Protein Evolution
Lecture 3

Lucy Colwell, ICTS Winter School on Quantitative Systems Biology, Dec 14 2017

Outline

Biology 101 – a quick tour of proteins

Biological information from covariance

- Protein structure and function
- Protein interaction specificity
- Ligand protein binding

Theoretical analysis of data limitations

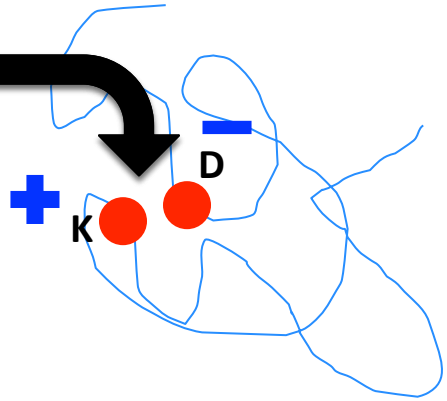
– when can we predict reliably?
– Noise due to finite sampling, samples are not independent.
– central role played by Random Matrix Theory.

Exploit correlation structure of protein sequences

What constraints on the mutations that can be accepted are induced by the requirements of 3D structure and function of the protein?

	1	2	3	4	5	6
H.s	D	A	I	L	V	K
M.m	D	A	I	L	A	K
B.t	R	A	I	M	V	D
G.g	R	V	I	L	V	D

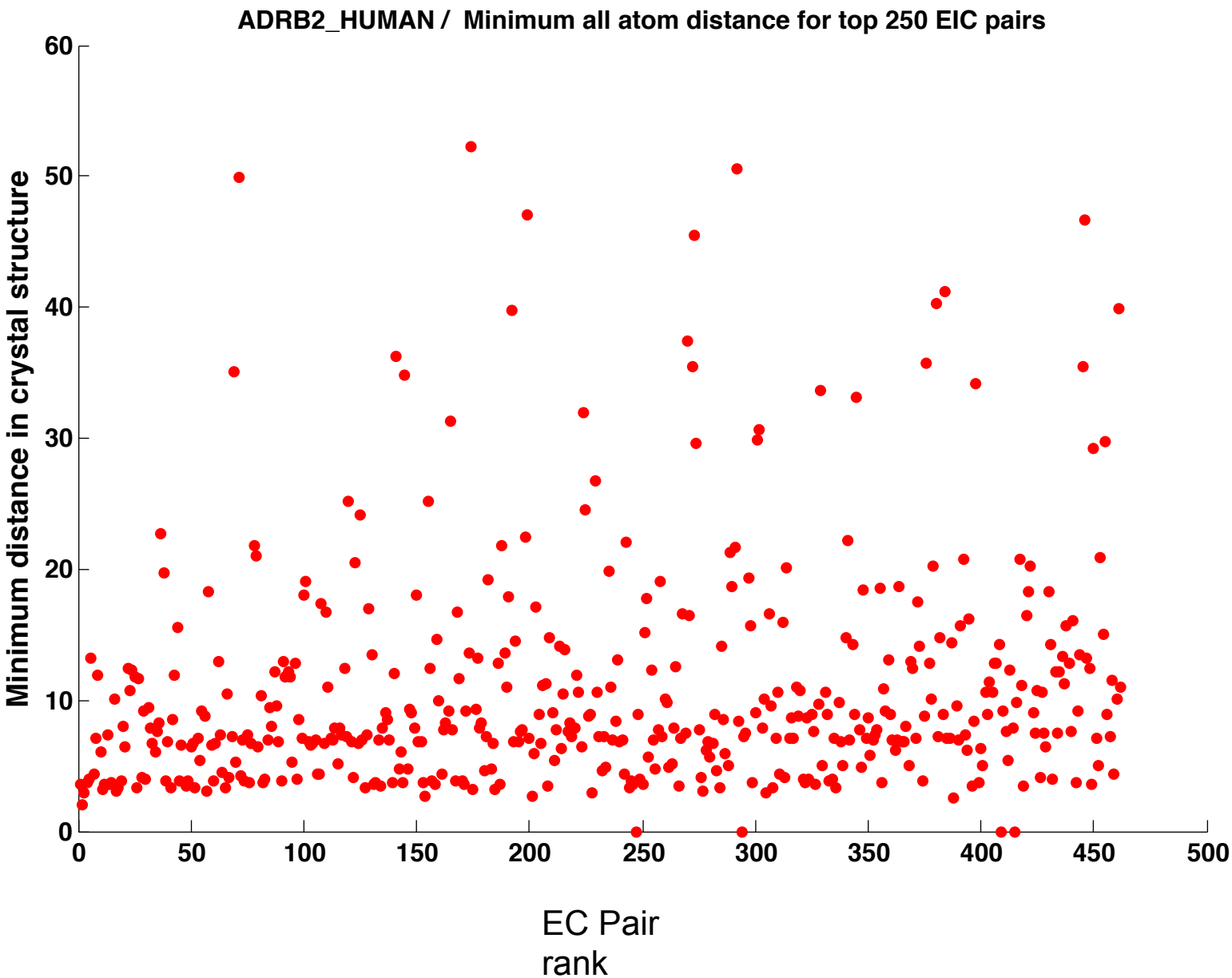
correlated



Wild Type Protein
(active)

Can we **exploit correlations** in the sequence data to build a model that can predict **the 3D protein structure?**

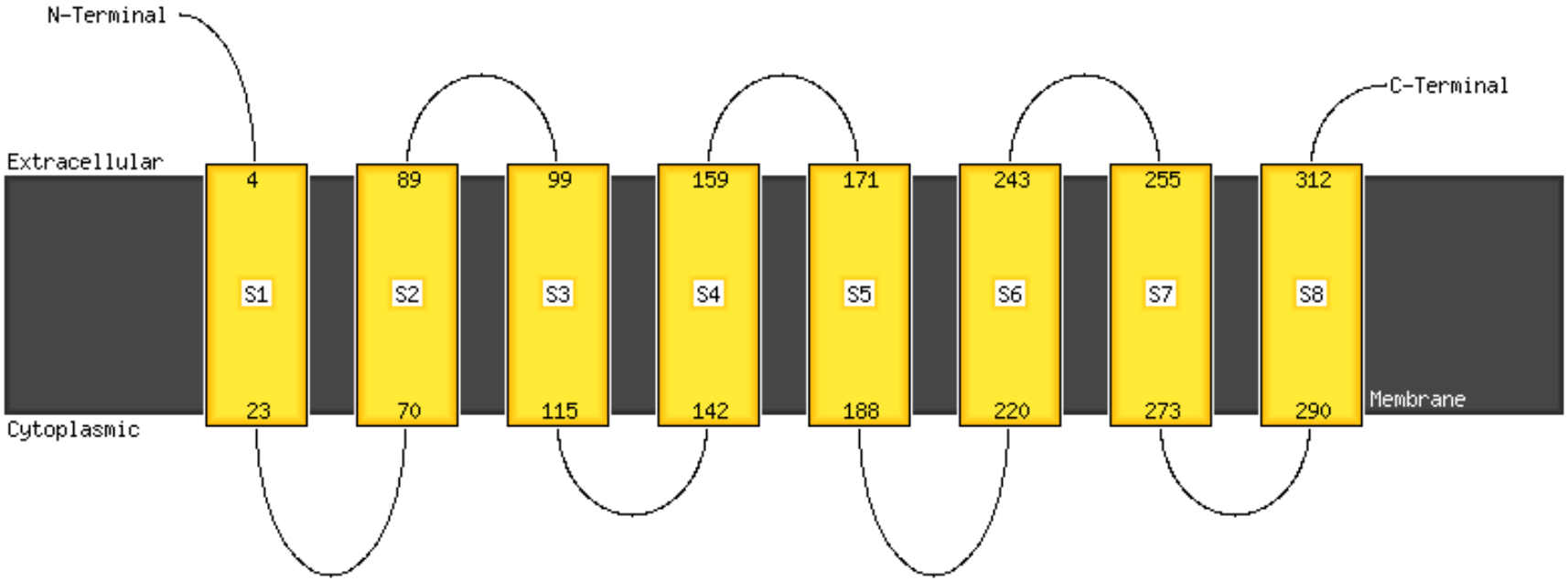
Evolutionary constraints for adrenergic beta receptor 2

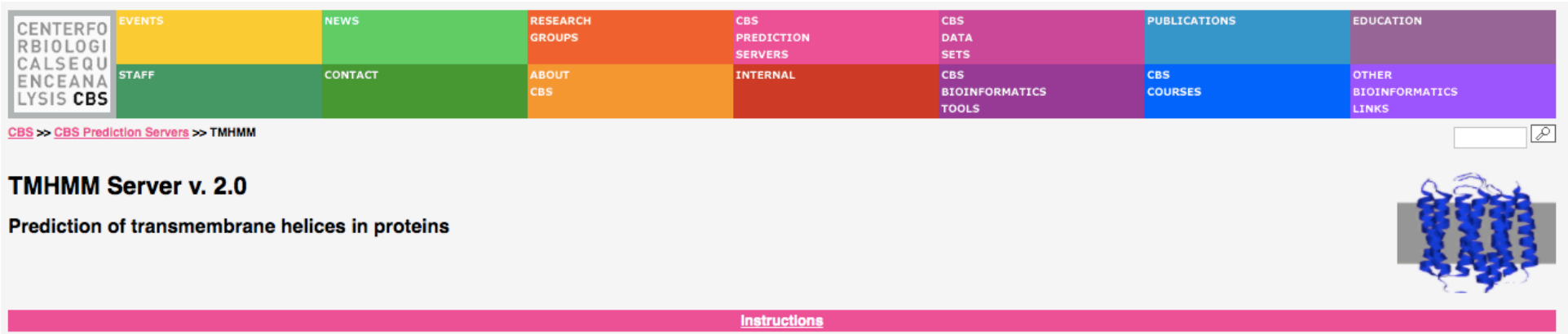


Transmembrane protein predicted topology

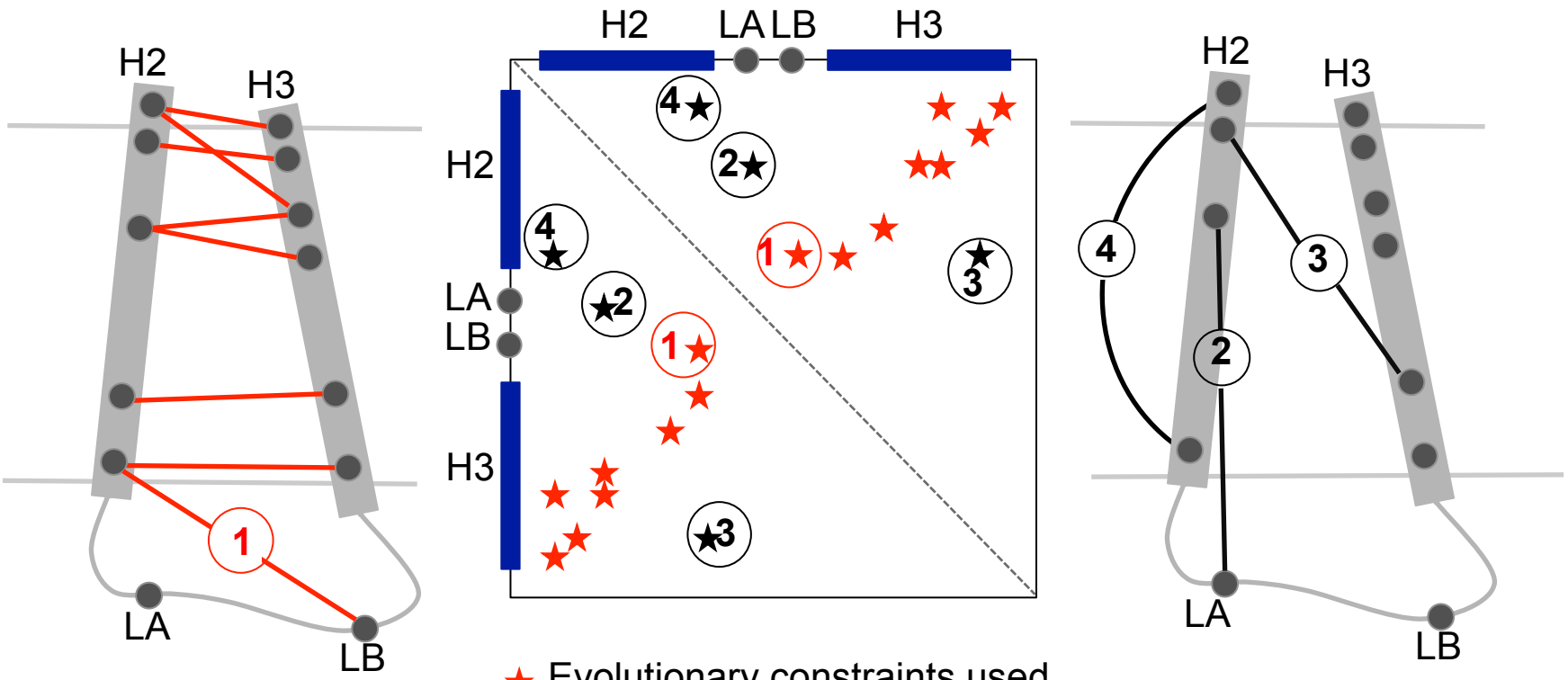
15e71e73-ae04-493e-a3ca-254c7341265c.mem

Membrane helical secondary structure





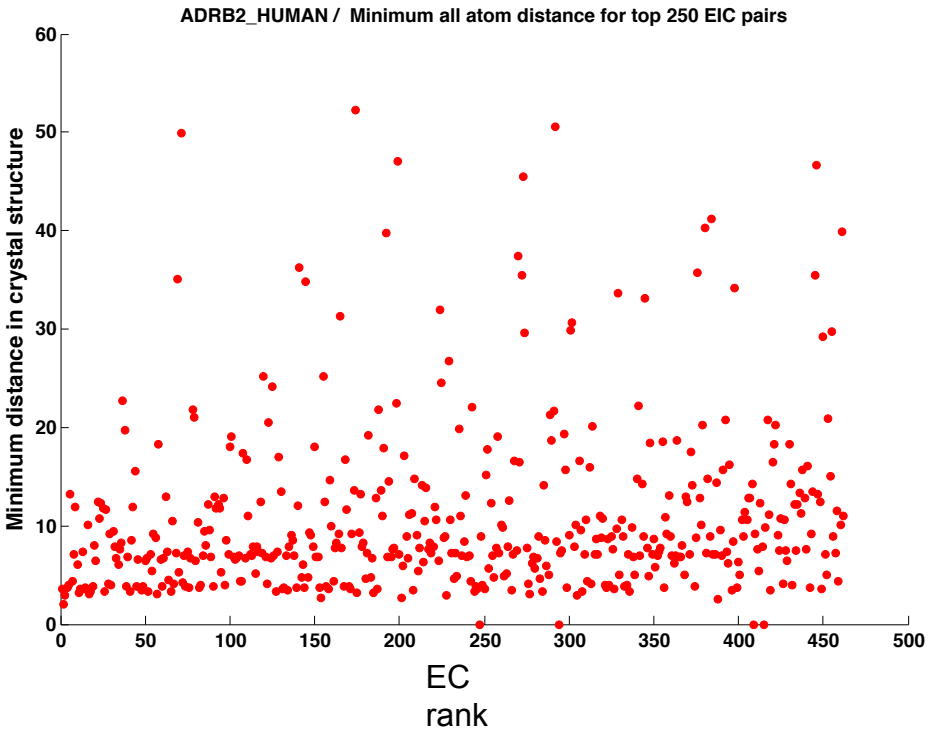
Transmembrane protein predicted topology

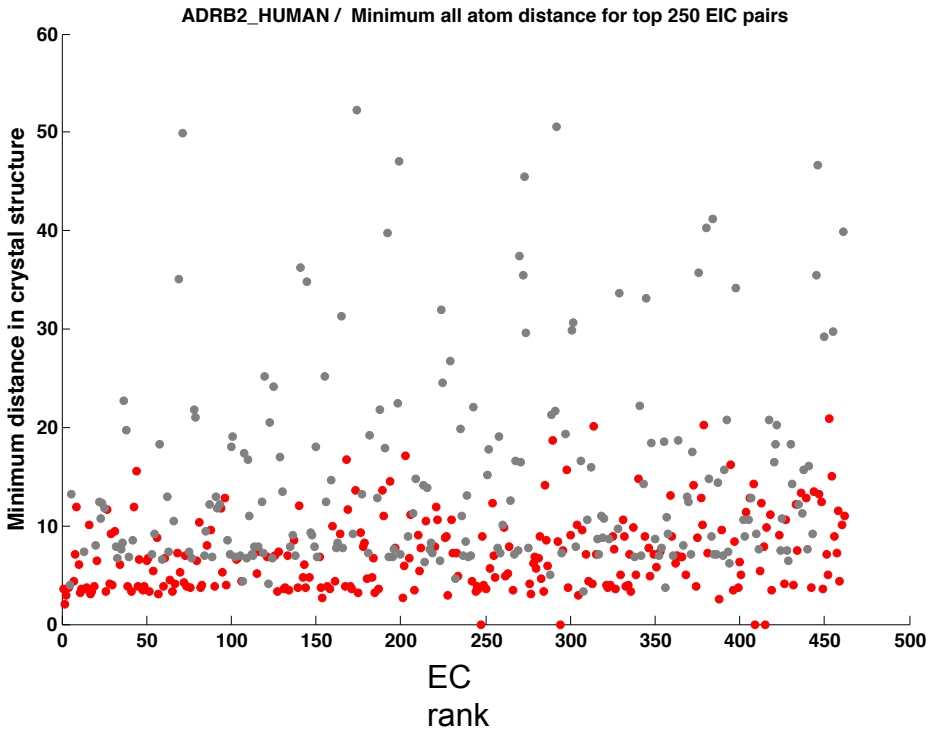


★ Evolutionary constraints used as distance constraints

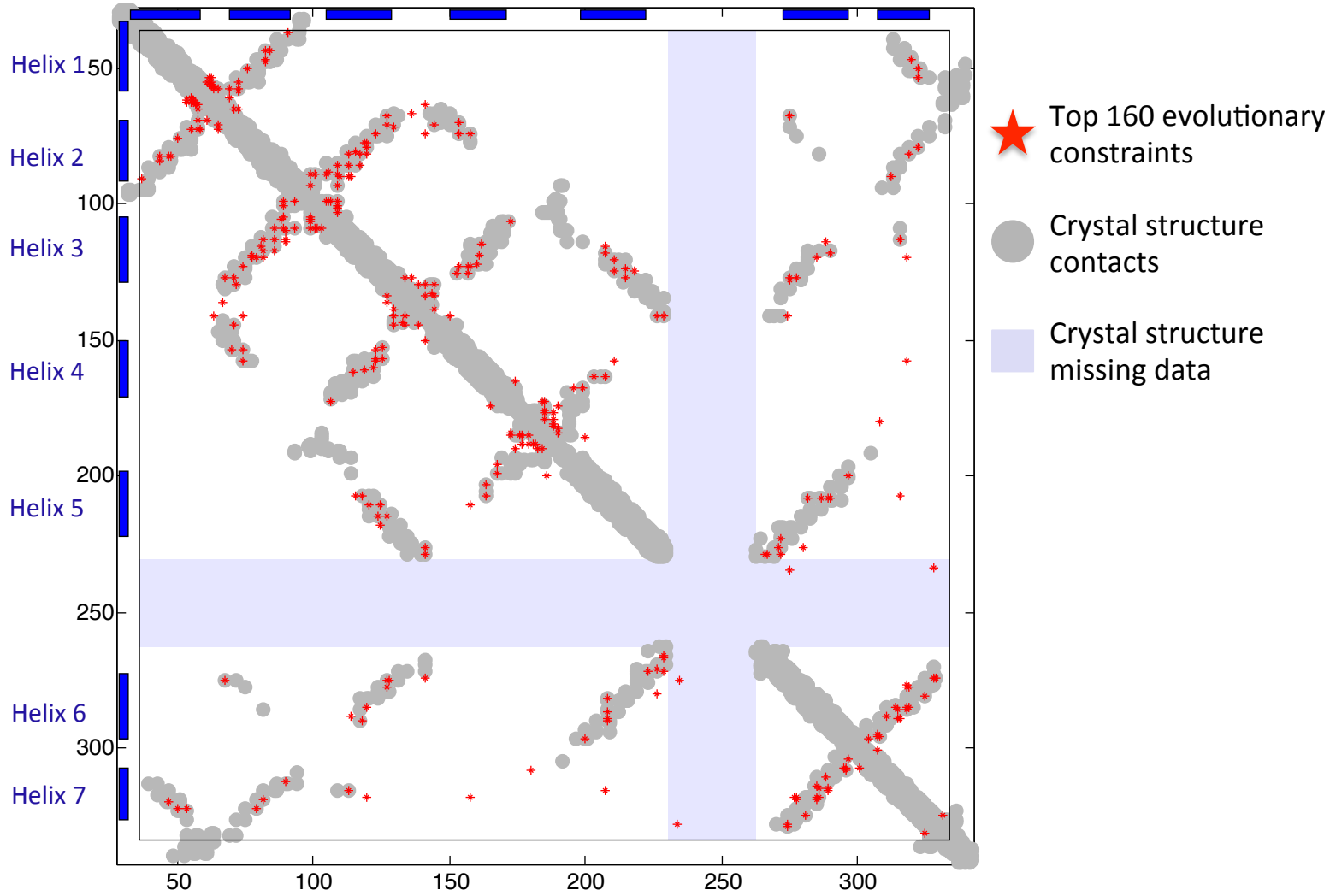
★ Constraints conflicting with predicted transmembrane topology

Evolutionary constraints for adrenergic beta receptor 2





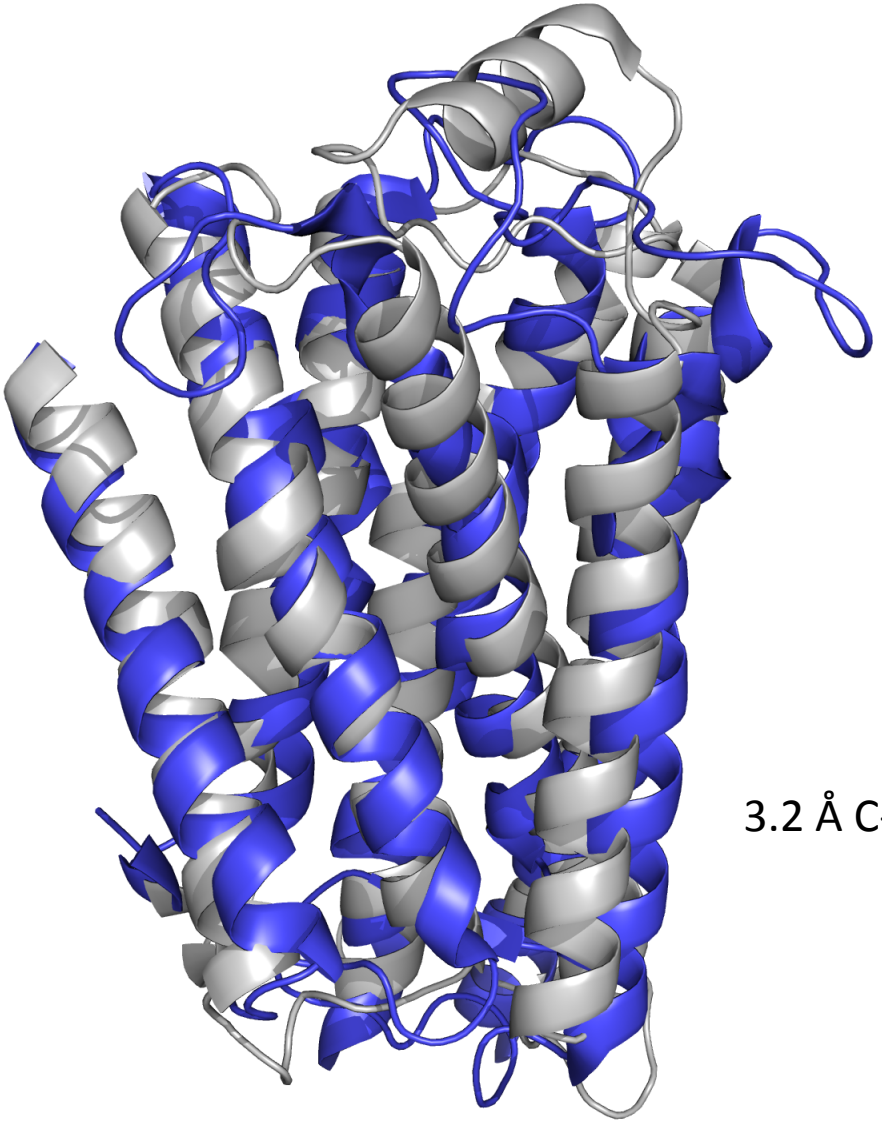
β2 adrenergic receptor : evolutionary constraints



The β2 adrenergic receptor, a seven transmembrane helix GPCR family member.

Hopf, Colwell *et al*, Cell 2012

Blind prediction for β2 adrenergic receptor



Crystal structure*

Predicted structure

3.2 Å C-α rmsd over 200 residues

*Crystal Structure: Rasmussen SG, DeVree BT, Zou Y, Kruse AC, Chung KY, Kobilka TS, Thian FS, Chae PS, Pardon E, Calinski D, Mathiesen JM, Shah ST, Lyons JA, Caffrey M, Gellman SH, Steyaert J, Skiniotis G, Weis WI, Sunahara RK, Kobilka BK. Nature. 2011 Jul 19.

