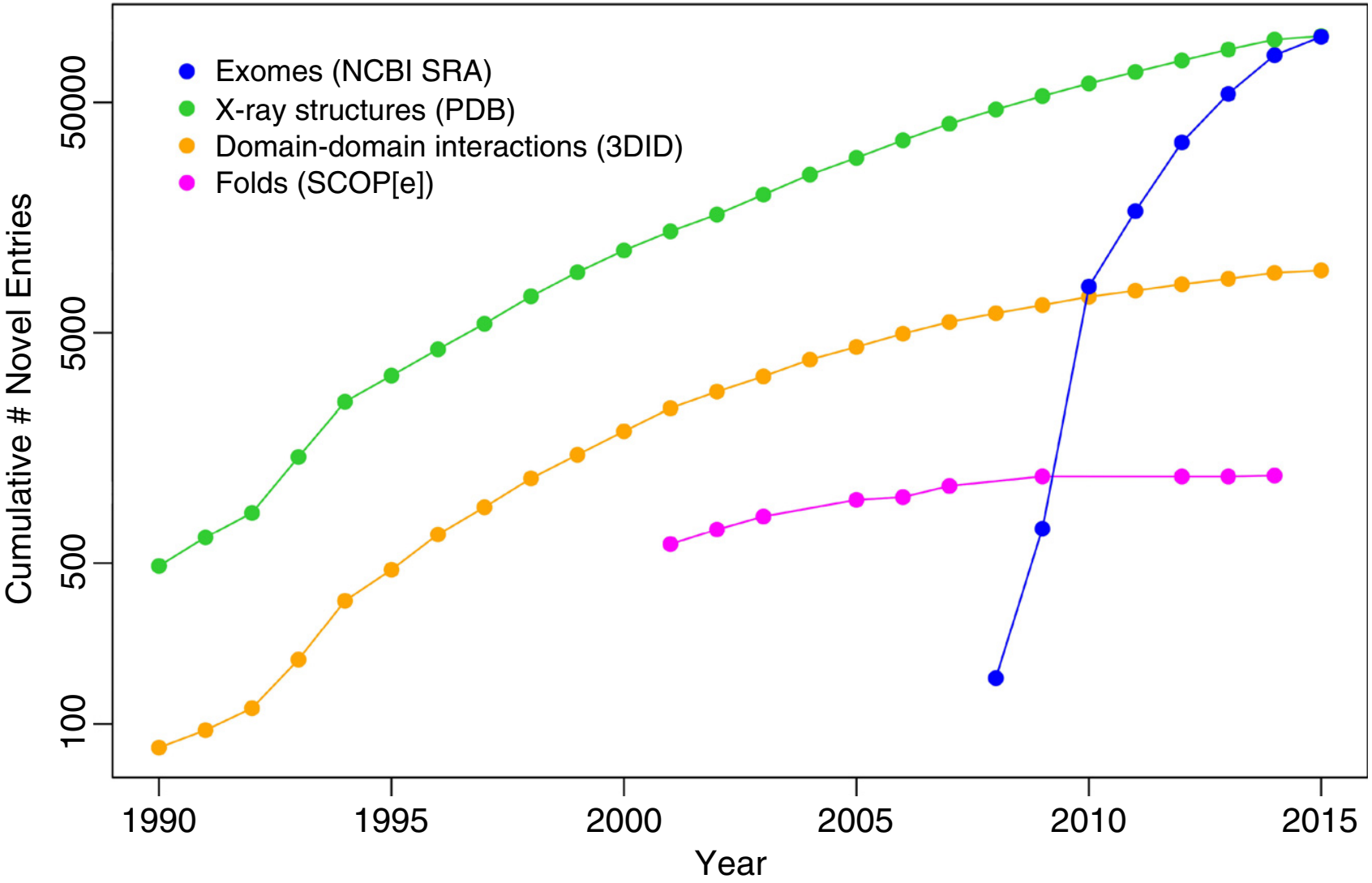

variable
conserved
N
C
?



Adapted from Echave et al, 2016







Sethi et al, 2015. *Curr. Opin Struct Biol.*

How do we make sense of these "genomic shadows"?

The approach must be generalizable and it must apply to many (most?) proteins.

We're only given the structures as starting points -- we'd ideally like some property of the proteins which 'bridges' both structural and functional constraints.

How do we make sense of these "genomic shadows"?

The approach must be generalizable and it must apply to many (most?) proteins.

We're only given the structures as starting points -- we'd ideally like some property of the proteins which 'bridges' both structural and functional constraints.

→ ***Allostery* often provides the missing conceptual link**

How do we make sense of these "genomic shadows"?

If allostery brings us further toward elucidating these signatures, then how can we identify the residues that are most important allosterically?

Experimental studies with site-directed mutagenesis on each protein?

How do we make sense of these "genomic shadows"?

If allostery brings us further toward elucidating these signatures, then how can we identify the residues that are most important allosterically?

Experimental studies with site-directed mutagenesis on each protein?

→ ***Mathematical models*** **can provide the means**

1. Models for predicting allosteric hotspots

2. Speed optimization & web server to predict allosteric sites on a large scale

3. Identifying alternative conformations throughout large protein datasets

4. Signatures of conservation

1. **Models for predicting allosteric hotspots**

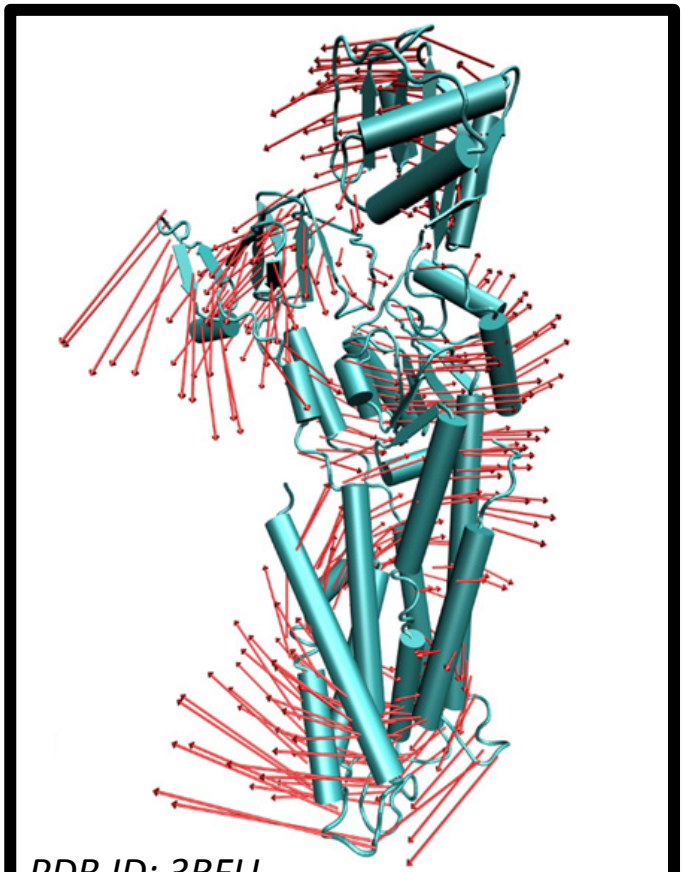
2. Speed optimization & web server to predict allosteric sites on a large scale

3. Identifying alternative conformations throughout large protein datasets

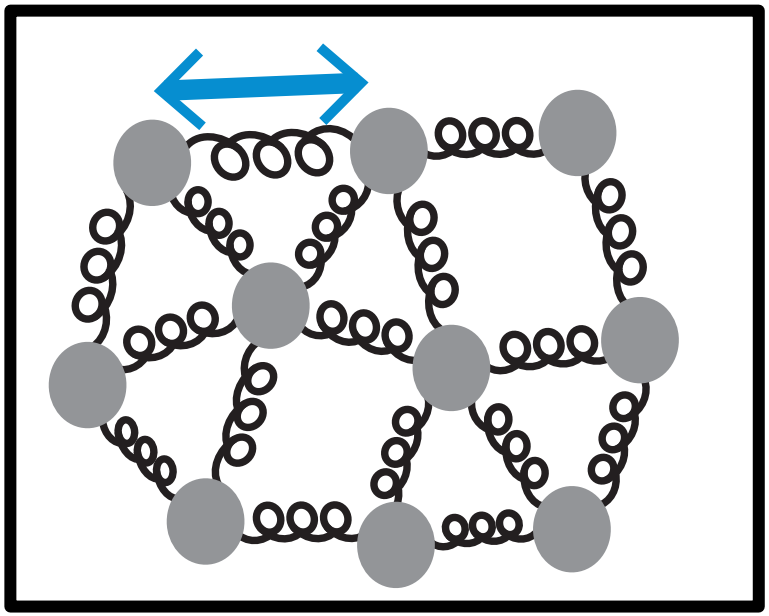
4. Signatures of conservation

Models of Protein Conformational Change

Motion Vectors from Normal Modes (ANMs)

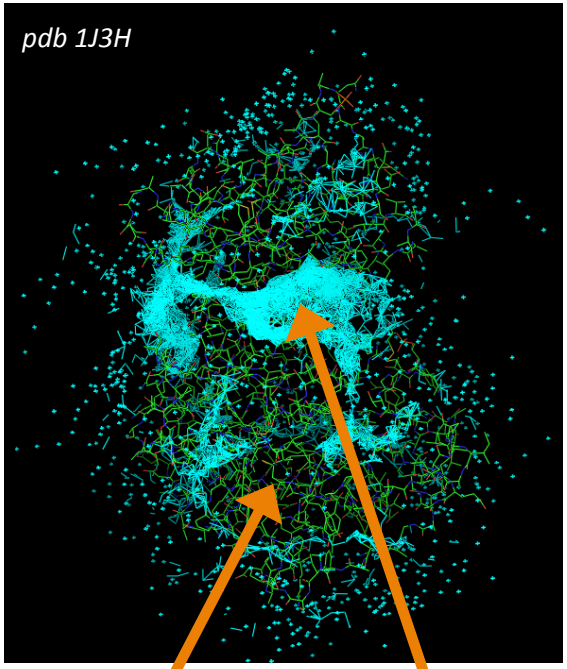


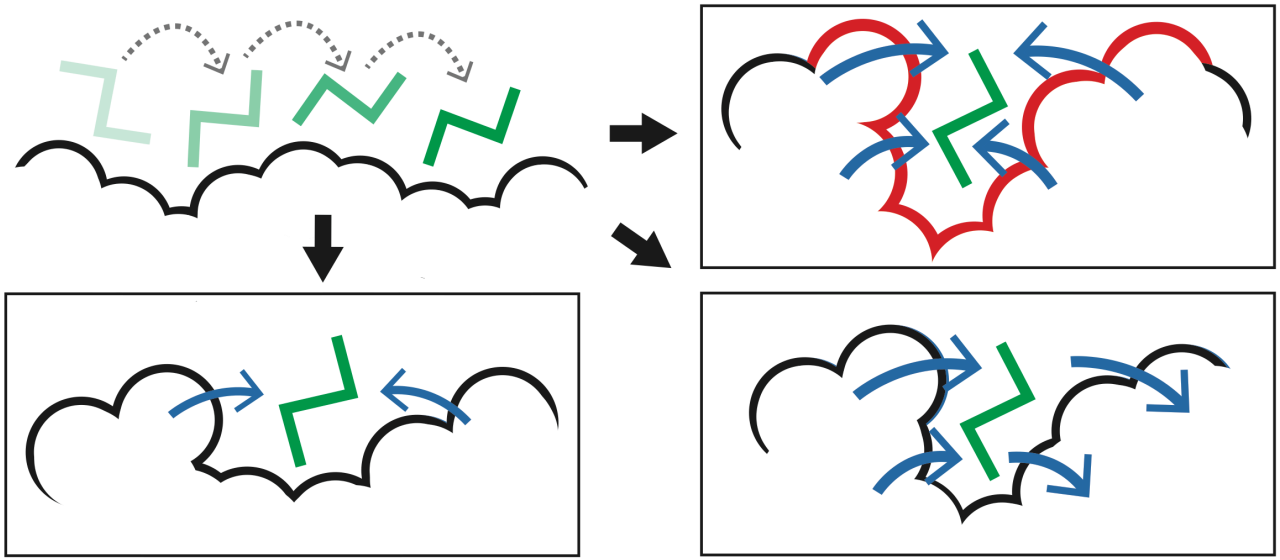
PDB ID: 3RFU
Adapted from Fuglebakk et al, 2014



Predicting Allosterically-Important Residues at the Surface

1. MC simulations generate a large number of candidate sites
2. Score each candidate site by the degree to which it perturbs large-scale motions
3. Prioritize & threshold the list to identify the set of high confidence-sites





Surface region with high density of candidate sites

Surface region with low density of candidate sites

$$\textit{binding leverage} = \sum_{m=1}^{10}\left(\sum_i\sum_j \Delta d^2_{ij(m)}\right)$$

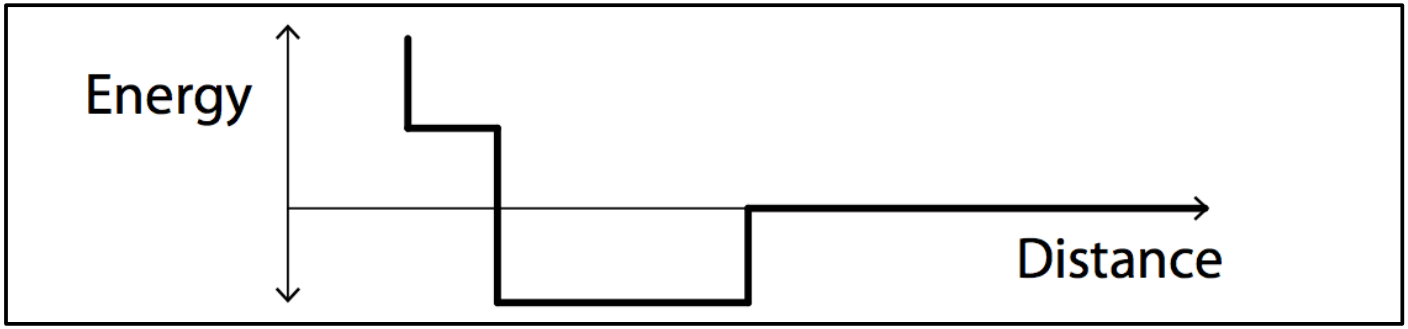
Adapted from Clarke*, Sethi*, et al *(in press)*

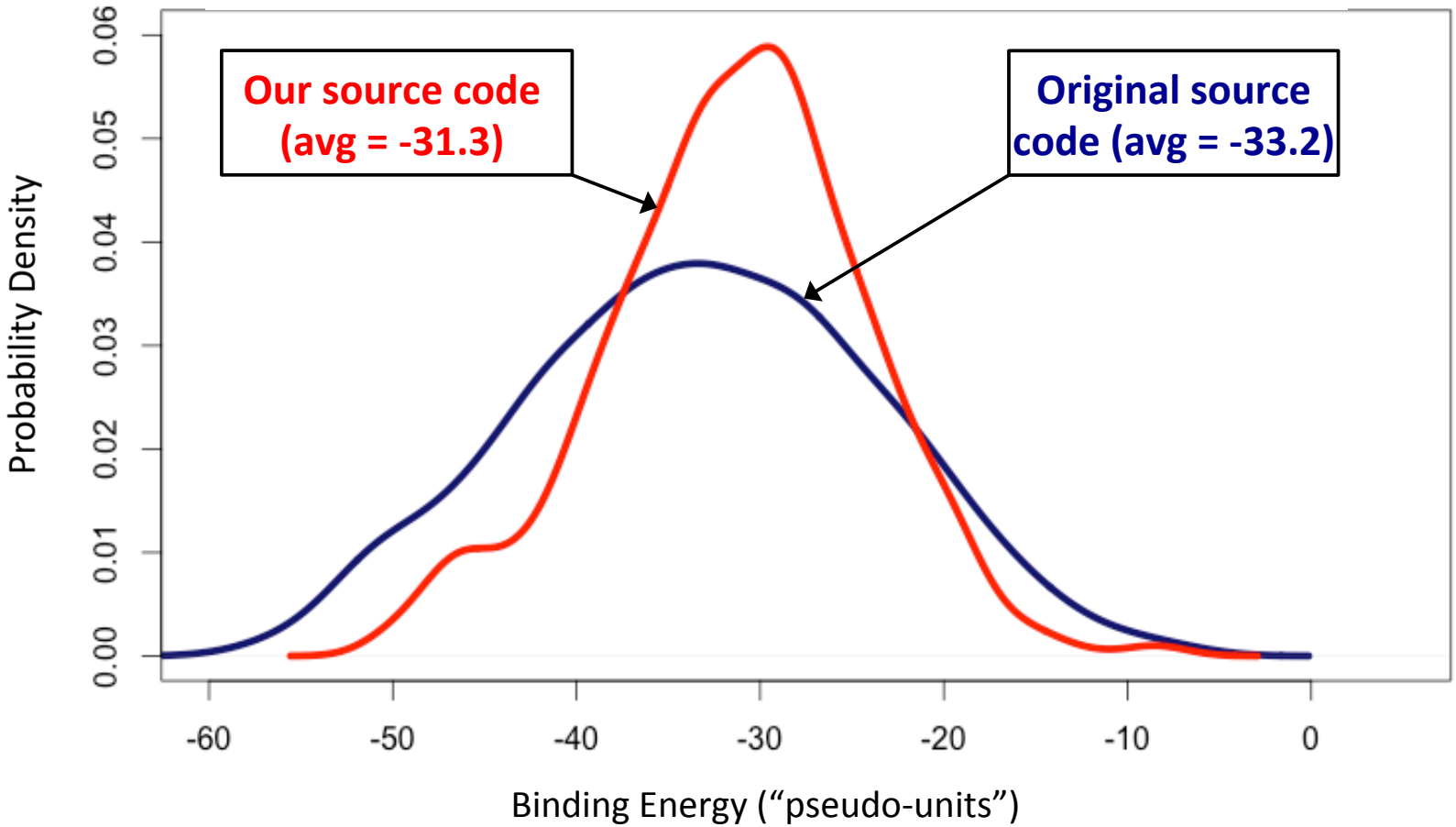
Testing the improvement of this method on a gold standard set

Known binding sites constitute a subset of the allosteric sites on the protein surface – to what degree can they be found?

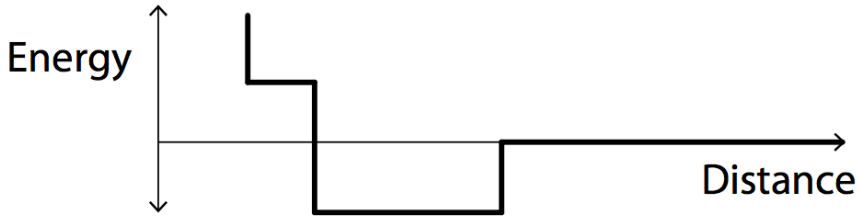
HOLO	***APO***
1ake (**AP5**)	4ake
3cep (**G3P, IDM, PLP**)	1bks (**PLP**)
1hor (**AGP**, *PO4*, [& **16G** in pdb 1HOT])	1cd5
2c2b (**SAM**, [& **LLP** in pdb 2c2g])	1e5x
1gz3 (**ATP, FUM**, *OXL*)	1efk (MAK)
1atp (**ATP**)	1j3h
1hwz (**GLU**, **GTP, NDP** [& **ADP** in PDB 1NQT])	1nr7
1xtu (**CTP, U5P**)	1xtt (*ACY*, **U5P**)
1aax **(BPM** [& **892** in PDB 1T49])	2hnp
7at1 (**ATP**, **MAL**, **PCT** [& **CTP** in PDB 1RAC], [& **PAL** in PDB 1D09])	3d7s
3ju6 (**ANP**, **ARG**)	3ju5
6pfk (PGA [& **F6P + ADP** in PDB 4PFK])	3pfk (*PO4*)

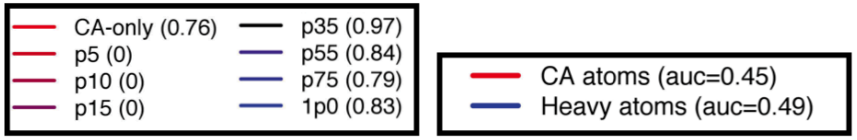
First step: faithful reproduction of the originally published* formalism

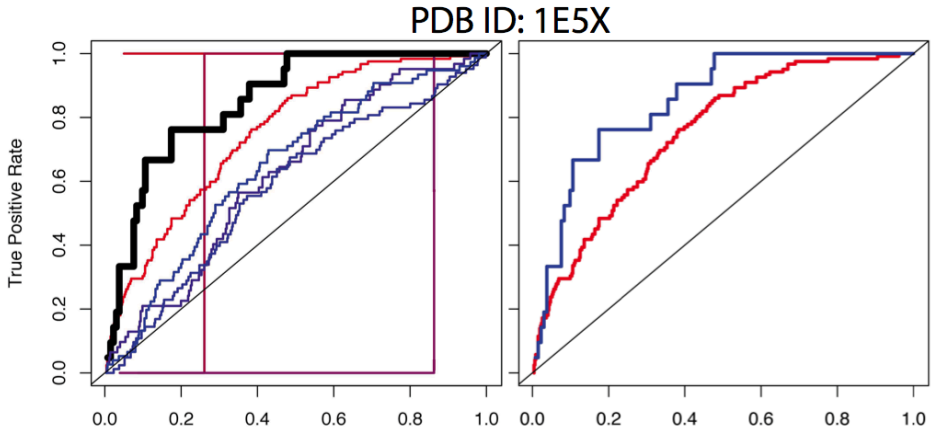


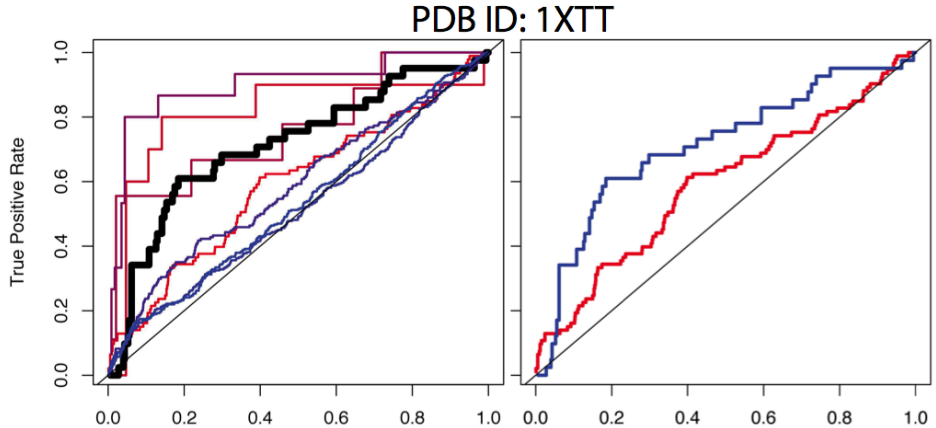


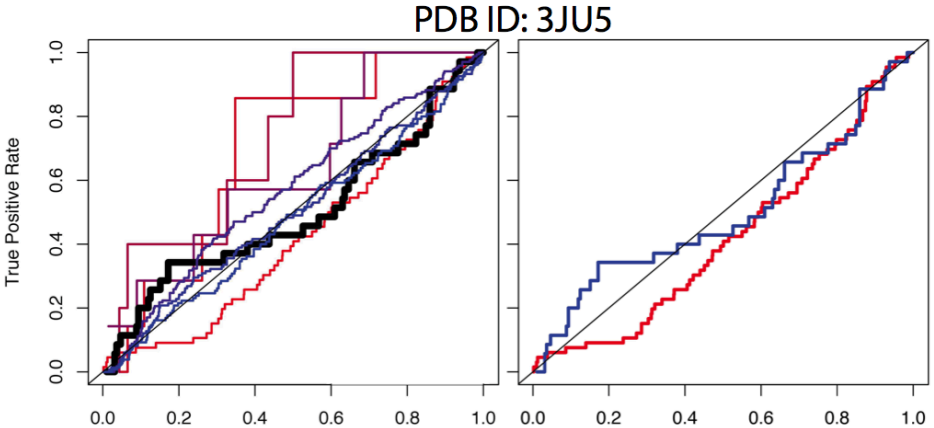
*Original source code from Mitternacht, S and Berezovsky, I (2011). *PLoS Comput Biol*

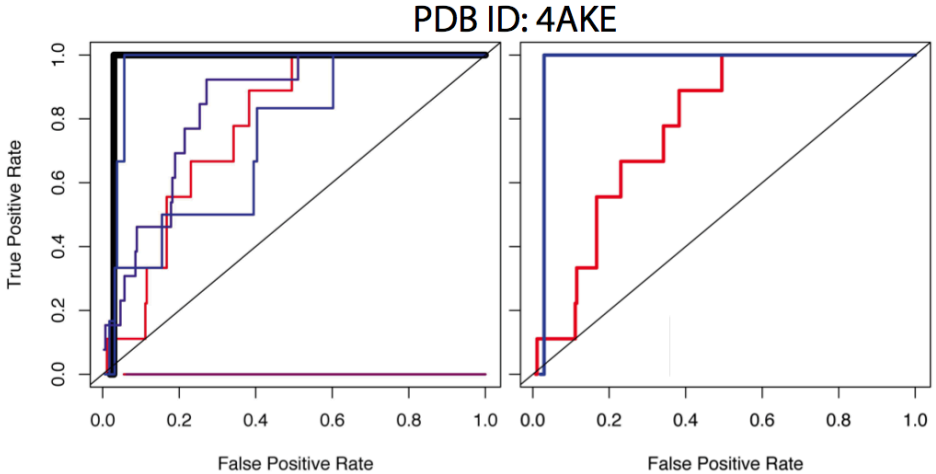

Energy
Distance

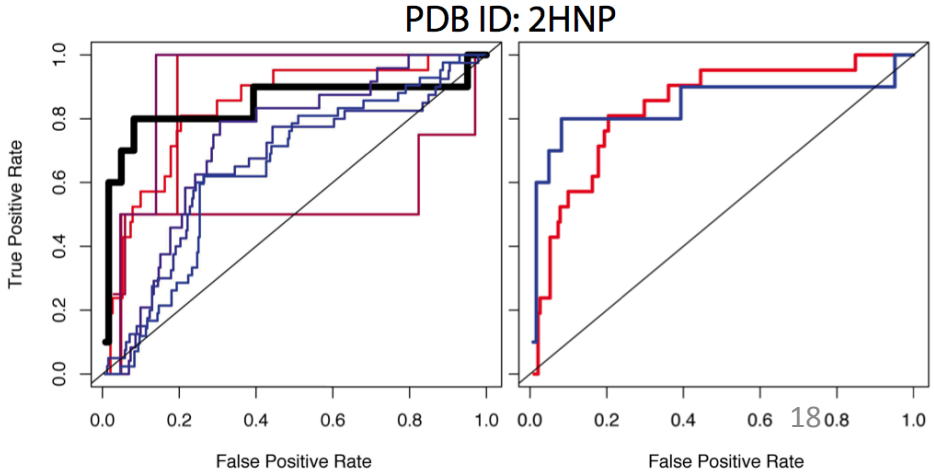

CA-only (0.76)
p5 (0)
p10 (0)
p15 (0)
p35 (0.97)
p55 (0.84)
p75 (0.79)
1p0 (0.83)
CA atoms (auc=0.45)
Heavy atoms (auc=0.49)