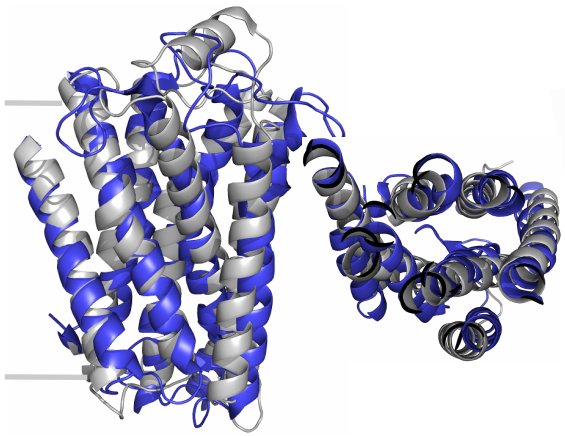
Hopf, Colwell *et al*, Cell 2012

Sequences to structure:

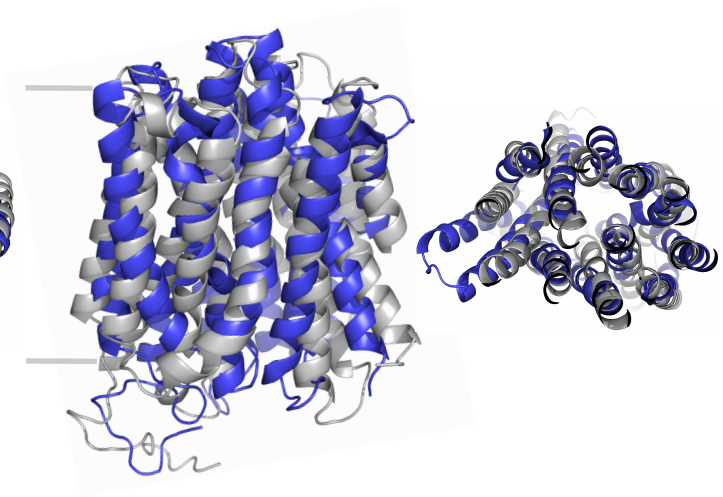
Blind structure prediction.

observed predicted

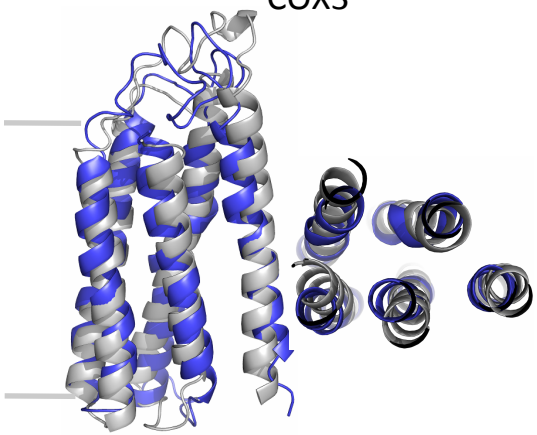
β2 adrenergic receptor



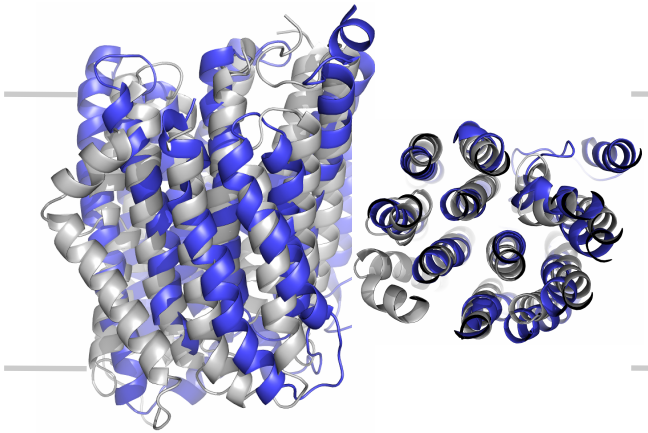
G-3-P transporter GlpT



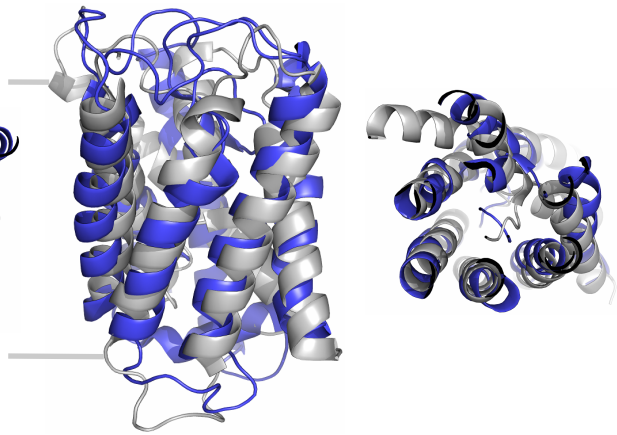
COX3



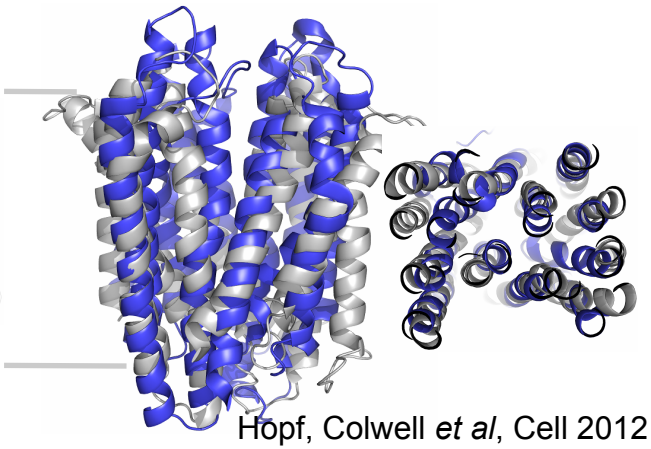
NADH-quinone oxidoreductase subunit N 1



Aquaporin-0



Multidrug resistance protein norM



Hopf, Colwell *et al*, Cell 2012

Predict *de novo* structures for transmembrane proteins for which no homologous structure has been solved experimentally

Adiponectin receptor 1

Adiponectin receptor 1

- An essential hormone secreted by adipocytes that acts as an antidiabetic.
- 375 amino acids
- 7 TM helices

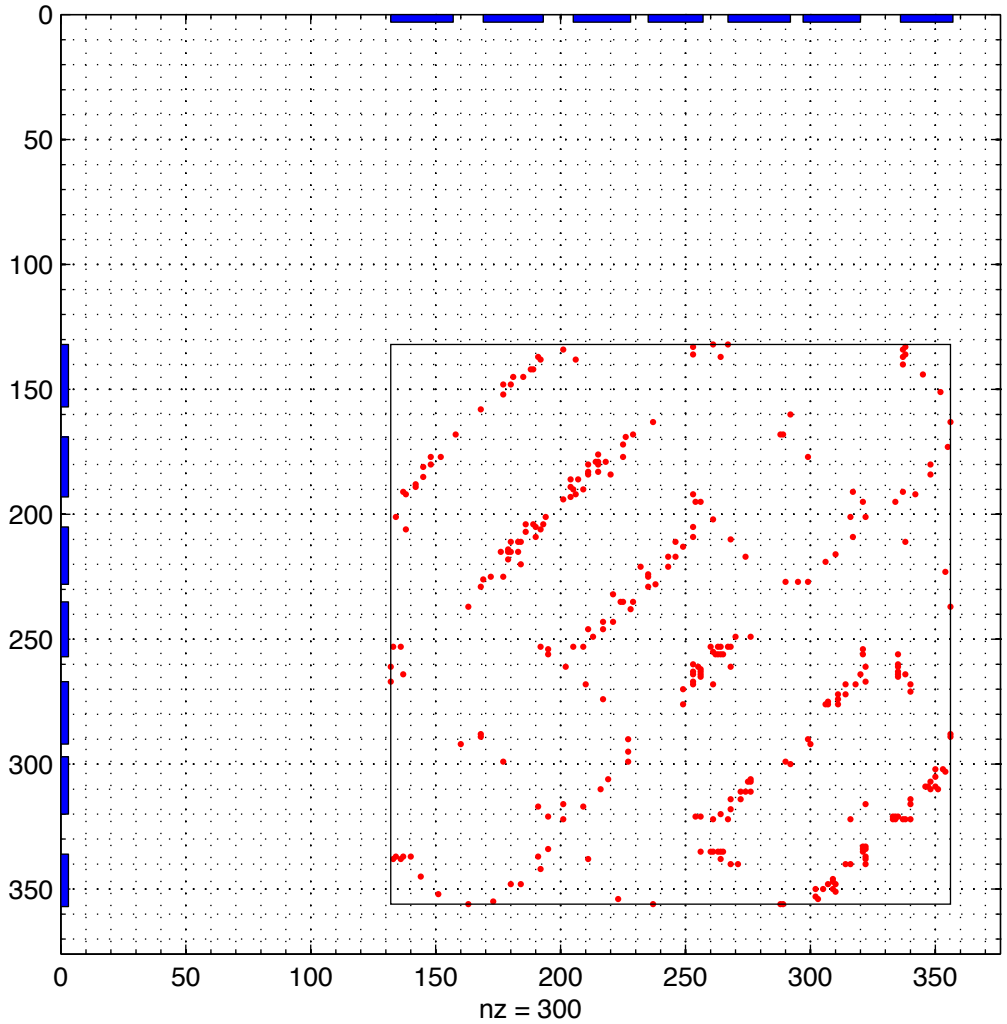
Cloning of adiponectin receptors that mediate antidiabetic metabolic effects

Toshimasa Yamauchi*†‡, Junji Kamon*‡§, Yusuke Ito*, Atsushi Tsuchida*, Takehiko Yokomizo||, Shunbun Kita*, Takuya Sugiyama¶, Makoto Miyagishi#, Kazuo Hara*†, Masaki Tsunoda☆, Koji Murakami☆, Toshiaki Ohteki**†, Shoko Uchida*, Sato Takekawa*, Hironori Waki*†, Nelson H. Tsuno††, Yoichi Shibata††, Yasuo Terauchi*†, Philippe Froguel‡‡, Kazuyuki Tobe*†, Shigeo Koyasu**†, Kazunari Taira#, Toshio Kitamura¶, Takao Shimizu||, Ryozo Nagai* & Takashi Kadowaki*†

Levels of adiponectin in the blood are decreased under conditions of obesity, insulin resistance and type 2 diabetes

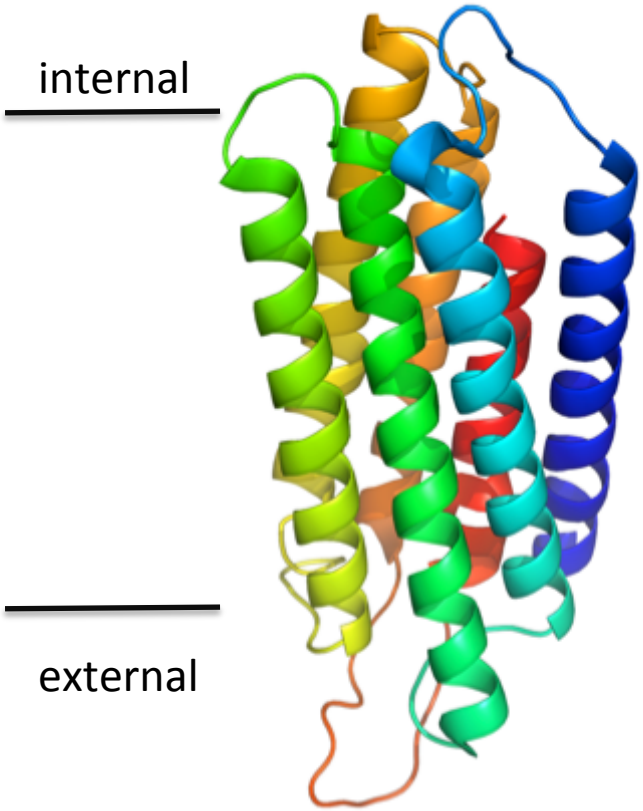
De novo predicted 3D structure of adiponectin receptor 1

predicted contact map



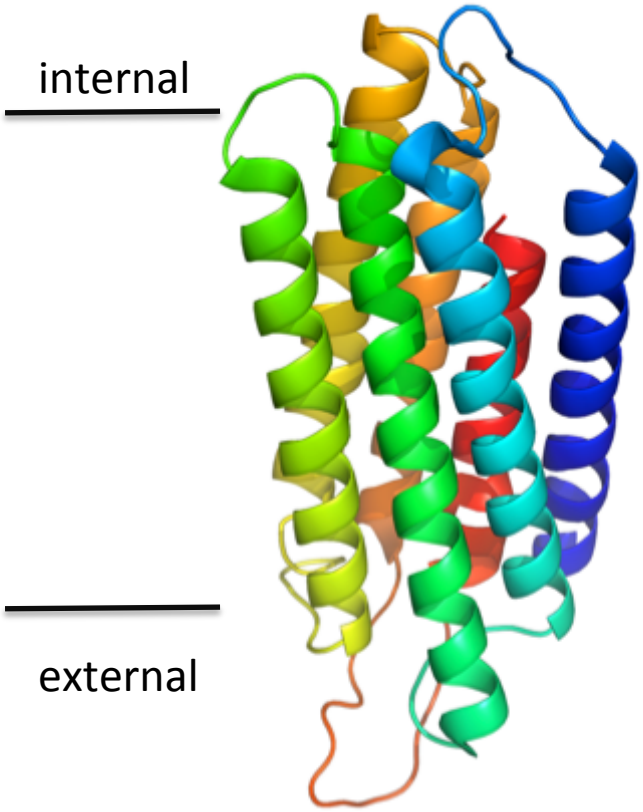
De novo predicted 3D structure of adiponectin receptor 1

3. Predicted structure

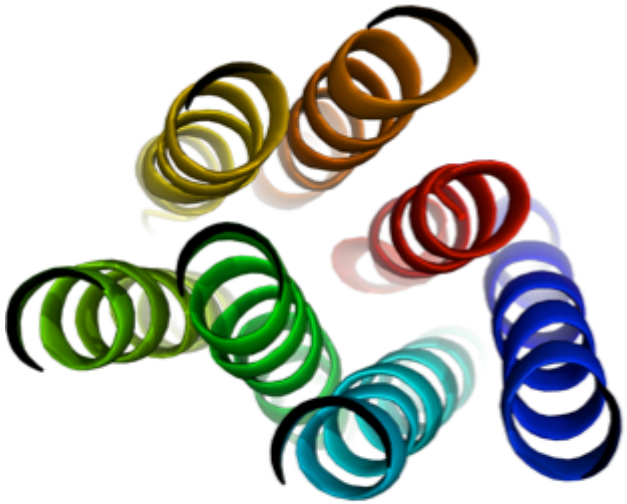


Top ranked predicted 3D structure of Adiponectin receptor 1 from sequences alone

De novo predicted 3D structure of adiponectin receptor 1



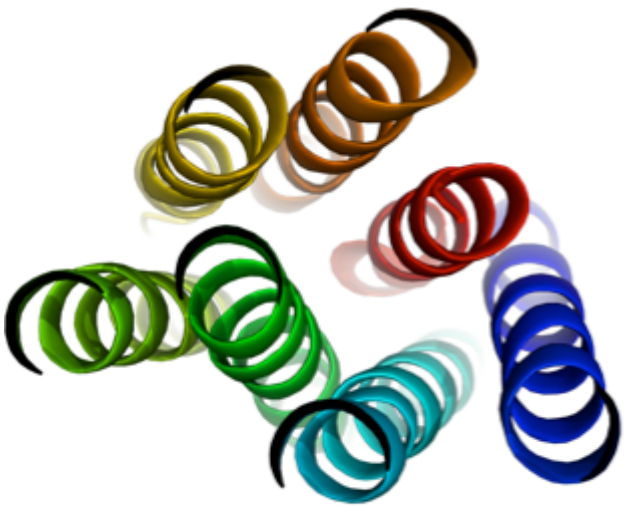
Viewed into membrane from external



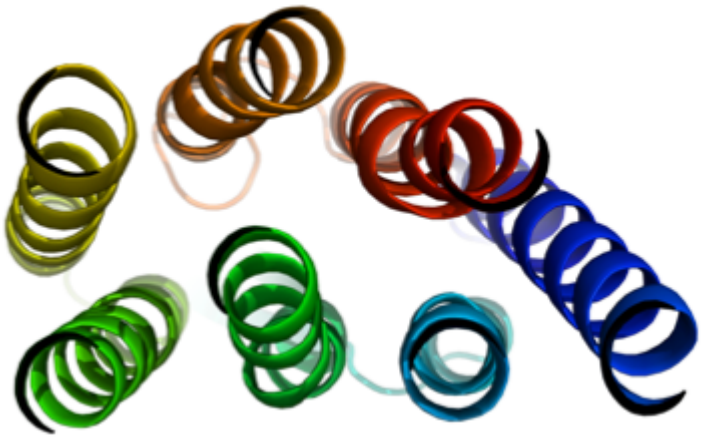
Top ranked predicted 3D structure of Adiponectin receptor 1 from sequences alone

De novo predicted 3D structure of adiponectin receptor 1

ADR1 prediction



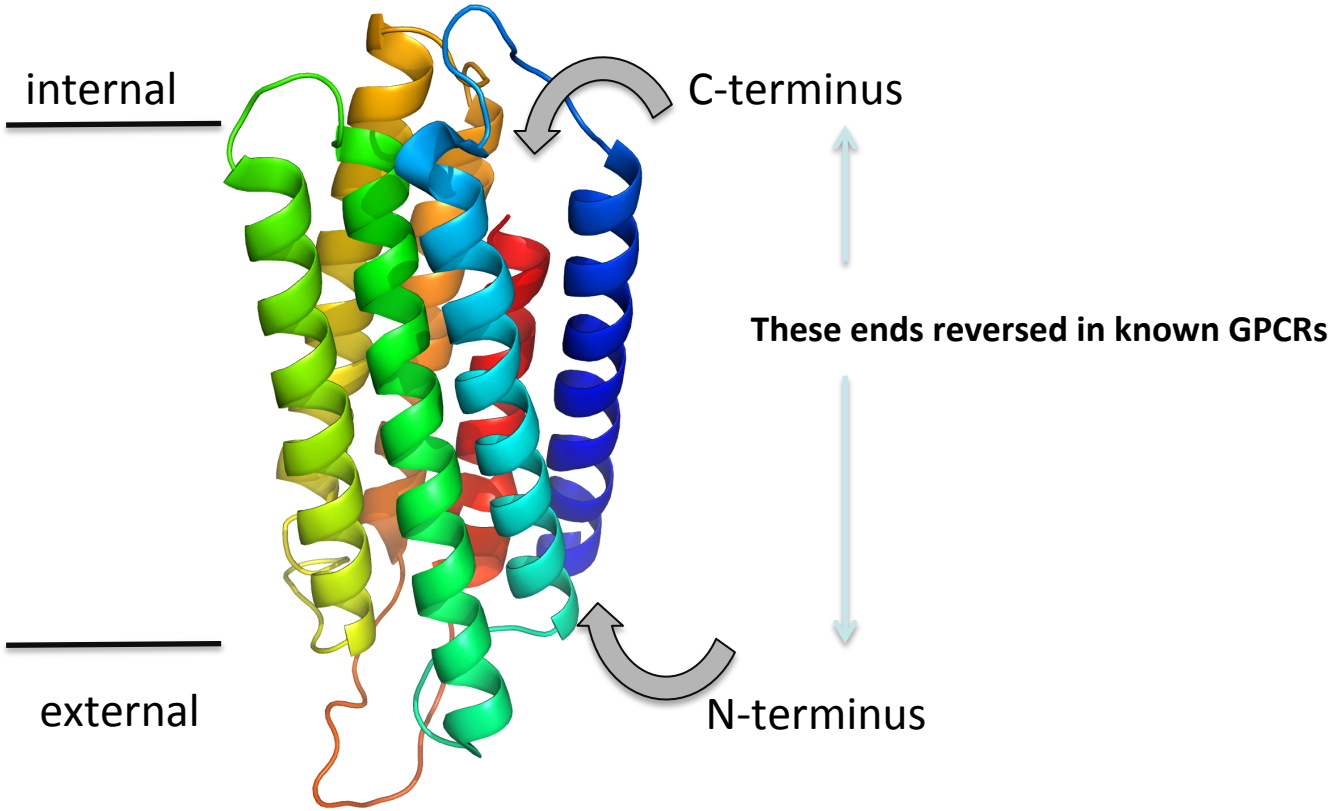
Bacteriorhodopsin



Is the predicted 3D structure of Adiponectin receptor 1 similar to any known TM protein structures ?

Yes – 4.5 A rmsd over >200 residues

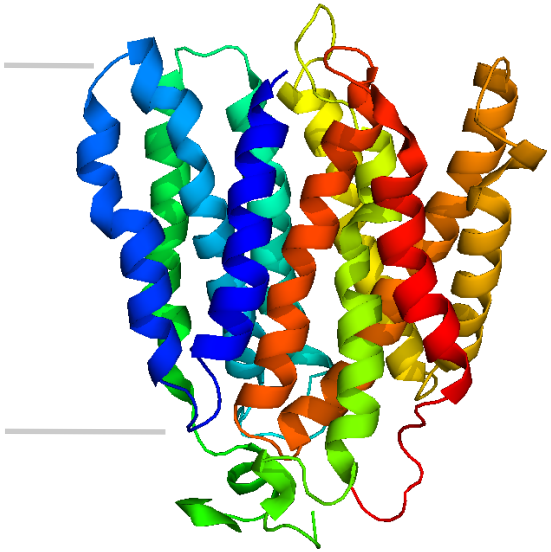
De novo predicted 3D structure of adiponectin receptor 1



Predicted Adiponectin receptor 1 is flipped in the membrane compared to GPCRs

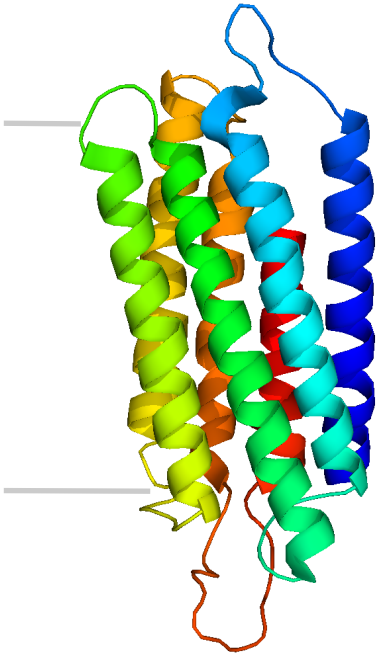
11 medically important membrane proteins of unknown structure predicted

OCTN1



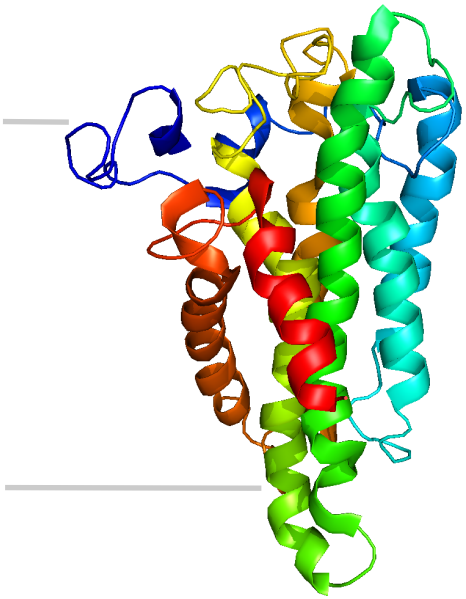
Crohn's disease,
rheumatoid arthritis

Adiponectin receptor 1



diabetes, obesity,
cancer

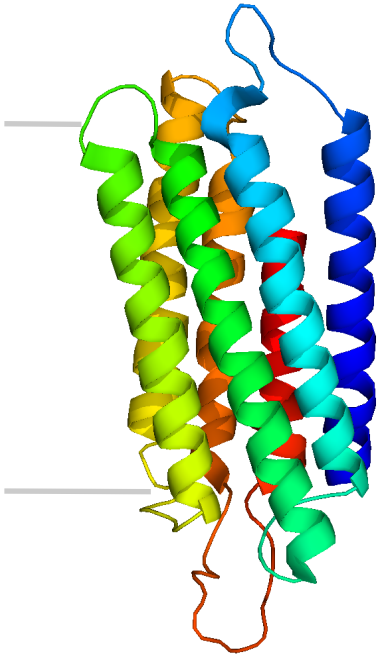
MT-ND1



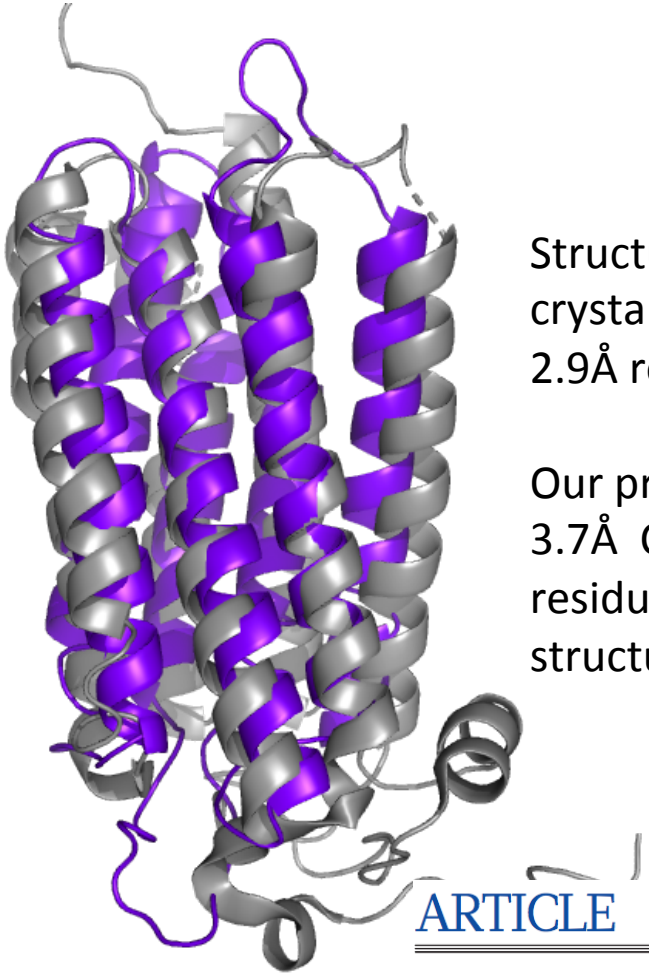
LHON, MELAS,
Alzheimer, Parkinson

11 medically important membrane proteins of unknown structure predicted

Adiponectin receptor 1



diabetes, obesity, cancer



Structure solved by X-ray crystallography in 2015 at 2.9Å resolution.

Our predicted structure is 3.7Å C-α rmsd over 186 residues from this crystal structure 3WXV.

ARTICLE

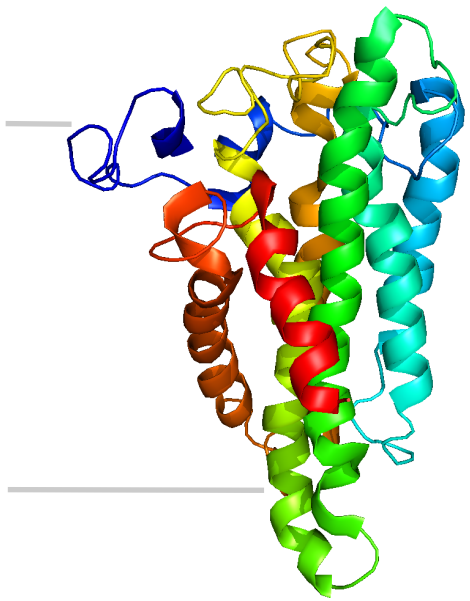
doi:10.1038/nature14301

Crystal structures of the human adiponectin receptors

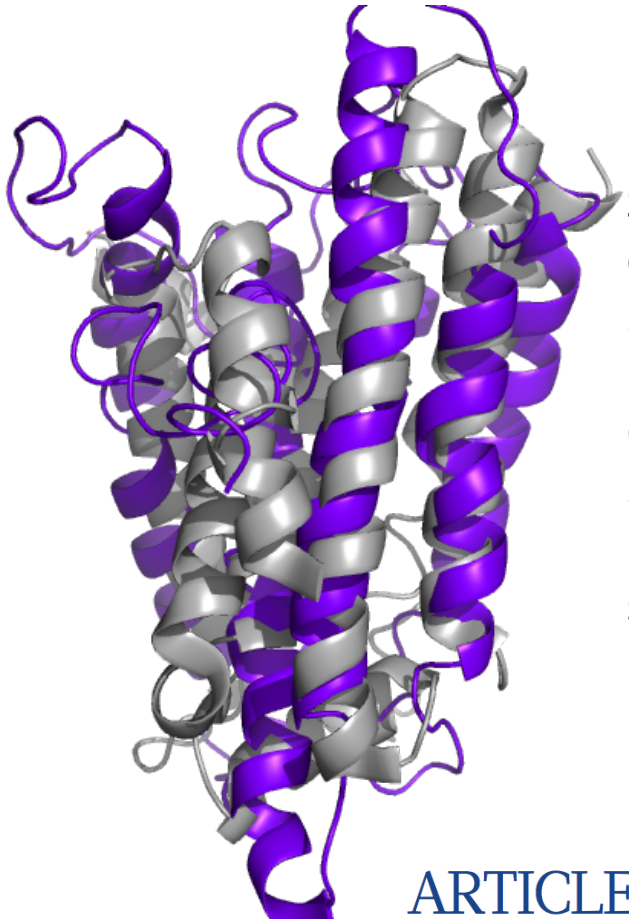
Hiroaki Tanabe[1,2,3,4], Yoshifumi Fujii[1,4], Miki Okada-Iwabu[5,6]*, Masato Iwabu[5,6,7]*, Yoshihiro Nakamura[1,3,4]*, Toshiaki Hosaka[1,3], Kanna Motoyama[1], Mariko Ikeda[1,3], Motoaki Wakiyama[1,3], Takaho Terada[1,4], Noboru Ohsawa[1,3], Masakatsu Hato[1,3], Satoshi Ogasawara[8], Tomoya Hino[8,9], Takeshi Murata[1,8,9,10], So Iwata[1,8,9,11,12,13], Kunio Hirata[13], Yoshiaki Kawano[13], Masaki Yamamoto[13], Tomomi Kimura-Someya[1,3], Mikako Shirouzu[1,3], Toshimasa Yamauchi[5,6,14], Takashi Kadowaki[5,6] & Shigeyuki Yokoyama[1,2,4]

11 medically important membrane proteins of unknown structure predicted

MT-ND1



LHON, MELAS,
Alzheimer, Parkinson



Structure solved by X-ray crystallography in 2013 at 3.3Å resolution.

Our predicted structure is 3.67Å C-α rmsd over 253 residues from the crystal structure 4HE8.

ARTICLE

doi:10.1038/nature11871

Crystal structure of the entire respiratory complex I

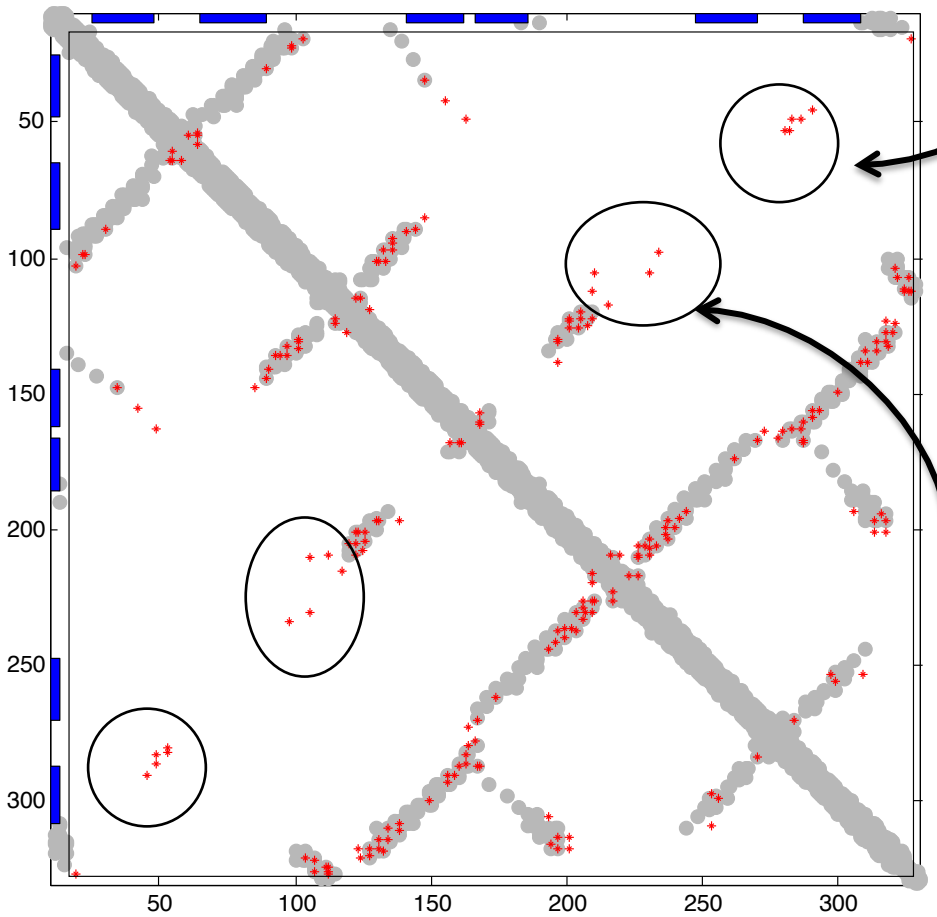
Rozbeh Baradaran[1], John M. Berrisford[1]†, Gurdeep S. Minhas[1] & Leonid A. Sazanov[1]

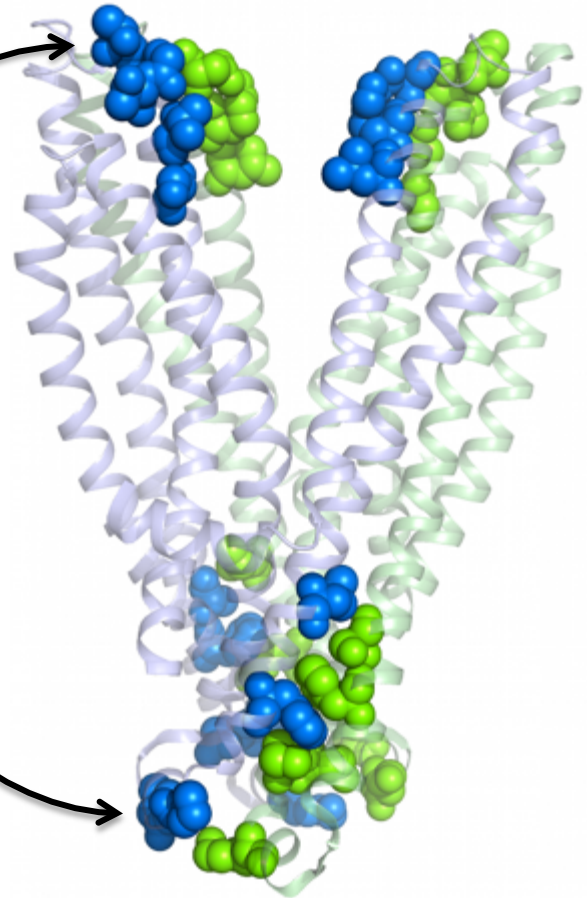
Beyond structure prediction

Constraints reflect multimerisation

Beyond structure prediction

Constraints reflect multimerisation





S. typhimurium ABC transporter
(antibiotic transporter)

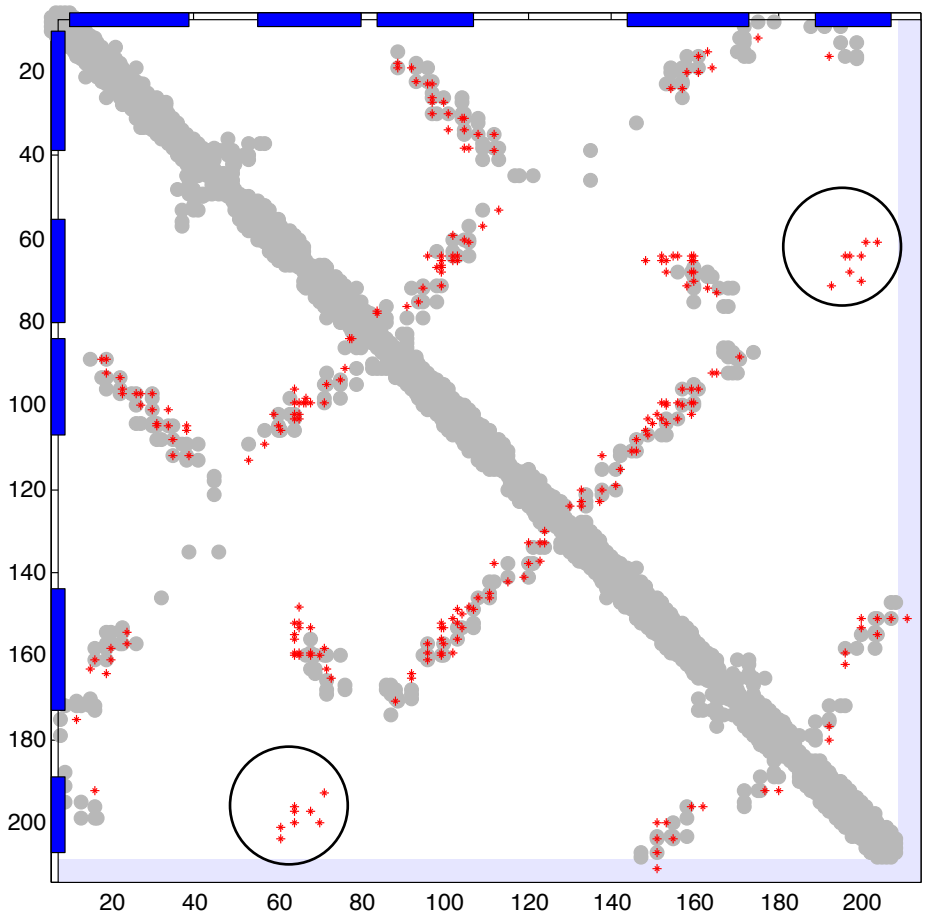
★ Top ranked evolutionary constraints (ECs)

○ Multimer ECs

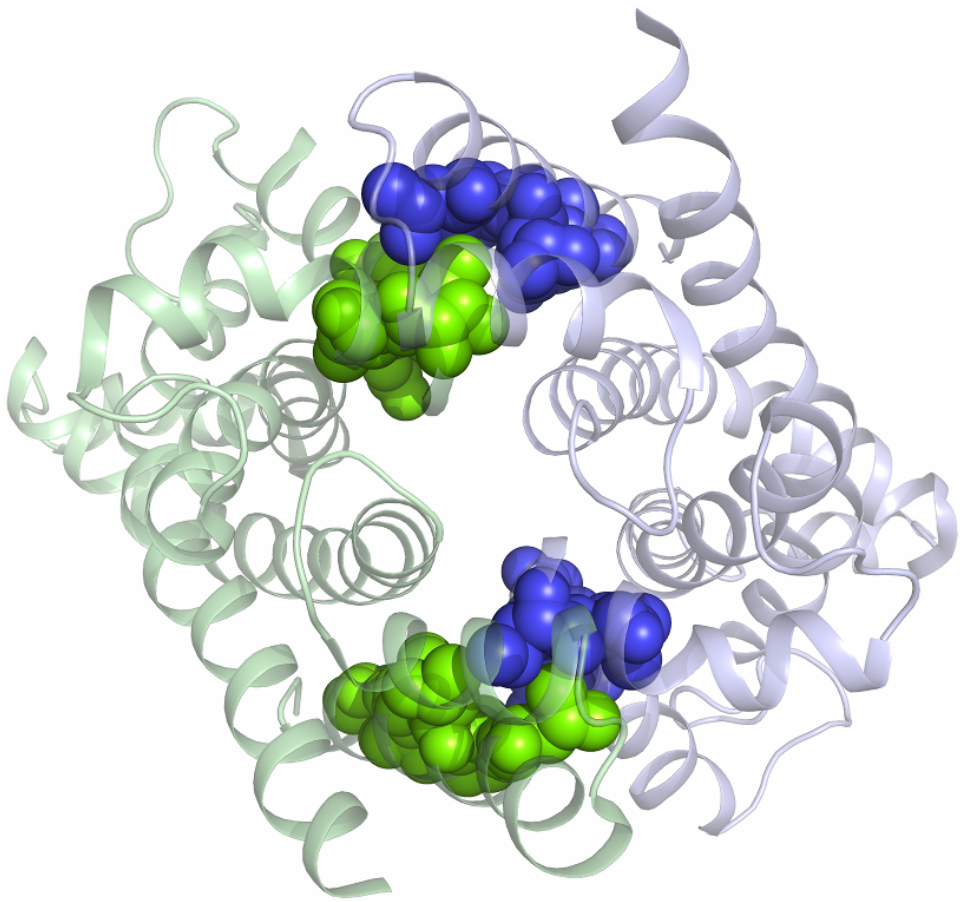
● Crystal structure contacts

Hopf, Colwell *et al*, Cell 2012

Evolutionary constraints identify dimer contacts

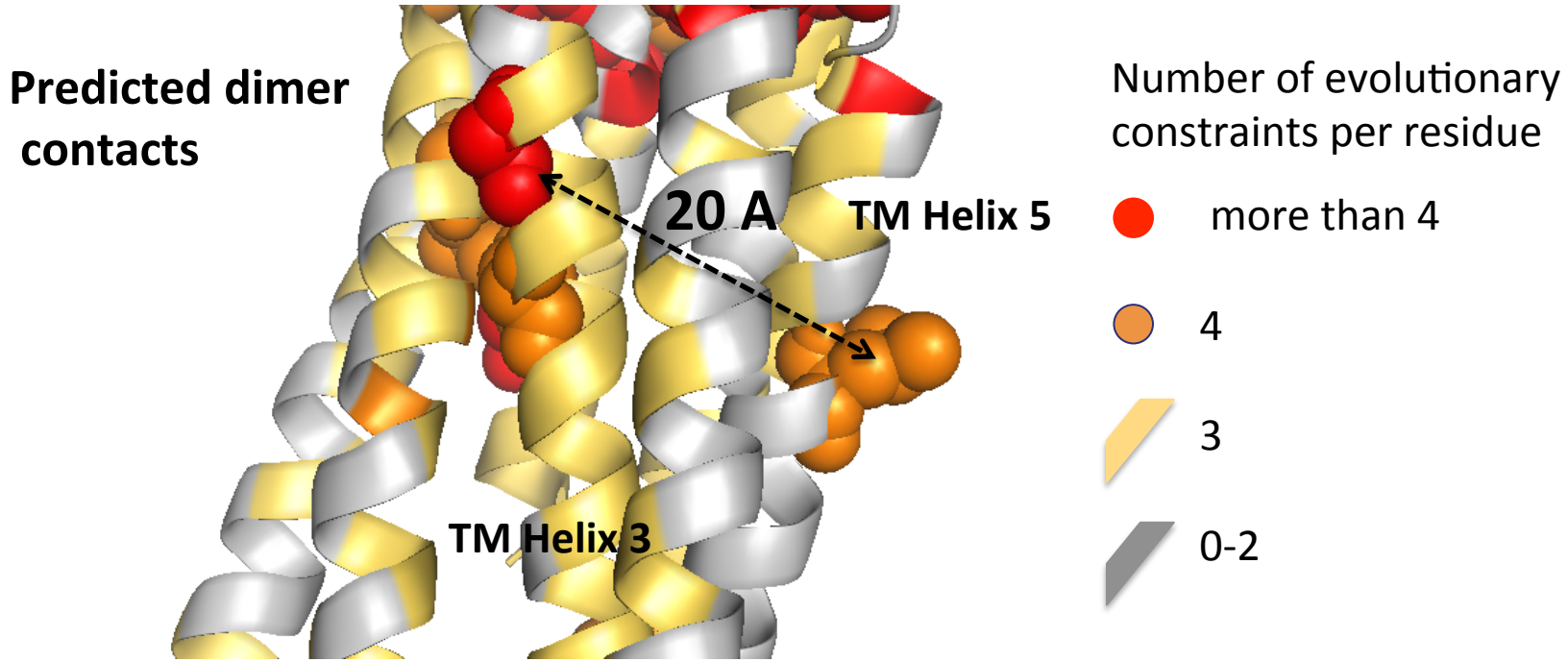


E. coli methionine importer MetI



E. coli methionine importer MetI

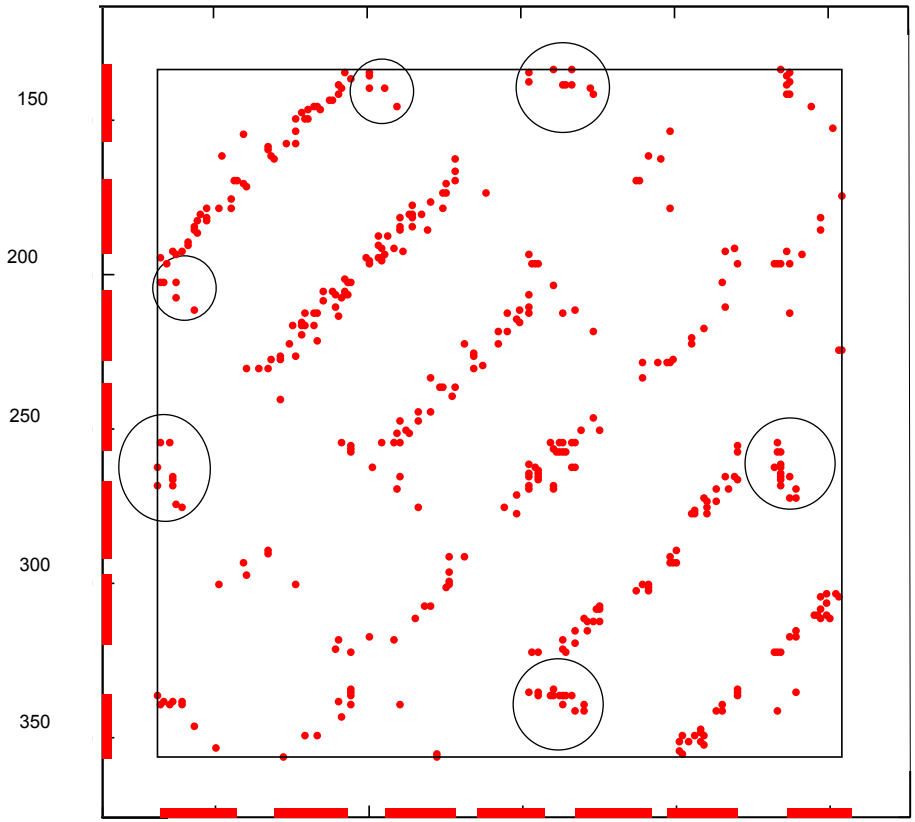
3D complexes for proteins of unknown monomer structure



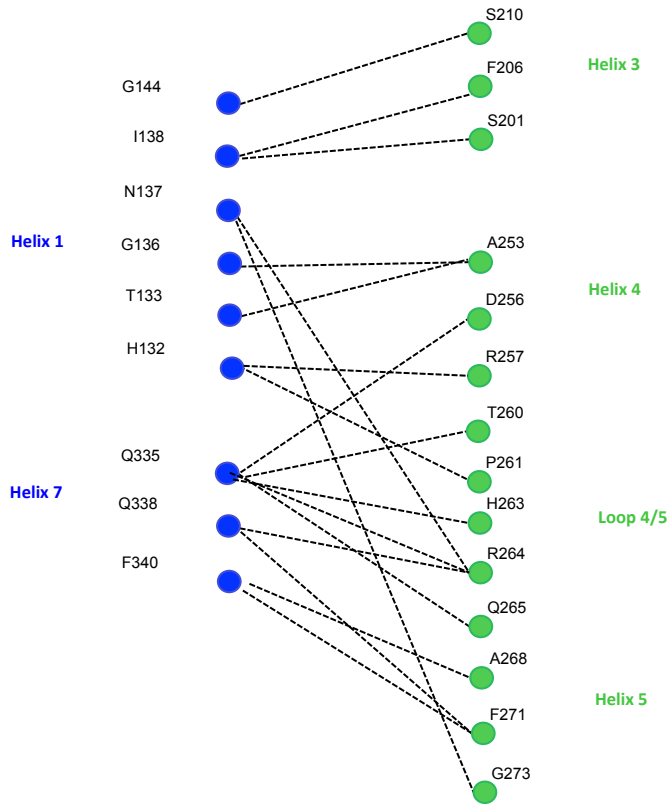
adiponectin receptor model – violated strong constraints are predicted dimer contacts

Predicted dimer interface, adiponection receptor

ADIPOR1 (200 ECs)



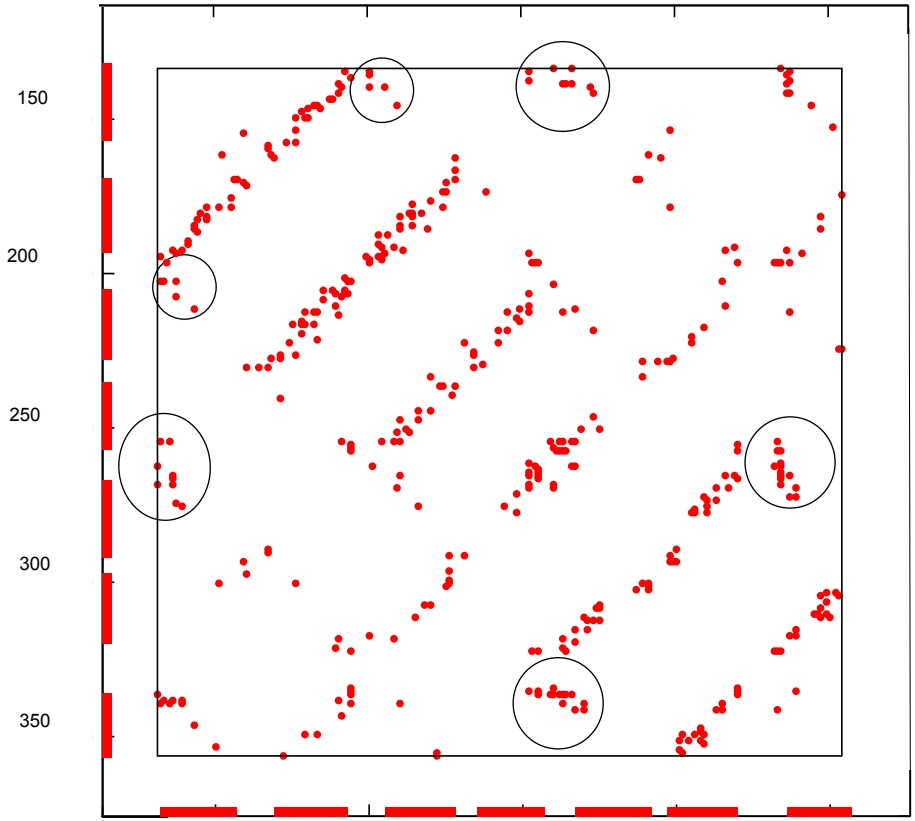
Predicted dimer interface contacts



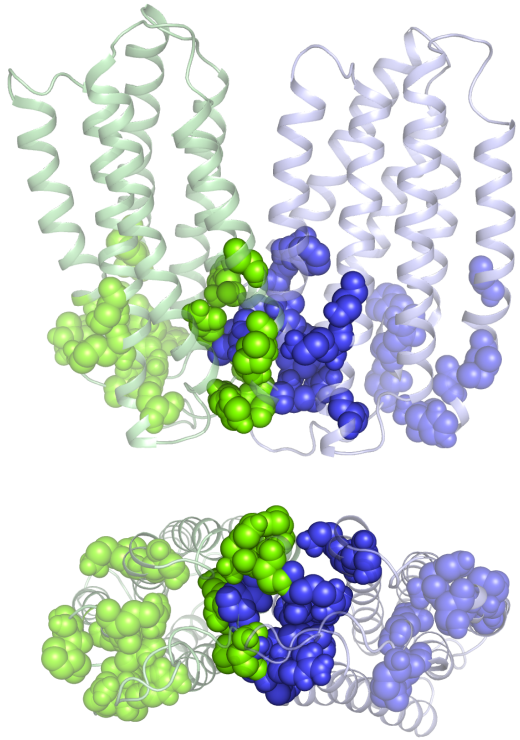
Hopf, Colwell *et al*, Cell 2012

Predicted dimer interface, adiponection receptor

ADIPOR1 (200 ECs)