PDB ID: 1E5X
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

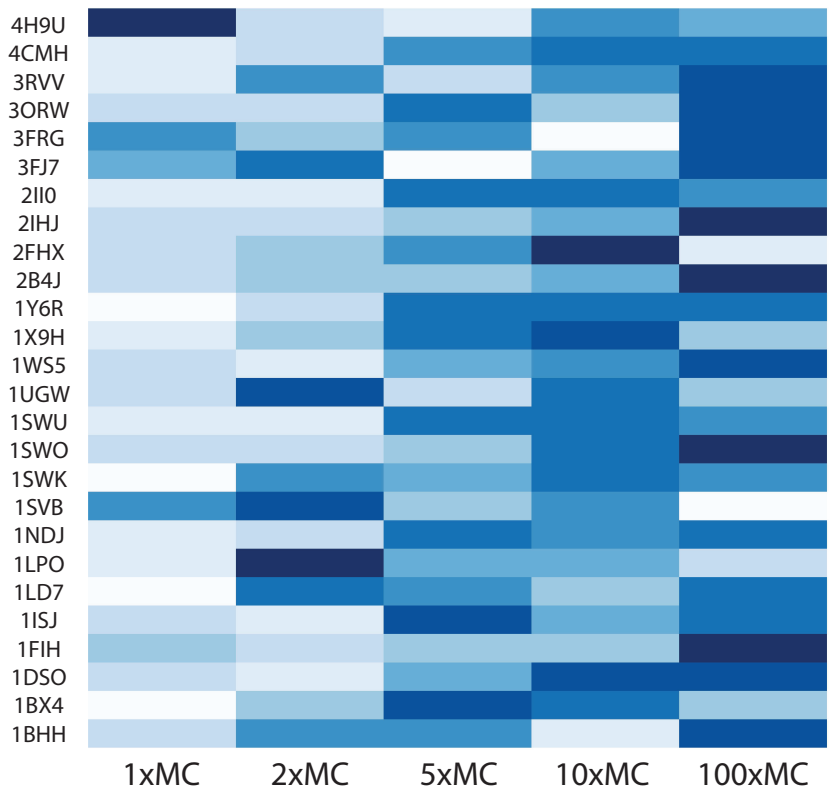

PDB ID: 1XTT
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

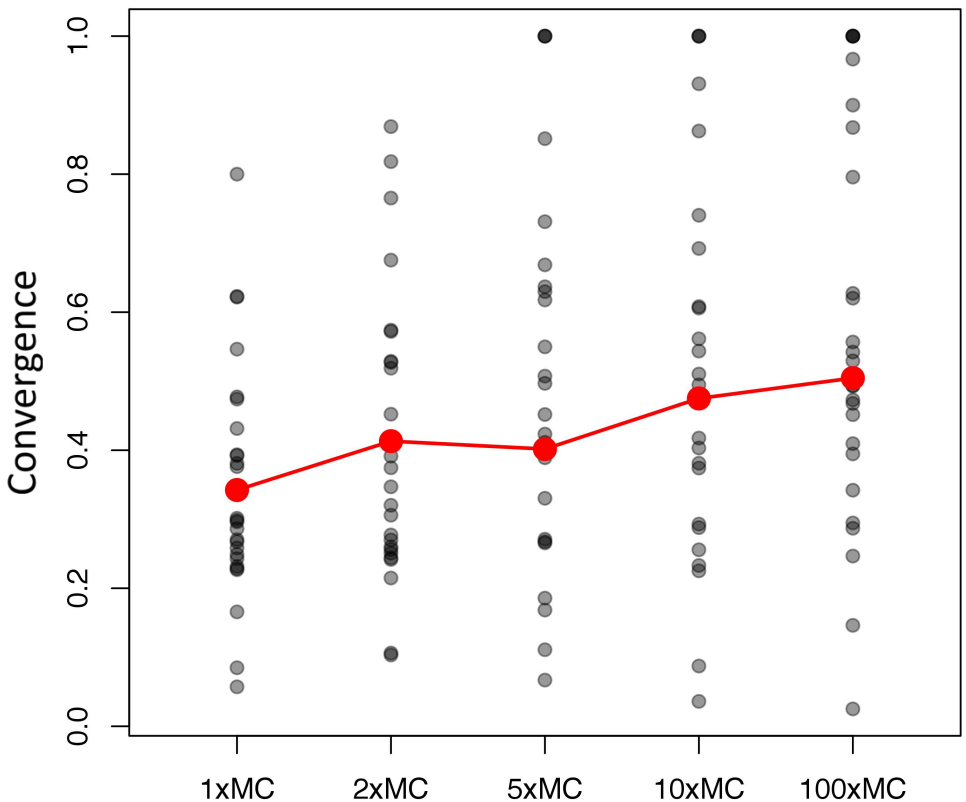

PDB ID: 3JU5
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


PDB ID: 4AKE
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0


PDB ID: 2HNP
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

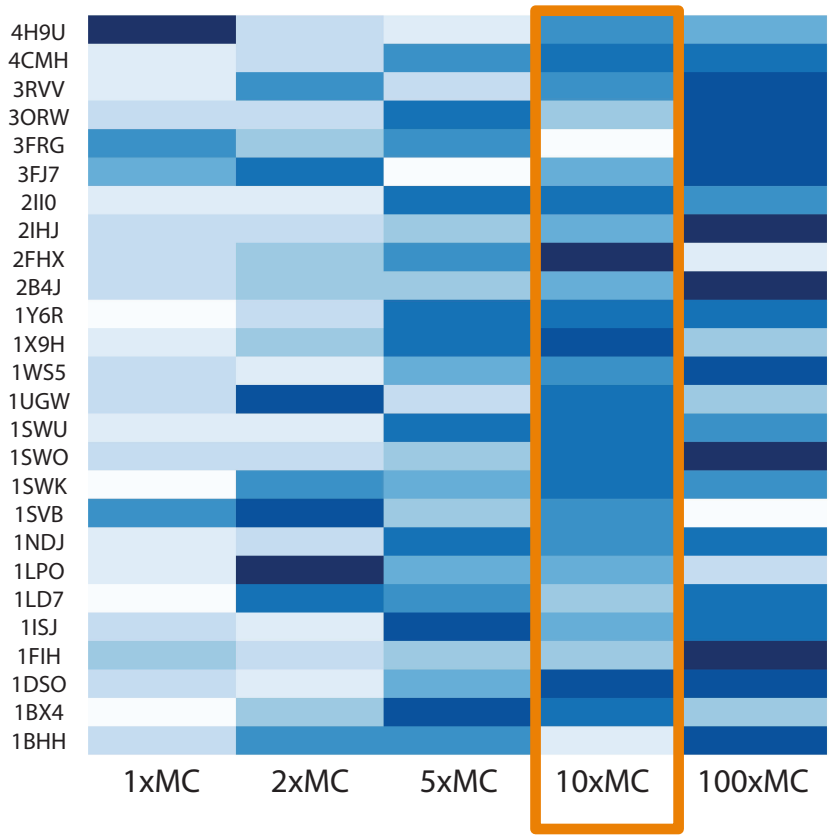
Measures of convergence using different scaling factors for the number of steps in each MC simulation

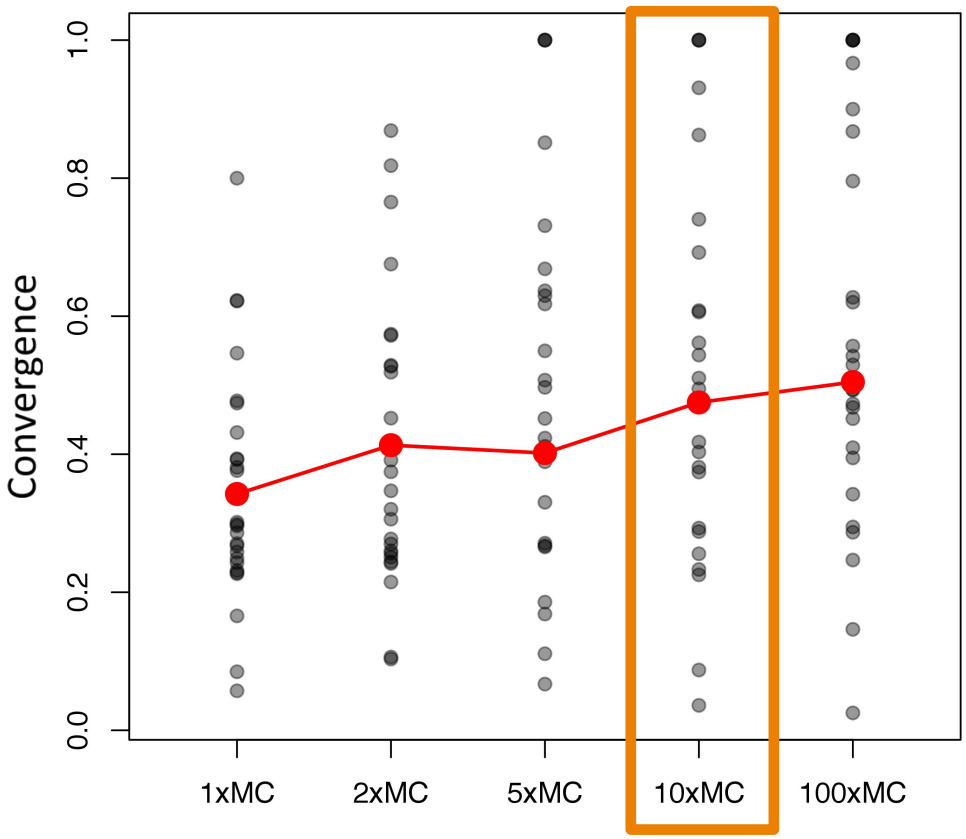




"**1xMC**": The number of MC steps in each run is set to **1x1000** times the "size of the simulation box"

Measures of convergence using different scaling factors for the number of steps in each MC simulation



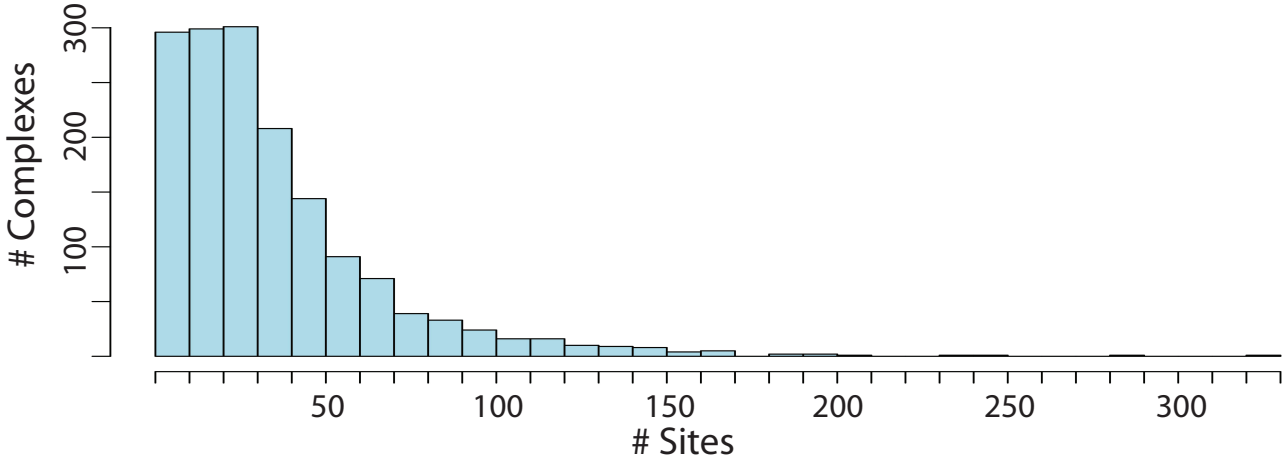


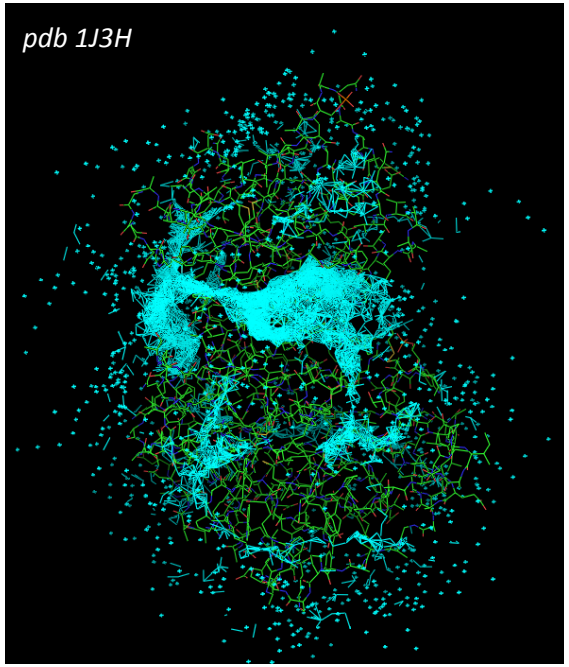
"**1xMC**": The number of MC steps in each run is set to **1x1000** times the "size of the simulation box"

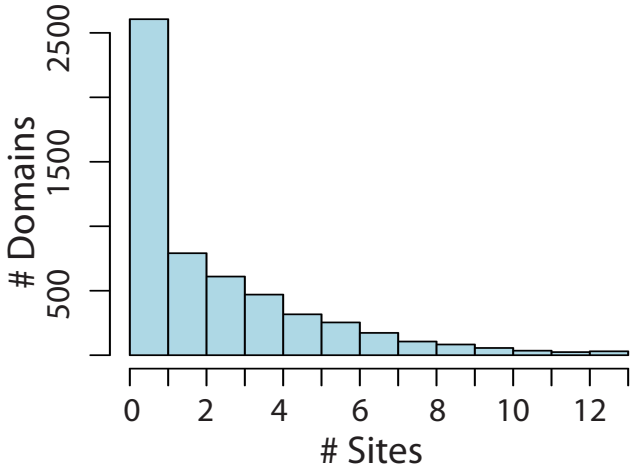
Predicting Allosterically-Important Residues at the Surface

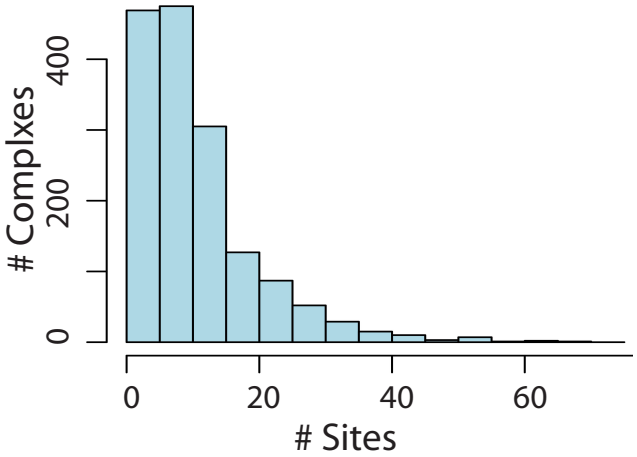
Heavy Atom Inclusion & Energy Gap Framework to Generate Prioritized Sites

Why Apply Automated Thresholding?



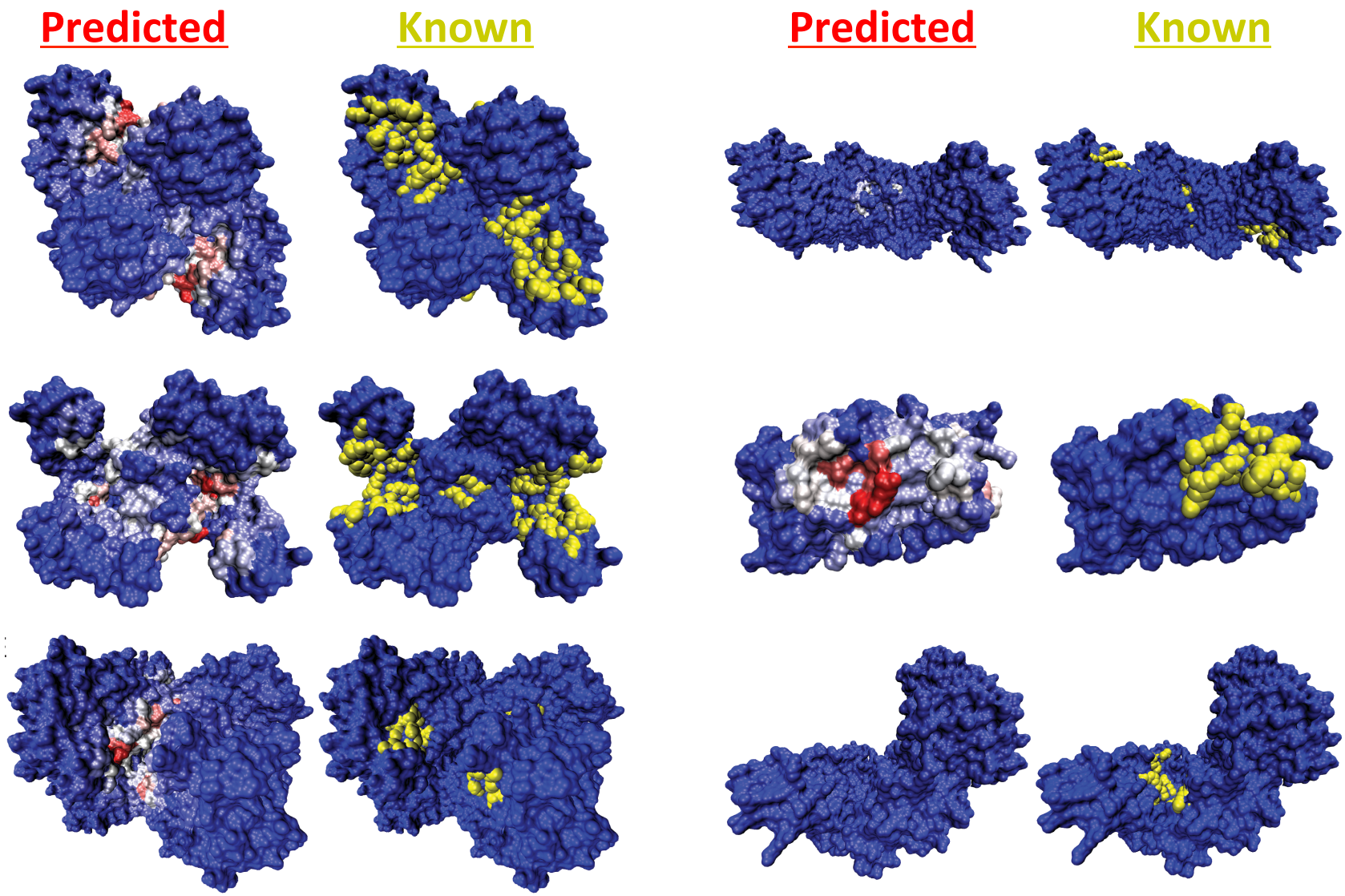






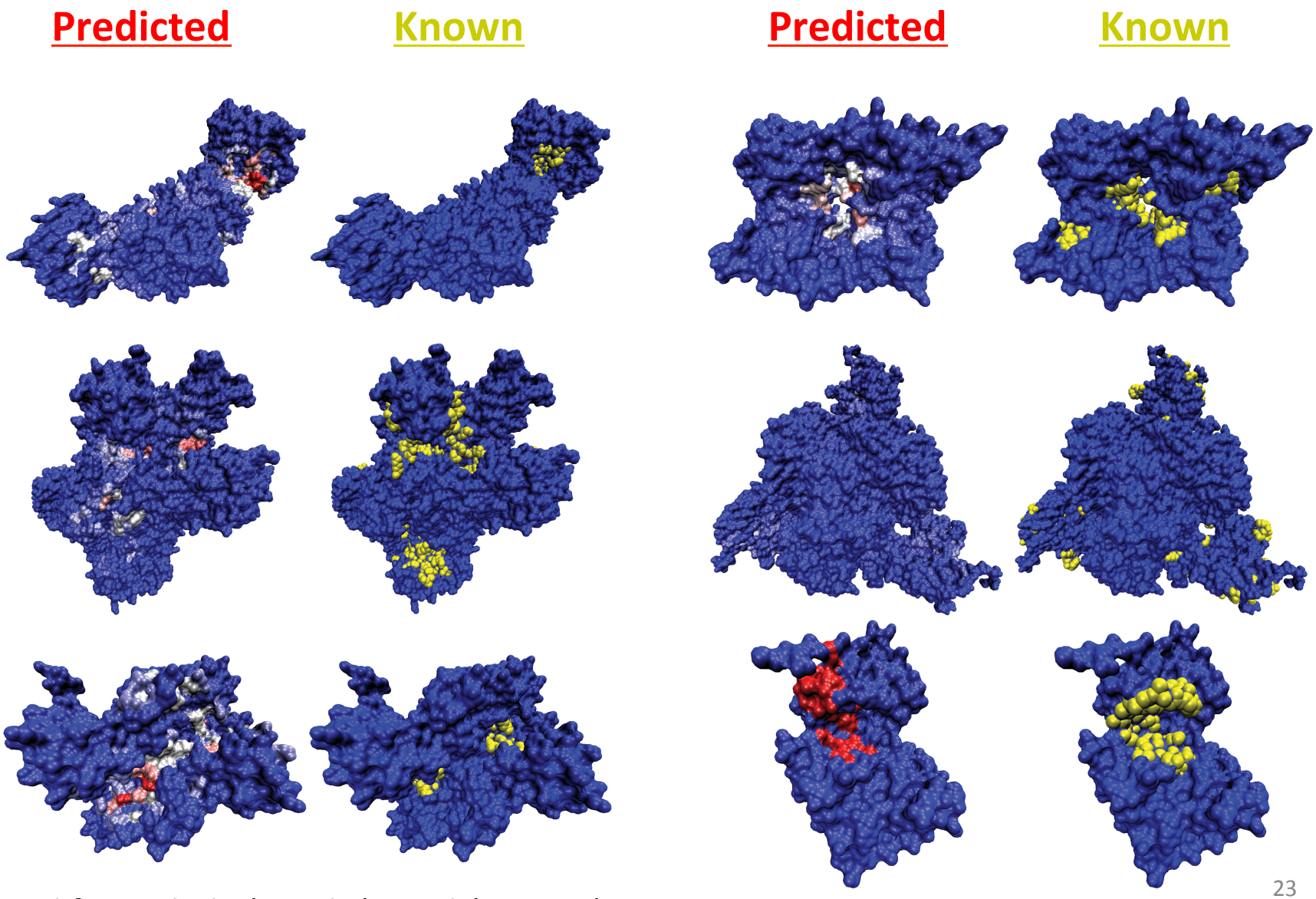
Adapted from Clarke*, Sethi*, et al *(in press)*

Predicting Allosterically-Important Residues at the Surface within the Canonical Set



Adapted from Clarke*, Sethi*, et al *(in press)*

Predicting Allosterically-Important Residues at the Surface within the Canonical Set



Adapted from Clarke*, Sethi*, et al *(in press)*

Predicting Allosterically-Important Residues at the Surfaces within the Canonical Set

Statistics on the surfaces of *apo* structures

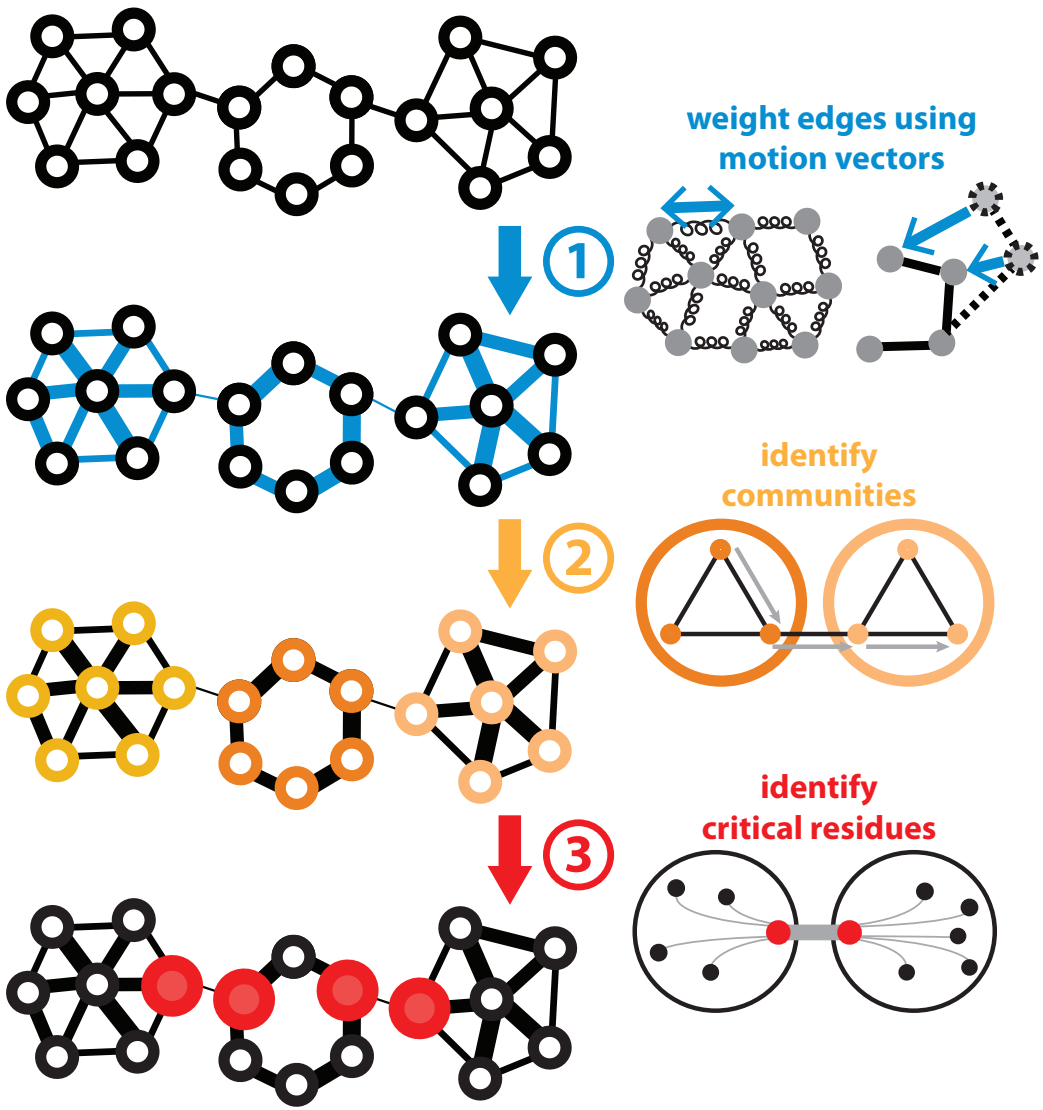
Protein name (pdb ID)	% Surf (SC res)	% Surf (LB res)	SC-LB overlap	# SC sites	# LB sites	# Overlapping sites	% LB sites identified
Phosphofructokinase (3pfk)	51.0	20.4	0.255 (0.155)	19	3	3	100.0
Adenylate kinase (4ake)	45.4	17.8	0.274 (0.154)	29	2	2	100.0
G-6-P deaminase (1cd5)	58.9	10.0	0.153 (0.096)	24	2	1	50.0
cAMP-dep. prot. kin. (1j3h)	6.6	8.0	0.25 (0.041)	2	1	1	100.0
Trp synthase (1bks)	34.3	9.7	0.079 (0.079)	24	4	1	25.0
Thr synthase (1e5x)	20.7	9.3	0.139 (0.077)	17	3	2	66.7
Hum. malic enzyme (1efk)	5.5	8.6	0.03 (0.036)	10	10	0	0.0
Glu dehydrogenase (1nr7)	14.9	17.5	0.187 (0.102)	45	24	6	25.0
P-ribosyltransferase (1xtt)	29.8	19.6	0.295 (0.154)	31	5	5	100.0
Tyr phosphatase (2hnp)	73.9	13.3	0.16 (0.134)	25	2	2	100.0
Asp transcarbamoylase (3d7s)	26.7	13.7	0.054 (0.064)	26	9	0	0.0
Arg kinase (3ju5)	1.6	3.9	0 (0.013)	1	2	0	0.0
mean	30.8	12.7	0.156 (0.092)	21.083	5.583	1.917	55.6