AdipoR1 predicted dimer



Hopf, Colwell *et al*, Cell 2012

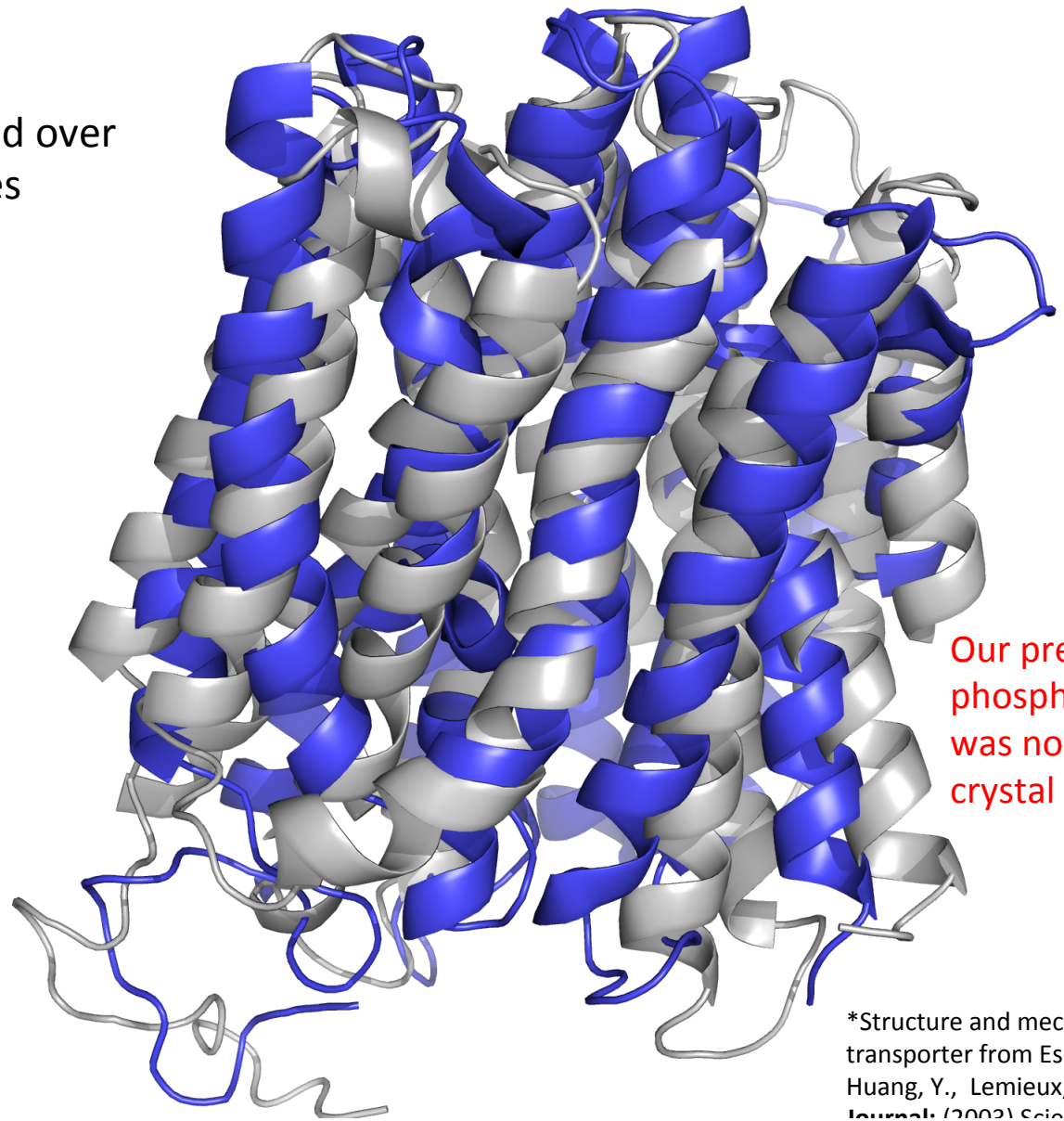
Beyond structure prediction

Constraints reflect conformational change

Beyond structure prediction

Constraints reflect conformational change

4 Å C-α rmsd over
300 residues



Crystal structure*

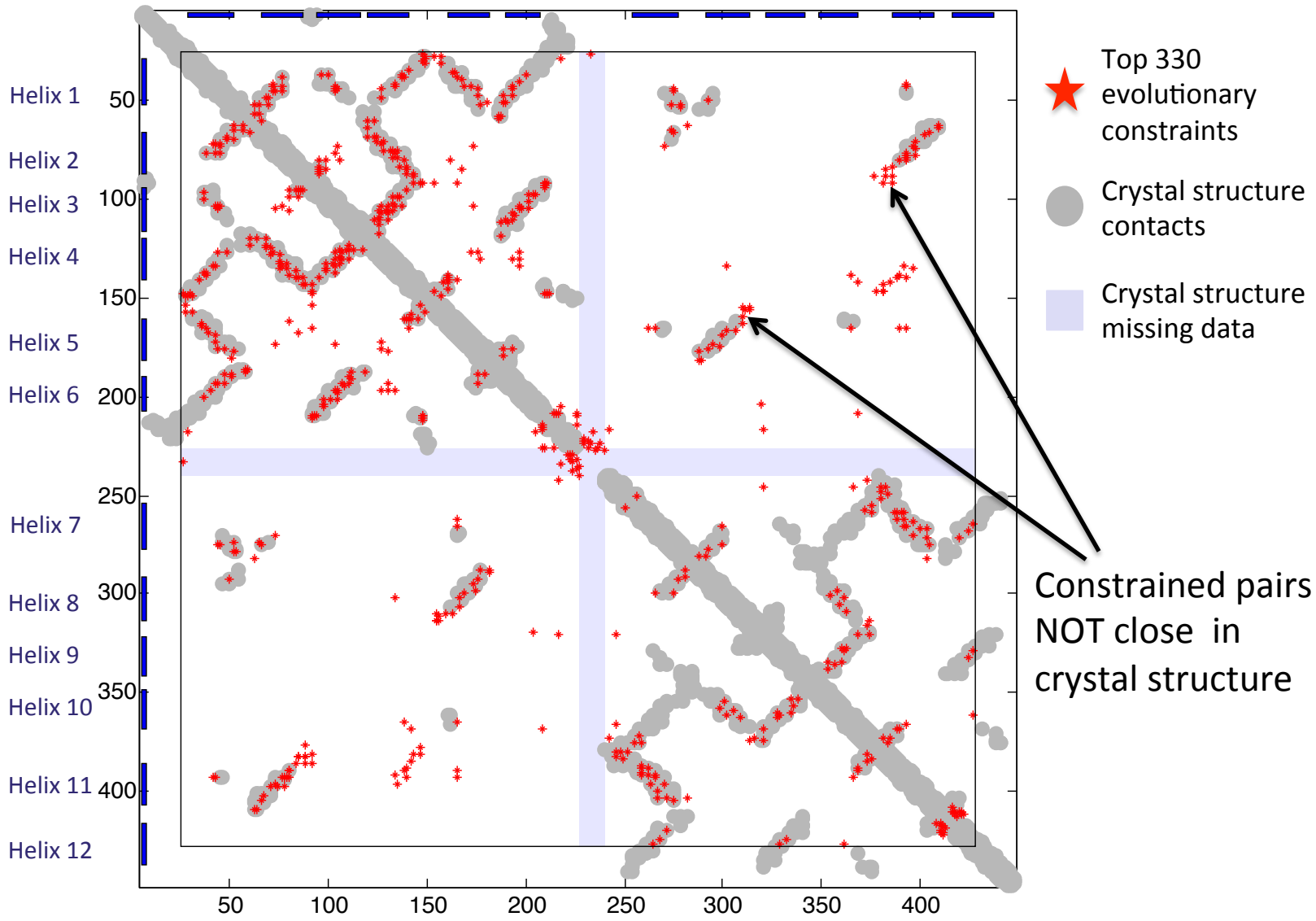
Predicted structure

Our prediction for the Glycerol-3-phosphate transporter
was not as open to cytoplasm as crystal structure

*Structure and mechanism of the glycerol-3-phosphate transporter from Escherichia coli.
Huang, Y., Lemieux, M.J., Song, J., Auer, M., Wang, D.N., **Journal:** (2003) Science **301:** 616-620

Glycerol-3-phosphate transporter: evolutionary constraints

Constraints reflect conformational change

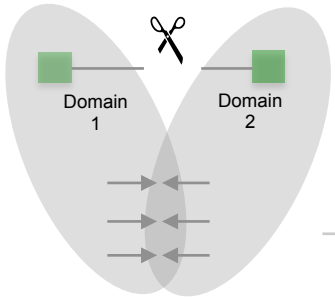


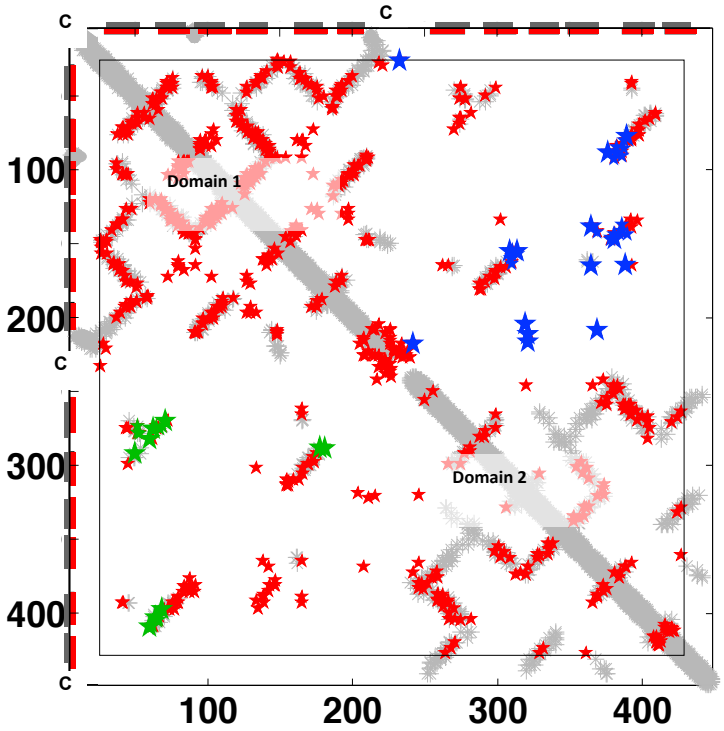
Hopf, Colwell *et al*, Cell 2012

Glycerol-3-phosphate transporter: evolutionary constraints

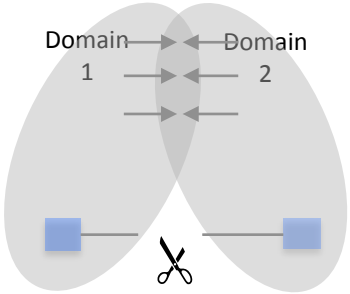
Constraints reflect conformational change

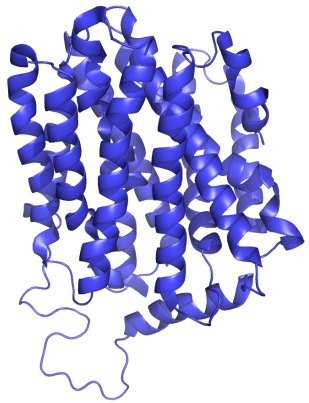
cytoplasmic side closed





cytoplasmic side open





Blue stars at intersection: pairs in **closed** conformation

Green stars at intersection: pairs in **open** conformation

Crystal structure residues less than 5Å apart

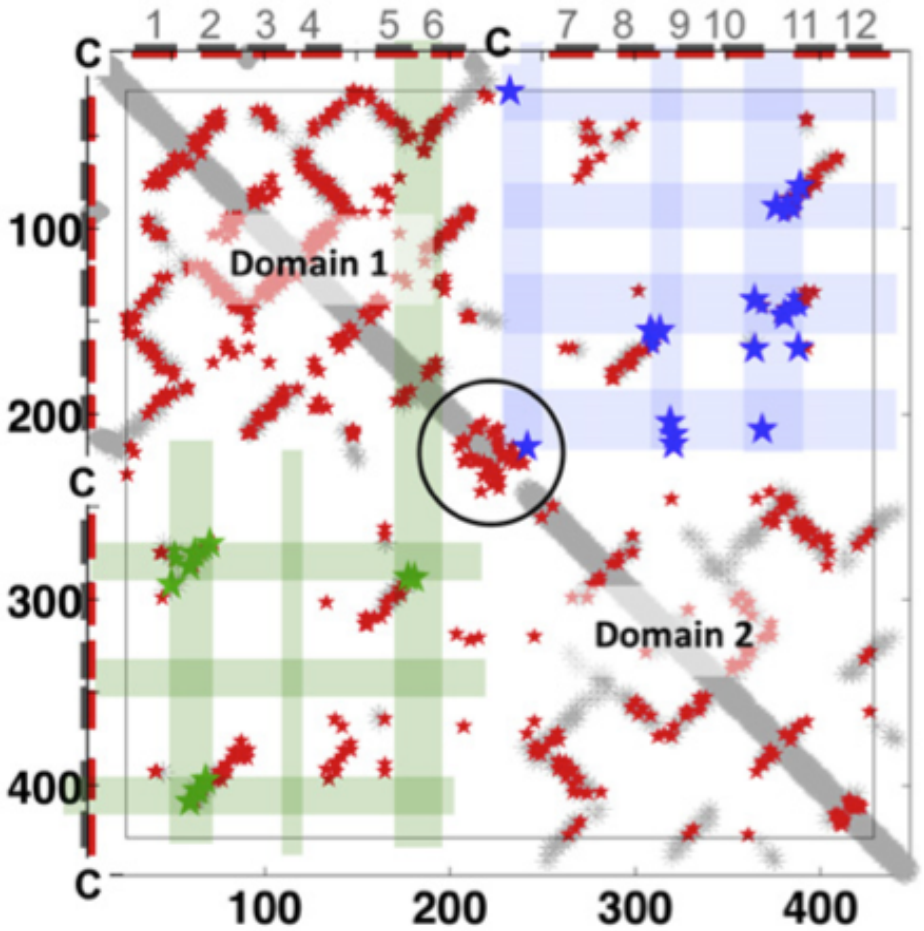
Top ranked pairs

Glycerol-3-phosphate transporter, a secondary active membrane transporter.

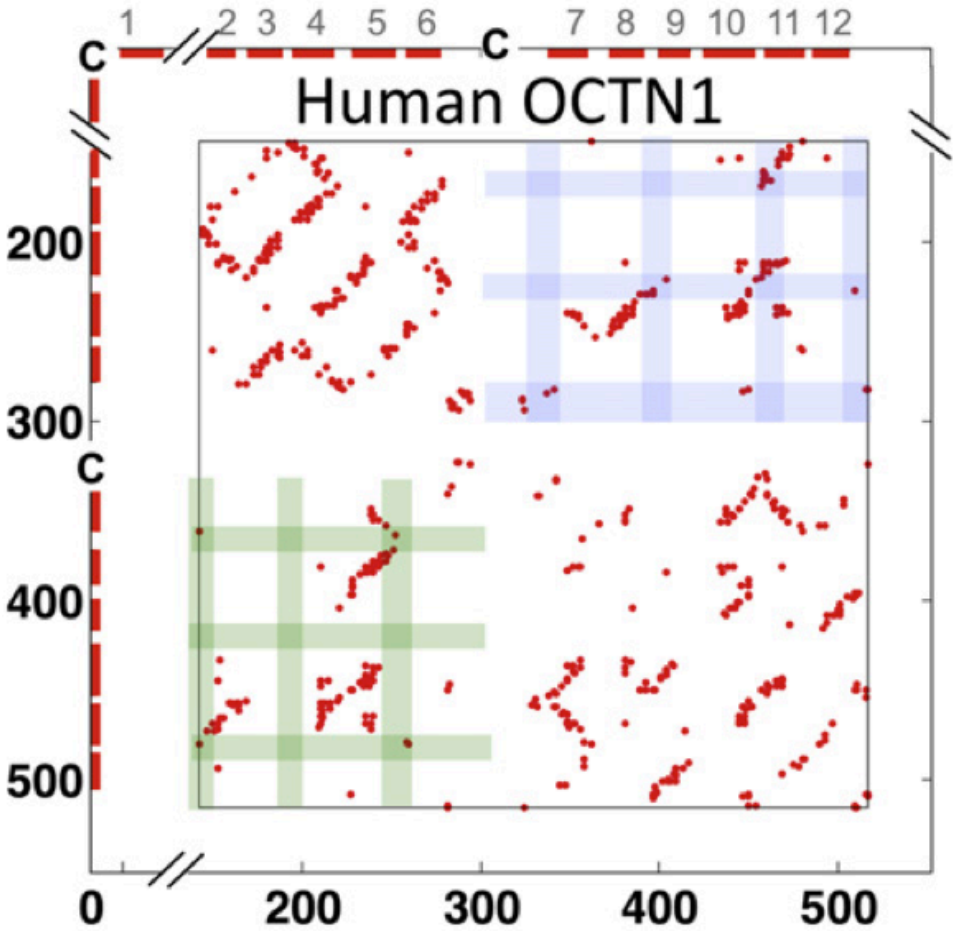
Changes conformation via a hypothesized 'rocker switch' mechanism from cytoplasmic open to periplasmic open.

Hopf, Colwell *et al*, Cell 2012

Method can be generalized to other membrane proteins to explore conformational flexibility



Glycerol-3-phosphate transporter, a secondary active membrane transporter.

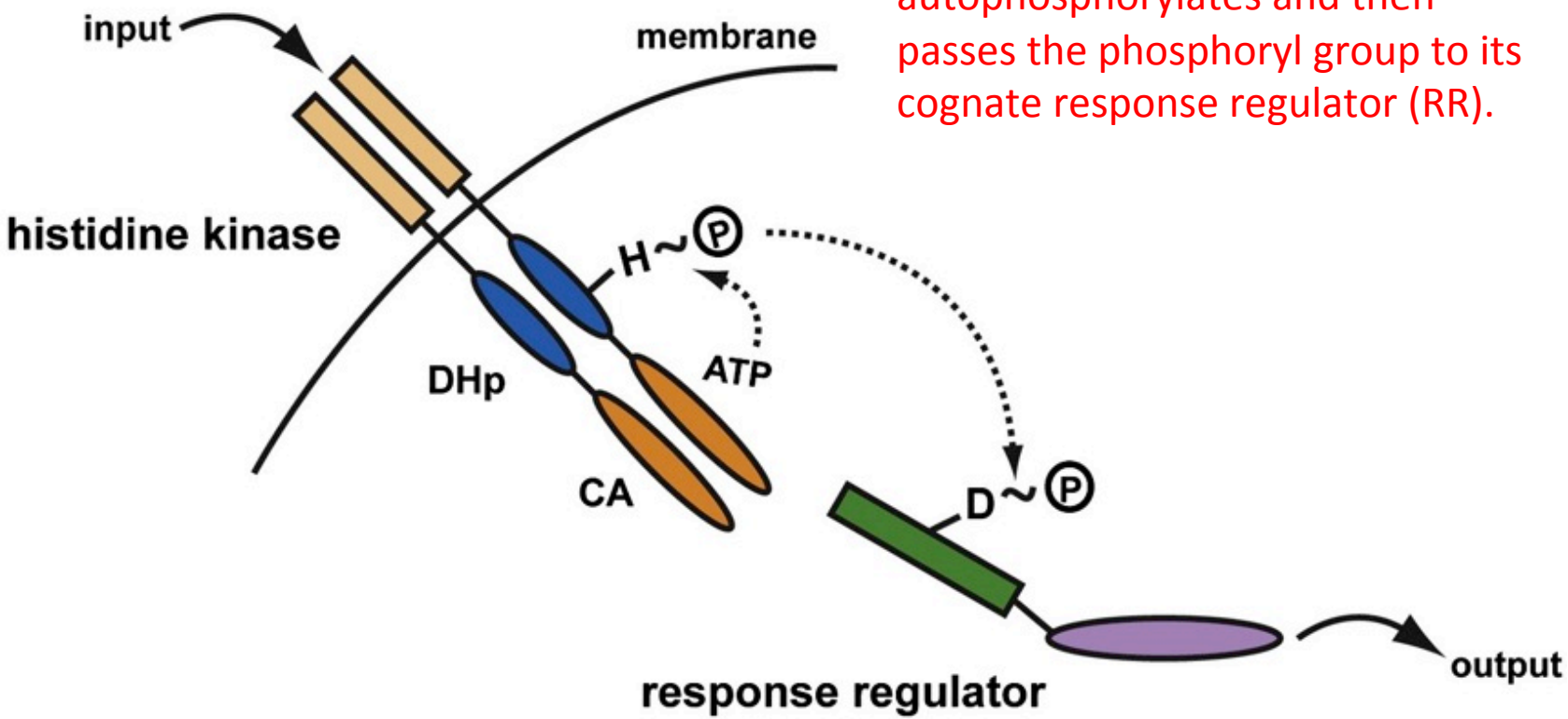


Protein responsible for the cotransport of sodium ions and ergothioneine, which is an antioxidant, into cells

Suggests 'helix rocking' typical of alternating access membrane transporters

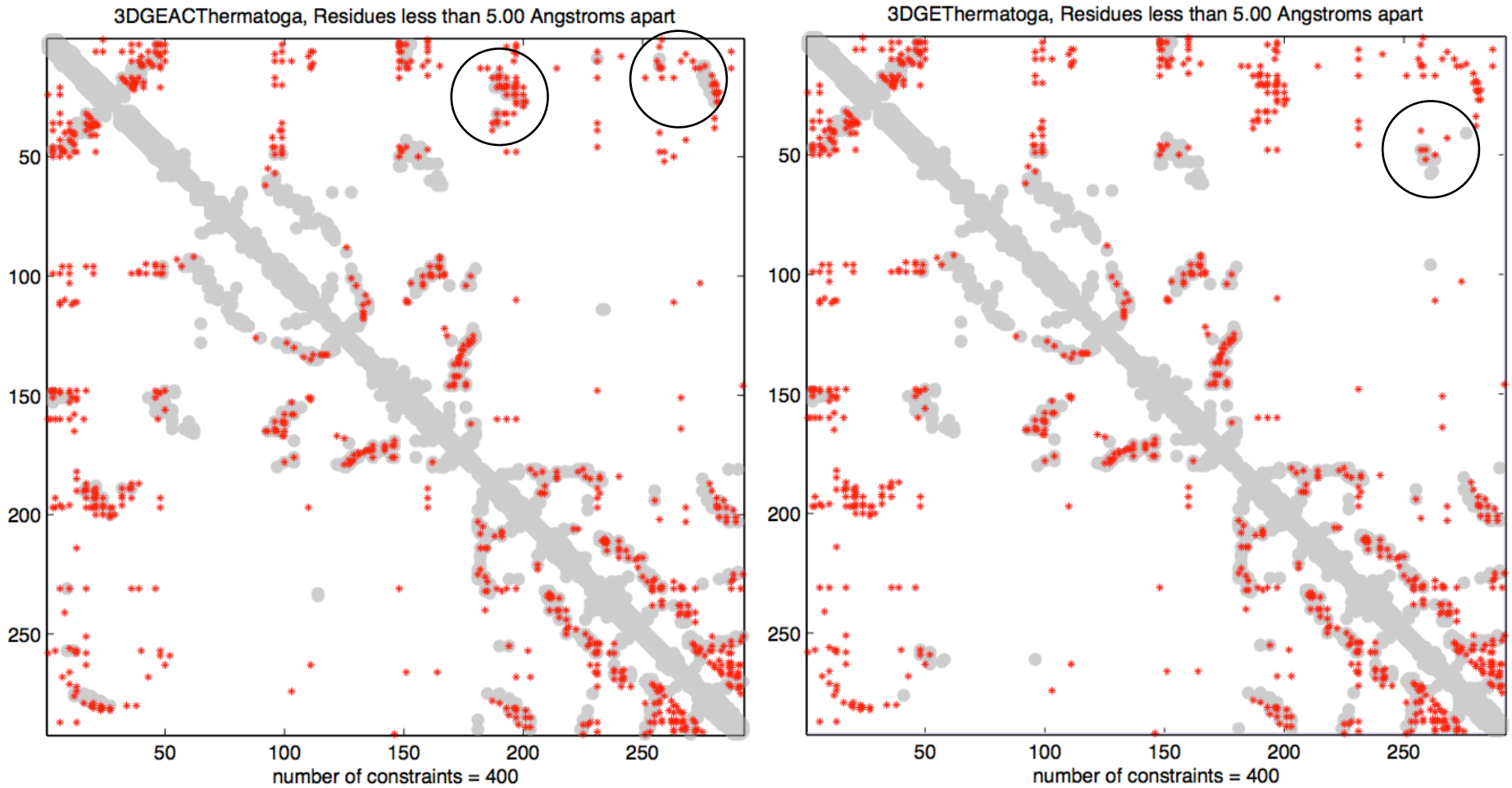
Conformational change

Histidine kinase (HK) autophosphorylates and then passes the phosphoryl group to its cognate response regulator (RR).