Novel features

- heavy atom inclusion
- thresholding
- rigorous tests of convergence
- faster run times
- accessible server

Predicting Allosterically-Important Residues within the Interior

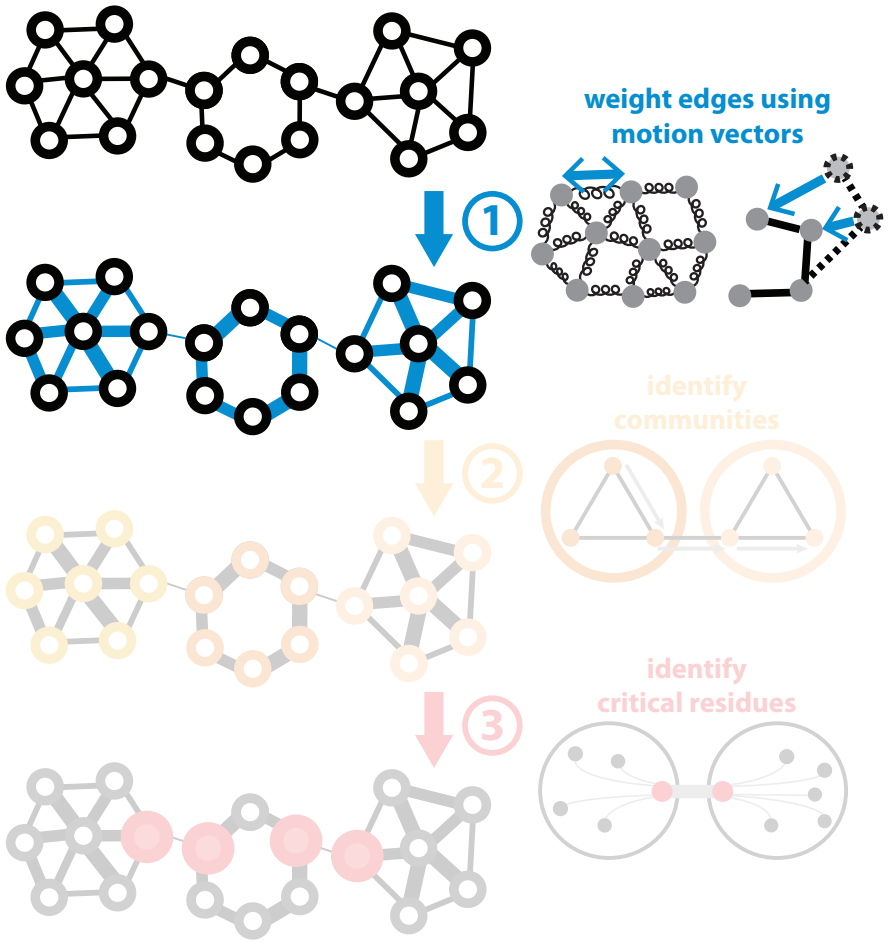
Sethi et al, (2009) *PNAS*



Adapted from Clarke*, Sethi*, et al *(in press)*

Predicting Allosterically-Important Residues within the Interior

Sethi et al, (2009) *PNAS*



$$Cov_{ij} = \langle \mathbf{r}_i \bullet \mathbf{r}_j \rangle$$

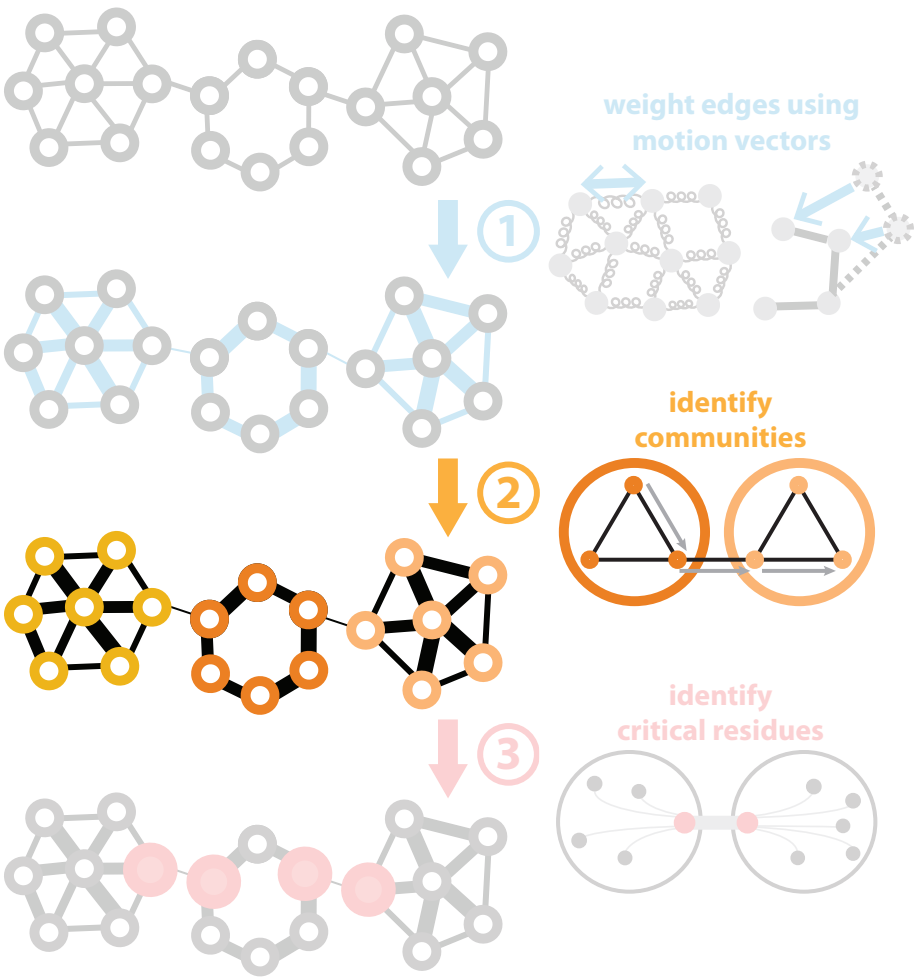
$$C_{ij} = Cov_{ij} \;/\; \sqrt{(\langle \mathbf{r}_i^2 \rangle \langle \mathbf{r}_j^2 \rangle)}$$

$$D_{ij} = -\log(|C_{ij}|)$$

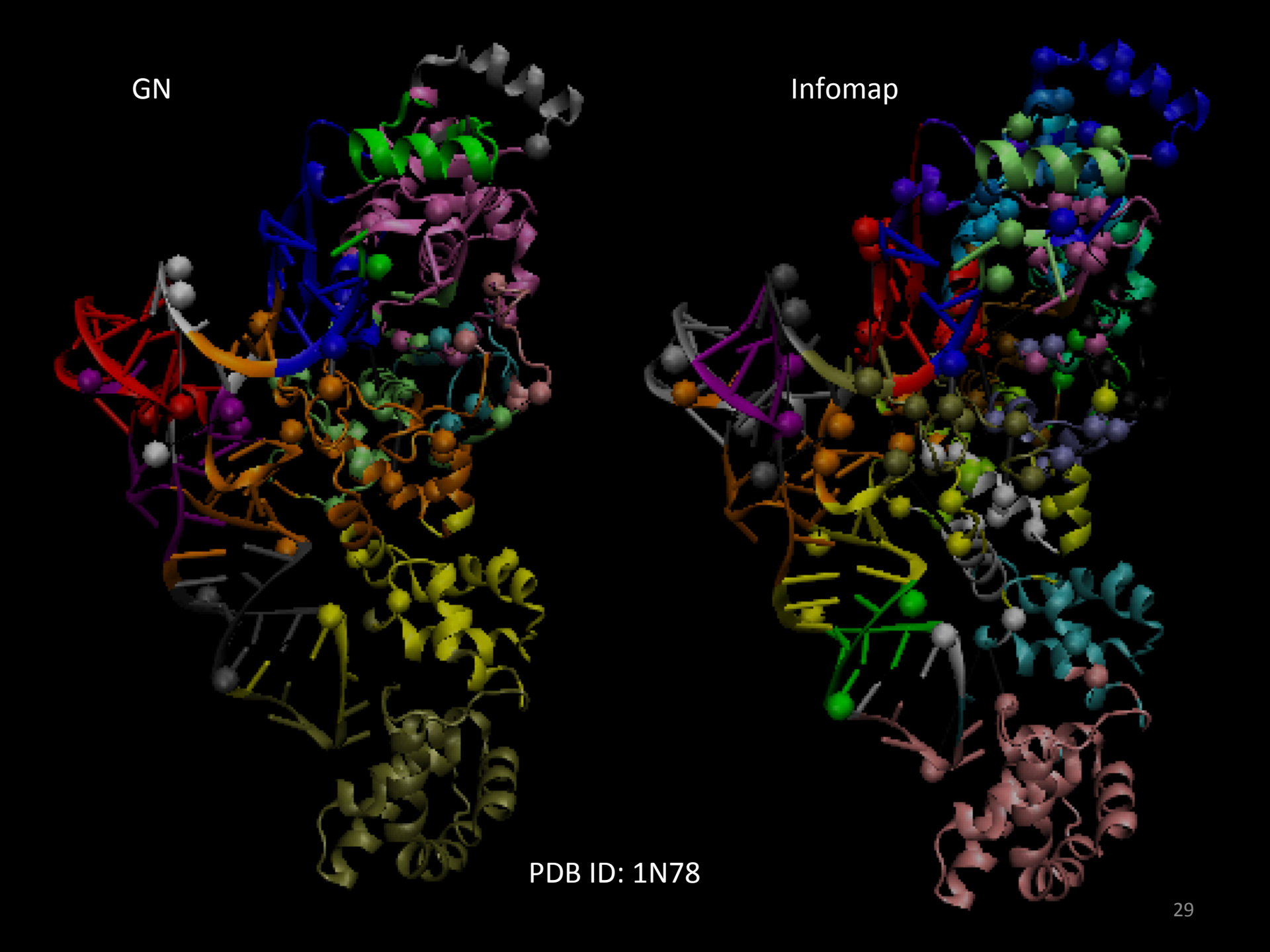
Adapted from Clarke*, Sethi*, et al *(in press)*

2 Network Community Algorithms

- **Girvan–Newman** -- Girvan, et al, (2002)
- **Infomap --** Rosvall et al, (2008)



Adapted from Clarke*, Sethi*, et al *(in press)*

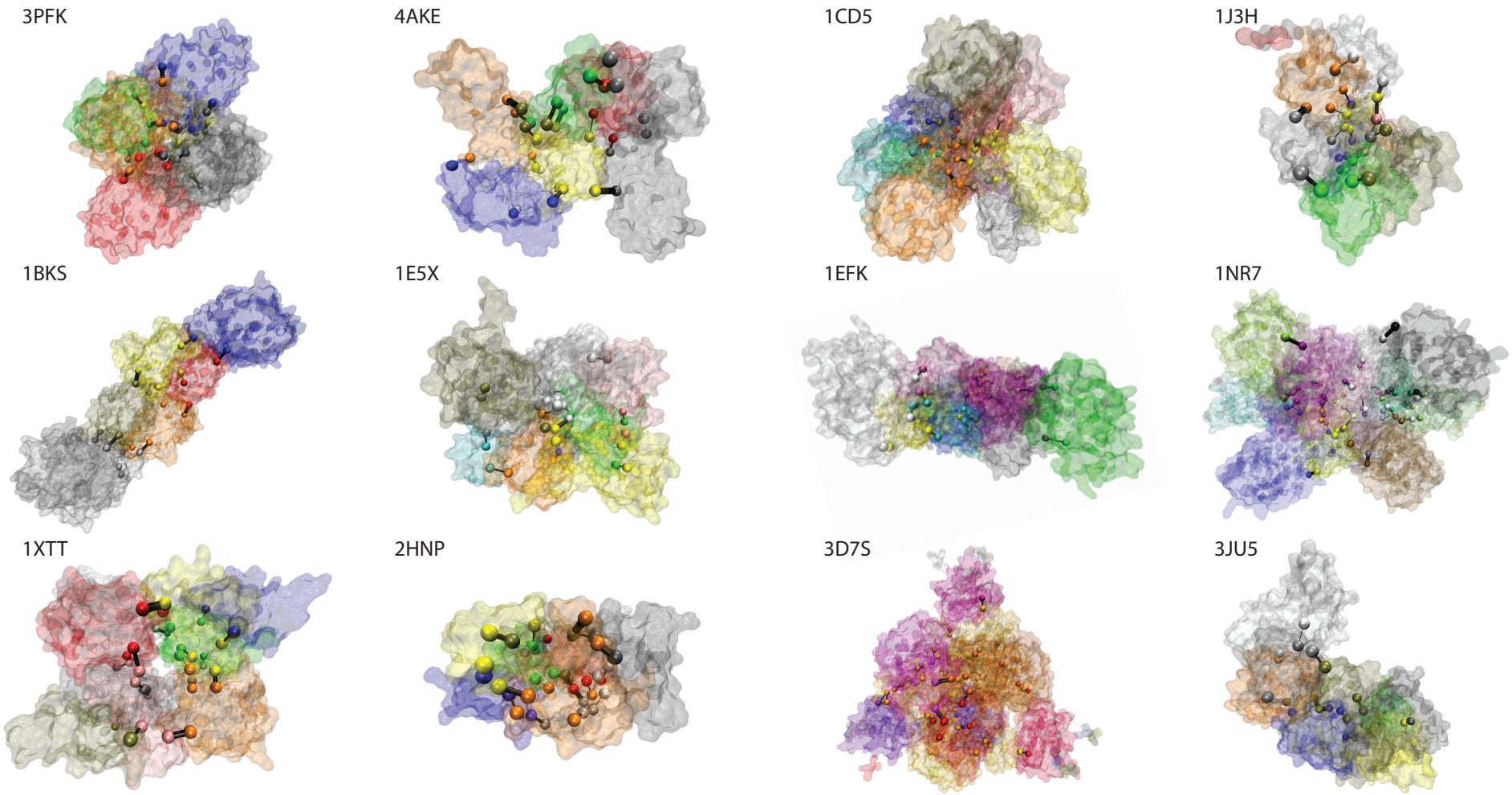

GN
Infomap
PDB ID: 1N78

Degree of Concordance Between Community Detection Methods: GN vs. Infomap

Protein (PDB, # residues)	Community Detection Method: GN | InfoMap			
	Modularity	# Comm.	# Critical Residues	% of GN critical residues which match those in Infomap (expected)
tRNA synthetase (1N78, 542)	0.71 | 0.68	14 | 25	47 | 109	0.28 (0.20)
Adenylate kinase (4AKE, 428)	0.73 | 0.70	11 | 20	39 | 82	0.90 (0.19)
Arginine Kinase (3JU5, 728)	0.72 | 0.69	12 | 28	41 | 142	0.22 (0.19)
Tyrosine Phosphatase (2HNP, 278)	0.59 | 0.59	7 | 15	27 | 70	0.26 (0.25)
Phosphoribosyltransferase (1XTT, 846)	0.72 | 0.68	9 | 32	36 | 174	0.22 (0.21)
cAMP-dep. PK (1J3H, 332)	0.66 | 0.64	11 | 19	36 | 78	0.33 (0.23)
Anthranilate synthase (1I7Q, 1418)	0.75 | 0.69	12 | 46	51 | 288	0.31 (0.20)
Malic enzyme (1EFK, 2212)	0.81 | 0.72	17 | 70	74 | 425	0.18 (0.19)
Threonine synthase (1E5X, 884)	0.73 | 0.69	13 | 36	43 | 192	0.28 (0.22)
G-6-P Deaminase (1CD5, 1596)	0.79 | 0.72	18 | 54	58 | 266	0.16 (0.17)
Phosphofructokinase (3PFK, 1276)	0.76 | 0.68	10 | 51	45 | 307	0.24 (0.24)
Tryptophan synthase (1BKS, 1294)	0.77 | 0.69	10 | 46	41 | 284	0.24 (0.22)
Means	***0.73 | 0.68***	***12.0 | 36.8***	***44.8 | 201.4***	***0.3***

Adapted from Clarke*, Sethi*, et al *(in press)*

Community Partitioning Using the Girvan-Newman Formalism



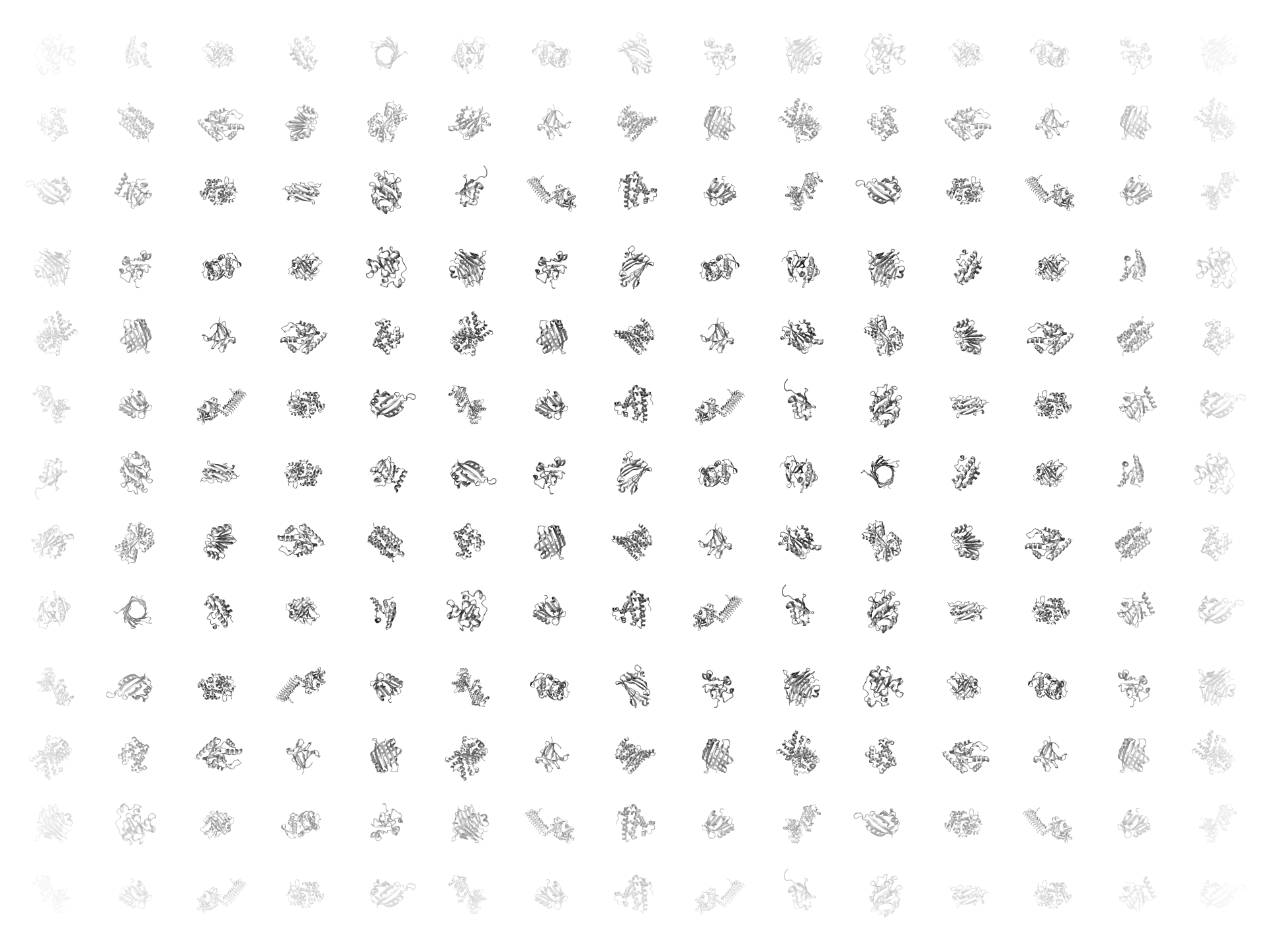
Taken from Clarke*, Sethi*, et al *(in press)*

1. Models for predicting allosteric hotspots

2. Speed optimization & web server to predict allosteric sites on a large scale

3. Identifying alternative conformations throughout large protein datasets

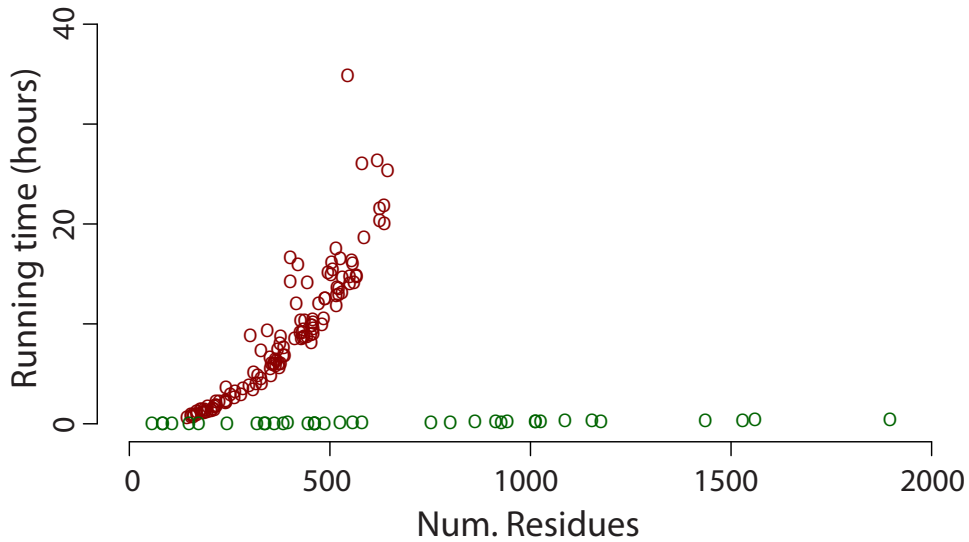
4. Signatures of conservation

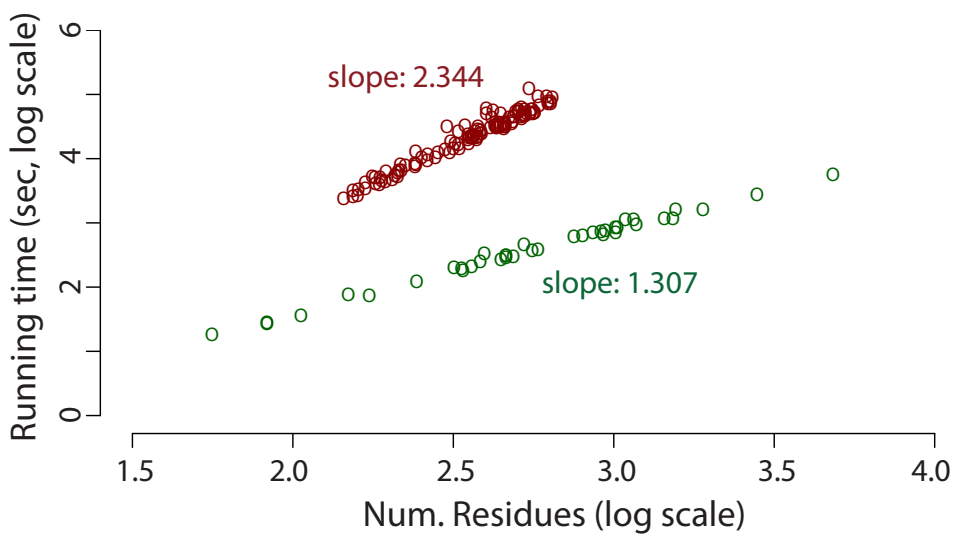


STRESS Server

Code Optimization for *Surface* Site Predictions: $O(n^3) \rightarrow O(n^2)$

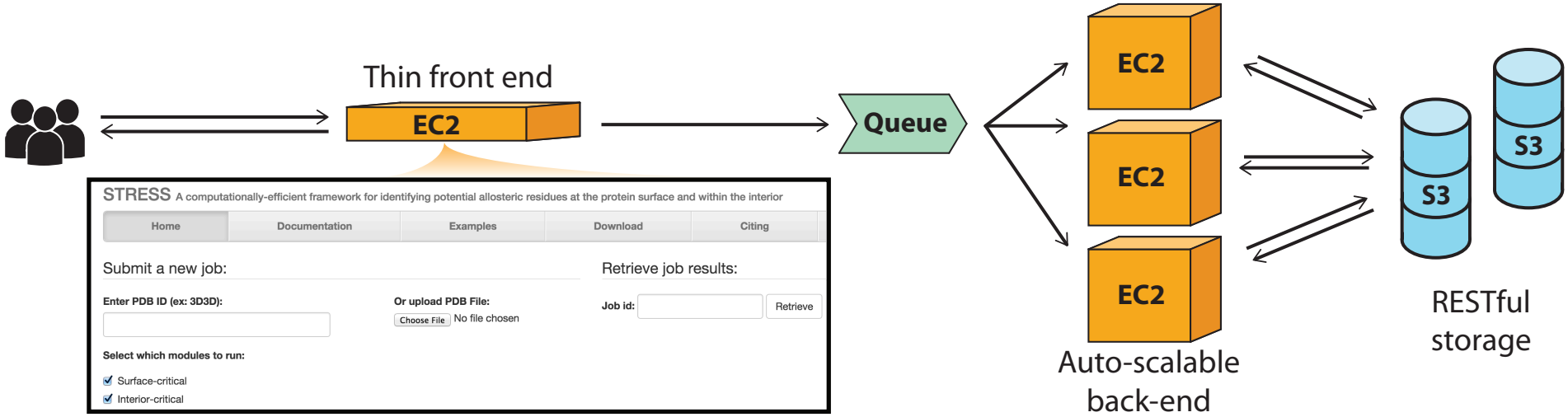
- Naïve implementation
- Optimized implementation

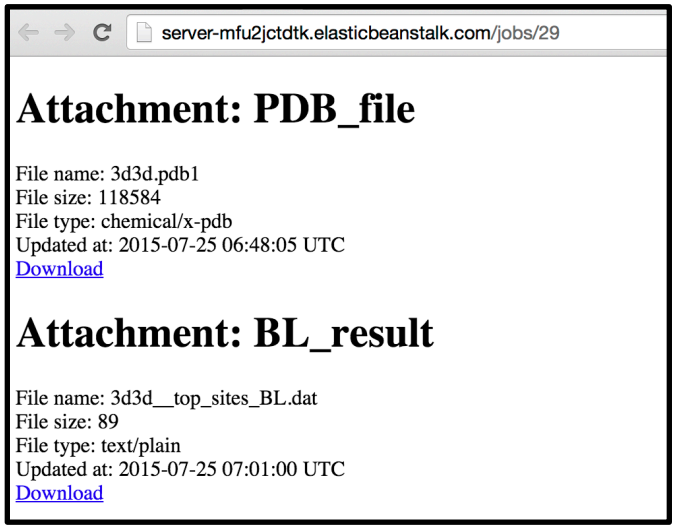




Adapted from Clarke*, Sethi*, et al *(in press)*

STRESS Server






```
11   LYS A
25   GLU A
136  ASP A
204  ASN A
.
.
```

Adapted from Clarke*, Sethi*, et al *(in press)*

1. Models for predicting allosteric hotspots

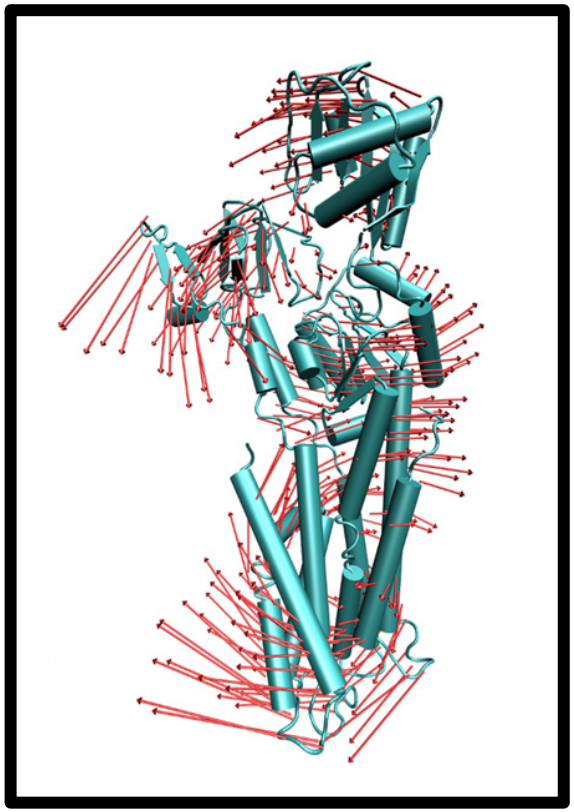
2. Speed optimization & web server to predict allosteric sites on a large scale

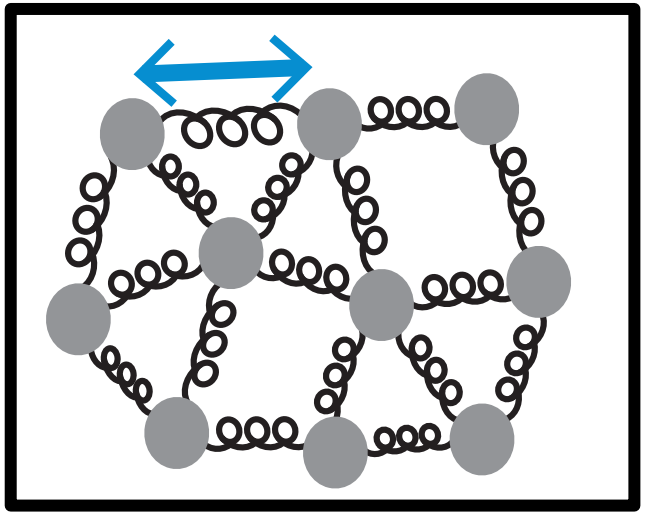
3. **Identifying alternative conformations throughout large protein datasets**

4. Signatures of conservation

Models of Protein Conformational Change

Motion Vectors from Normal Modes (ANMs)

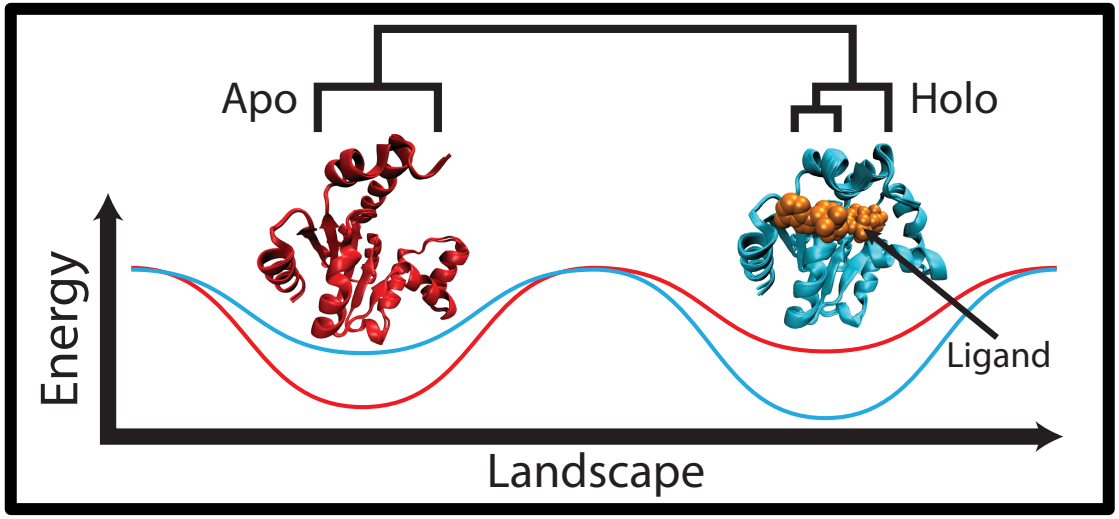


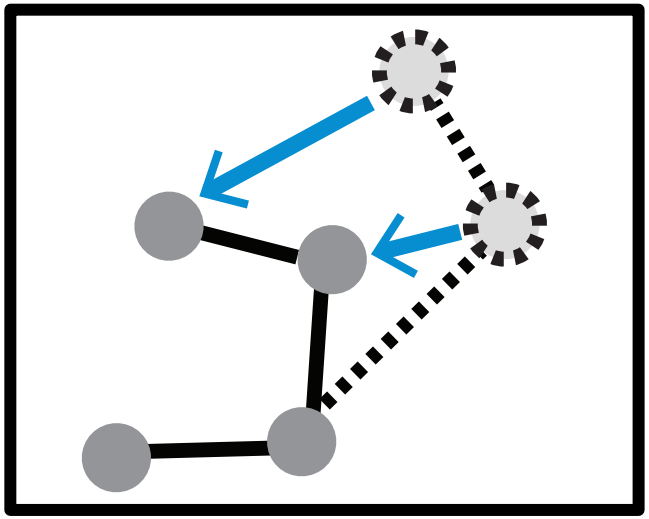


- harmonic approximations
- does not account for solvent damping
- no info regarding energy barriers/crossing events

Models of Protein Conformational Change

Motion Vectors from X-Ray Structures of Alternative Conformations (ACT)

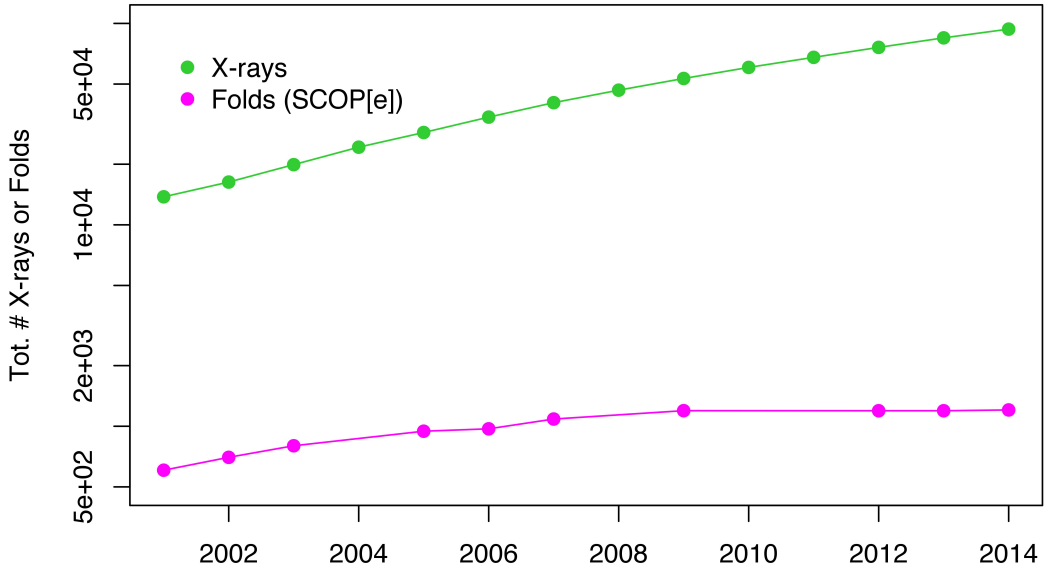


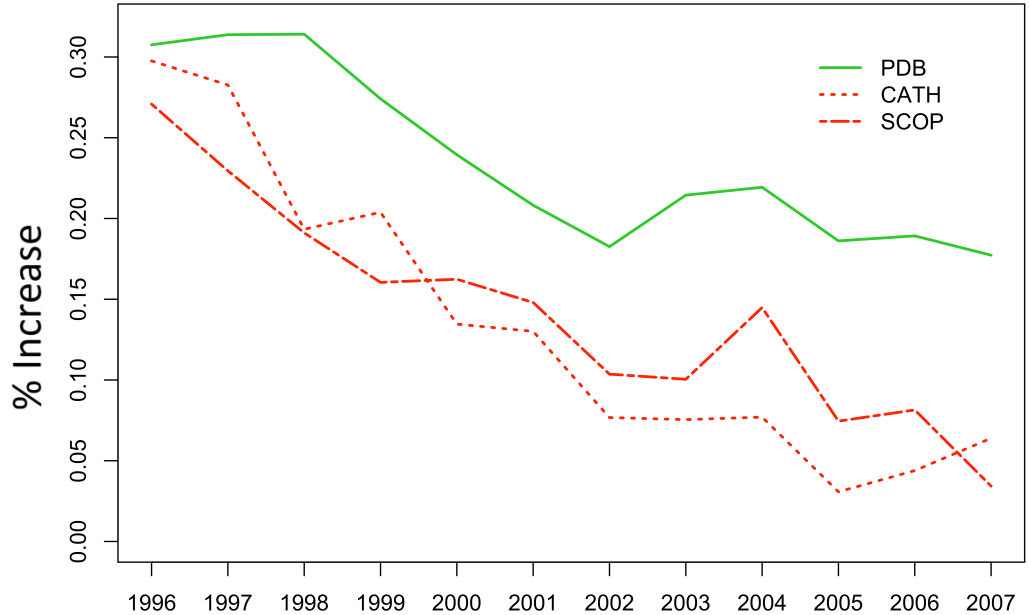


Adapted from Clarke*, Sethi*, et al *(in press)*

Identifying alternative conformations across the PDB

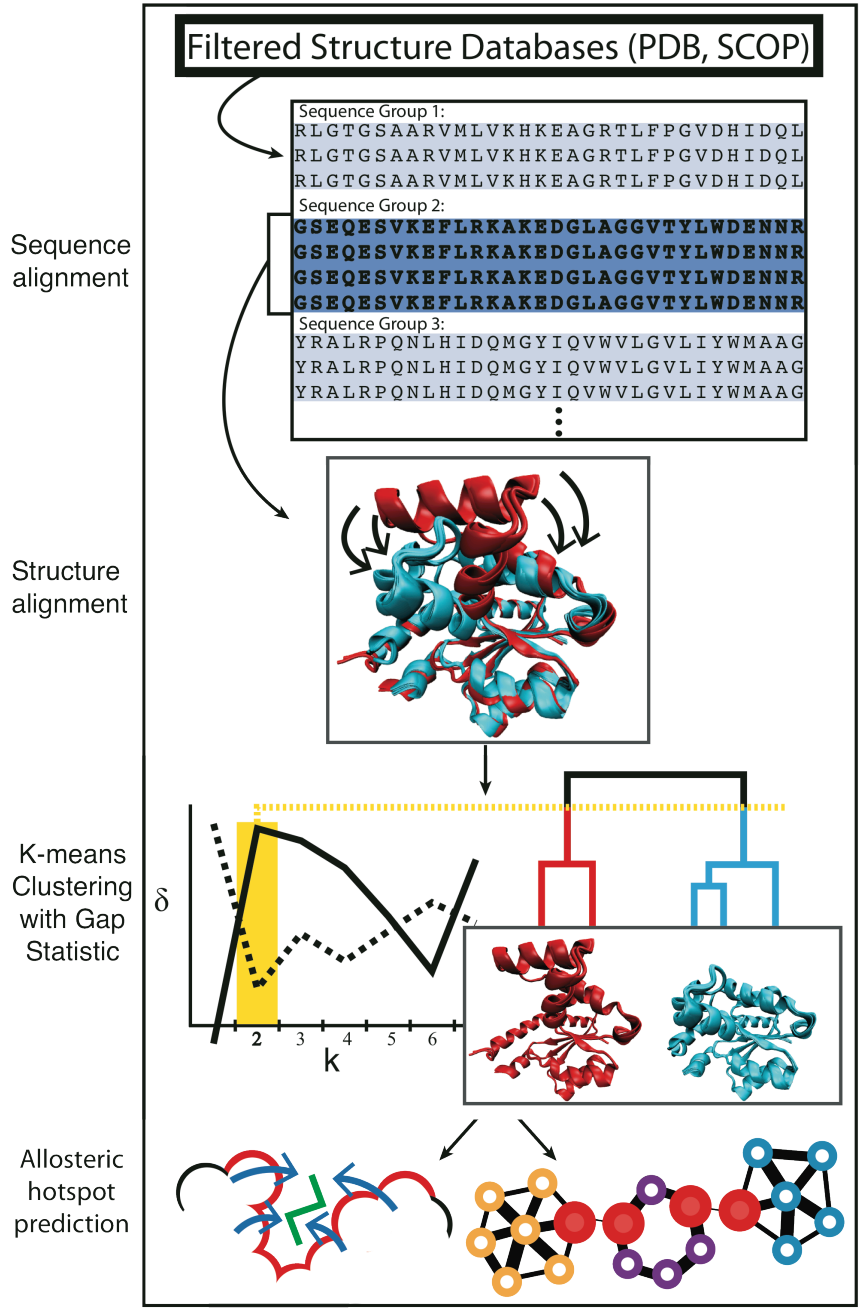
Growing sequence redundancy in the PDB (as evidenced by a reduced pace of novel fold discovery) offers a more comprehensive view of how such sequences occupy conformational landscapes





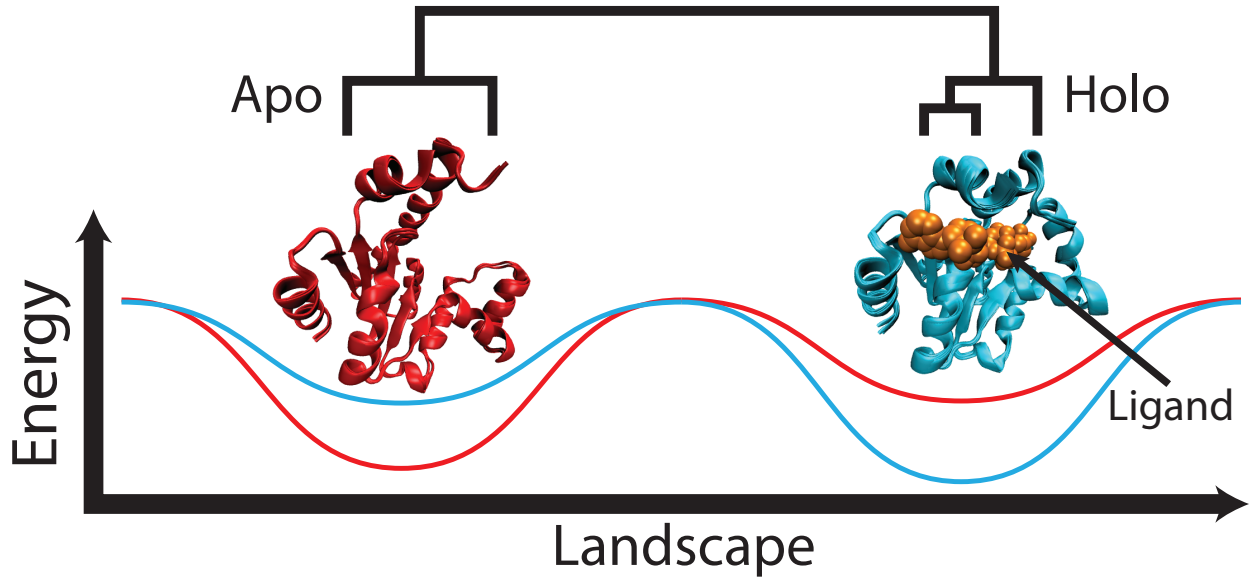
PDB: Berman HM, et al. NAR. (2000)
CATH: Sillitoe I, et al. NAR. (2015)
SCOP: Fox NK et al. NAR. (2014)

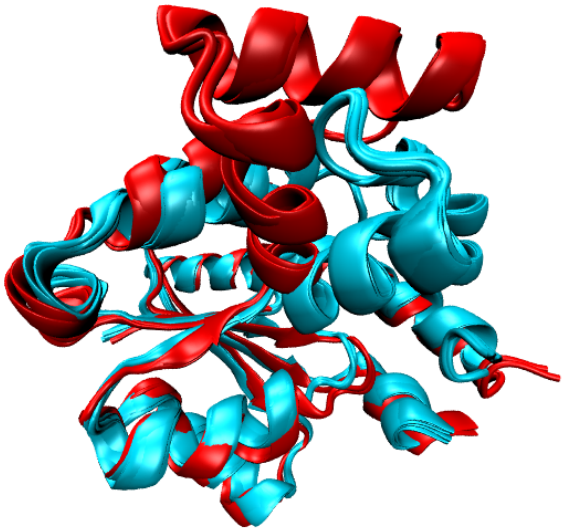
Identifying alternative conformations across the PDB



Adapted from Clarke*, Sethi*, et al *(in press)*

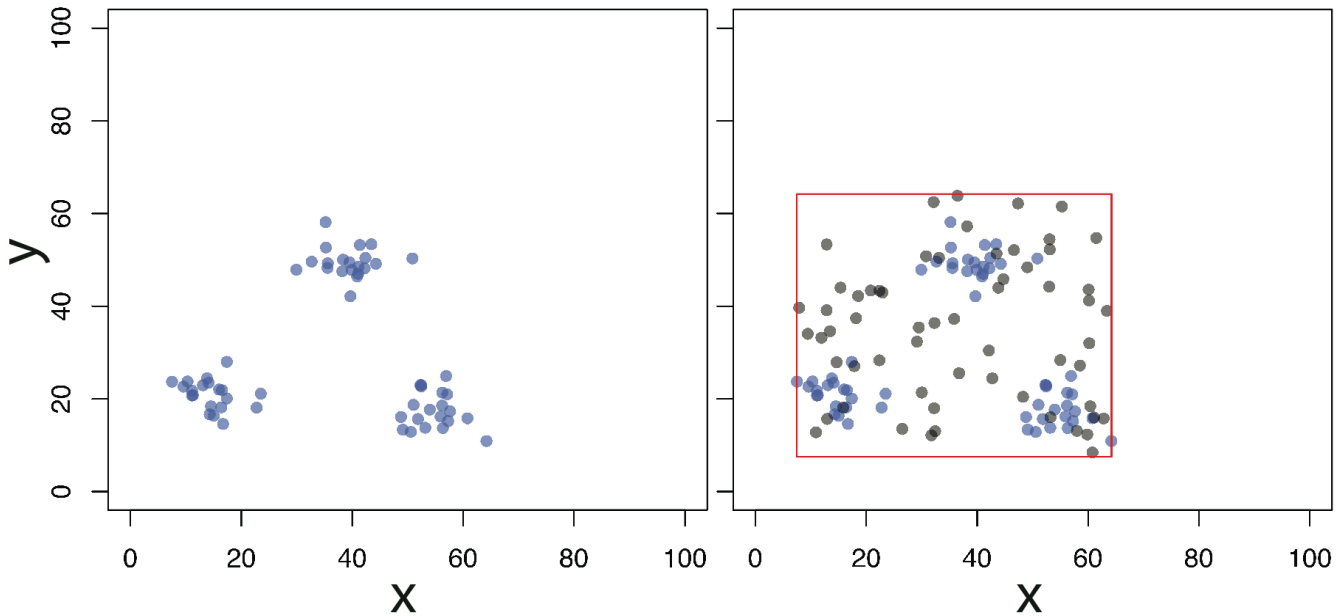
Identifying alternative conformations across the PDB





Matrix of RMSDs

	a	b	c	d
Domain a	0.0	0.1	2.2	2.1
Domain b	0.1	0.0	2.4	2.3
Domain c	2.2	2.4	0.0	0.1
Domain d	2.1	2.3	0.1	0.0



Dk: Measure to describe how compact cluster k is

$$D_k = \sum_{\mathrm{x}_i \in C_k} \sum_{\mathrm{x}_j \in C_k} ||\mathrm{x}_i - \mathrm{x}_j||^2$$

Wk: Normalized sum of these measures for a given 'partition'

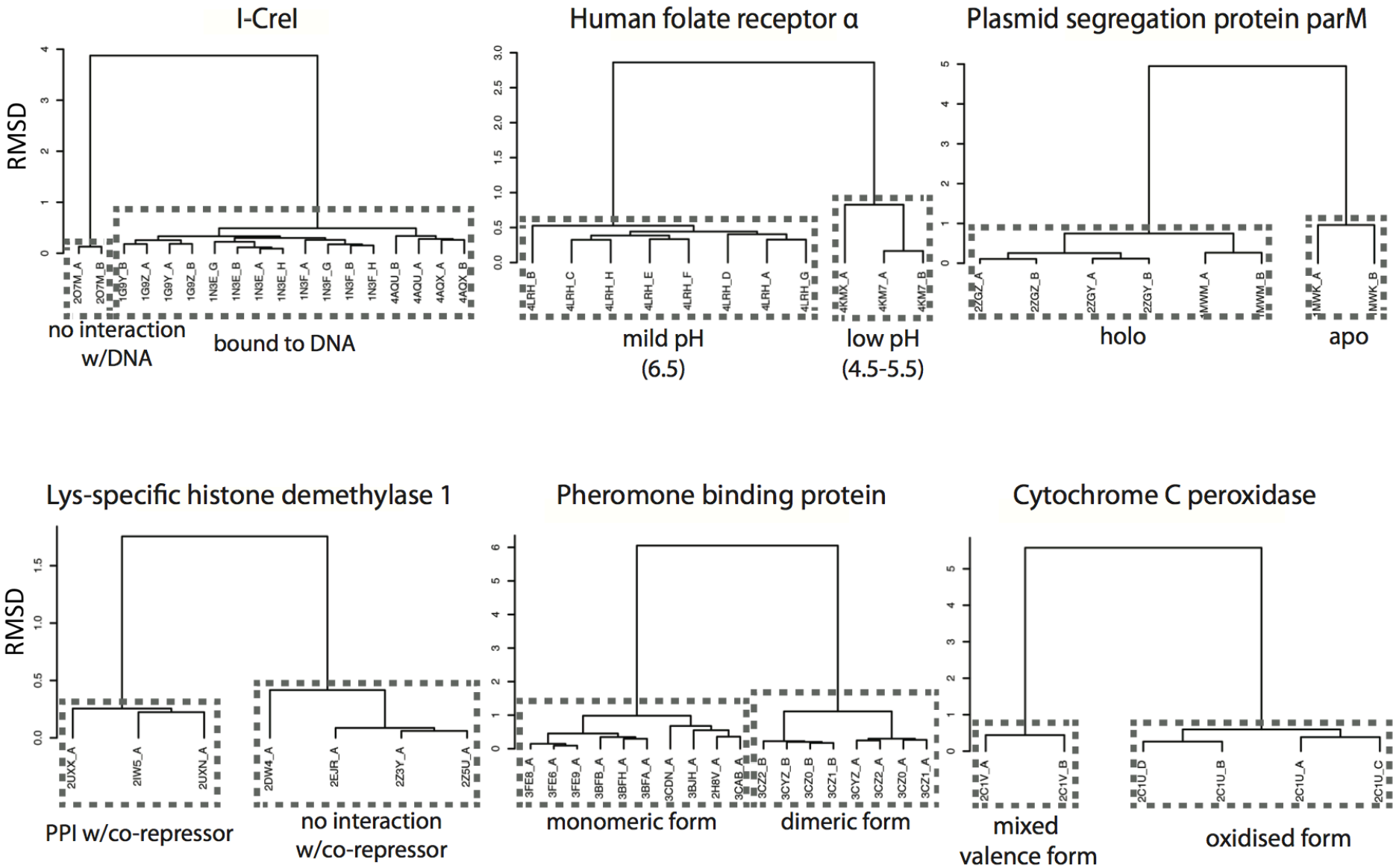
$$W_k = \sum_{k=1}^{K} \frac{1}{2n_k} D_k$$

How much does this score differ from that in a randomized null?