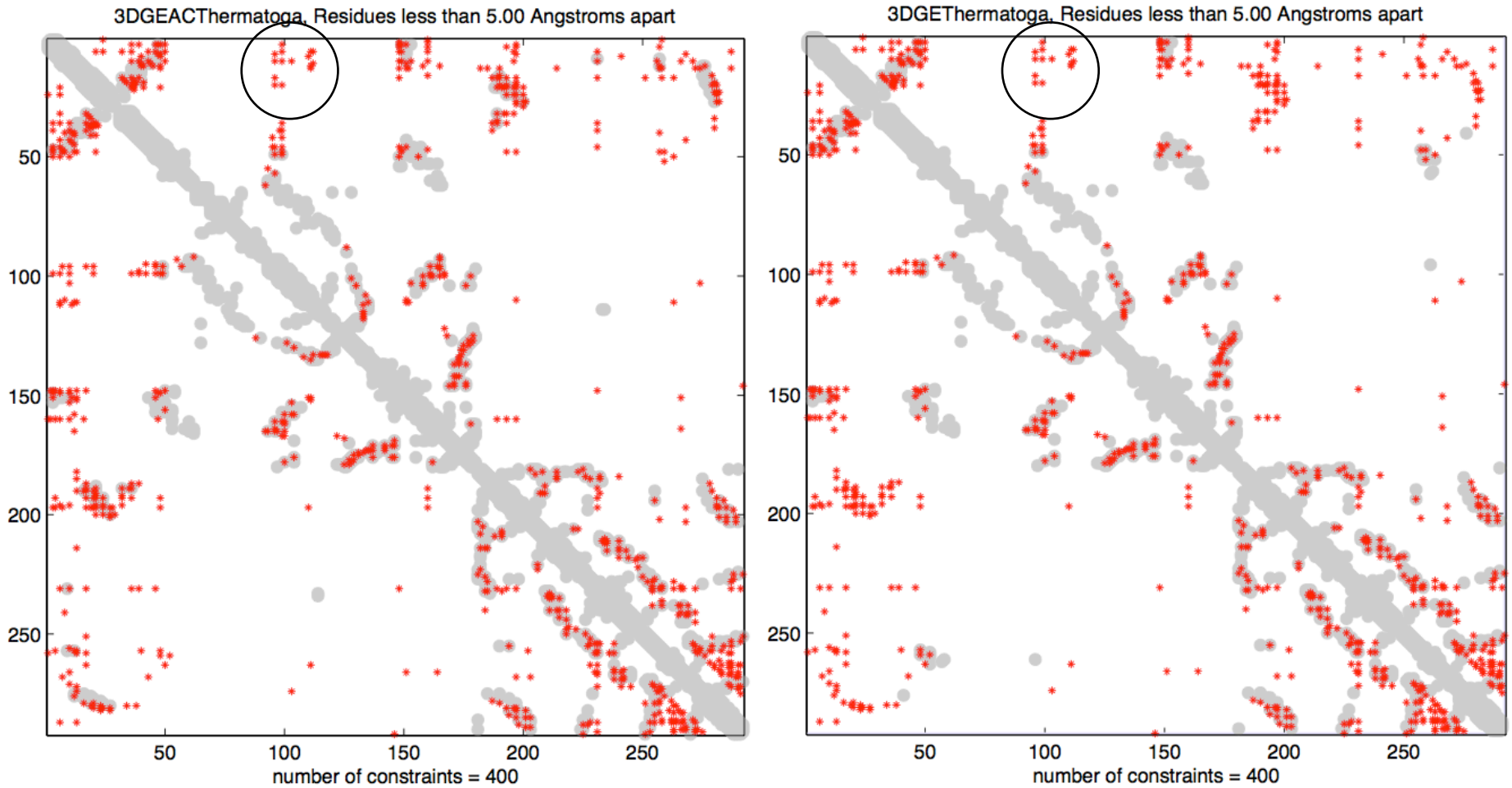


Rewiring the Specificity of Two-Component Signal Transduction Systems, Skerker *et al,* Cell, June 12 2008

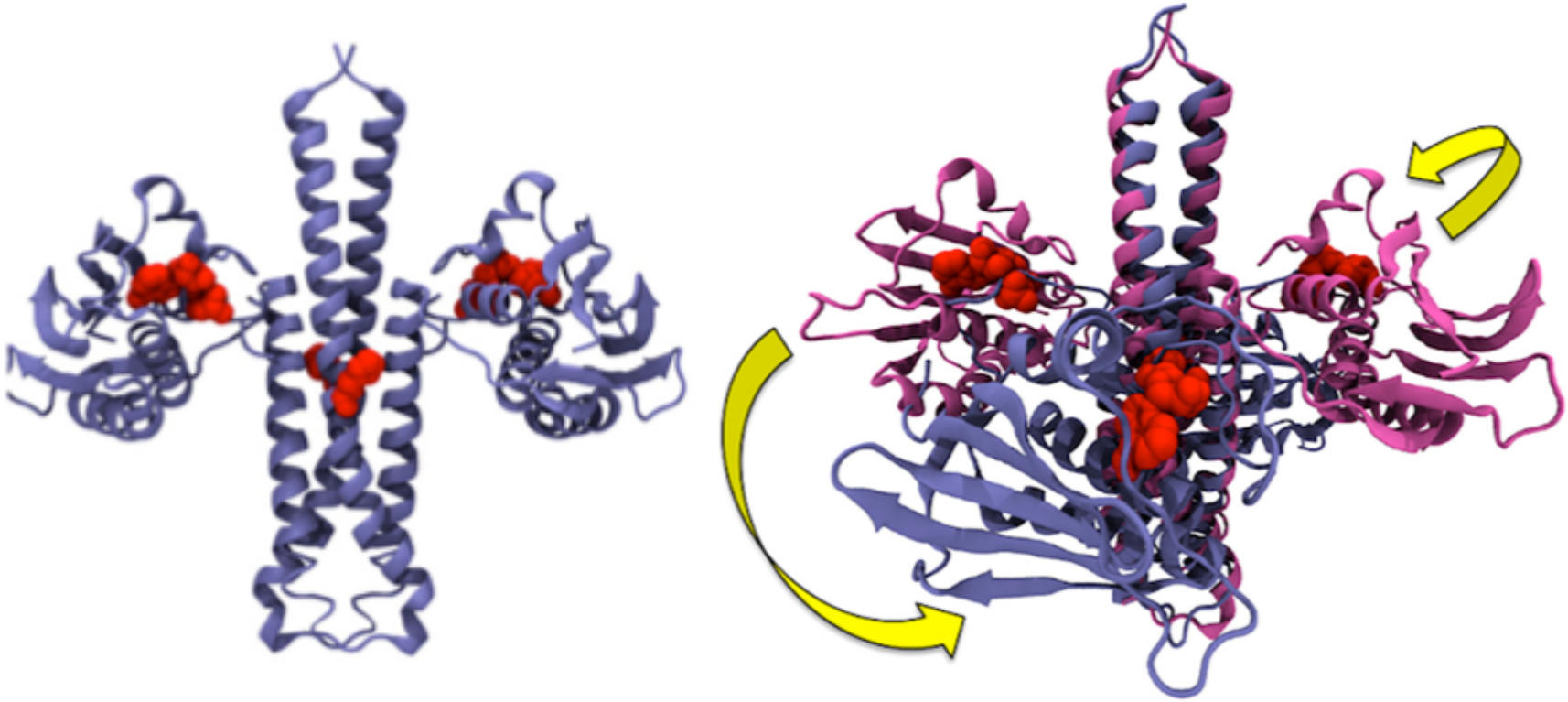
Protein protein interactions

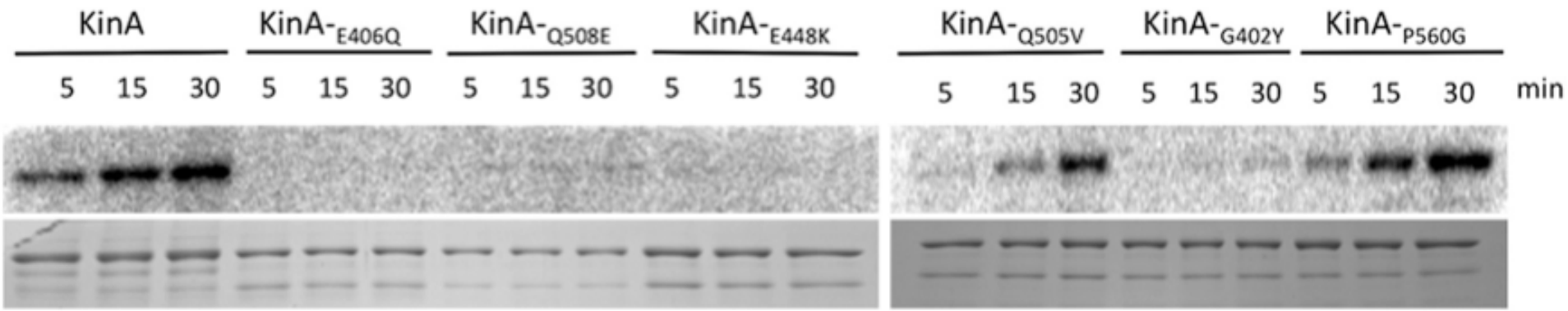


Predicting alternative conformations



Protein protein interactions



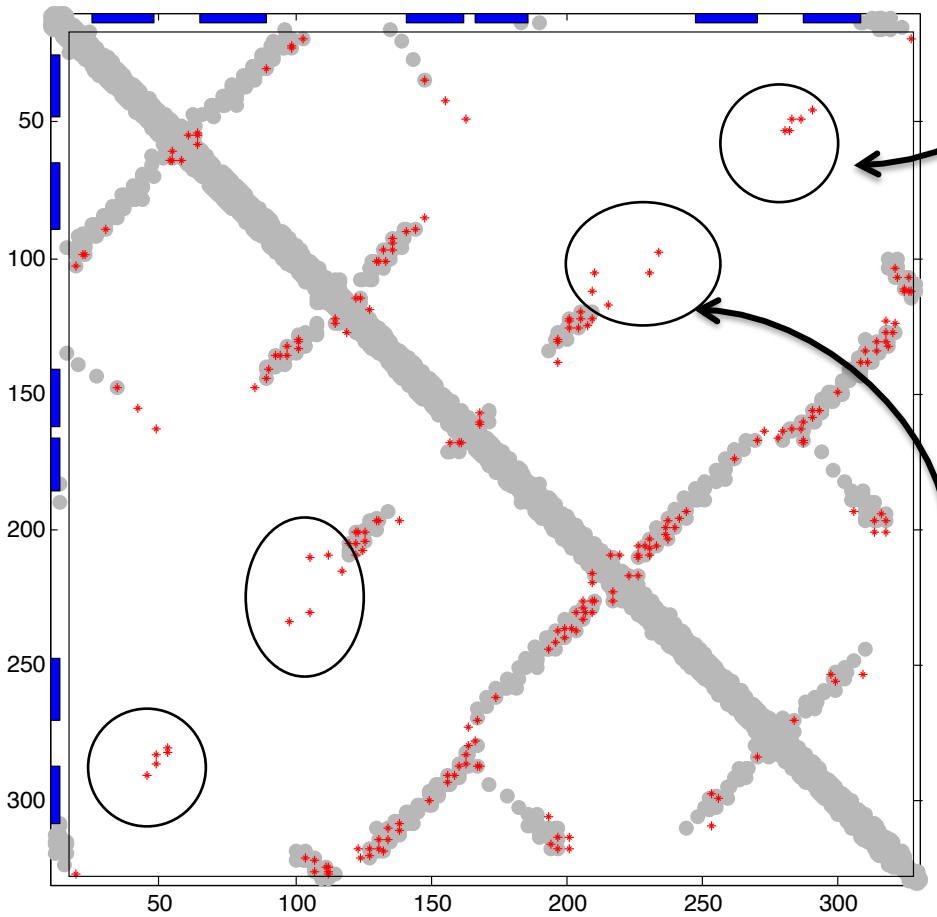


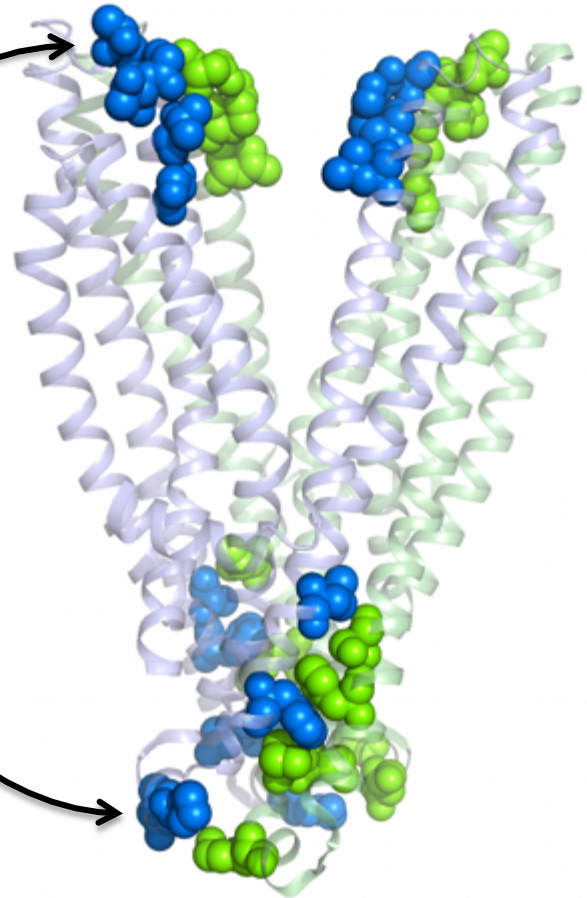
Beyond structure prediction

Constraints reflect multimerisation

Beyond structure prediction

Constraints reflect multimerisation





S. typhimurium ABC transporter
(antibiotic transporter)

★ Top ranked evolutionary constraints (ECs)

○ Multimer ECs

● Crystal structure contacts

Hopf, Colwell *et al*, Cell 2012

Outline

Biology 101 – a quick tour of proteins

Biological information from covariance

- Protein structure and function
- Protein interaction specificity
- Ligand protein binding

Theoretical analysis of data limitations

– when can we predict reliably?
– Noise due to finite sampling, samples are not independent.
– central role played by Random Matrix Theory.

Accurately predict protein interaction partners

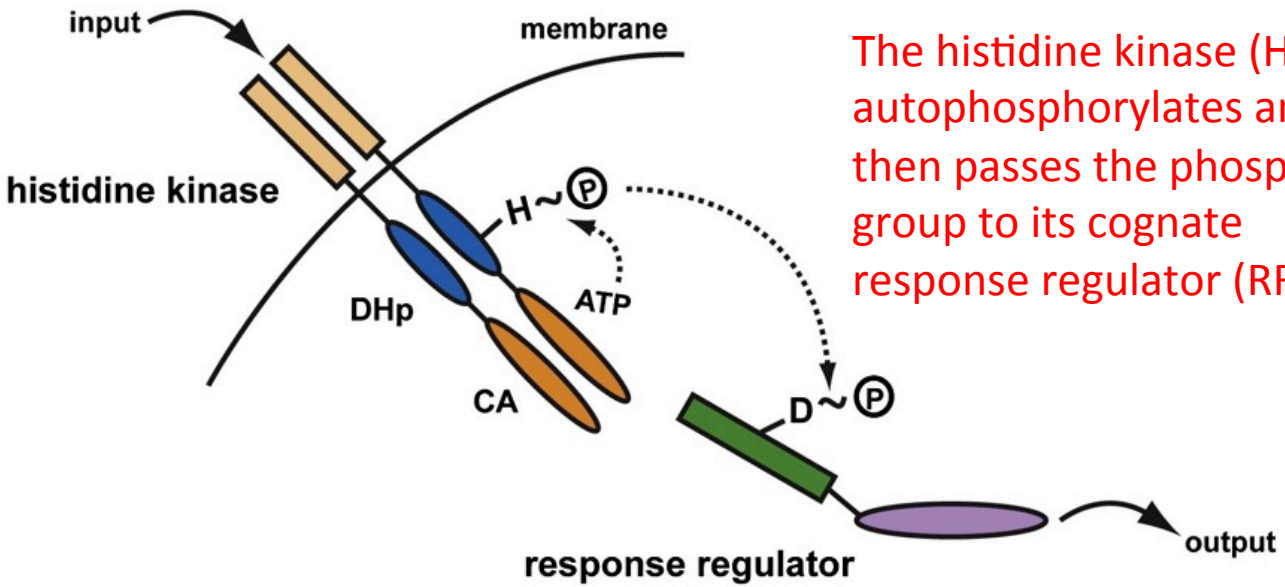
Covariance calculated from the concatenated sequences of protein interaction partners accurately predicts the protein complex structure.

BUT….

To apply this analysis, we need a starting set of known interaction partners, so we can form the concatenated sequence alignment.

This is particularly difficult to obtain if genomes contain multiple paralogs (versions) of either of the two proteins, as is frequently the case – PFAM A-30 families have on average 3.9 paralogs per species.

Protein interaction partners – HKs and RRs



The histidine kinase (HK) autophosphorylates and then passes the phosphoryl group to its cognate response regulator (RR).

	a1	a2	a3	a4	a5	a6		b1	b2	b3	b4	b5	b6	b7
E. c	K	D	I	L	V	D		T	I	F	D	F	C	L
S. a	K	D	I	L	V	D		T	I	F	D	P	C	I
T. t	D	D	I	K	V	D		T	I	F	H	P	C	L
S. e	D	E	I	L	V	H		S	I	F	H	P	C	I

Rewiring the Specificity of Two-Component Signal Transduction Systems, Skerker *et al,* Cell, June 12 2008

Protein interaction partners – HKs and RRs

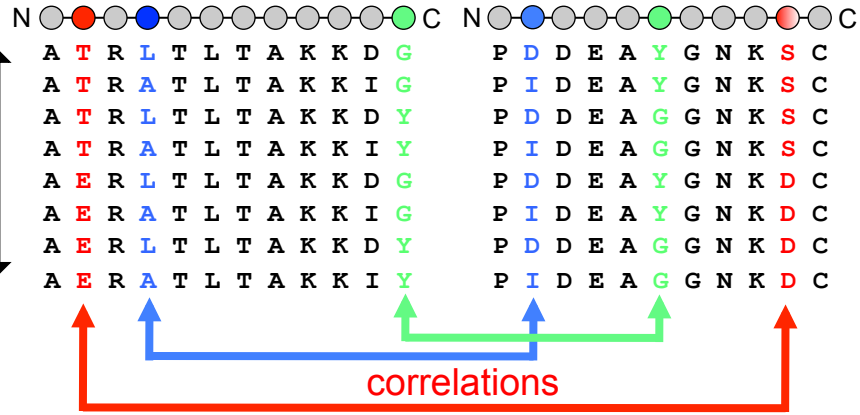
Can we invent an algorithm that does not need a set of known interaction partners?

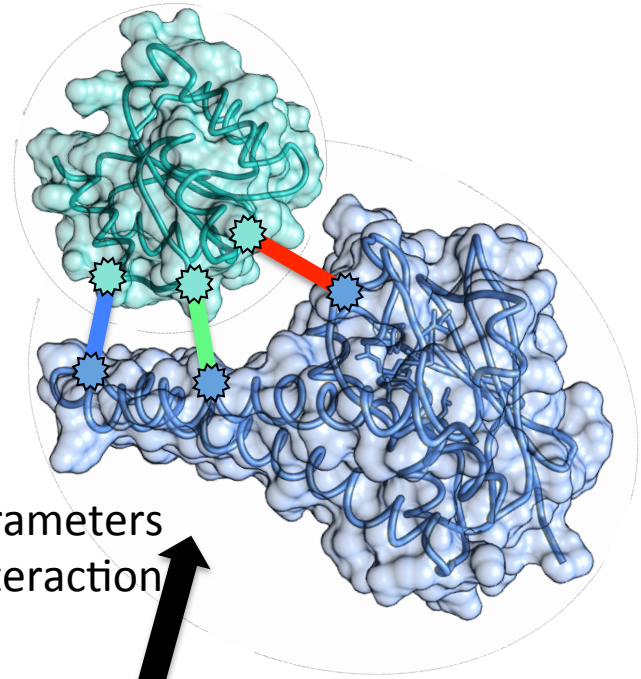
Anne-Florence Bitbol

Ned Wingreen

Use the same approach to predict interactions between proteins

Protein sequence homologs





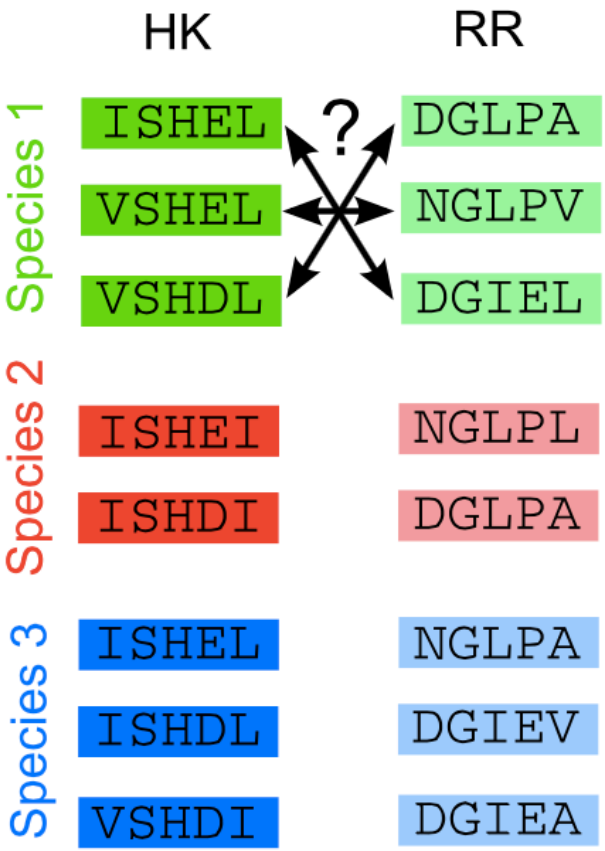
Use sequence data to build model

Model parameters predict interaction partners

$$P(A_1,\ldots,A_L) = \frac{1}{Z}\exp\left\{-\sum_i h_i A_i - \sum_{(i,j)} e_{ij}\left(A_i, A_j\right)\right\}$$

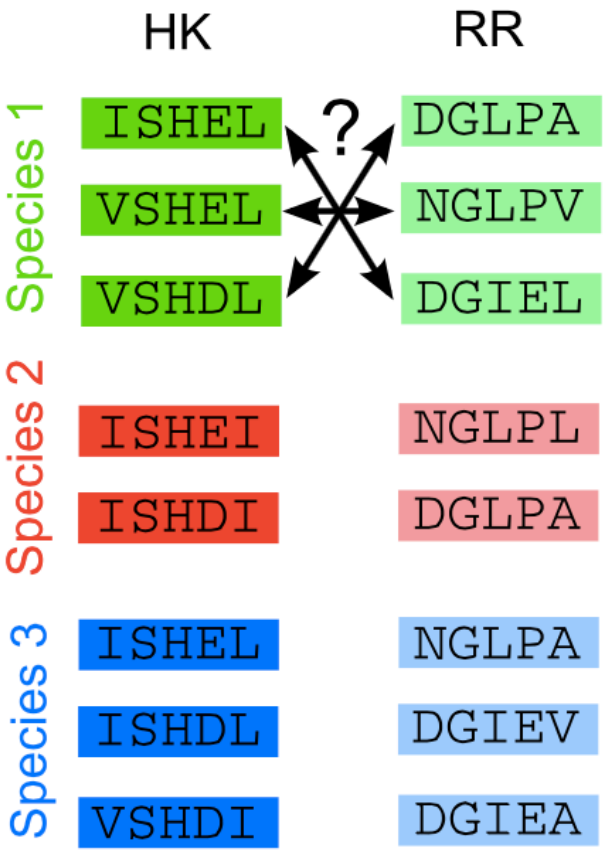
Bitbol *et al*, PNAS 2016

Protein interaction partners – HKs and RRs

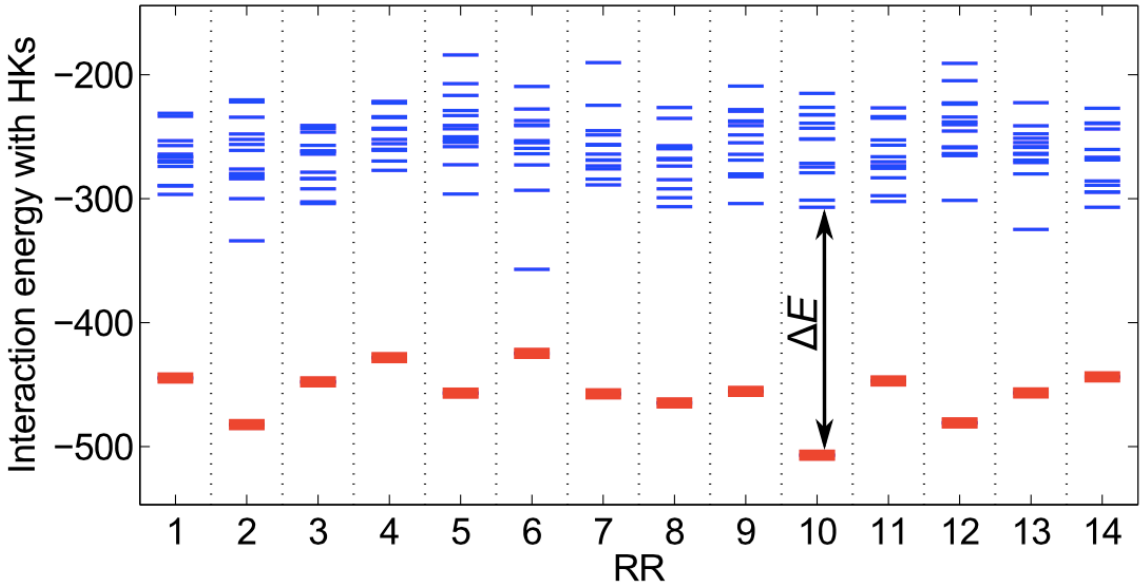


	a 1	a 2	a 3	a 4	a 5	a 6		b 1	b 2	b 3	b 4	b 5	b 6	b 7
E. c	K	D	I	L	V	D		T	I	F	D	F	C	L
S. a	K	D	I	L	V	D		T	I	F	D	P	C	I
T. t	D	D	I	K	V	D		T	I	F	H	P	C	L
S. e	D	E	I	L	V	H		S	I	F	H	P	C	I

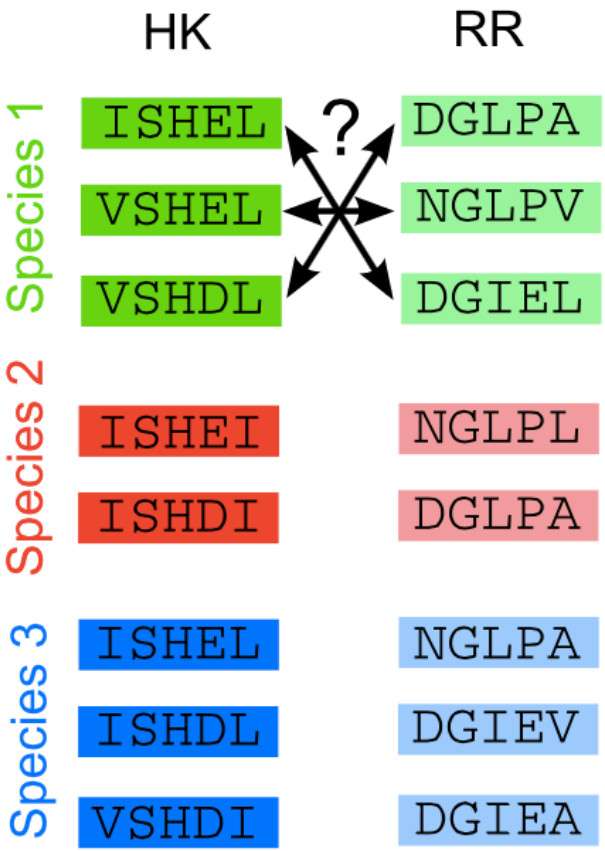
Protein interaction partners – HKs and RRs

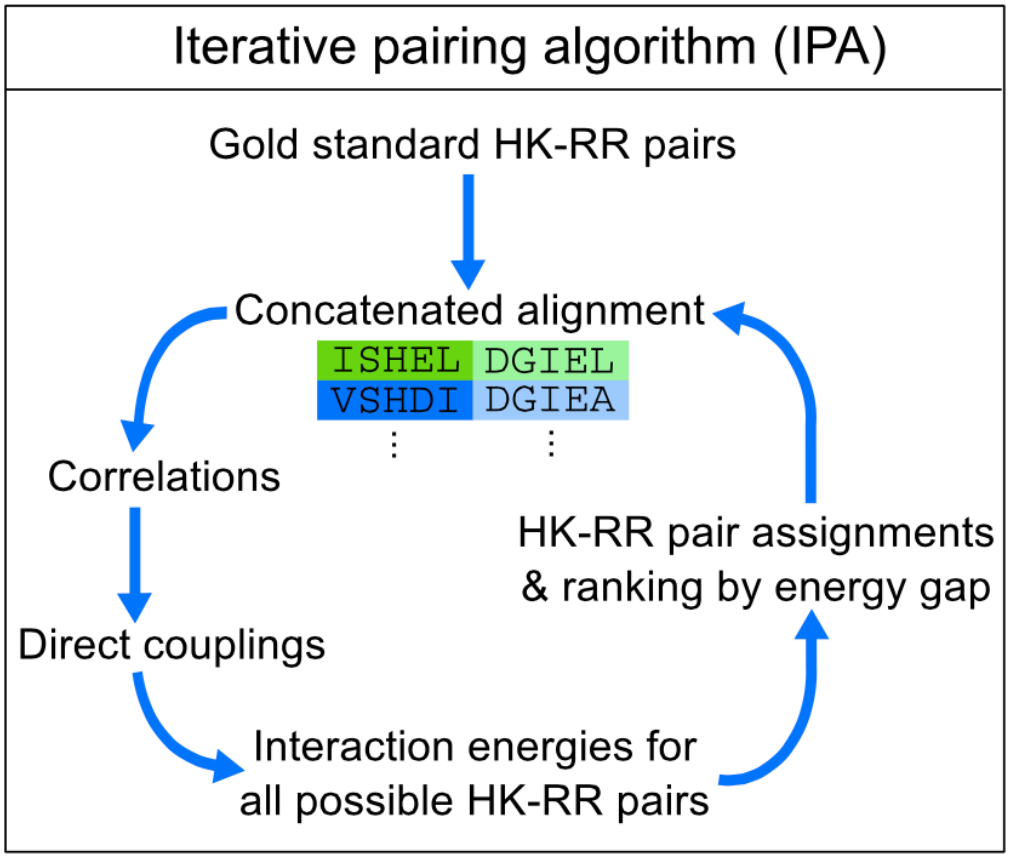


Use the model built from training data to score each possible pairing, within each species.