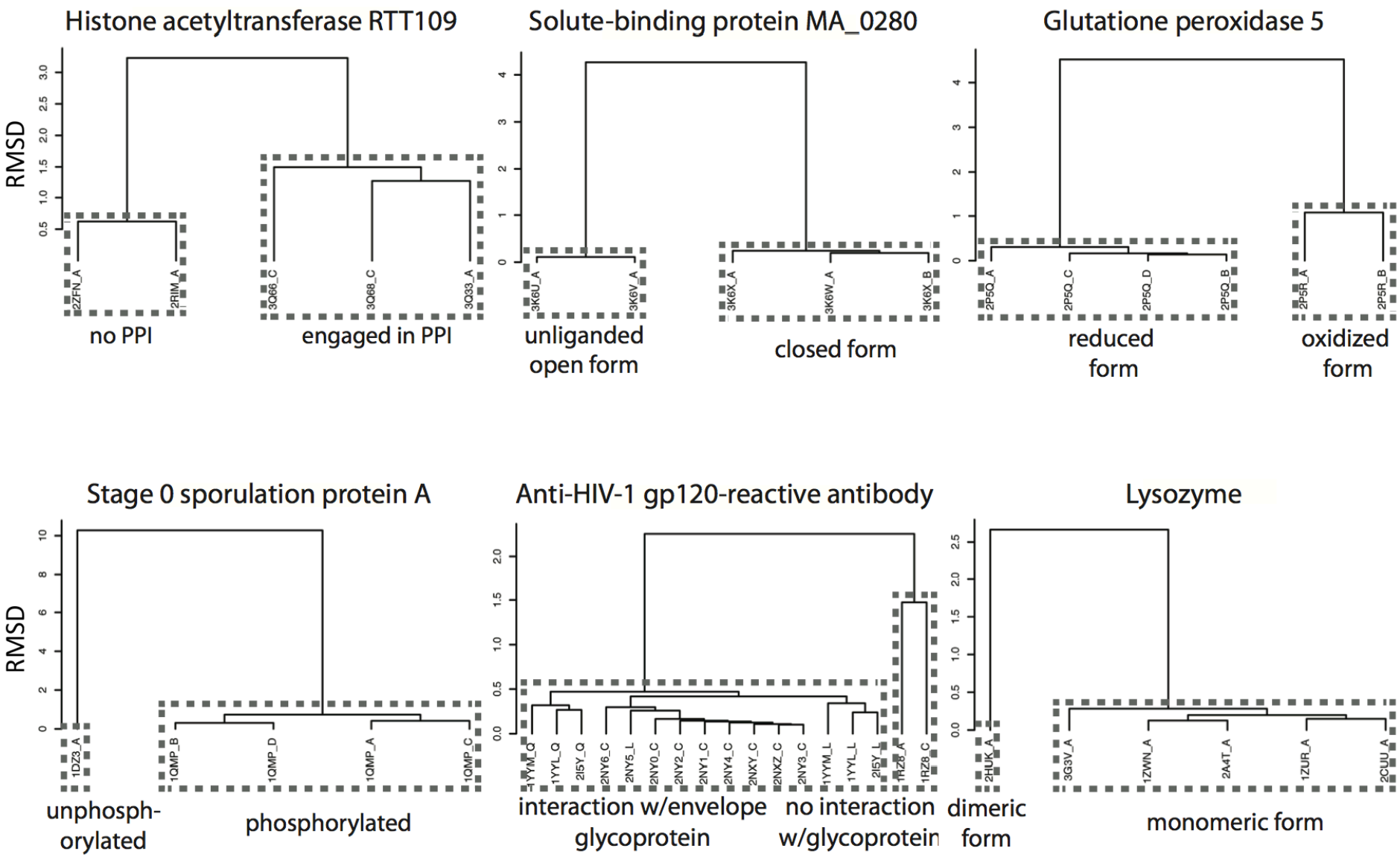
$$\mathrm{Gap}_n(k) = E_n^*\{\log W_k\} - \log W_k$$

Tibshirani et al., 2001

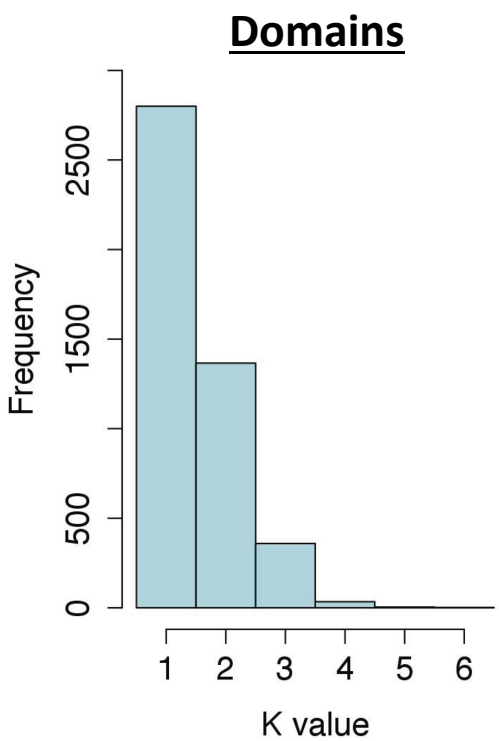
Identification of Alternative Bio States in Diverse Biological Contexts

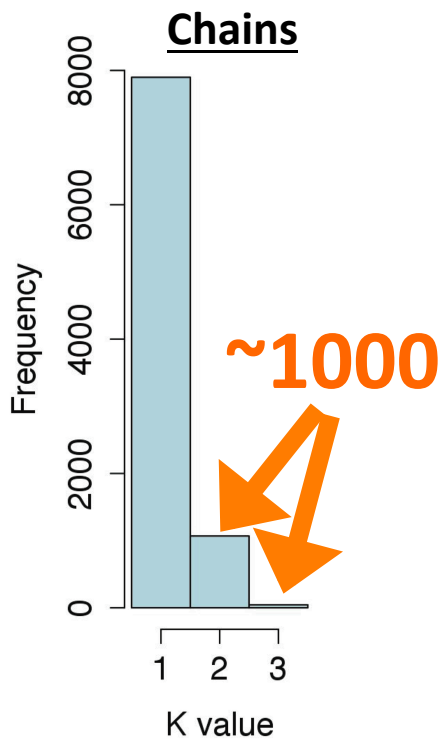


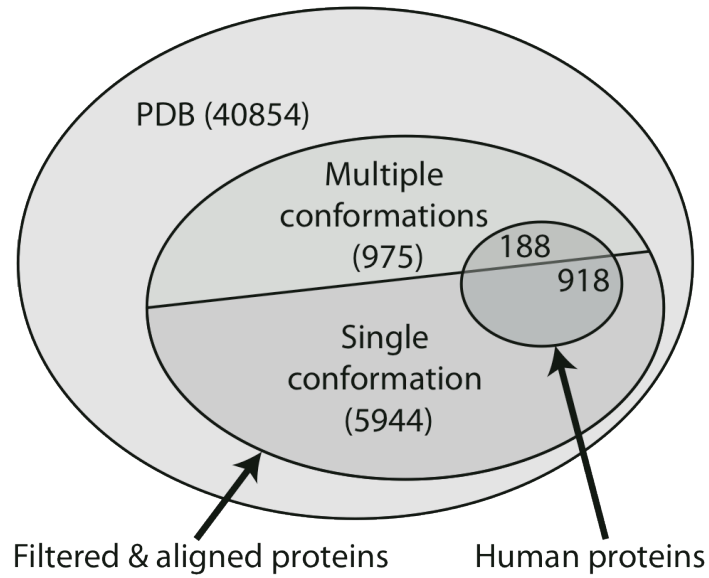
Identification of Alternative Bio States in Diverse Biological Contexts



Clustering Results







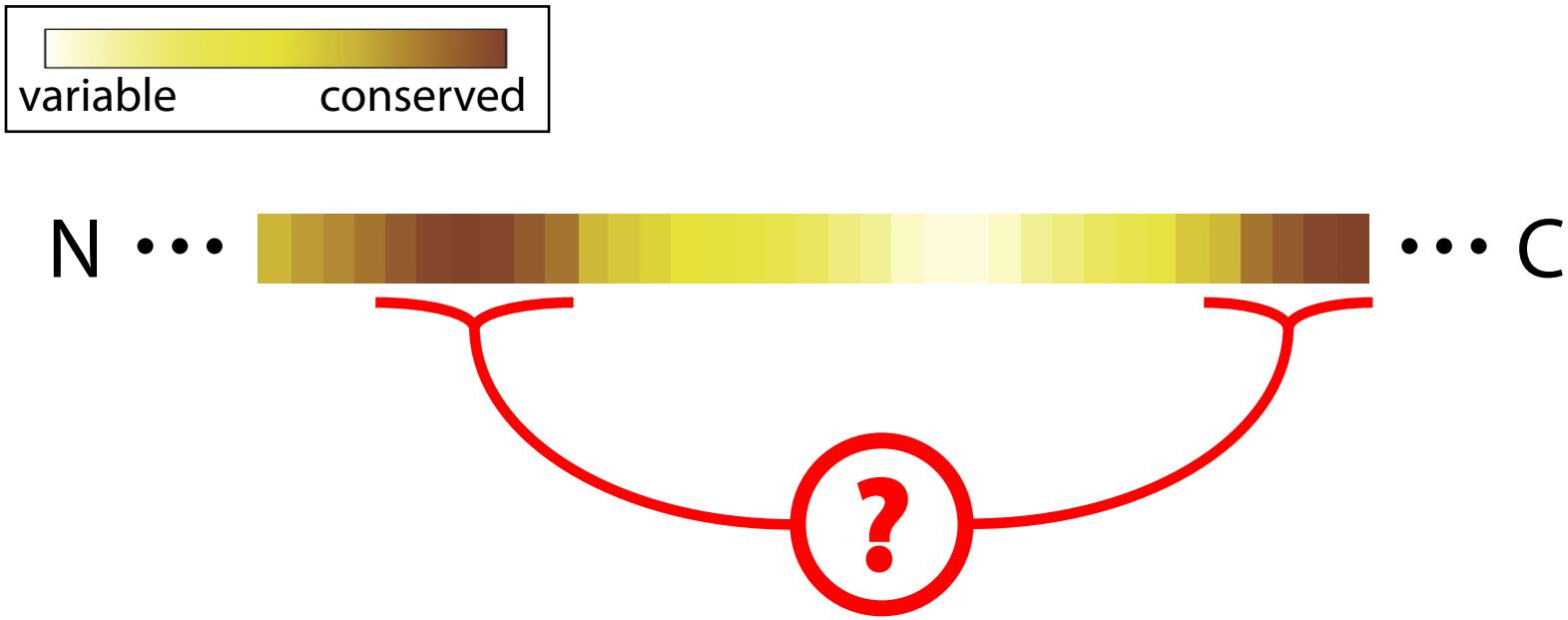
Adapted from Clarke*, Sethi*, et al *(in press)*

1. Models for predicting allosteric hotspots

2. Speed optimization & web server to predict allosteric sites on a large scale

3. Identifying alternative conformations throughout large protein datasets

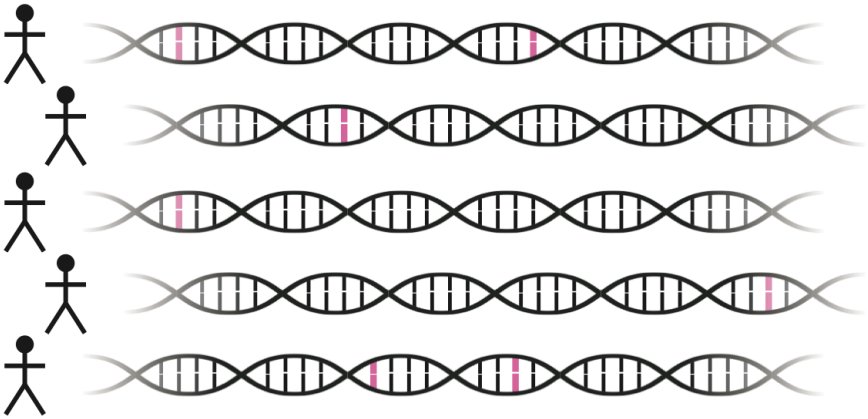
4. **Signatures of conservation**


variable
conserved
N
C
?

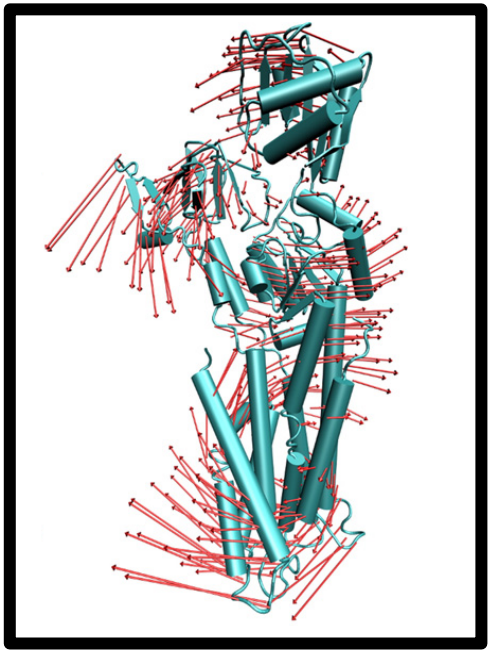
How to measure conservation?

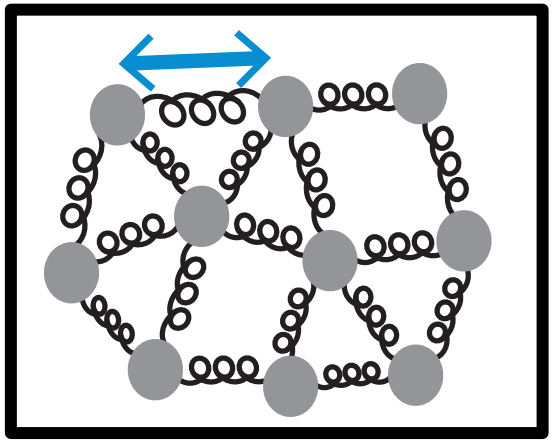
across species

amongst humans



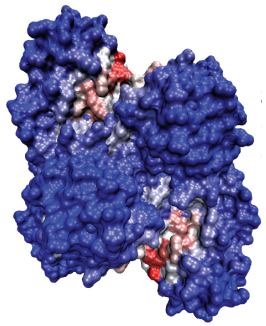
Conservation of predicted allosteric residues (using ANMs)

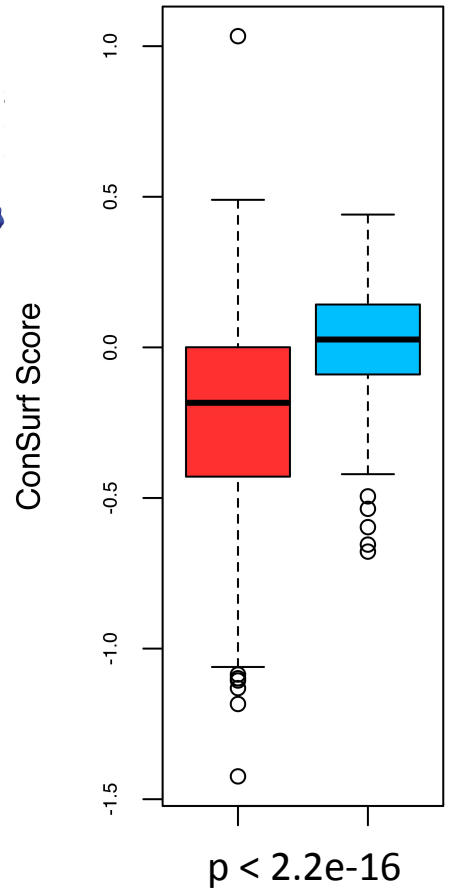


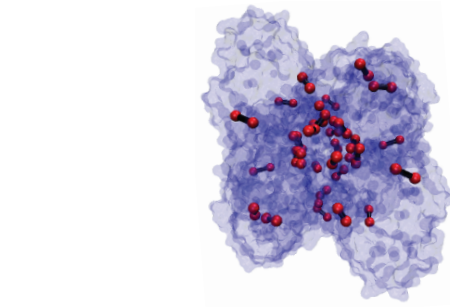


Cross-species conservation of predicted allosteric residues

Surface



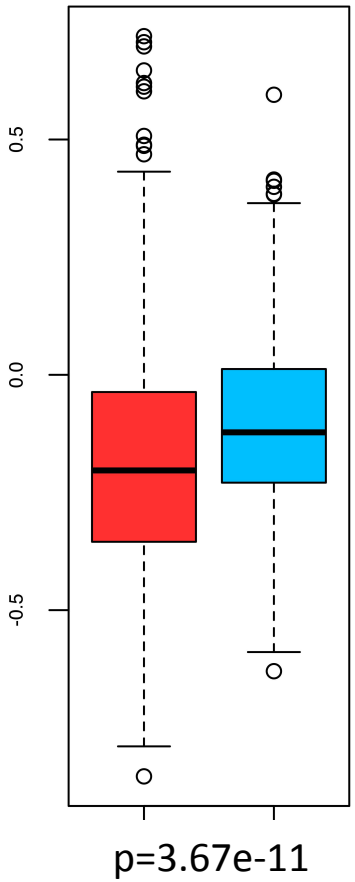




critical

non-critical

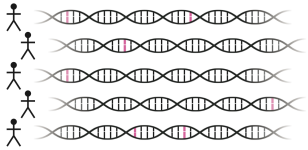
Interior



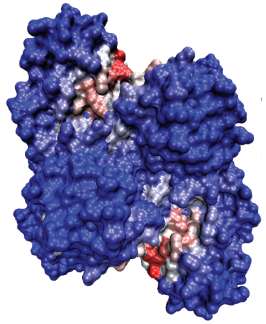
Adapted from Clarke*, Sethi*, et al *(in press)*

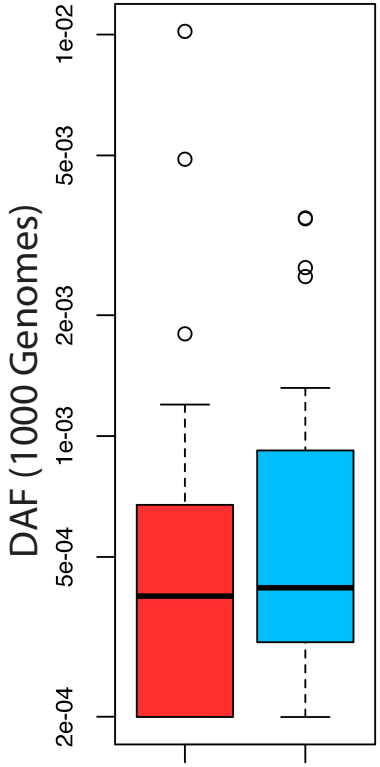
Intra-species conservation of predicted allosteric residues

1000 Genomes

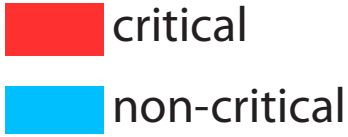


Surface

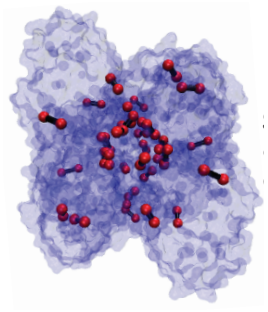


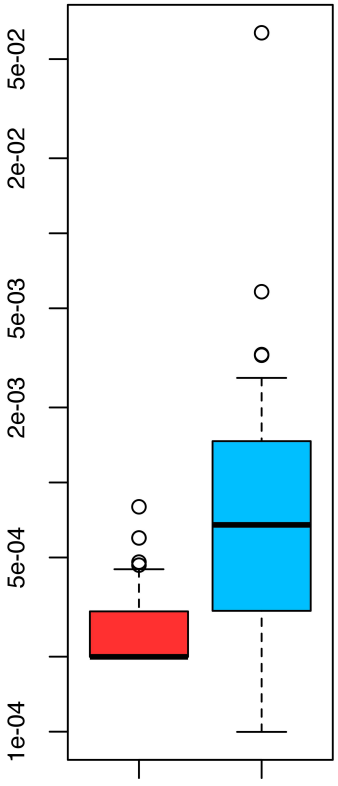


p=0.309



Interior



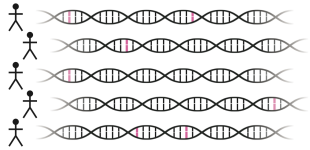


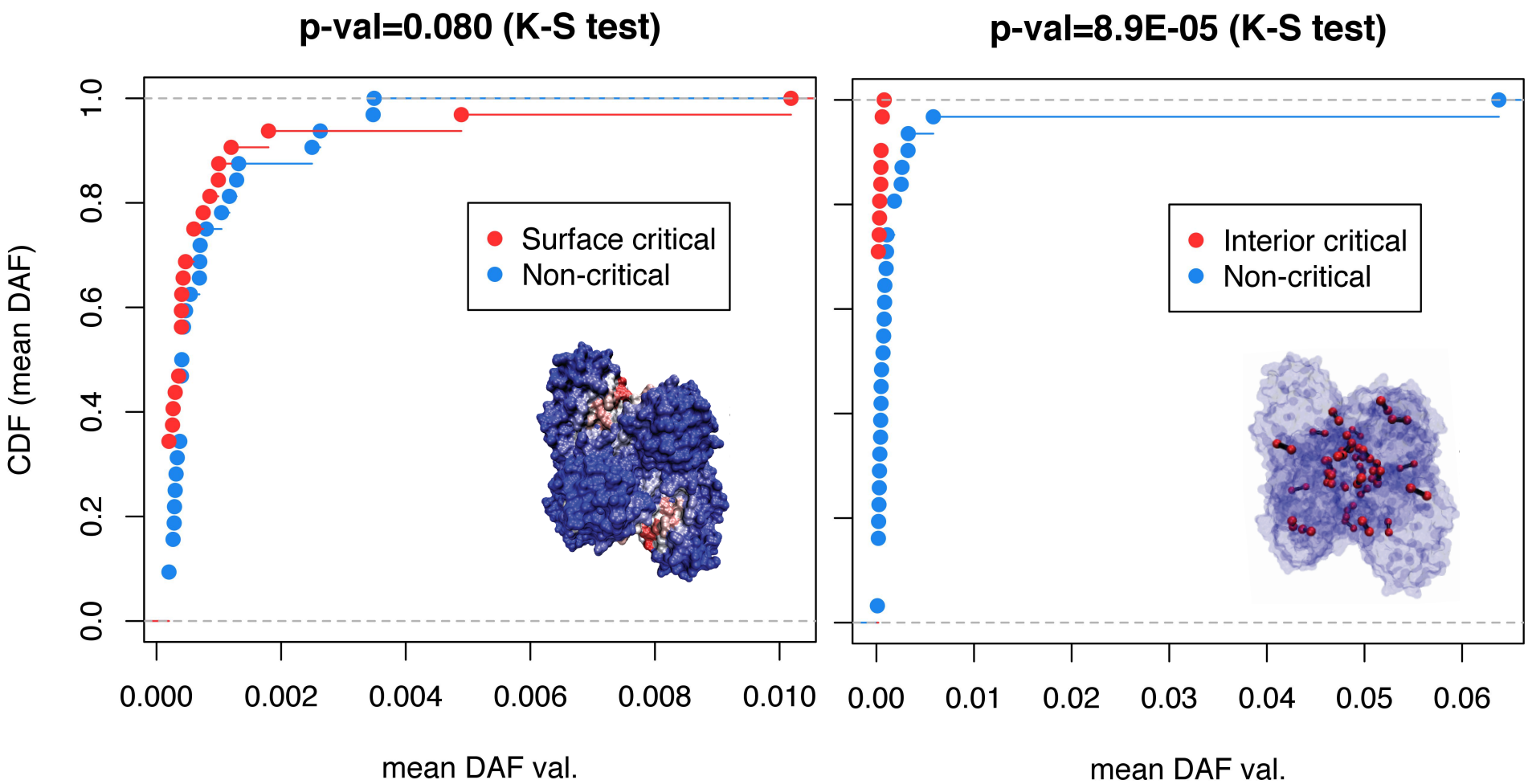
p=1.80e-05

Adapted from Clarke*, Sethi*, et al *(in press)*

Intra-species conservation of predicted allosteric residues

1000 Genomes

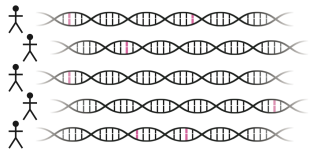




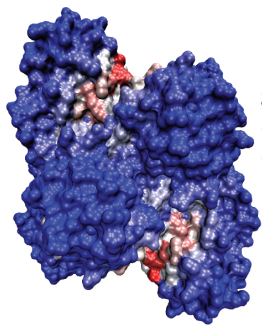
Adapted from Clarke*, Sethi*, et al *(in press)*

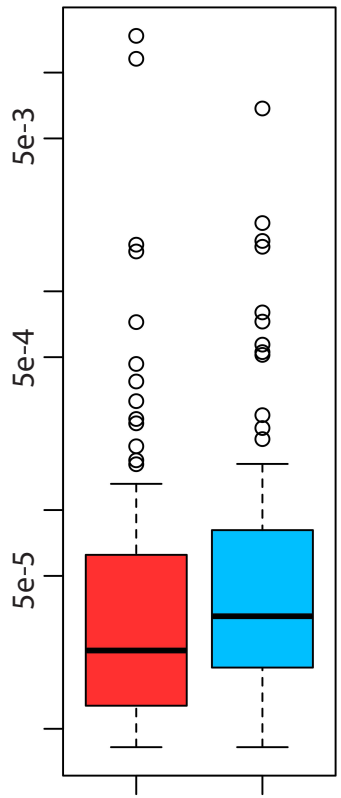
Intra-species conservation of predicted allosteric residues

ExAC

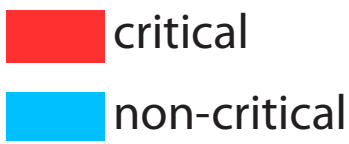


Surface

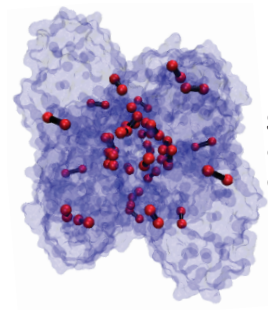


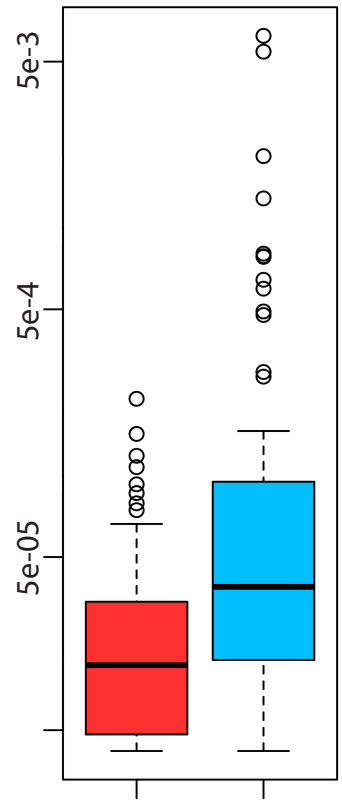


p=1.49e-3



Interior



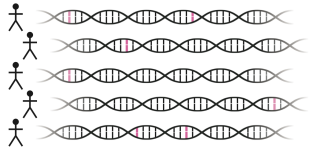


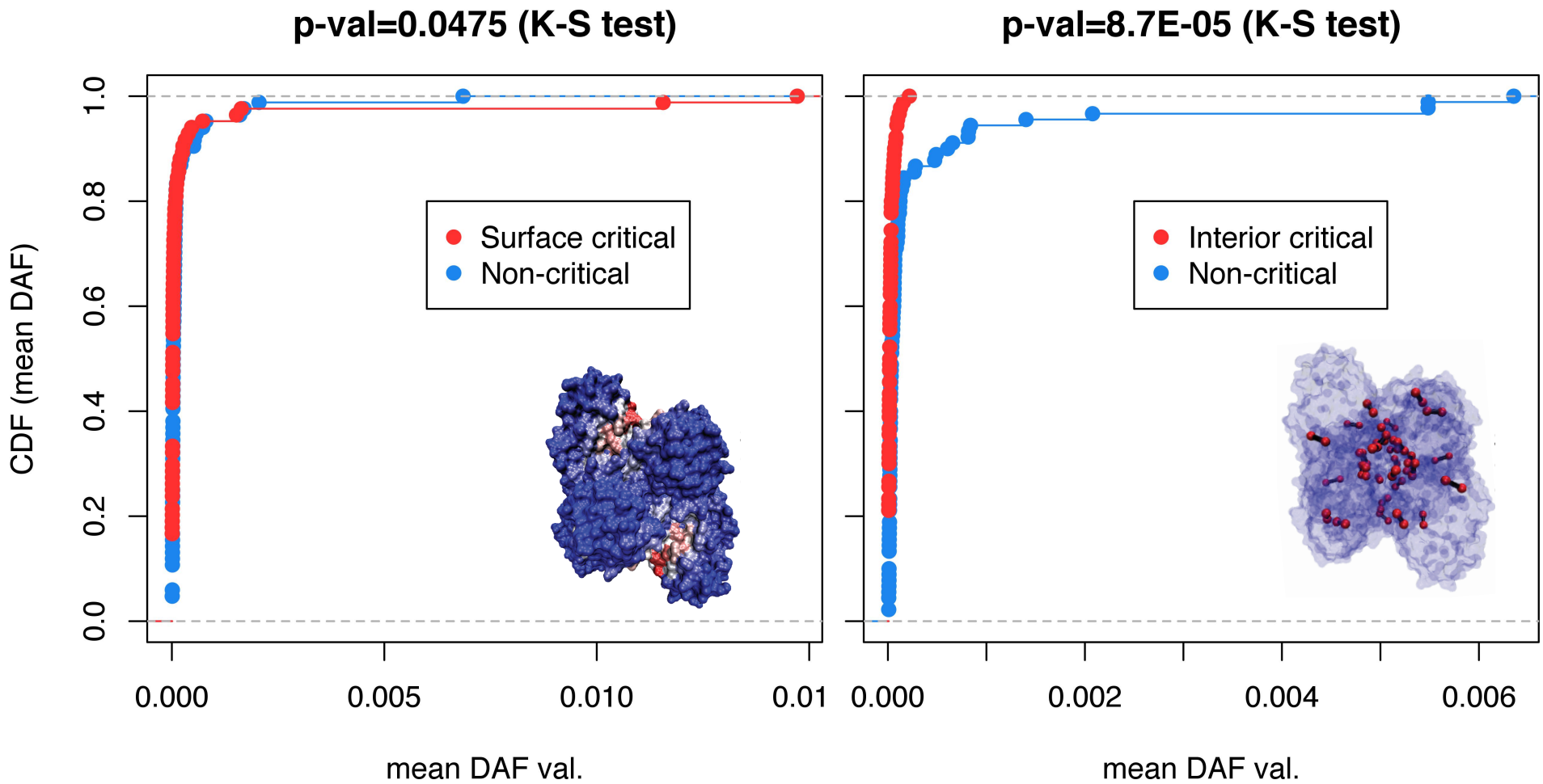
p=7.98e-09

Adapted from Clarke*, Sethi*, et al *(in press)*

Intra-species conservation of predicted allosteric residues

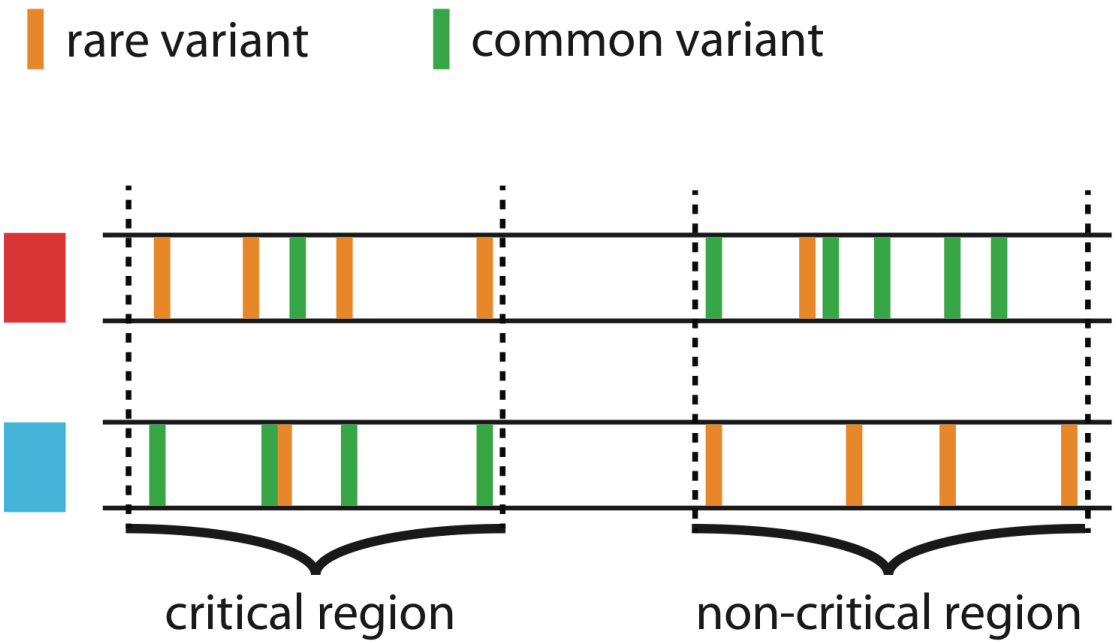
ExAC



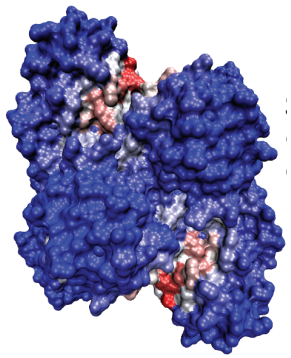


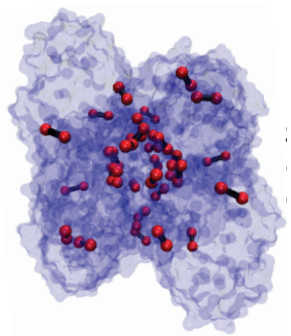
Adapted from Clarke*, Sethi*, et al *(in press)*

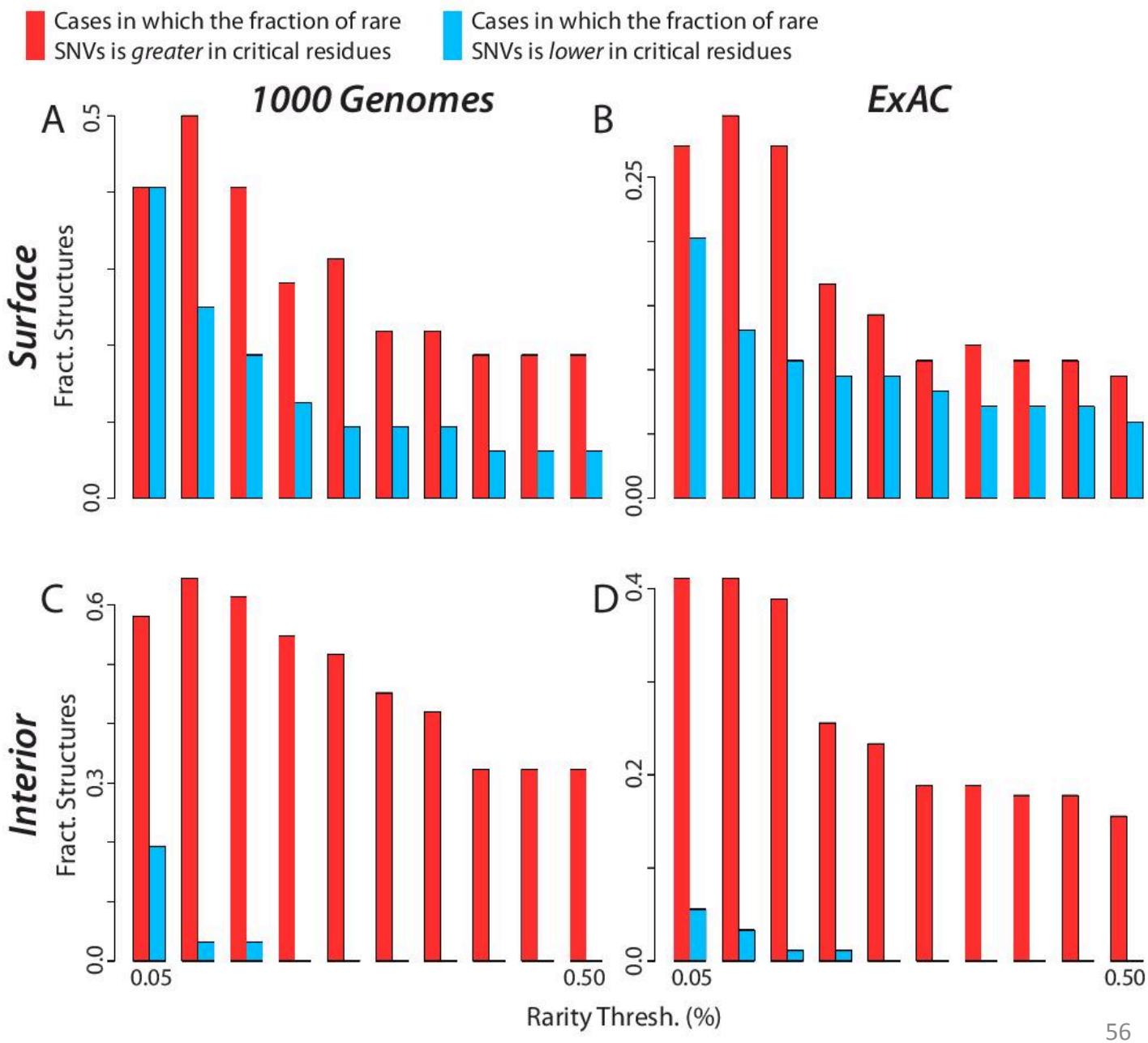
Using the *fraction of rare alleles* a conservation metric



Using the *fraction of rare alleles* a conservation metric

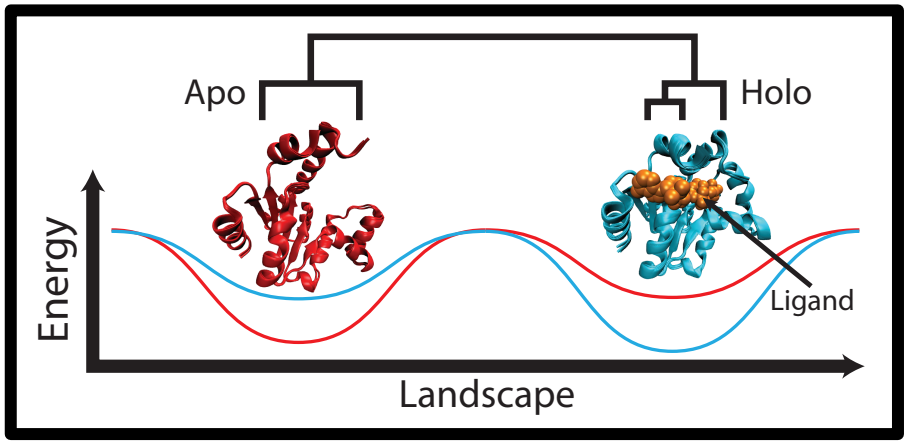


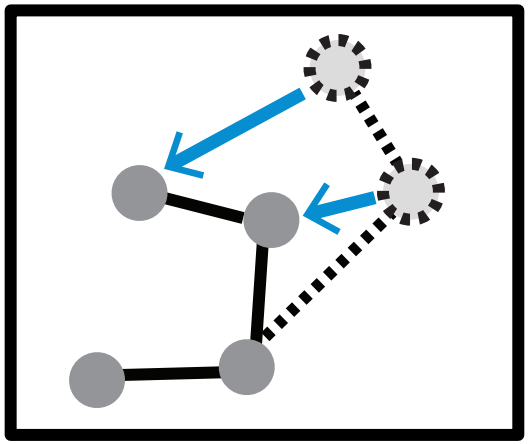




Adapted from Clarke*, Sethi*, et al *(in press)*

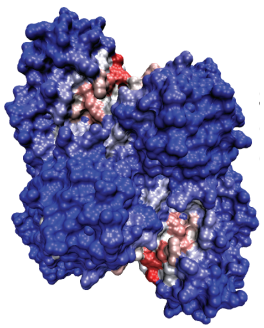
Conservation of predicted allosteric residues using alternative crystal structures ("ACT")

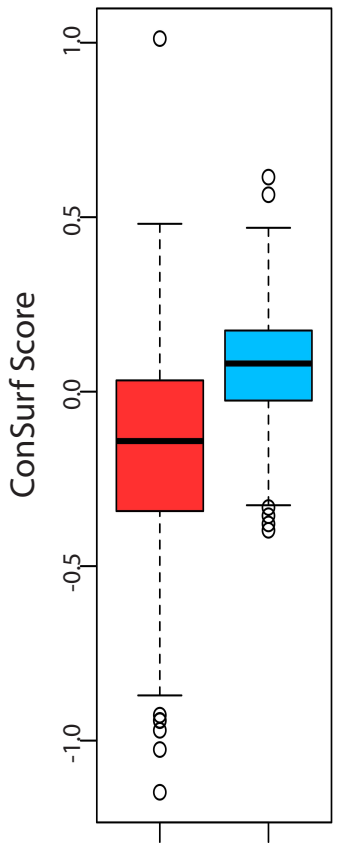




Cross-species conservation of predicted allosteric residues

Surface



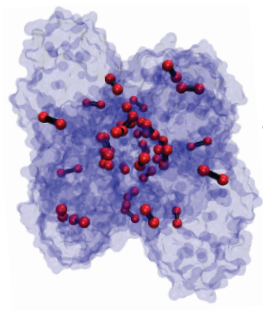


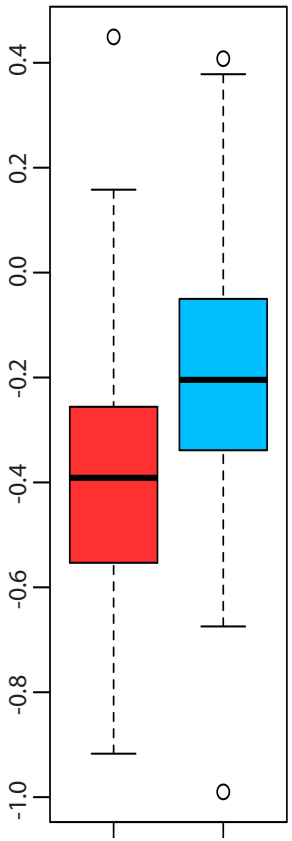
p=8.26e-14

critical

non-critical

Interior



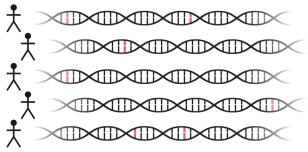


p=4.94e-13

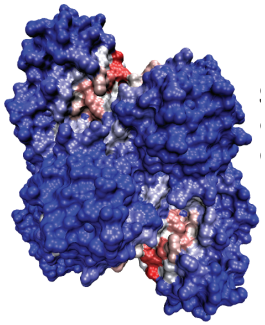
Adapted from Clarke*, Sethi*, et al *(in press)*

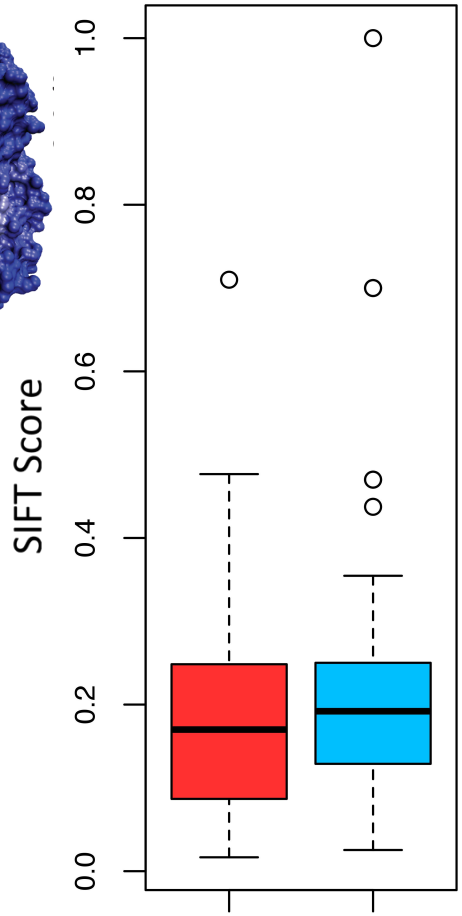
Predicted allosteric residues in the context of human health & disease

SIFT Scores on ExAC Variants



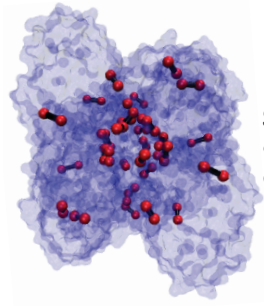
Surface

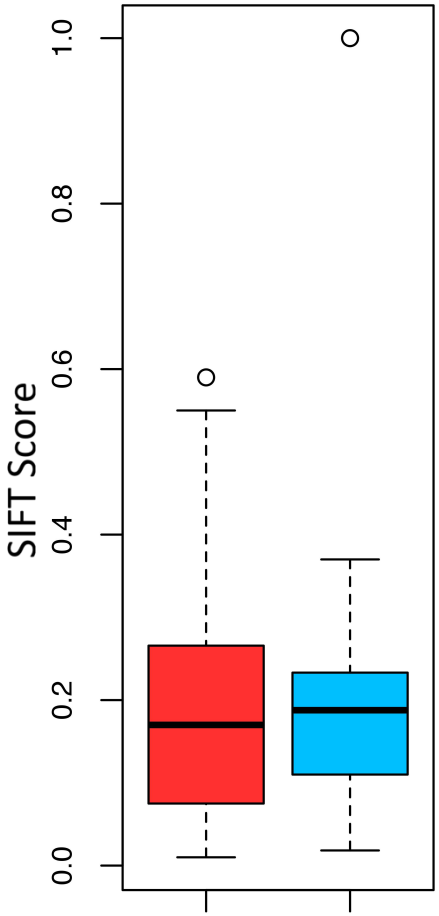




p-value = 0.089

Interior





p-value = 0.901

critical

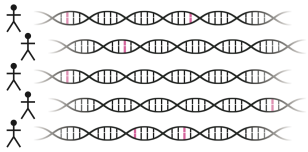
non-critical

High score: benign

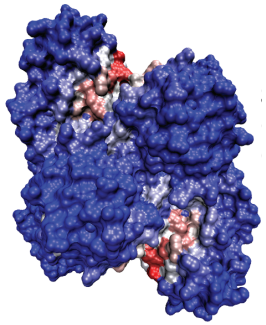
Low score: damaging

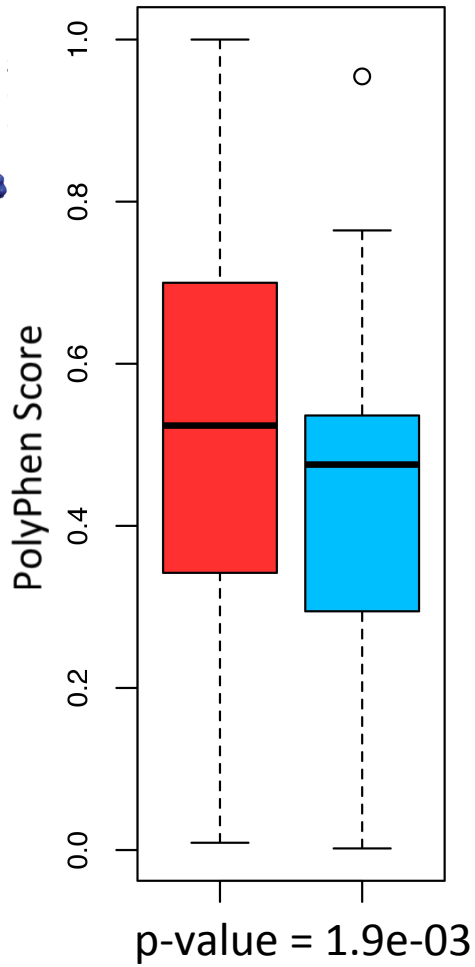
Adapted from Clarke*, Sethi*, et al *(in press)*

PolyPhen Scores on ExAC Variants

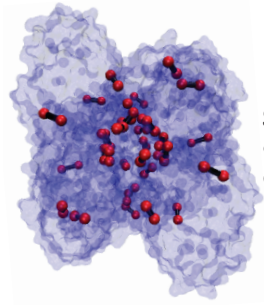


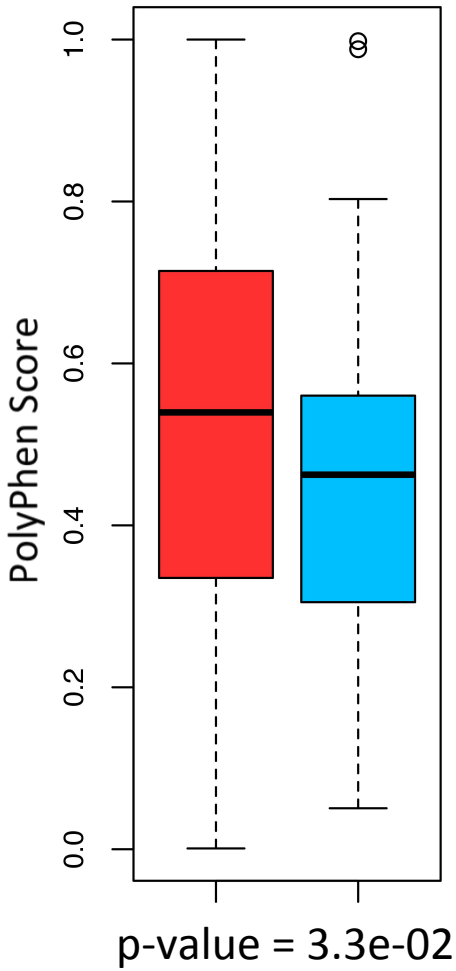
Surface





Interior





critical

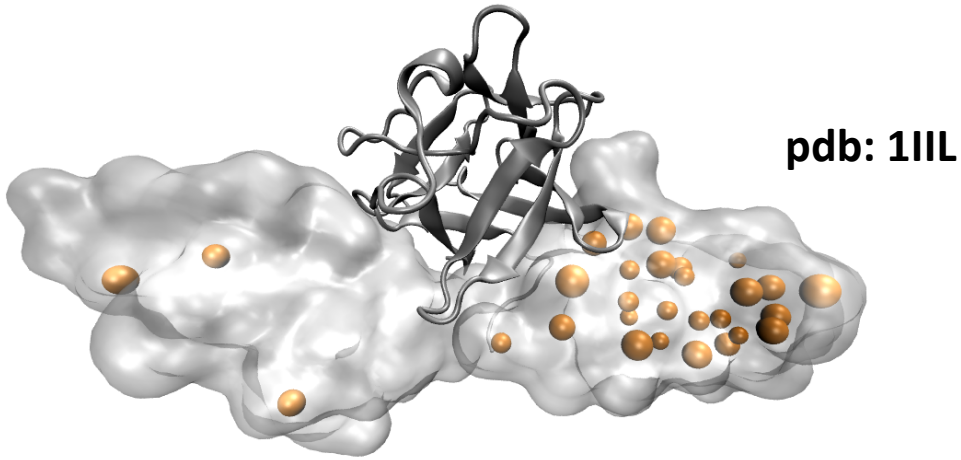
non-critical

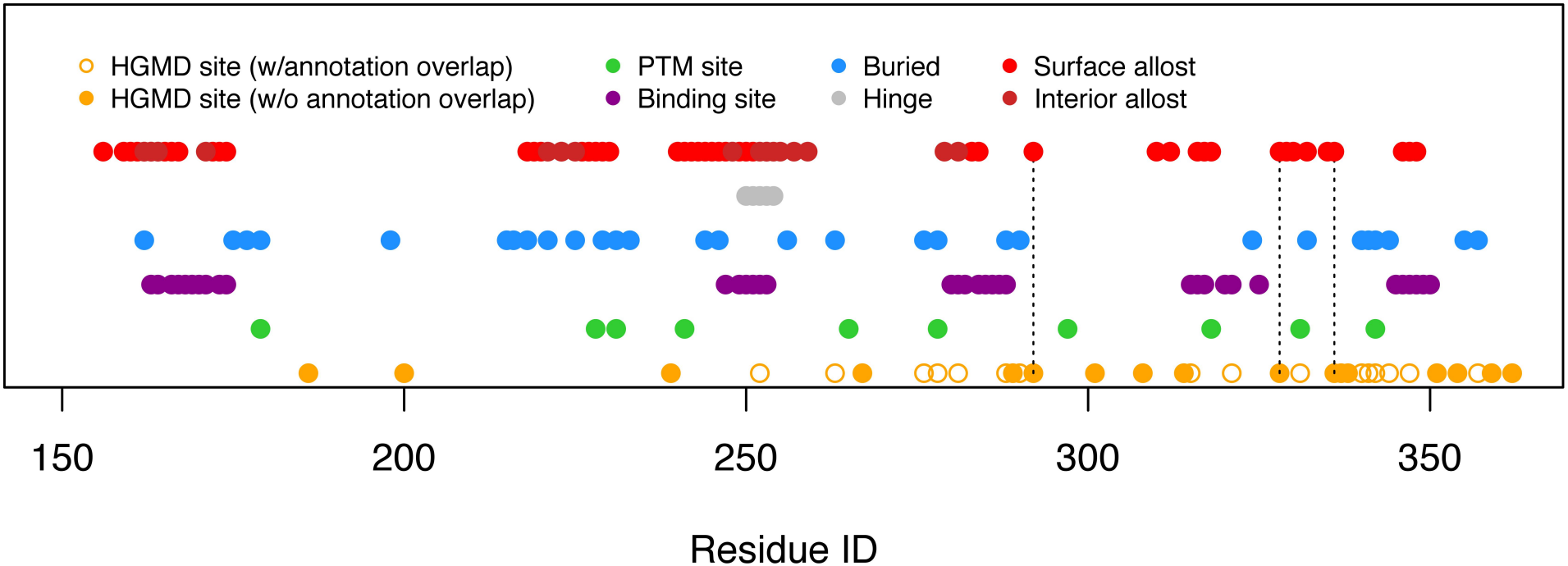
High score: damaging

Low score: benign

Adapted from Clarke*, Sethi*, et al *(in press)*

Rationalizing Disease Variants in the Context of Allosteric Behavior





Sethi et al, 2015. *Curr. Opin Struct Biol.*

Summaries

Improvements made to existing models (specifically the surface module) - including changes that enable applications to large protein datasets in a computationally tractable manner