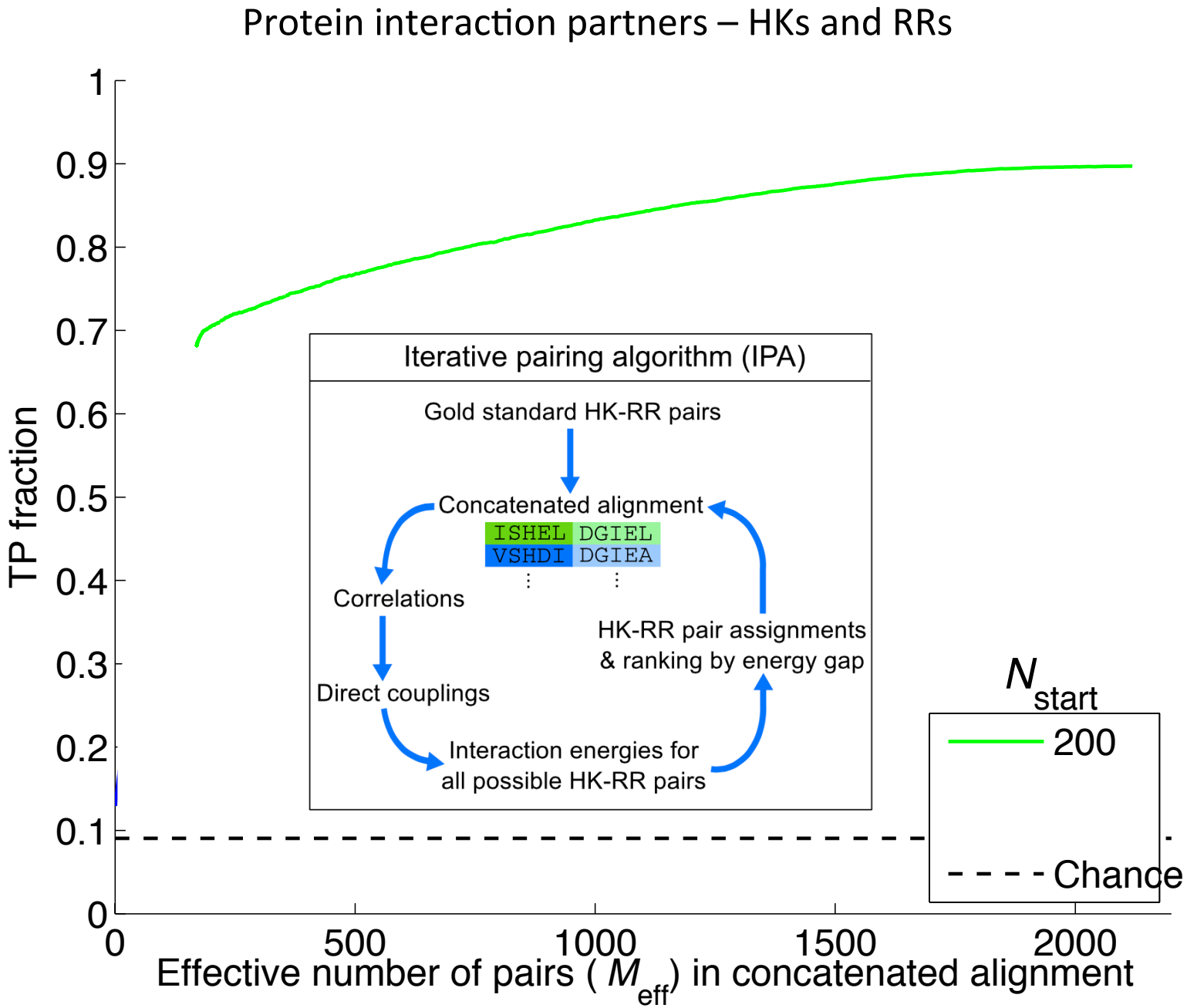


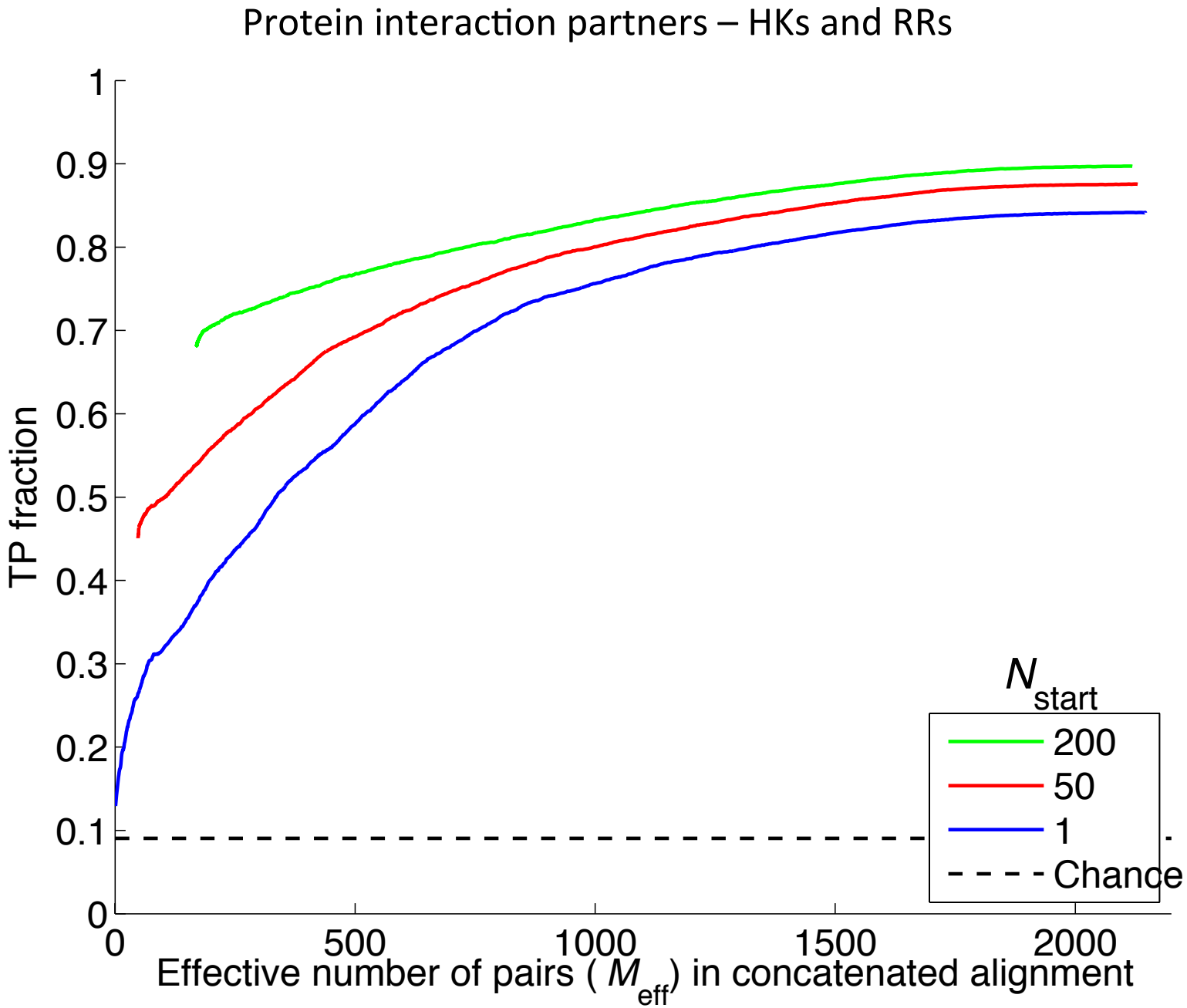
Accurately predict protein interaction partners – HKs and RRs



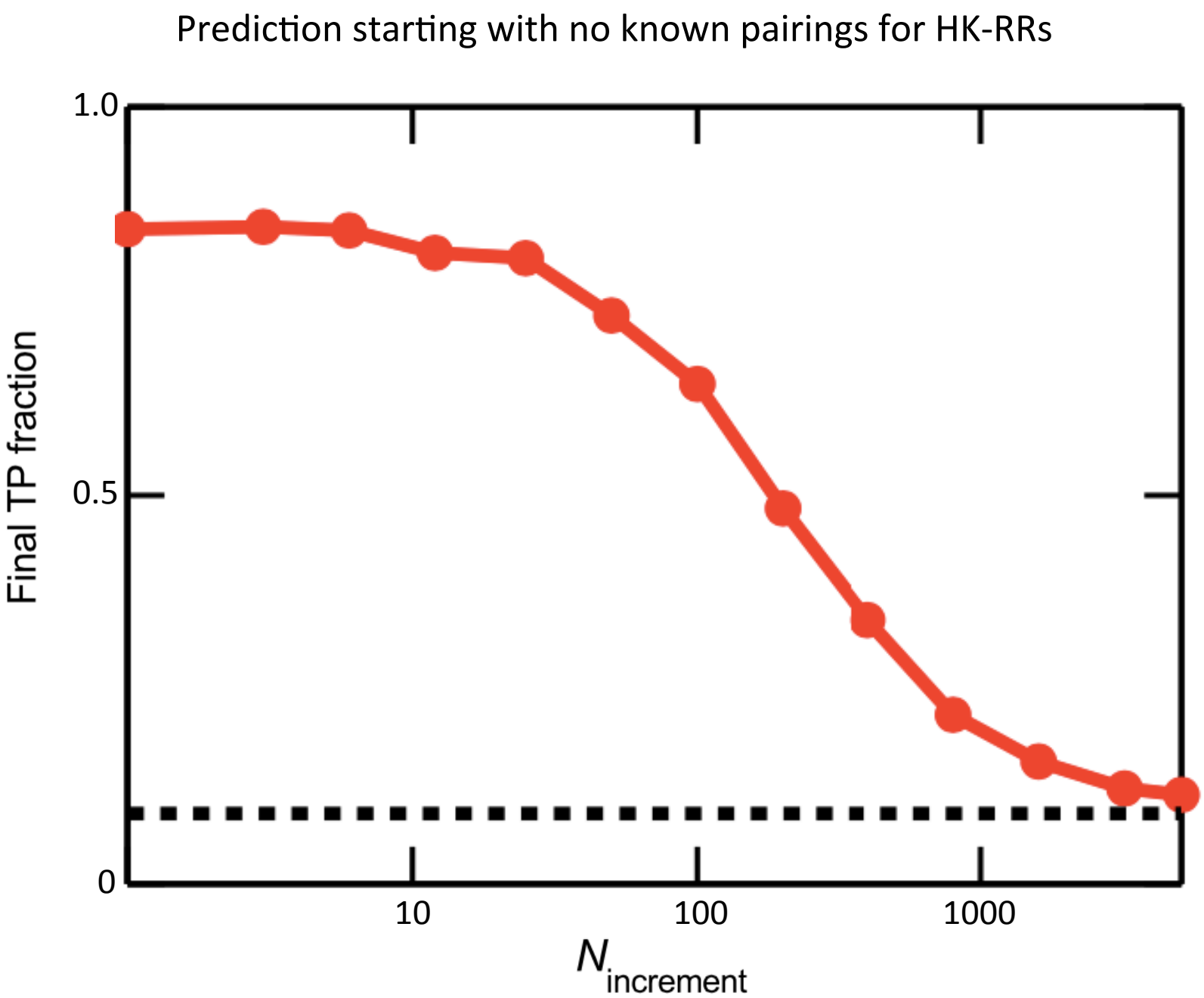


Bitbol *et al*, PNAS 2016

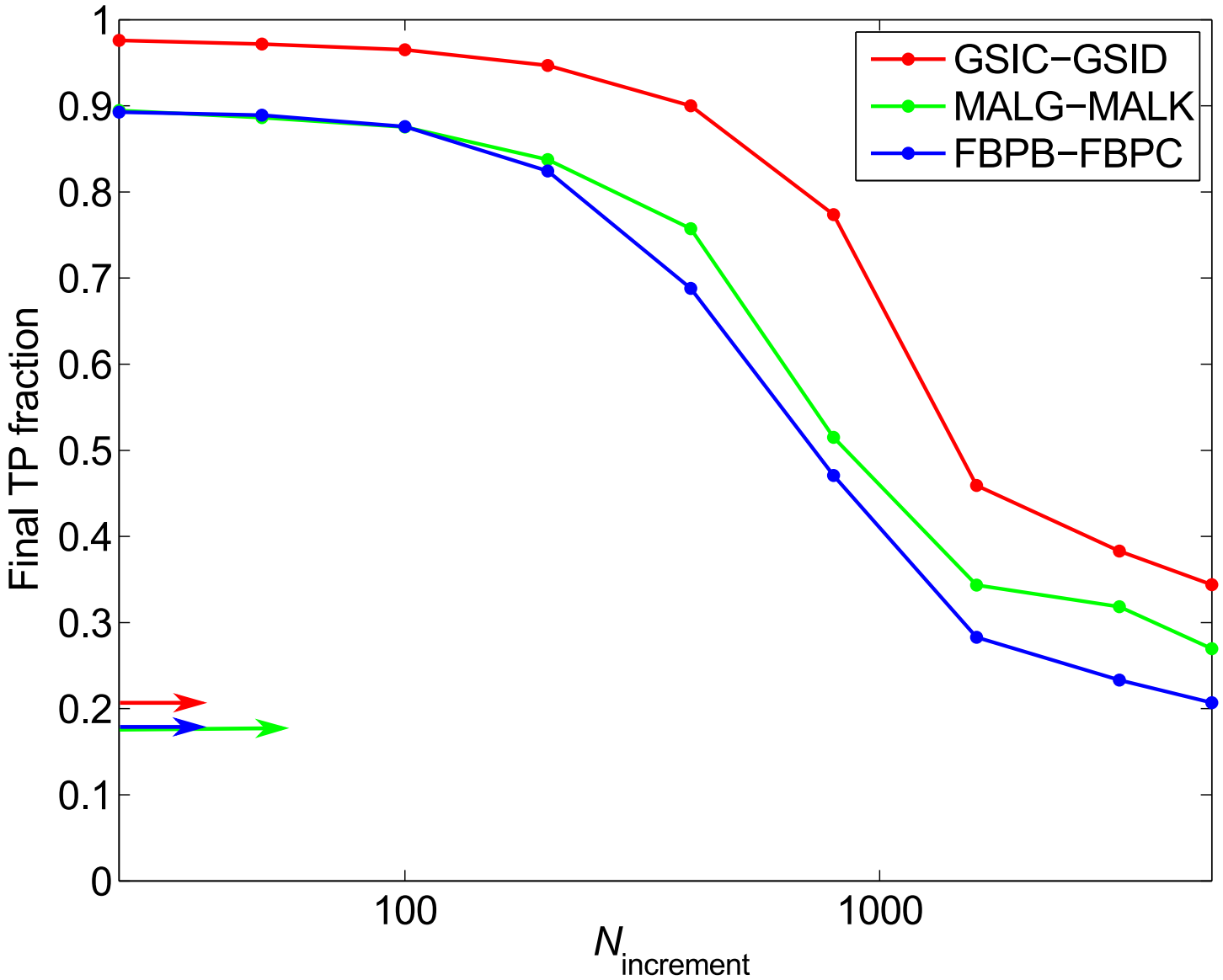

Protein interaction partners – HKs and RRs
TP fraction
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
500
1000
1500
2000
Effective number of pairs (M_{eff}) in concatenated alignment
Iterative pairing algorithm (IPA)
Gold standard HK-RR pairs
Concatenated alignment
ISHEL DGIEL
VSHDI DGIEA
Correlations
Direct couplings
Interaction energies for
all possible HK-RR pairs
HK-RR pair assignments
& ranking by energy gap
N_{start}
200
Chance



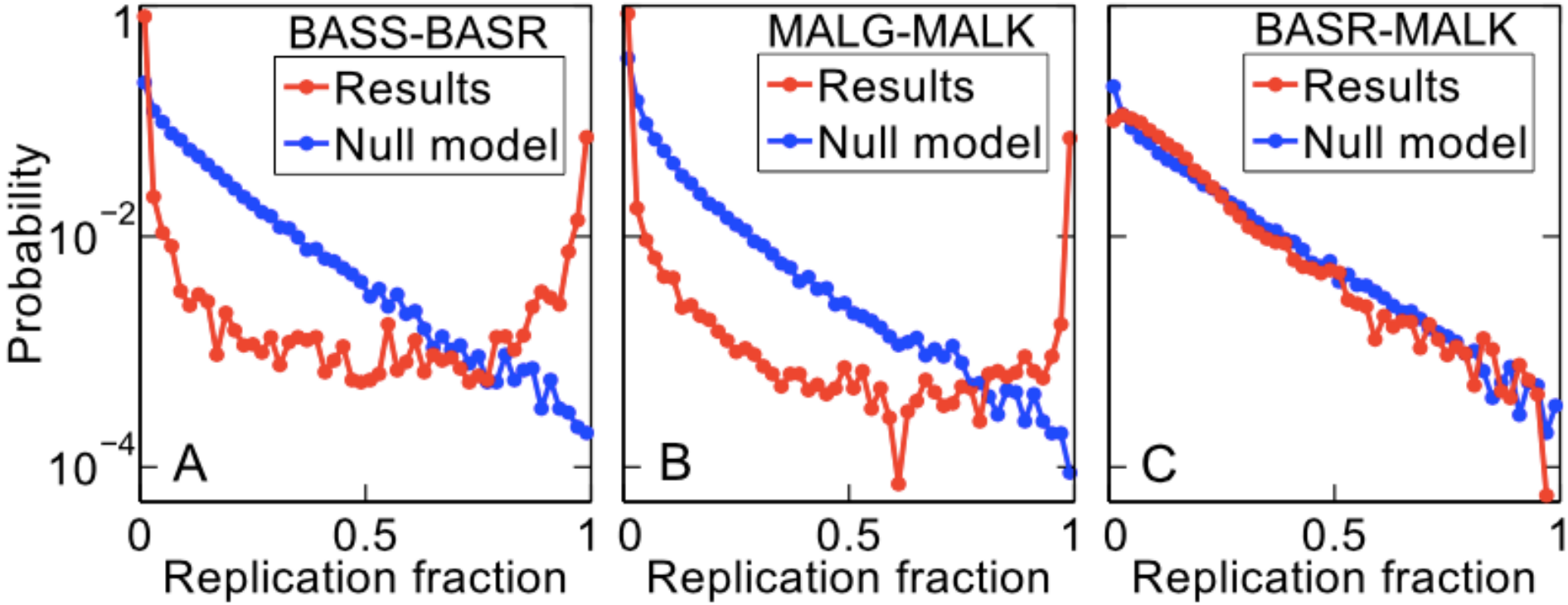
Bitbol *et al*, PNAS 2016


Prediction starting with no known pairings for HK-RRs
1.0
0.5
0
Final TP fraction
10
100
1000
$N_{increment}$

Prediction starting with no known pairings for other protein pairs



We can also predict whether two proteins interact or not!

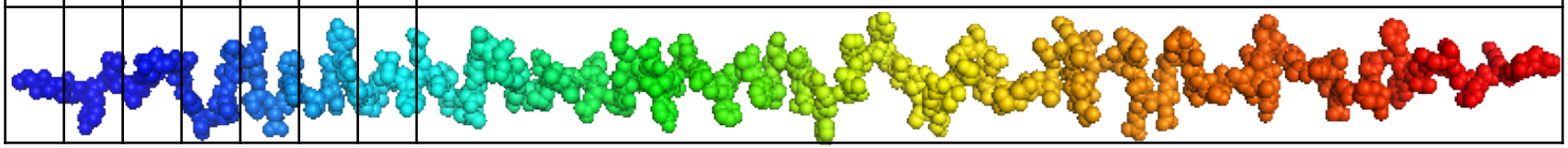


Bitbol *et al*, PNAS 2016

How to extract useful information from protein sequences?

Each protein is constructed as a specific chain or string of amino acids.

In a specific protein of interest, such as hemoglobin, the order of amino acids is highly important and contains all the information necessary to produce the folded, functional molecule.

1	2	3	4	5	6	7	8 …….
M	V	L	S	P	A	D	K ……
							

We would like to find a probability model for the sequence of amino acids that corresponds to each protein of interest.

$P(A_1,\ldots,A_L)$ = Probability that a sequence produces a folded, functional hemoglobin molecule.

What information is stored in sequences?

Sequences record the outcome of **millions of evolutionary experiments** which are **constrained** by the requirements of protein structure and function.

Ligand binding

Structure

Oligerimization

Enzyme Kinetics

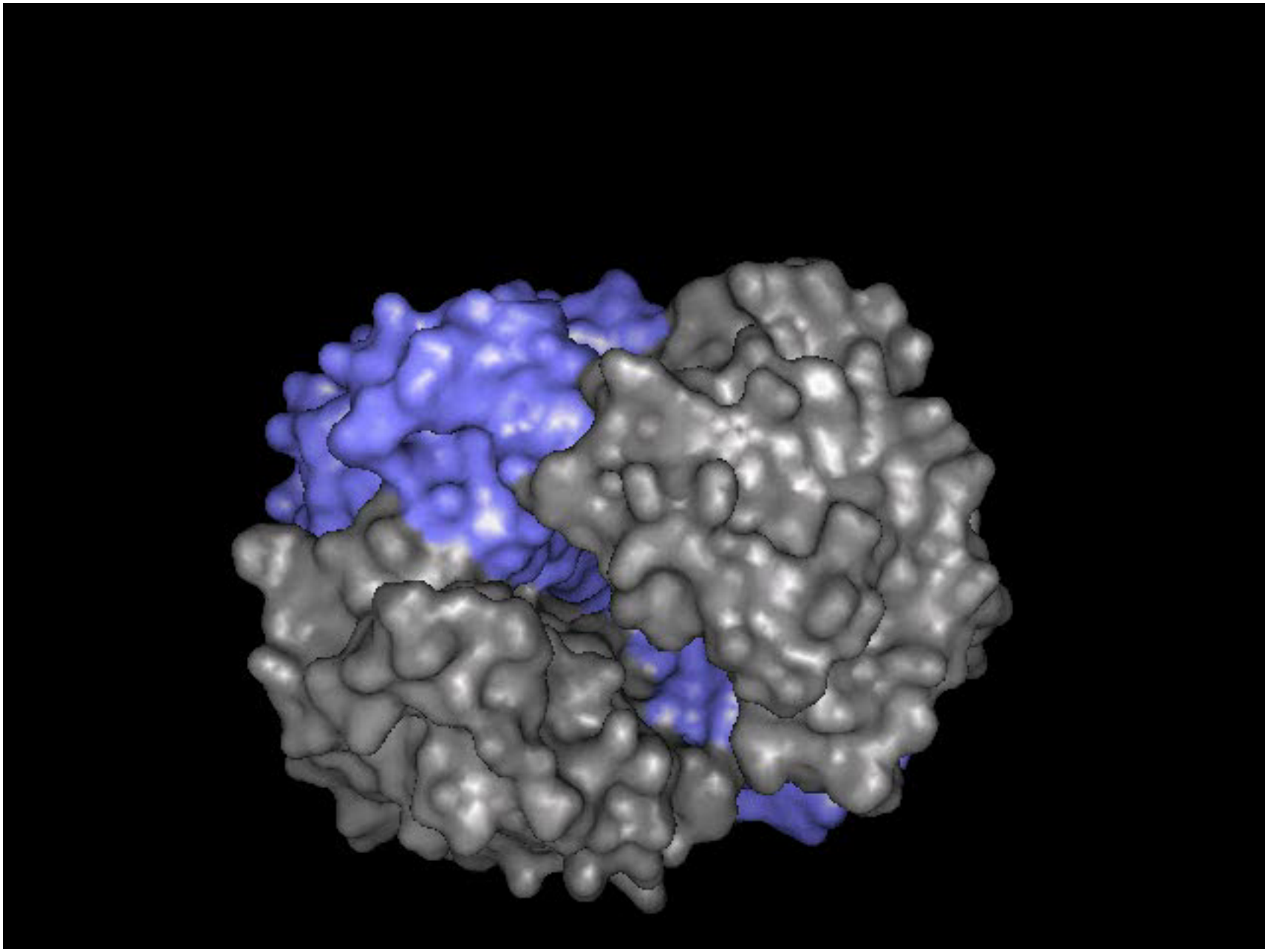
Interaction specificity

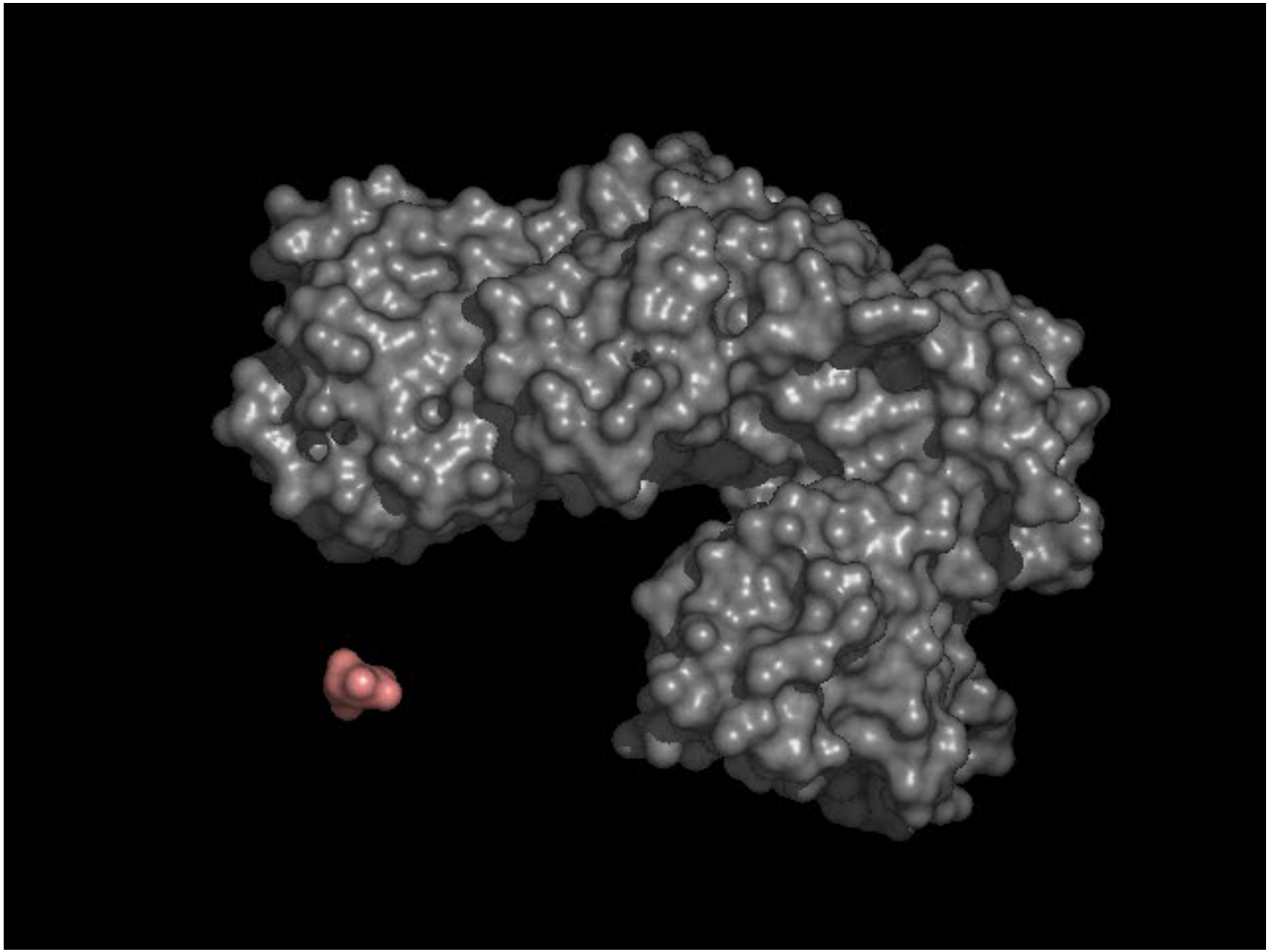
Indirect constraints from RNA function

-- how are these encoded?