A combination of both models as complementary approaches for predicting allosteric residues throughout the entire protein (surface and interior) within one unified study

A newly-introduced piece of software (which may either be accessed as a web server or downloaded as source code) that makes both methods more easily available to the scientific public

A downloadable database/atlas of allosteric sites within many proteins, as well as a dataset of the culled alternative conformations

The application of these models to large datasets produced through next-generation sequencing initiatives, and the finding that the predicted sites are conserved across diverse evolutionary timescales, as measured using multiple metrics and sources of data

Acknowledgements

Mark Gerstein
Gary Brudvig
Patrick Loria

Anurag Sethi
Shantao Li
Sushant Kumar
Richard Chang
Jieming Chen

Lori Iannicelli
Anne Nicotra

Koon-Kiu Yan
Arif Harmanci
Nitin Bhardwaj
Mihali Felipe
Lukas Habegger
Raymond Auerbach
Jinrui Xu
William Meyerson
Gang Fang
Mengting Gu
Suganthi Balasubramanian
Alexej Abyzov
Michael R. Schoenberg
Bo Wang
Fabio Navarro
Roger Alexander

Lucas Lochovsky
Timur Galeev
Donghoon Lee
Shaoke Lou
Xiaotong Li
Paul Muir
Yao Fu
Leonidas Salichos
Dan Spakowicz
Shuang Liu
Daifeng Wang
Yan Zhang
Baikang Pei
Jing Zhang
Joel Rozowsky
Rob Kitchen

Yale Dept. of Chemistry

NIH

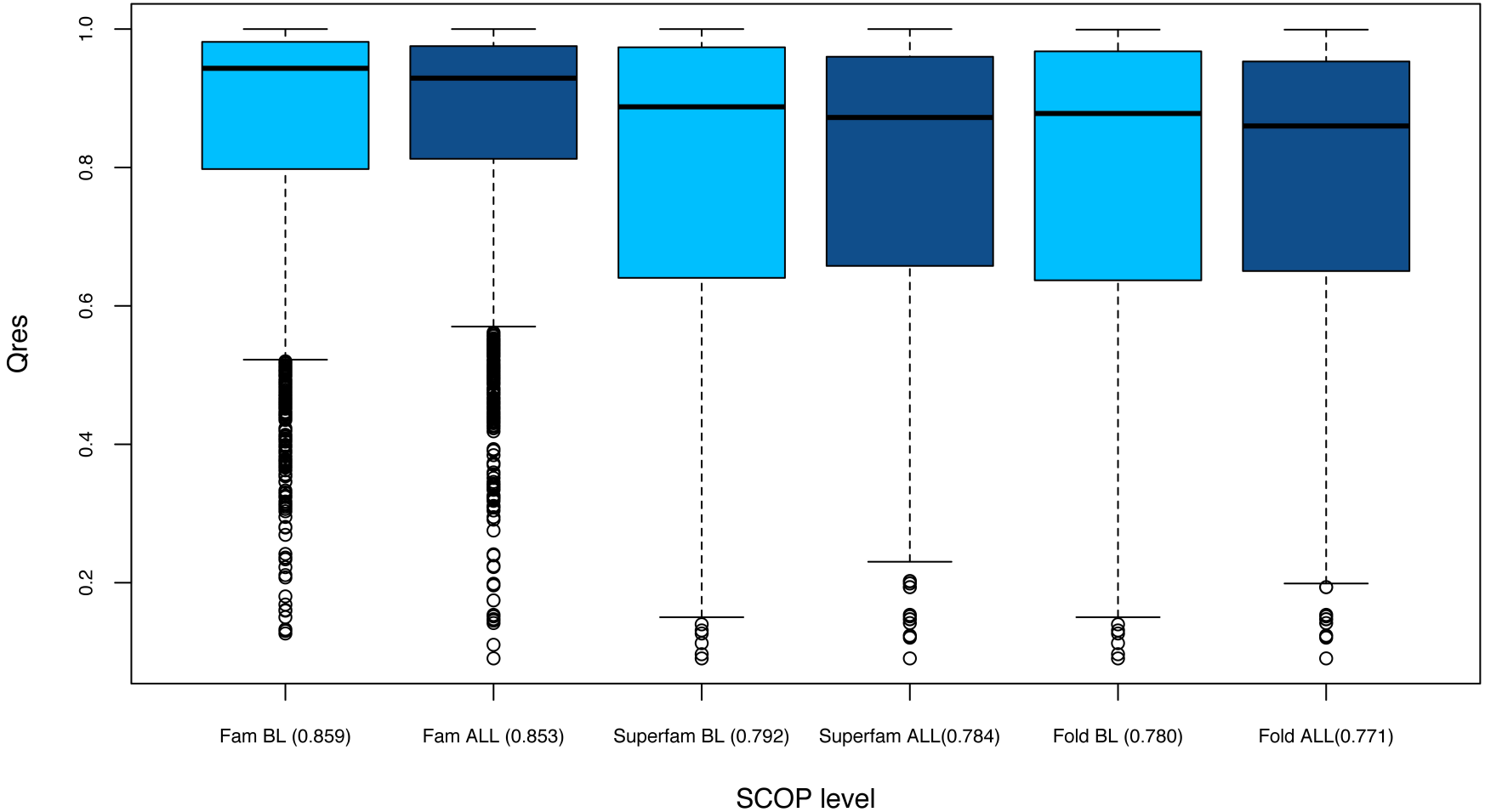
Family & Friends

NIH National Institutes of Health
Turning Discovery Into Health

Supplementary slides

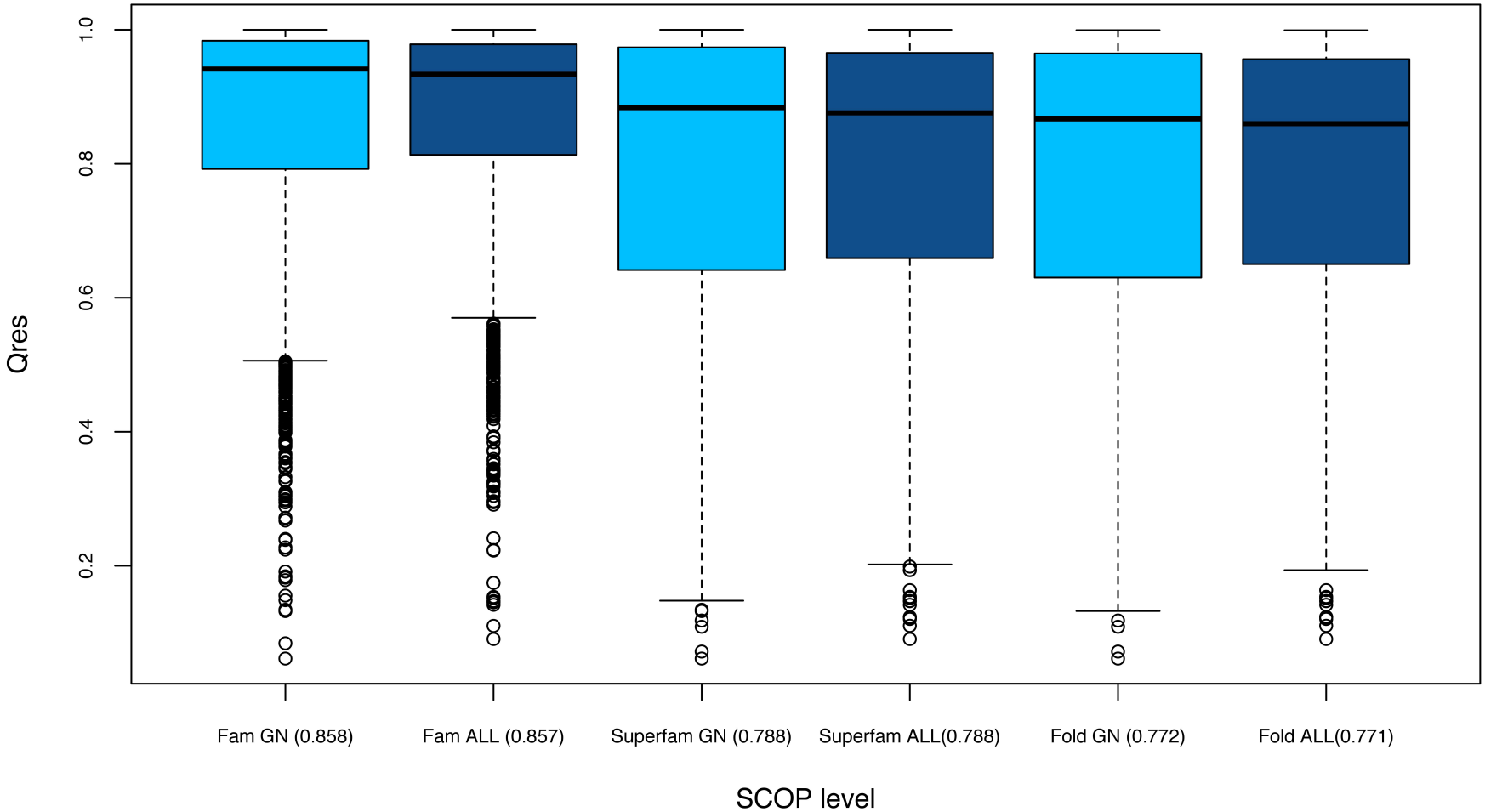
Structural Conservation

Surface



Structural Conservation

Interior



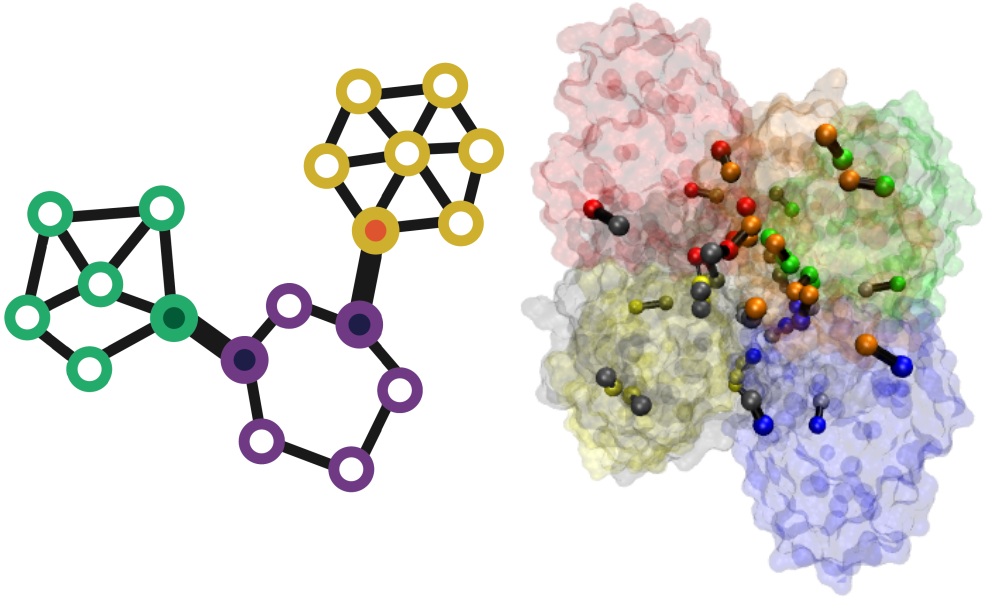
Predicting Allosterically-Important Residues within the Interior

Edge 'distance' between residues i & j is:

$$W_{ij} = -\ln(|C_{ij}|)$$

Cij is the correlation between the motions of residues i & j.

A *large* 'distance' (i.e., low correlated motion) *increases* the shortest path lengths between such residues.



Application of a gap statistic for the determination of an optimal K value in K-means clustering

$$D_k = \sum_{\mathrm{x}_i \in C_k} \sum_{\mathrm{x}_j \in C_k} ||\mathrm{x}_i - \mathrm{x}_j||^2 = 2n_k \sum_{\mathrm{x}_i \in C_k} ||\mathrm{x}_i - \mu_k||^2$$

$$W_k = \sum_{k=1}^{K} \frac{1}{2n_k} D_k$$

$$\mathrm{Gap}_n(k) = E_n^*\{\log W_k\} - \log W_k$$

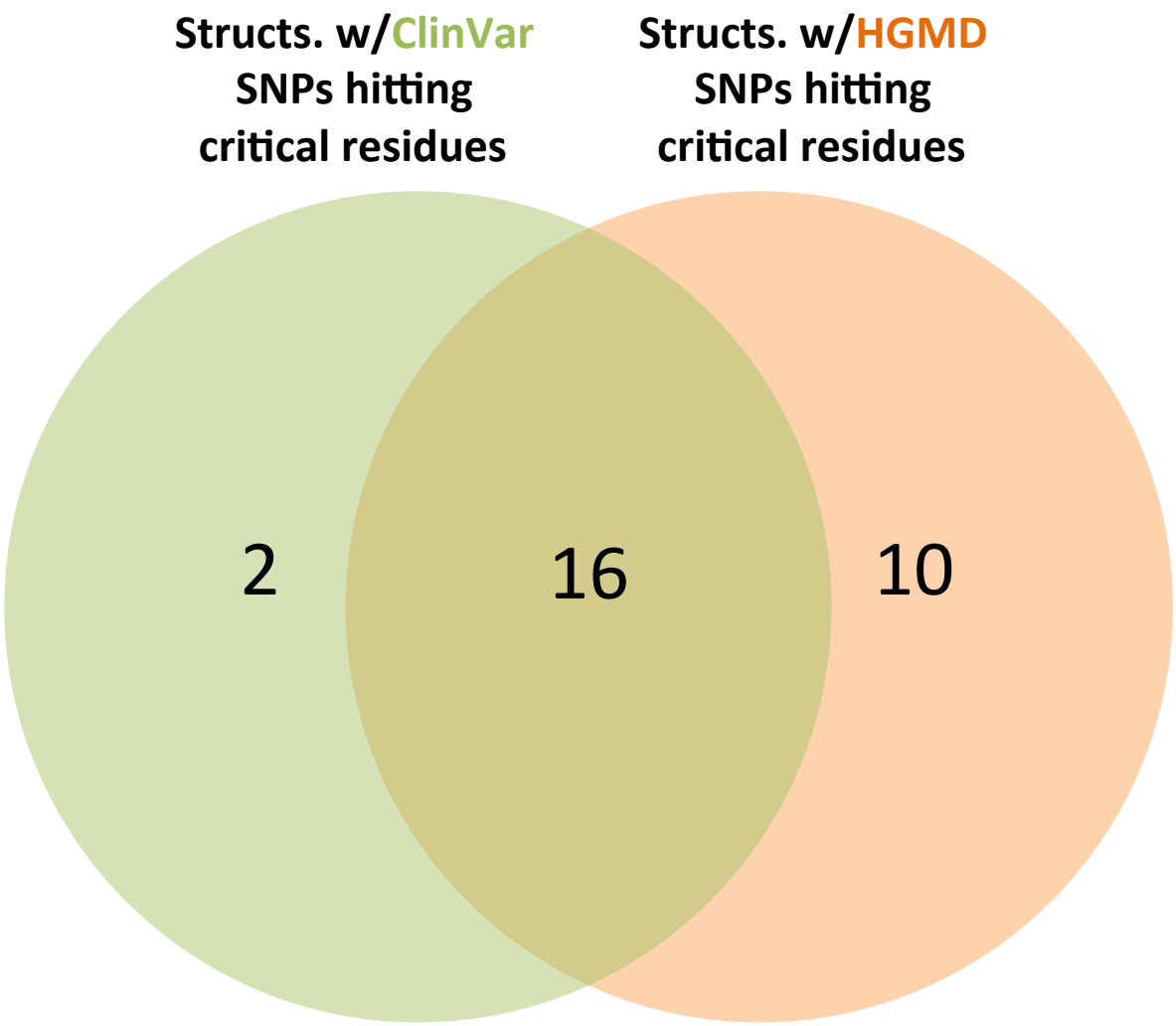
$$\mathrm{Gap}(k) \geq \mathrm{Gap}(k+1) - s_{k+1}$$

Tibshirani R. et al. Journal of the Royal Statistical Society: Series B (2001)

Binding Site Identification: GN vs. Infomap								
Protein (PDB, # residues)	**Community Detection Method: GN | InfoMap**							
	Modularity		**# Comm.**		**# Critical Residues**		**% Binding Sites Captured (expected)**	
tRNA synthetase (1N78, 542)	0.71	0.68	14	25	47	109	9.3* (8.7)	23.3 (20.1)
Adenylate kinase (4AKE, 428)	0.73	0.70	11	20	39	82	100 (99)	100 (100)
Arginine Kinase (3JU5, 728)	0.72	0.69	12	28	41	142	75 (41)	100 (86)
Tyrosine Phosphatase (2HNP, 278)	0.59	0.59	7	15	27	70	100 (43)	100 (78)
Phosphoribosyltransferase (1XTT, 846)	0.72	0.68	9	32	36	174	50 (36)	100 (90)
cAMP-dep. PK (1J3H, 332)	0.66	0.64	11	19	36	78	50 (54)	50 (70)
Anthranilate synthase (1I7Q, 1418)	0.75	0.69	12	46	51	288	25 (23)	50 (76)
Malic enzyme (1EFK, 2212)	0.81	0.72	17	70	74	425	25 (43)	100 (96)
Threonine synthase (1E5X, 884)	0.73	0.69	13	36	43	192	50 (31)	75 (69)
G-6-P Deaminase (1CD5, 1596)	0.79	0.72	18	54	58	266	8.3 (29)	100 (76)
Phosphofructokinase (3PFK, 1276)	0.76	0.68	10	51	45	307	37.5 (29)	87.5 (92)
Tryptophan synthase (1BKS, 1294)	0.77	0.69	10	46	41	284	-	
Means	***0.73***	***0.68***	***12.0***	***36.8***	***44.8***	***201.4***	***48.2 (39.7)***	***80.5 (77.6)***

* used only residues for 1N78

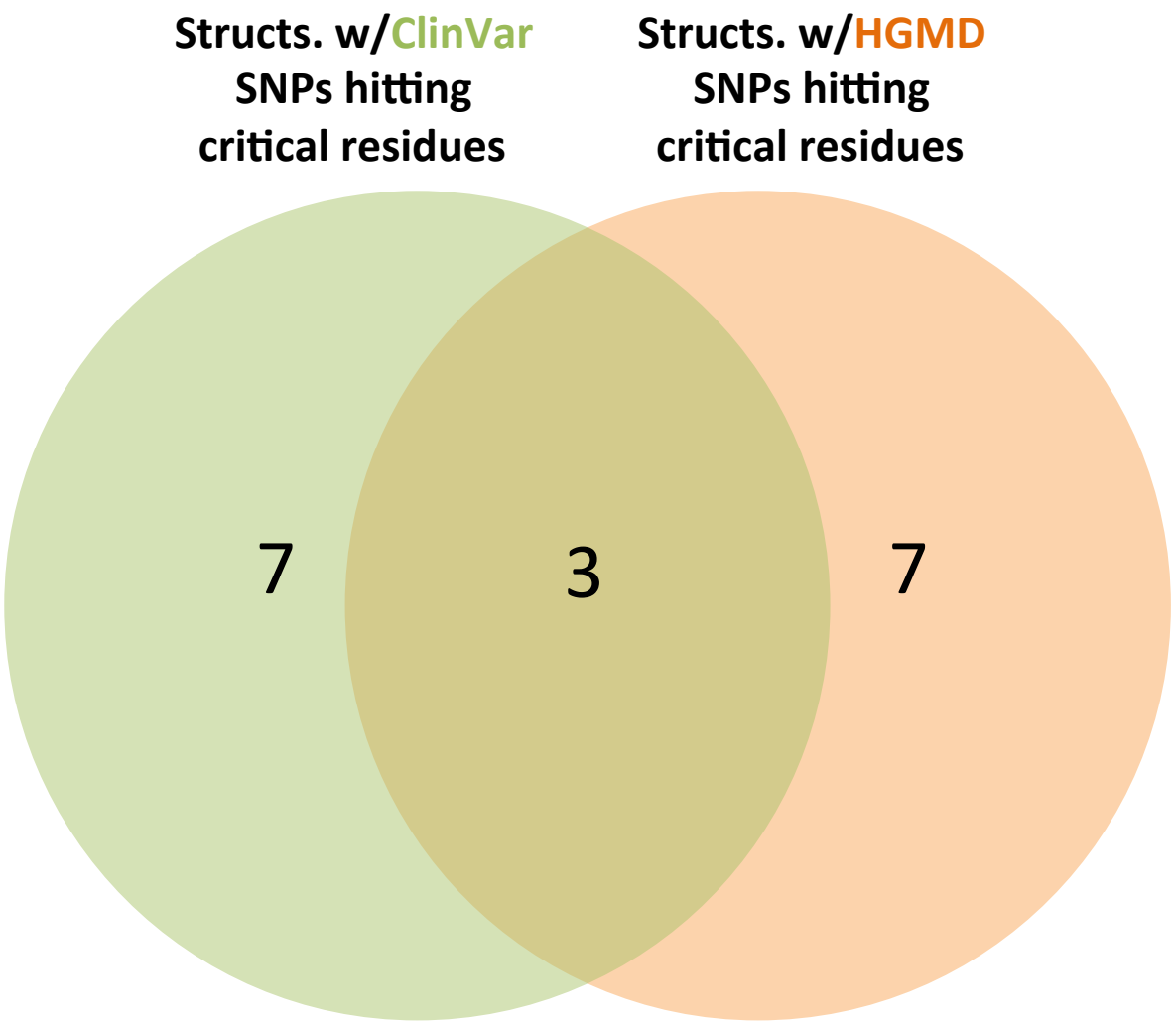
ClinVar vs. HGMD
BL Sites



ClinVar Annotations

0 - unknown
1 - untested
2 - non-pathogenic
3 - probable-non-pathogenic
4 - probable-pathogenic
5 - pathogenic
6 - drug-response
7 - histocompatibility
255 - other

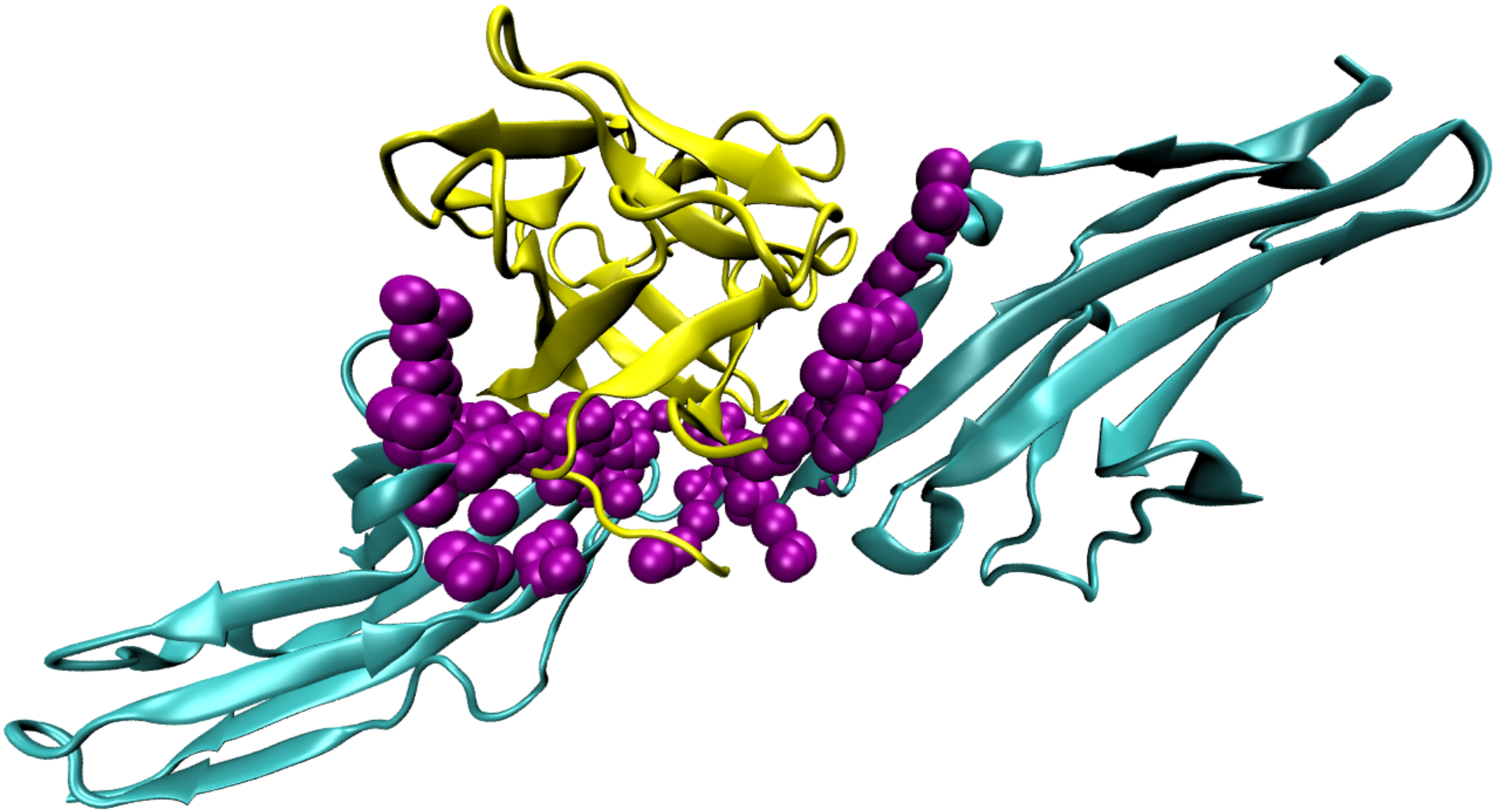
ClinVar vs. HGMD
GN Sites



ClinVar Annotations

0 - unknown
1 - untested
2 - non-pathogenic
3 - probable-non-pathogenic
4 - probable-pathogenic
5 - pathogenic
6 - drug-response
7 - histocompatibility
255 - other

pdbID: 1IIL



Predicting Allosterically-Important Residues at the Surface

"False positives" still catch some of the biological binding site real estate

n	Mean fract. Of ligand-binding sites captured
6	0.56
5	0.59
4	0.65
3	0.69
2	0.79
1	0.84