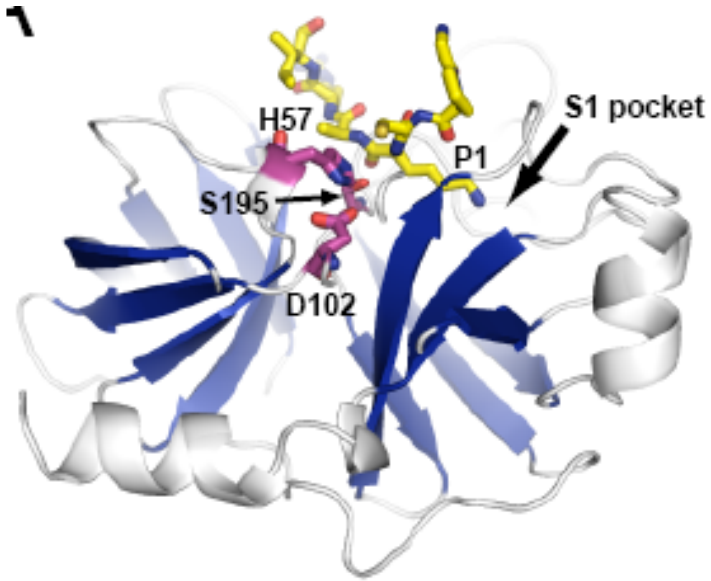
-- can we crack the code and make predictions?

Can we **invert sequence data** to reveal these constraints and so understand which sequence residues control phenotype?





Sectors – groups of residues that control distinct functional properties



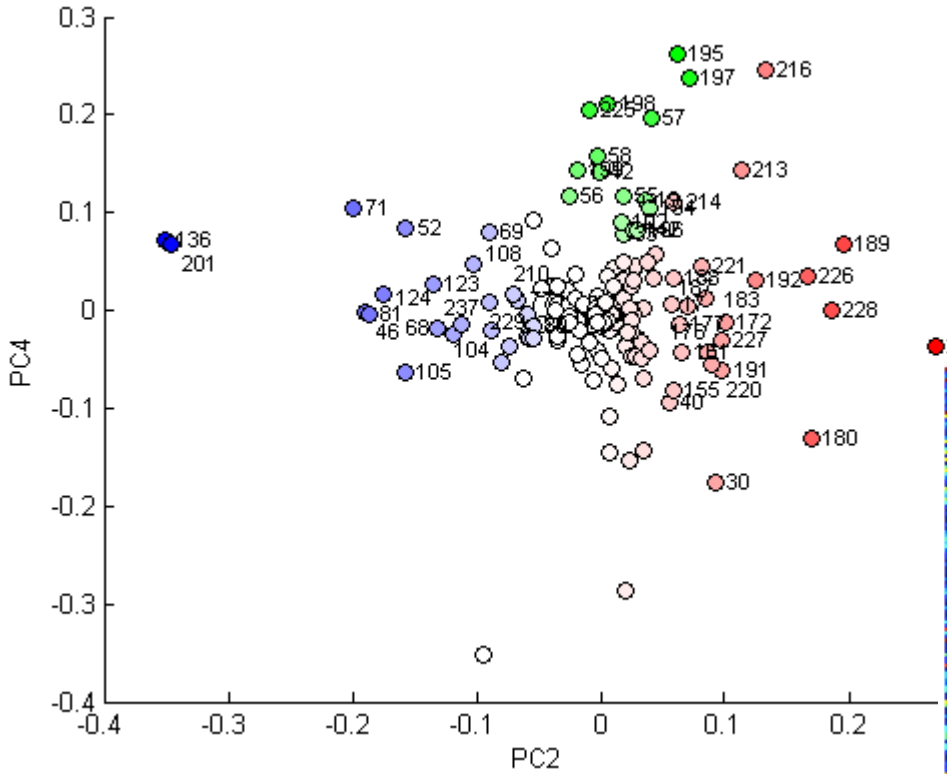
Halabi *et al* performed SCA correlated mutation analysis on an MSA of 1470 serine protease family sequences.

The eigenvectors with largest eigenvalues are localized on three groups of residues, defining 'sectors'.

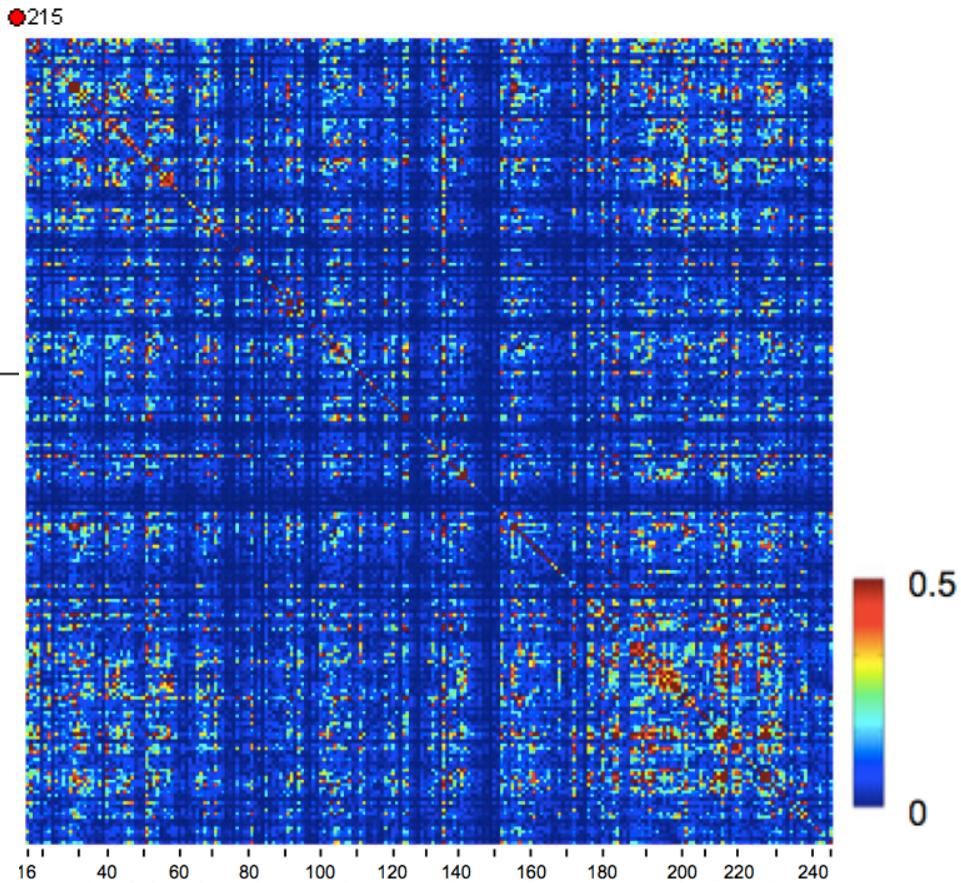
Mutational analysis found these sectors each control distinct functional properties of the proteases.

Protein sectors: functional units of tertiary structure, Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan

Serine Proteases – Sector Identification

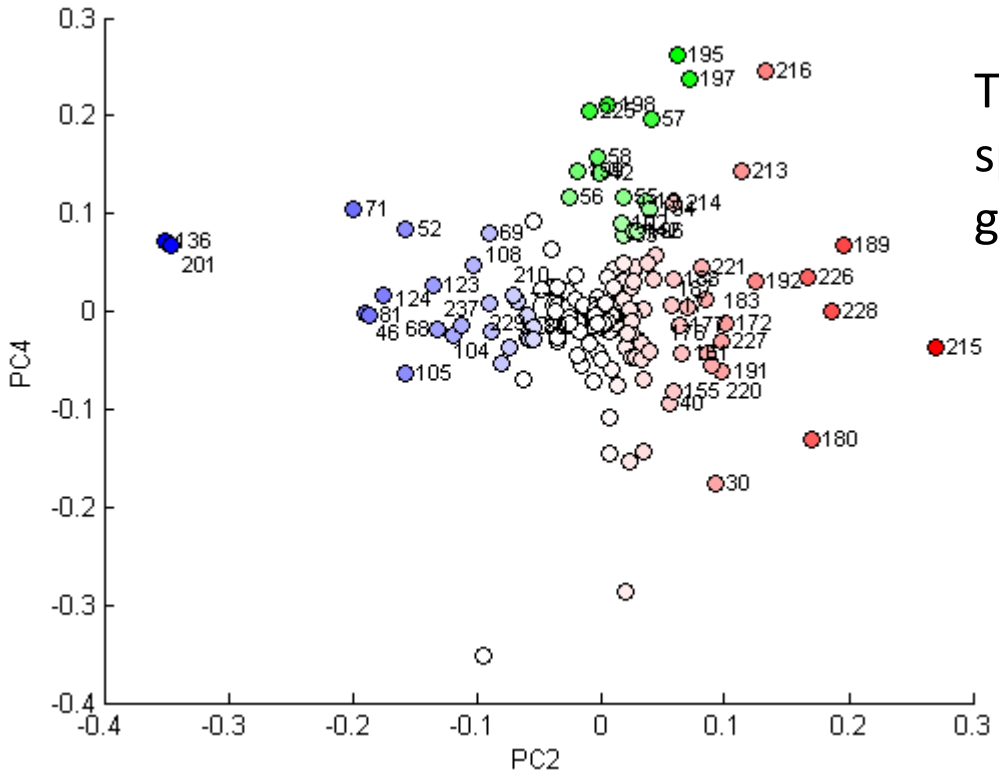


This work introduced the idea of using spectral analysis to find coevolving groups of residues.

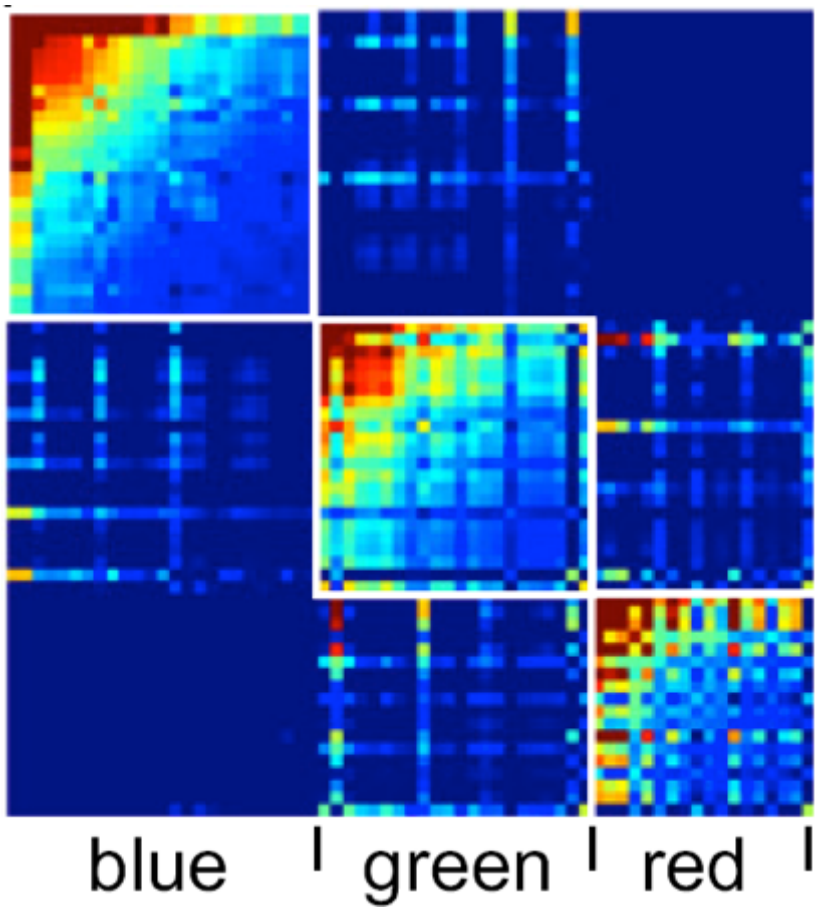


The localization of those eigenvectors with largest eigenvalues was used to identify sectors

Serine Proteases – Sector Identification

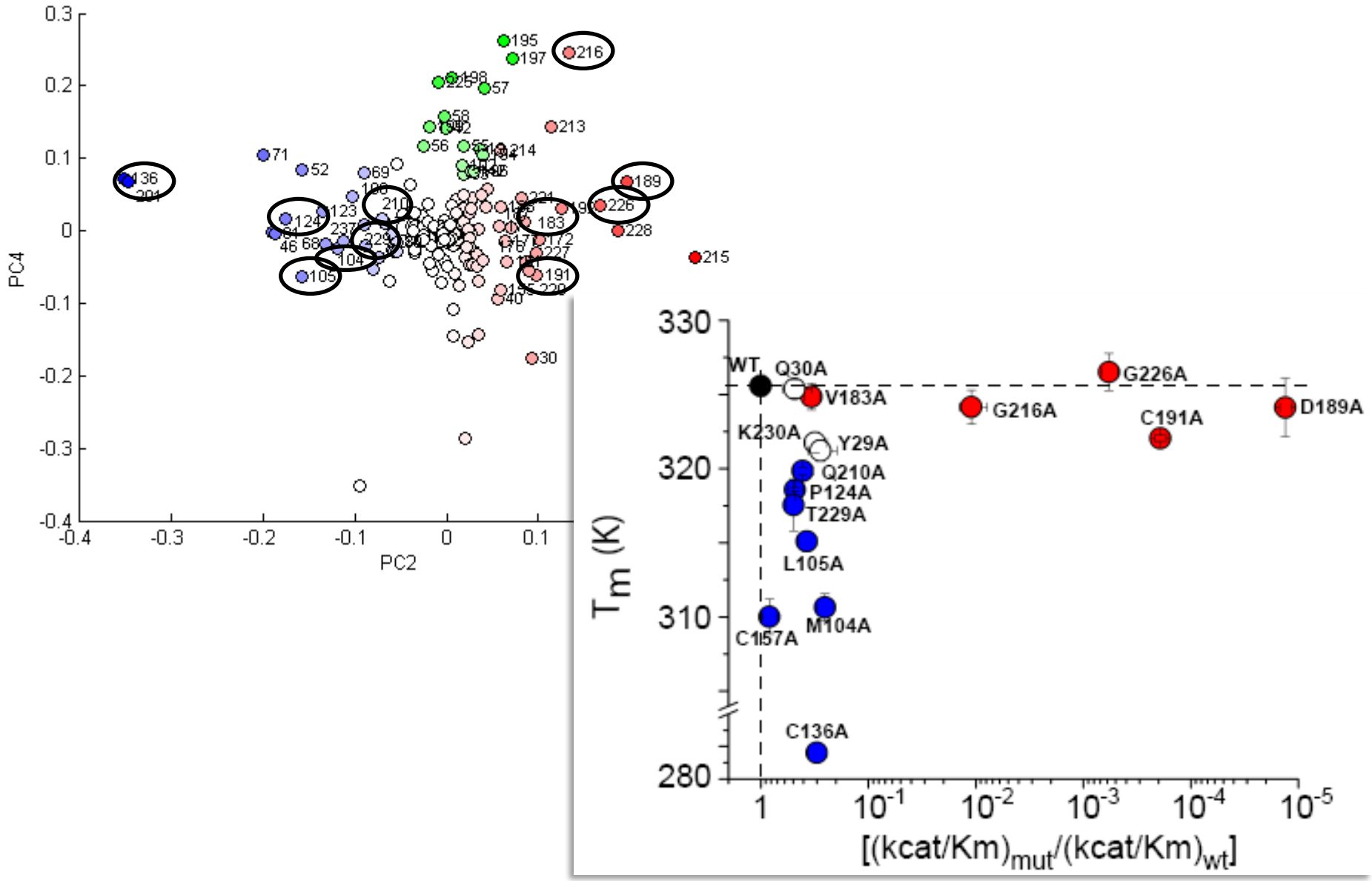


This work introduced the idea of using spectral analysis to find coevolving groups of residues.



The localization of those eigenvectors with largest eigenvalues was used to identify sectors

Serine Proteases – Experimental test of protein sectors



Outline

Biology 101 – a quick tour of proteins

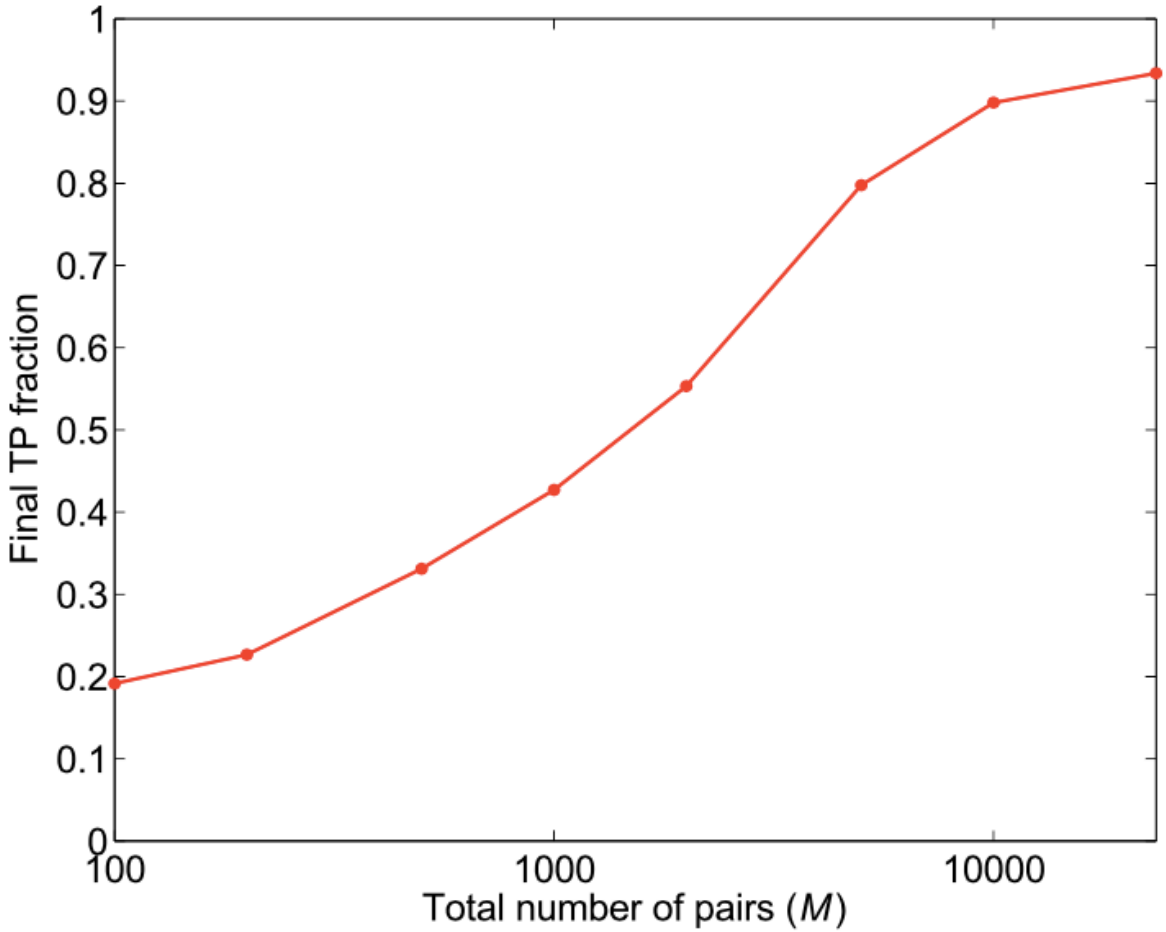
Biological information from covariance

- Protein structure and function
- Protein interaction specificity
- Ligand protein binding

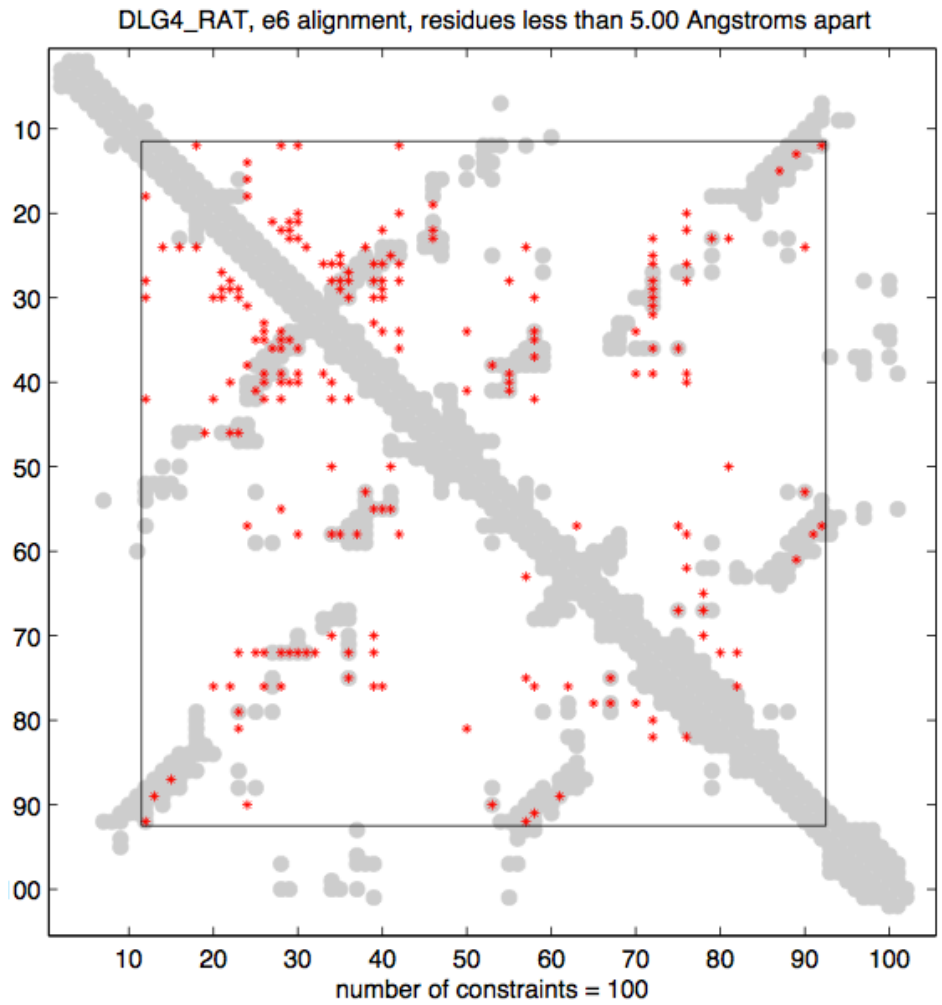
Theoretical analysis of data limitations

– when can we predict reliably?
– Noise due to finite sampling, also samples are not independent.
– central role played by Random Matrix Theory.