Identifying alternative conformations across the PDB

Number of distinct clusters

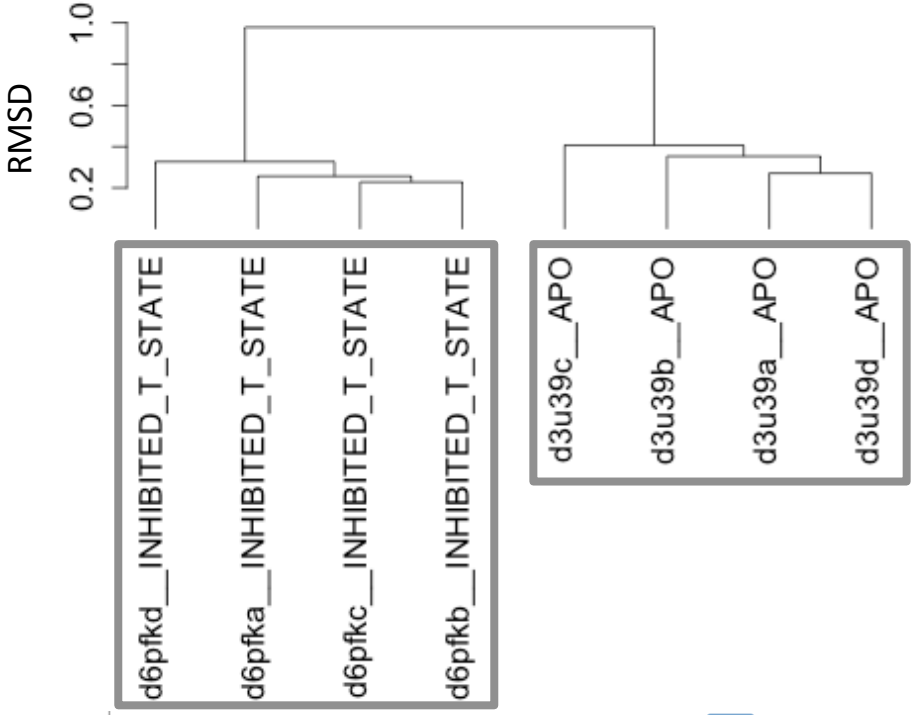
Statistics (from 1000 simulations) on the confidence of the number of clusters and cluster membership assigned

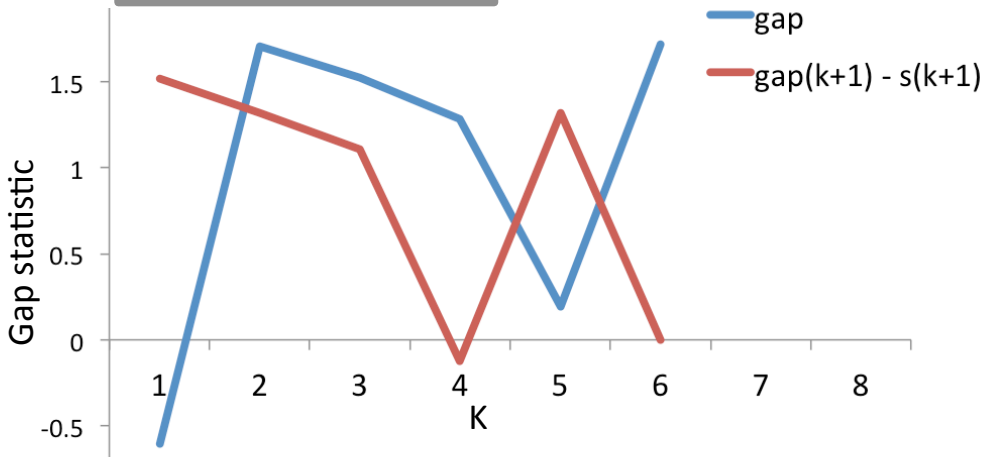
```
K=2   kf=0.969   n=6   pf=1.0   cl_1=1C9K_C::1C9K_A|1C9K_B|1C9K_C
cl_2=1CBU_C::1CBU_A|1CBU_B|1CBU_C   CL_1-CL_2::2.4150
```

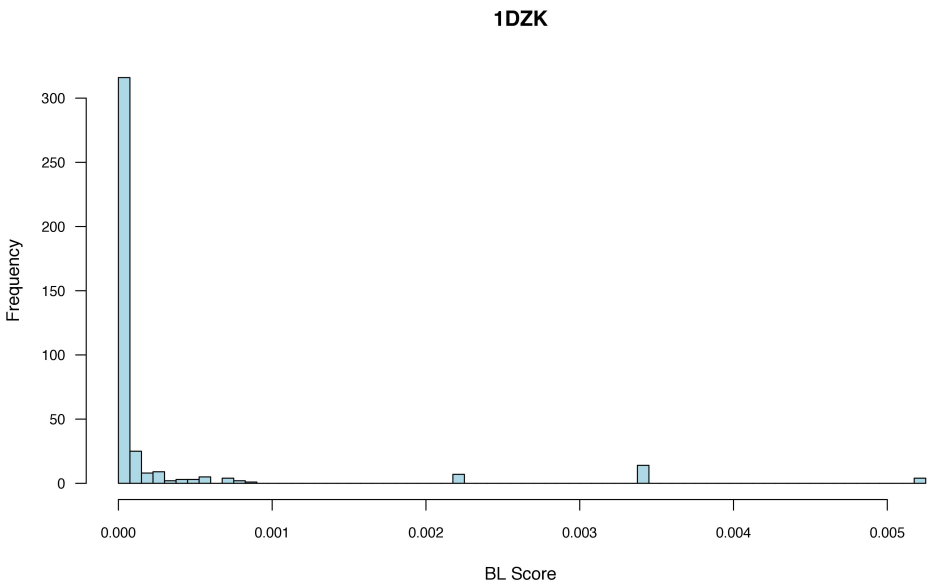
Cluster representatives and membership

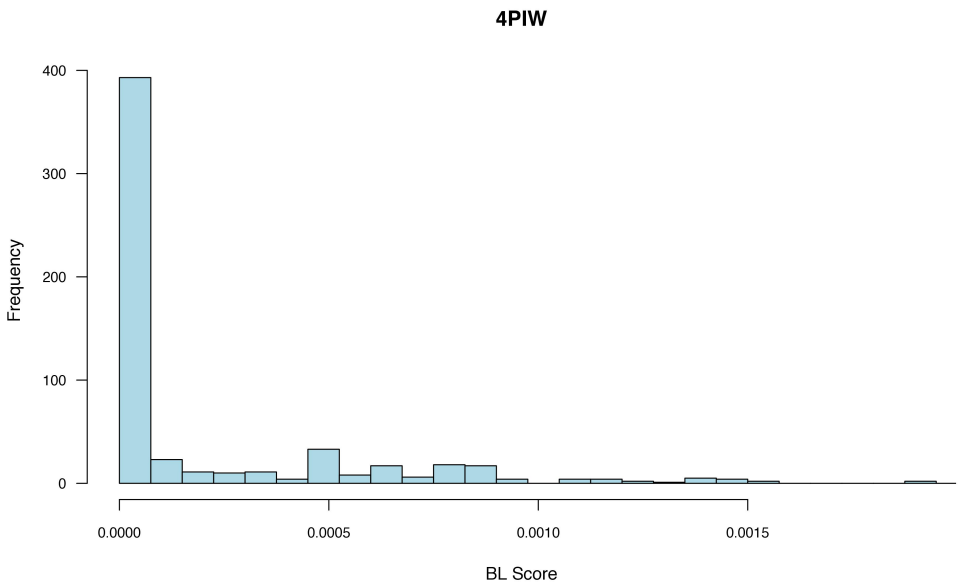
RMSD btwn cluster representatives

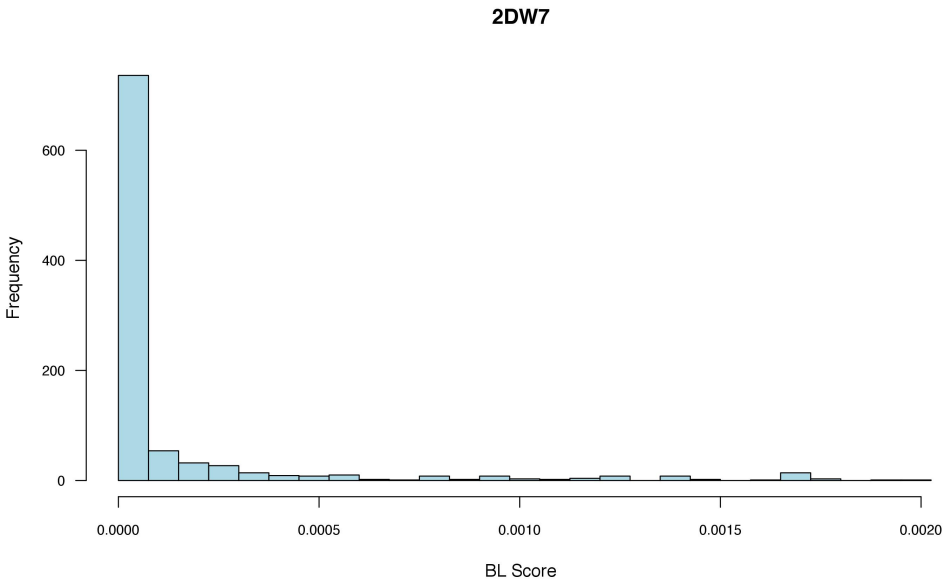
Clustering for Phosphfructokinase

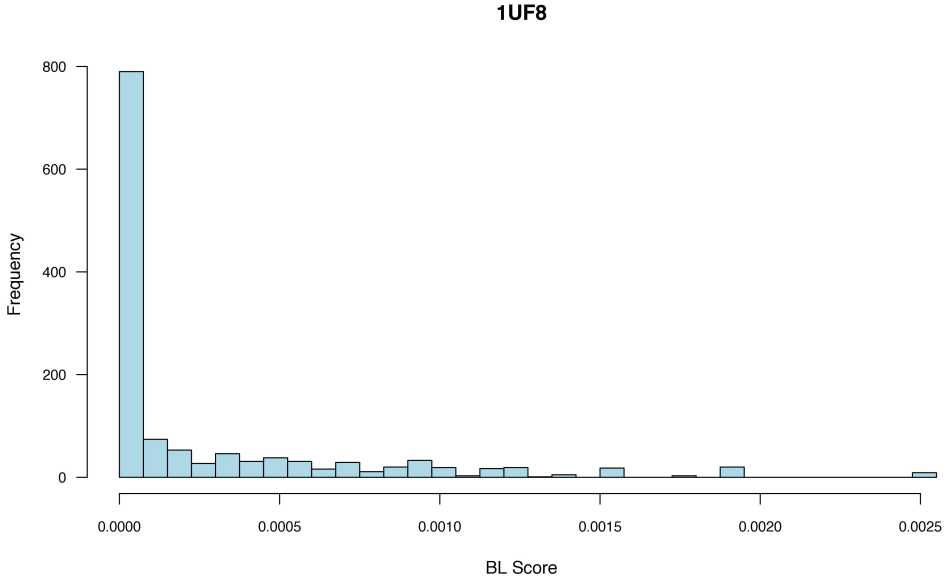




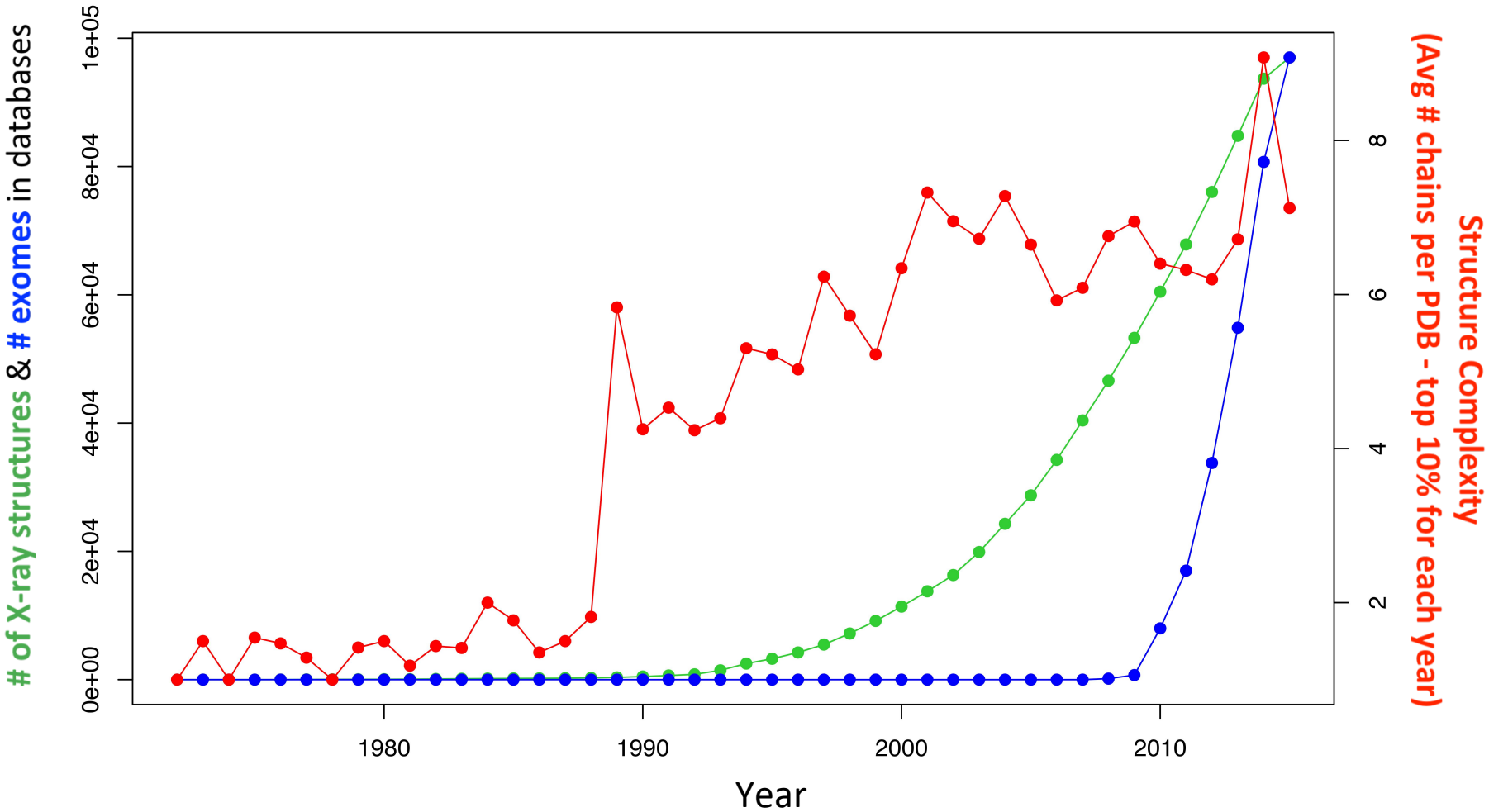

1DZK
Frequency
300
250
200
150
100
50
0
0.000
0.001
0.002
0.003
0.004
0.005
BL Score


4PIW
Frequency
400
300
200
100
0
0.0000
0.0005
0.0010
0.0015
BL Score


2DW7
Frequency
600
400
200
0
0.0000
0.0005
0.0010
0.0015
0.0020
BL Score

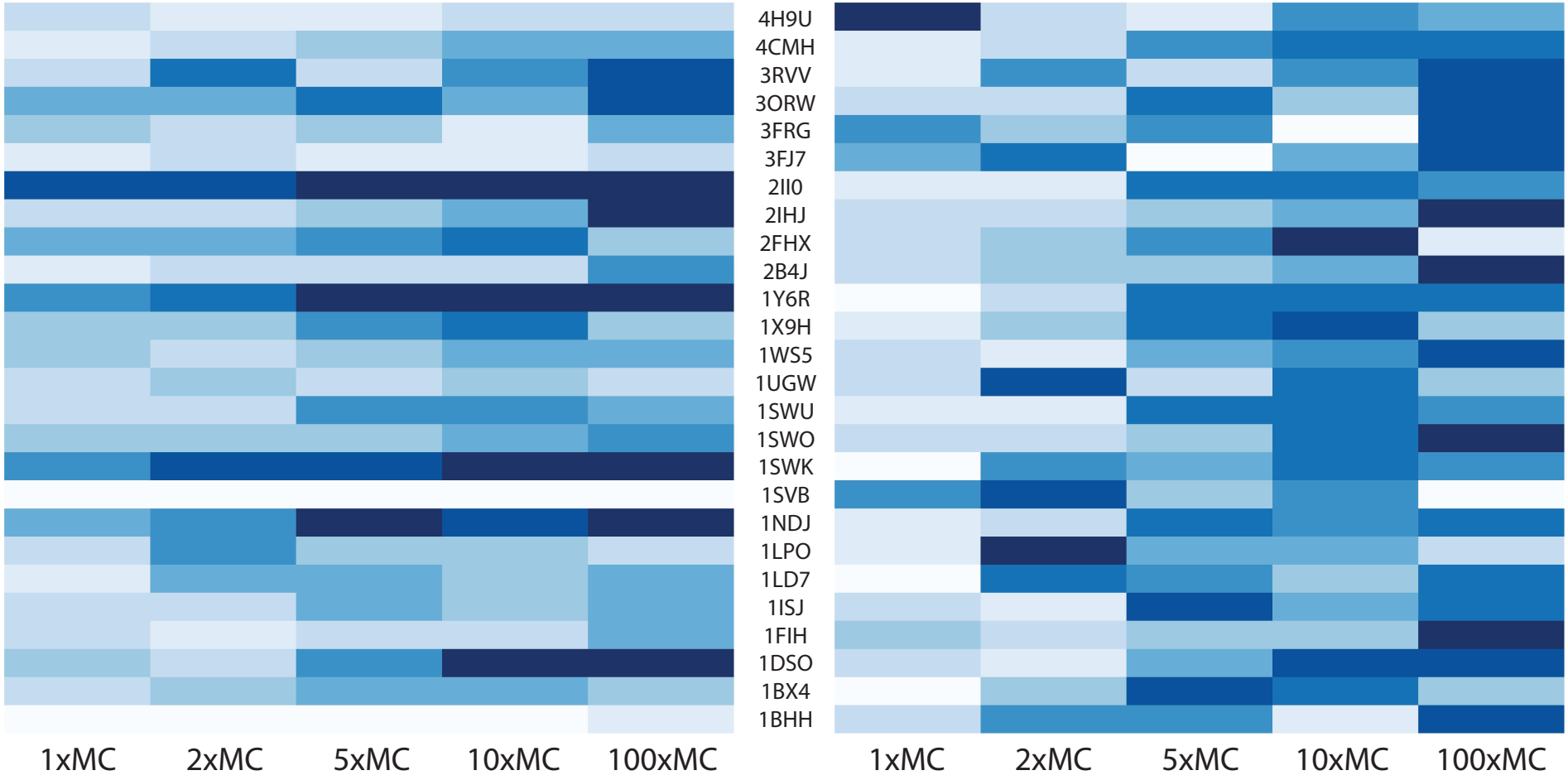

1UF8
Frequency
800
600
400
200
0
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
BL Score

The volume of sequenced exomes is outpacing that of structures, while solved structures have become more complex in nature.



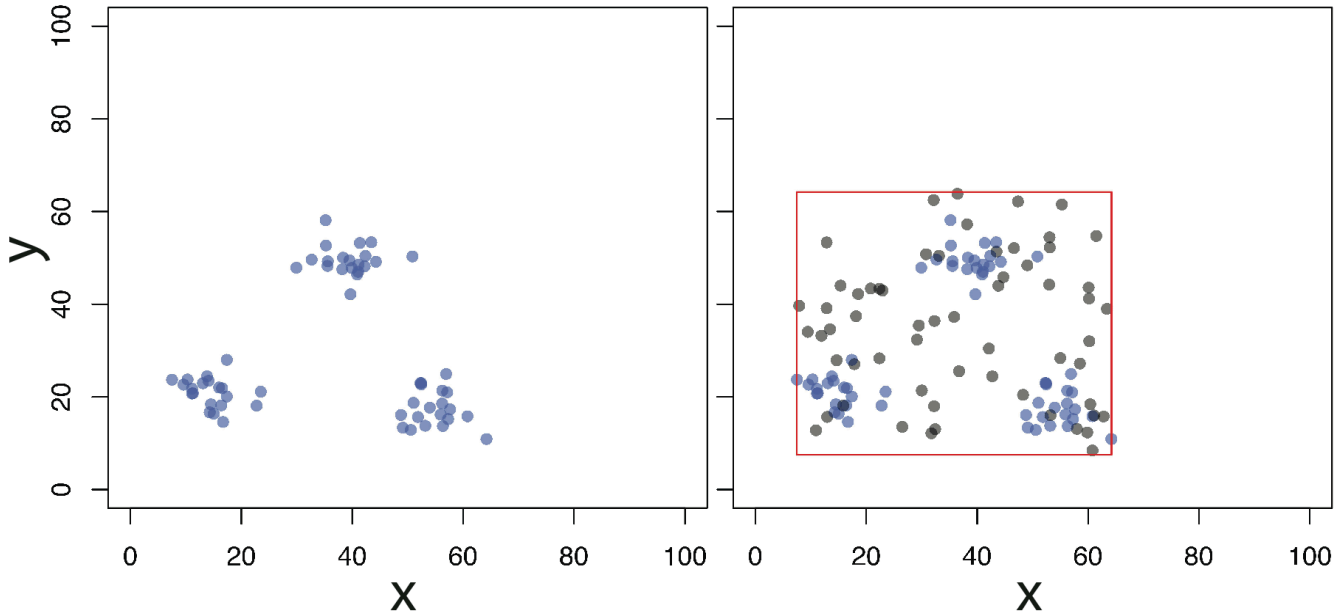
Sources: NCBI SRI, PDB, and SCOP

Measures of convergence using different scaling factors for the number of steps in each MC simulation



Predicting Allosterically-Important Residues within the Interior

Conservation of Critical Residues: GN vs. Infomap			
Protein (PDB, # residues)	**Community Detection Method: GN | InfoMap**		
	# Communities	**# Critical Residues**	**Conservation of CR (p-val)**
tRNA synthetase (1N78, 542)	14 | 25	47 | 109	-0.57 (2.0e-05) | -0.47 (1.3e-09)
Adenylate kinase (4AKE, 428)	11 | 20	39 | 82	-0.70 (3.2e-10) | -0.43 (8.9e-08)
Arginine Kinase (3JU5, 728)	12 | 28	41 | 142	**-0.21 (9.0e-02)** | -0.28 (4.4e-06)
Tyrosine Phosphatase (2HNP, 278)	7 | 15	27 | 70	-0.49 (4.2e-03) | -0.60 (3.1e-09)
Phosphoribosyltransferase (1XTT, 846)	9 | 32	36 | 174	-0.54 (2.1e-07) | -0.43 (5.9e-16)
cAMP-dep. PK (1J3H, 332)	11 | 19	36 | 78	-0.63 (5.1e-07) | -0.43 (4.0e-06)
Anthranilate synthase (1I7Q, 1418)	12 | 46	51 | 288	-0.44 (4.8e-07) | -0.45 (2.2e-16)
Malic enzyme (1EFK, 2212)	17 | 70	74 | 425	**0.22 (8.5e-01)** | -0.19 (5.6e-06)
Threonine synthase (1E5X, 884)	13 | 36	43 | 192	-0.53 (8.5e-07) | -0.32 (2.5e-08)
G-6-P Deaminase (1CD5, 1596)	18 | 54	58 | 266	-0.36 (4.1e-04) | **-0.08 (6.0e-02)**
Phosphofructokinase (3PFK, 1276)	10 | 51	45 | 307	-0.43 (1.7e-06) | -0.16 (4.2e-04)
Tryptophan synthase (1BKS, 1294)	10 | 46	41 | 284	-0.48 (3.0e-09) | -0.40 (2.0e-15)
Means	***12.0 | 36.8***	***44.8 | 201.4***	



Dk: Measure to describe how compact cluster k is

$$D_k = \sum_{\mathrm{x}_i \in C_k} \sum_{\mathrm{x}_j \in C_k} ||\mathrm{x}_i - \mathrm{x}_j||^2$$

Wk: Normalized sum of these measures for a given 'partition'

$$W_k = \sum_{k=1}^{K} \frac{1}{2n_k} D_k$$

How much does this score differ from that in a randomized null?

$$\mathrm{Gap}_n(k) = E_n^*\{\log W_k\} - \log W_k$$

Best K: value for which the gap value is higher than the next K

$$\mathrm{Gap}(k) \geq \mathrm{Gap}(k+1) - s_{k+1}$$

Tibshirani et al., 2001