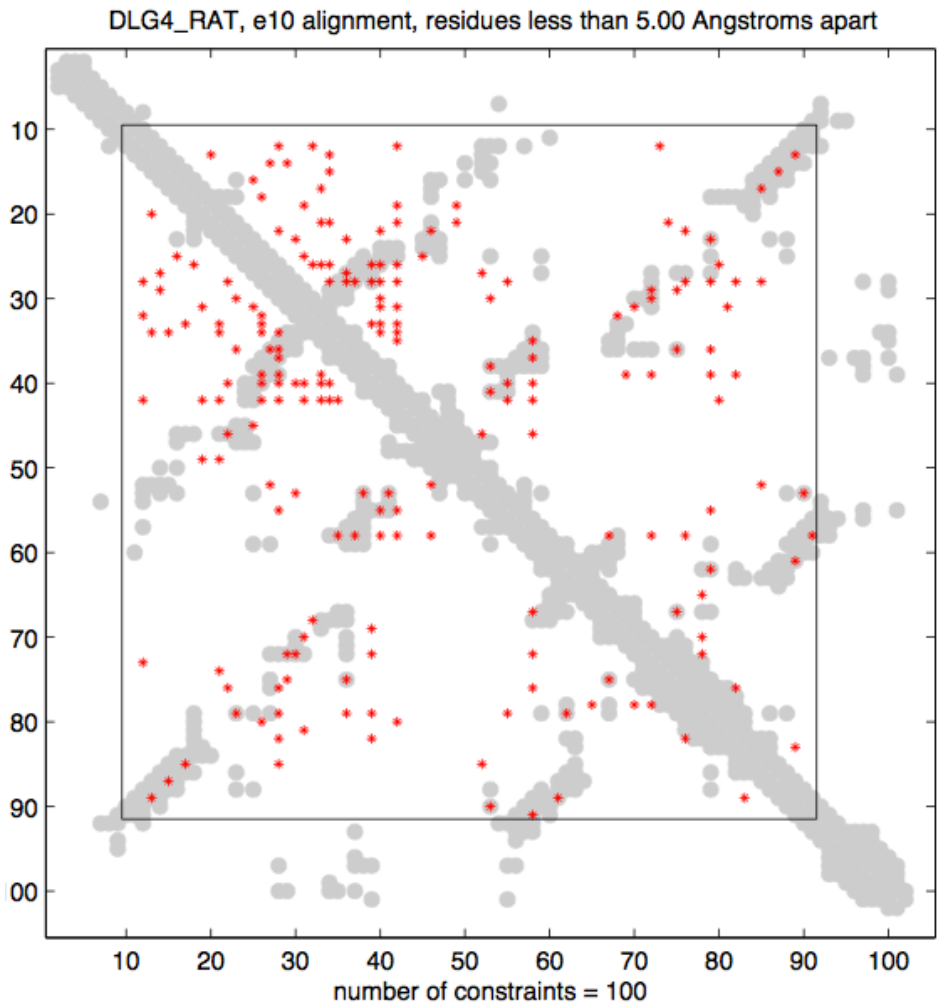
Prediction accuracy depends on the amount of input data available



Prediction accuracy depends on the input data available



14663 sequences



6724 sequences

Prediction accuracy depends on the input data available

We compute an empirical covariance matrix from a finite sample of data points.

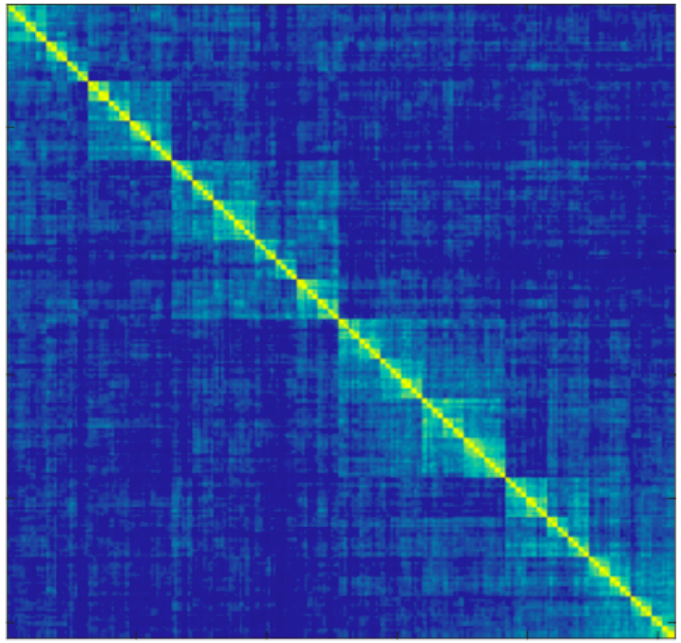
There are two major problems:

1. There will always be noise due to finite sampling.

2. Protein sequences are related by evolution, so they are not independent samples.

A natural framework to address the first issue is provided by Random Matrix Theory, and we will show how it can be extended to address the second.

Natural approach: Identify common causes (factors) that explain the joint dynamics of p variables



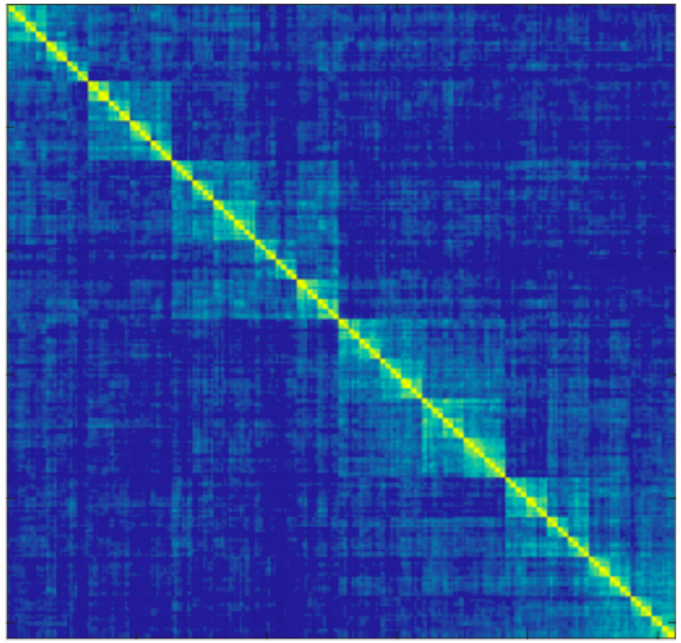
Consider a data matrix *X* which contains *N* samples of *p* variables.

Which variables interact?

Sample covariance matrix:

$$C = \frac{XX^T}{N}$$

Natural approach: Identify common causes (factors) that explain the joint dynamics of p variables



Consider a data matrix *X* which contains *N* samples of *p* variables.

Which variables interact?

Sample covariance matrix:

$$C = \frac{XX^T}{N}$$

But the data are highly undersampled – **due to finite sampling** there will be **spurious large entries.**

Problem:

These data contain multiple covariance signals – signal, batch effects, biased sampling…. that we need to de-convolve and understand. They are all obfuscated by noise due to finite sampling.

Noise due to finite sampling

Consider a set *X* of *n=300* samples, each of length *p=500* variables, drawn independently.

True covariance matrix

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Random matrix theory

Eigenvalues of sample covariance matrix

No signal – all variables independent.

Noise due to finite sampling

Consider a set *X* of *n=300* samples, each of length *p=500* variables, drawn independently.

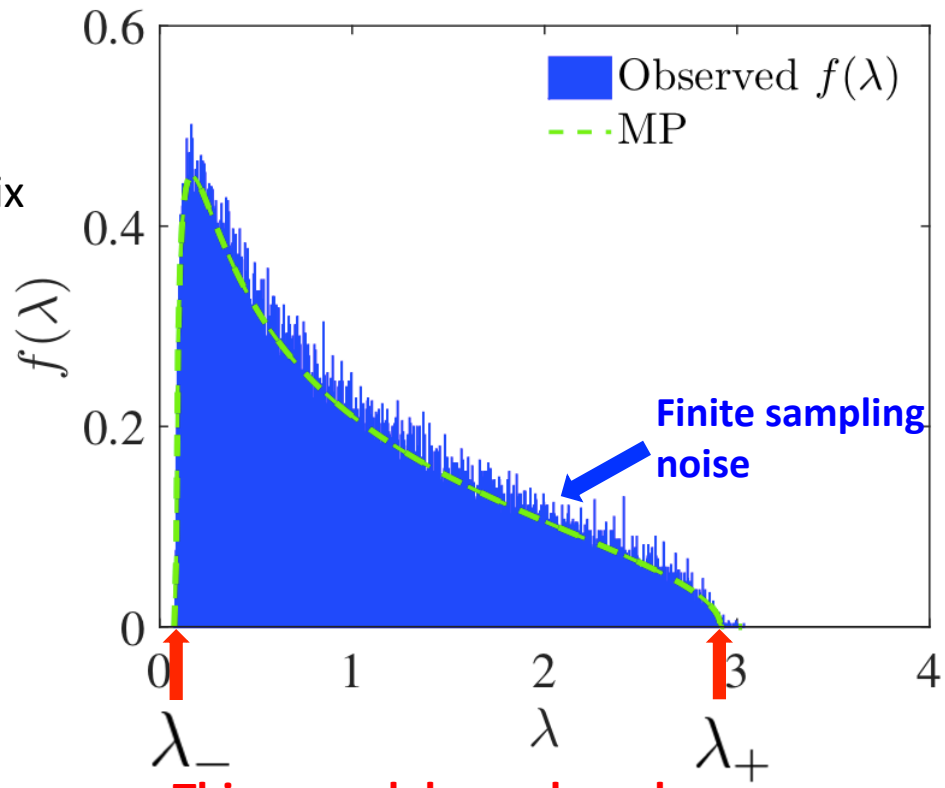
True covariance matrix

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

No signal – all variables independent.

Random matrix theory

Eigenvalues of sample covariance matrix



This spread depends on how well sampled the data is

Noise due to finite sampling

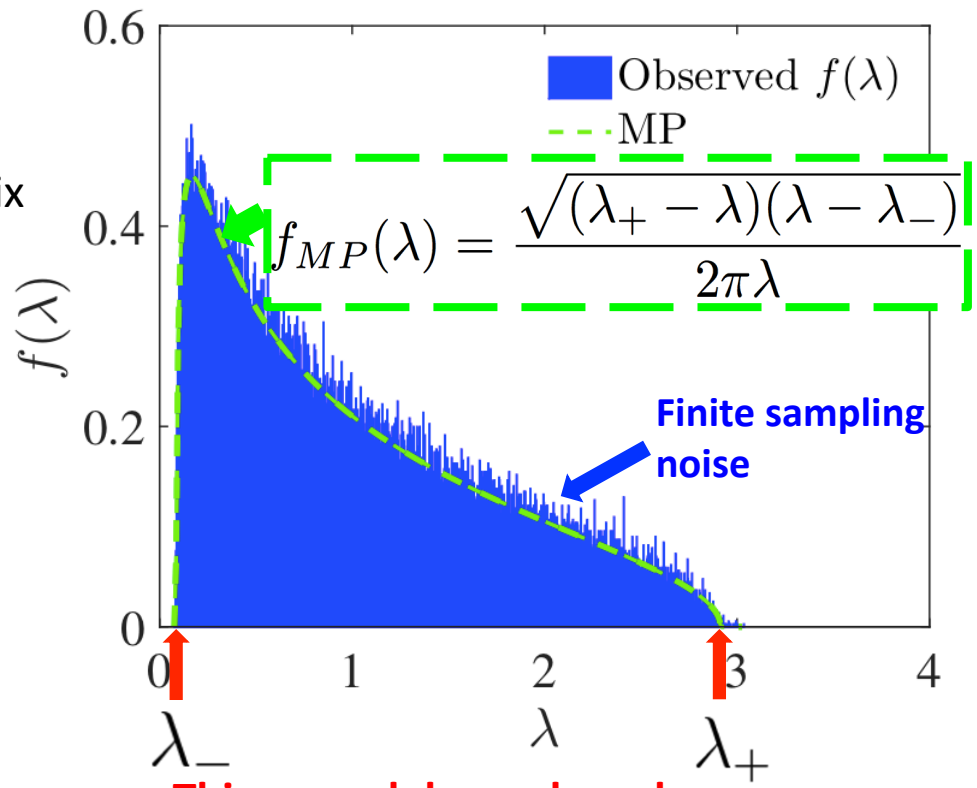
Consider a set *X* of *n=300* samples, each of length *p=500* variables, drawn independently.

True covariance matrix

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Random matrix theory

Eigenvalues of sample covariance matrix



This spread depends on how well sampled the data is

Random matrix theory:
The Marcenko-Pastur law describes the eigenvalue distribution of the sample covariance matrix of Gaussian IID noise.

Outline

Biology 101 – a quick tour of proteins

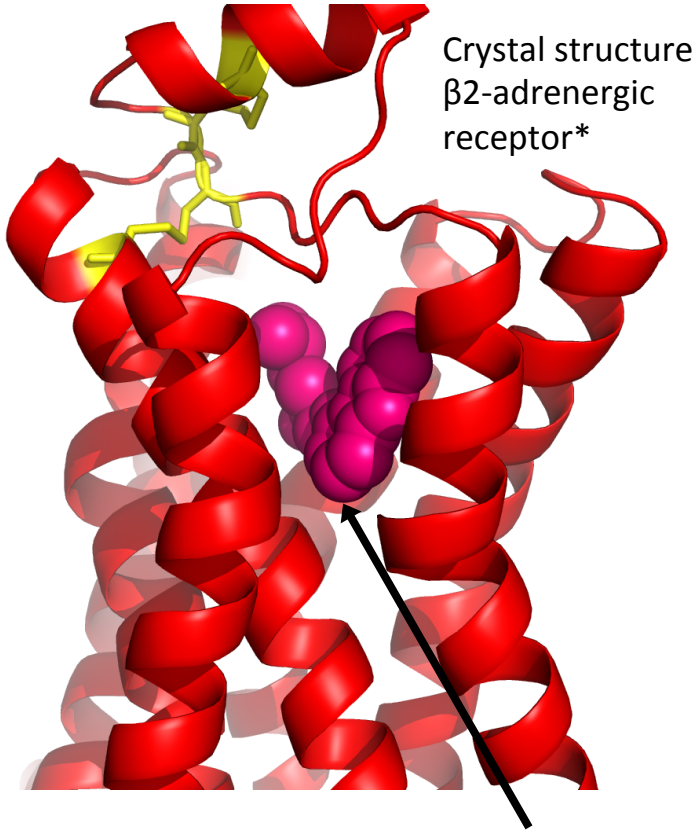
Biological information from covariance

- Protein structure and function
- Protein interaction specificity
- Ligand protein binding

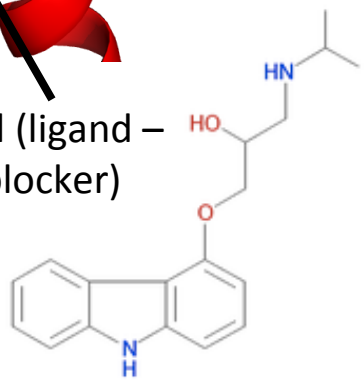
Theoretical analysis of data limitations

– when can we predict reliably?
– Noise due to finite sampling, also samples are not independent.
– central role played by Random Matrix Theory.

Example: protein-ligand binding

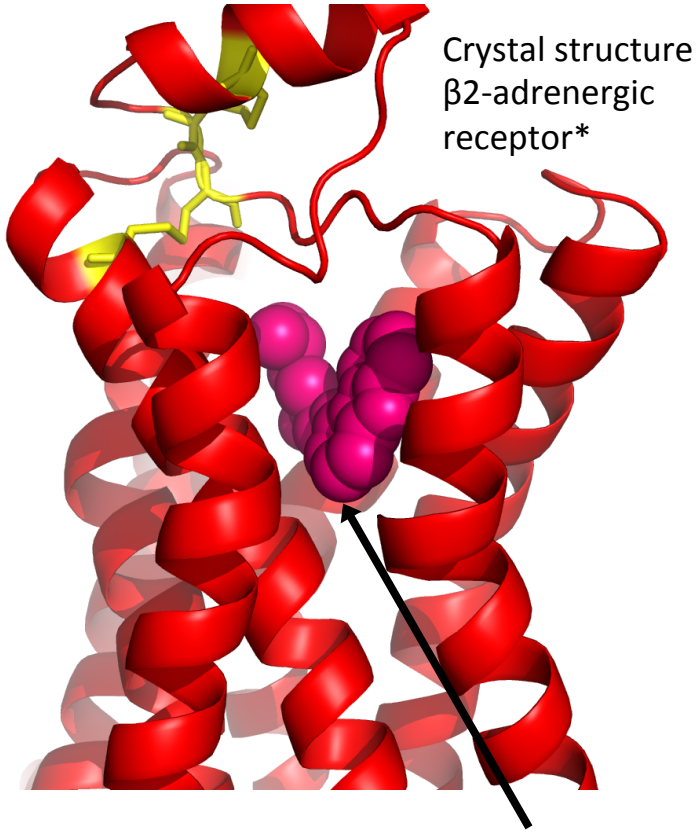


carazolol (ligand – beta blocker)



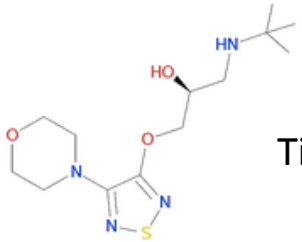
*Crystal Structure ADRB2: Rasmussen SG, DeVree BT, Zou Y, Kruse AC, Chung KY, Kobilka TS, Thian FS, Chae PS, Pardon E, Calinski D, Mathiesen JM, Shah ST, Lyons JA, Caffrey M, Gellman SH, Steyaert J, Skiniotis G, Weis WI, Sunahara RK, Kobilka BK. Nature. 2011 Jul 19.

Example: protein-ligand binding

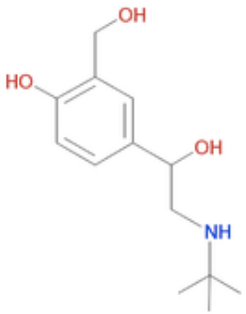


Crystal structure β2-adrenergic receptor*

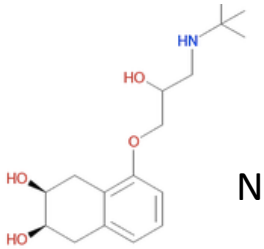
Many known ligands



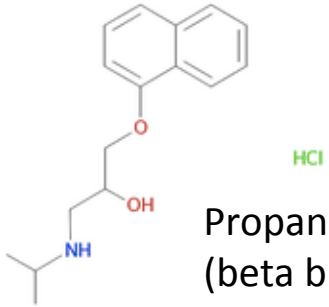
Timolol



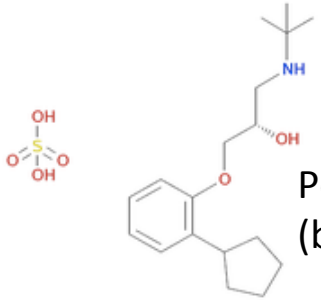
Albuterol



Nadalol

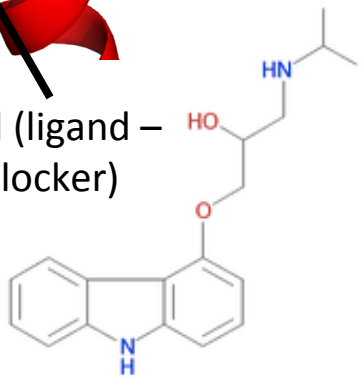


Propanolol Hcl (beta blocker)



Penbutolol Sulfate (beta blocker)

carazolol (ligand – beta blocker)

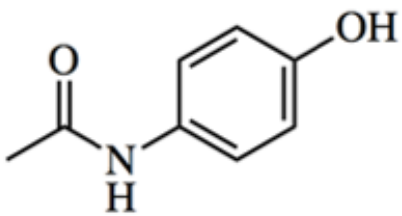


What other ligands will bind?

*Crystal Structure ADRB2: Rasmussen SG, DeVree BT, Zou Y, Kruse AC, Chung KY, Kobilka TS, Thian FS, Chae PS, Pardon E, Calinski D, Mathiesen JM, Shah ST, Lyons JA, Caffrey M, Gellman SH, Steyaert J, Skiniotis G, Weis WI, Sunahara RK, Kobilka BK. Nature. 2011 Jul 19.

Use molecular descriptors - fingerprints

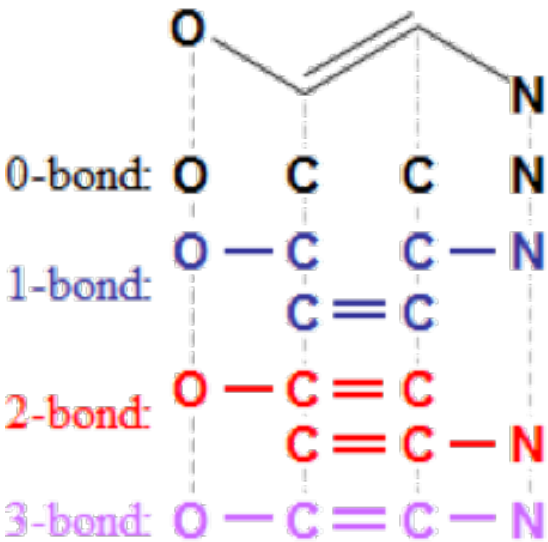
- Representation of a molecule as a string of bits (‘0’ and ‘1’)
- Usually the 2D structure, i.e. the molecular graph is used to generate the string.
- A specific chemical moiety gives rise to a pattern of bits



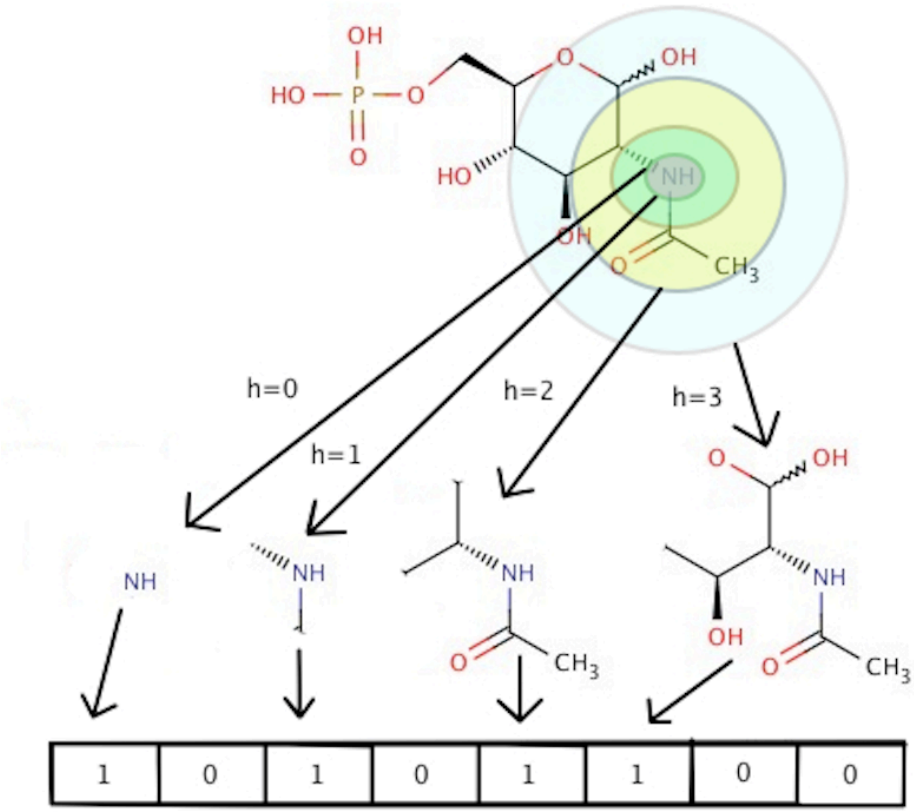
e.g.Morgan 3

0000000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000010000000000000000000000000
0000000000000000000010000000000000000000000000000000000000010000000000000001000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000010000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000010000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000010000000000000000000000000000
0000000000000100000000000000000000000000000000000000000000000010000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000001
0000000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000100
0000000000000000000000000000000000000010000000000000000000010000000000000000000000000
0000000000000000000000000000000000000000000000000010000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000010000000000000000000000000000000000000000000
0000000000000000000010000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000001000000000000
0000000000000000000000000000000000000000000000000010000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000
0001000000000000000000000000000000000000000000000010000000000000000000000000000000000
0000000000000000000000000000000000000000000000000000000000000000000000000000000000000
00000000

Linear, e.g. Daylight



Circular, e.g. Morgan 3, ECFP



What information is contained in the covariance matrix of ligand fingerprints?

Consider 1000 random molecules from Chembl

http://www.rsi.co.jp/kagaku/cs/daylight/fingerprint.html
https://chembioinfo.com/2011/10/30/revisiting-molecular-hashed-fingerprints/

Random matrix theory and covariance

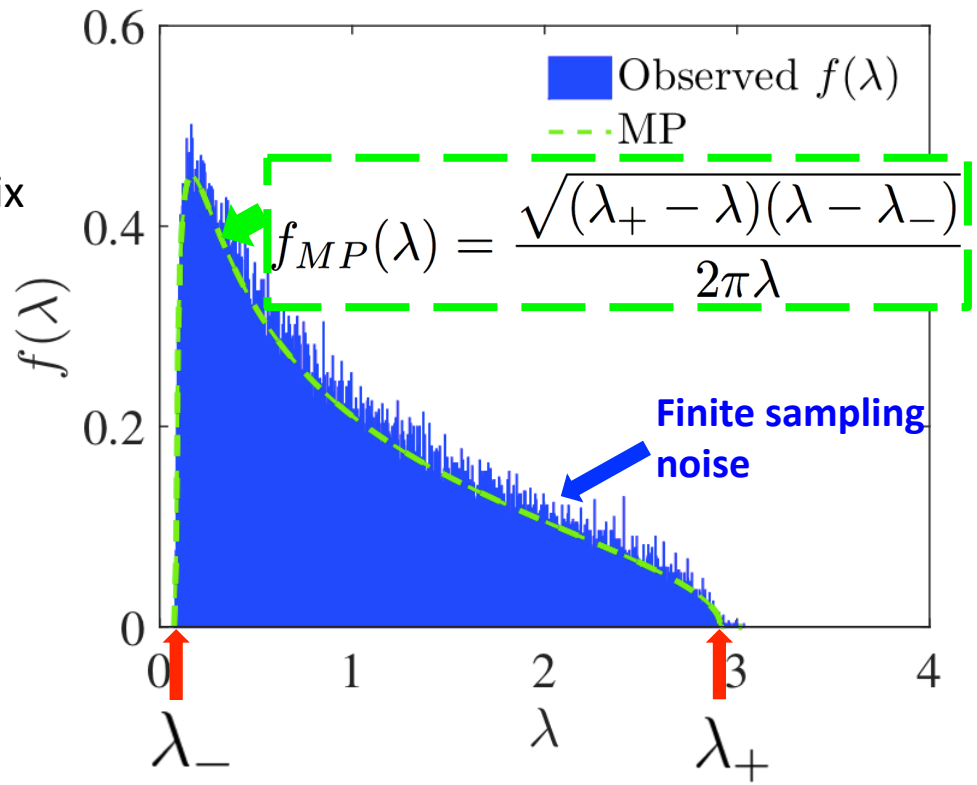
Consider a set *X* of *n=300* samples, each of length *p=500* variables, drawn independently.

True covariance matrix

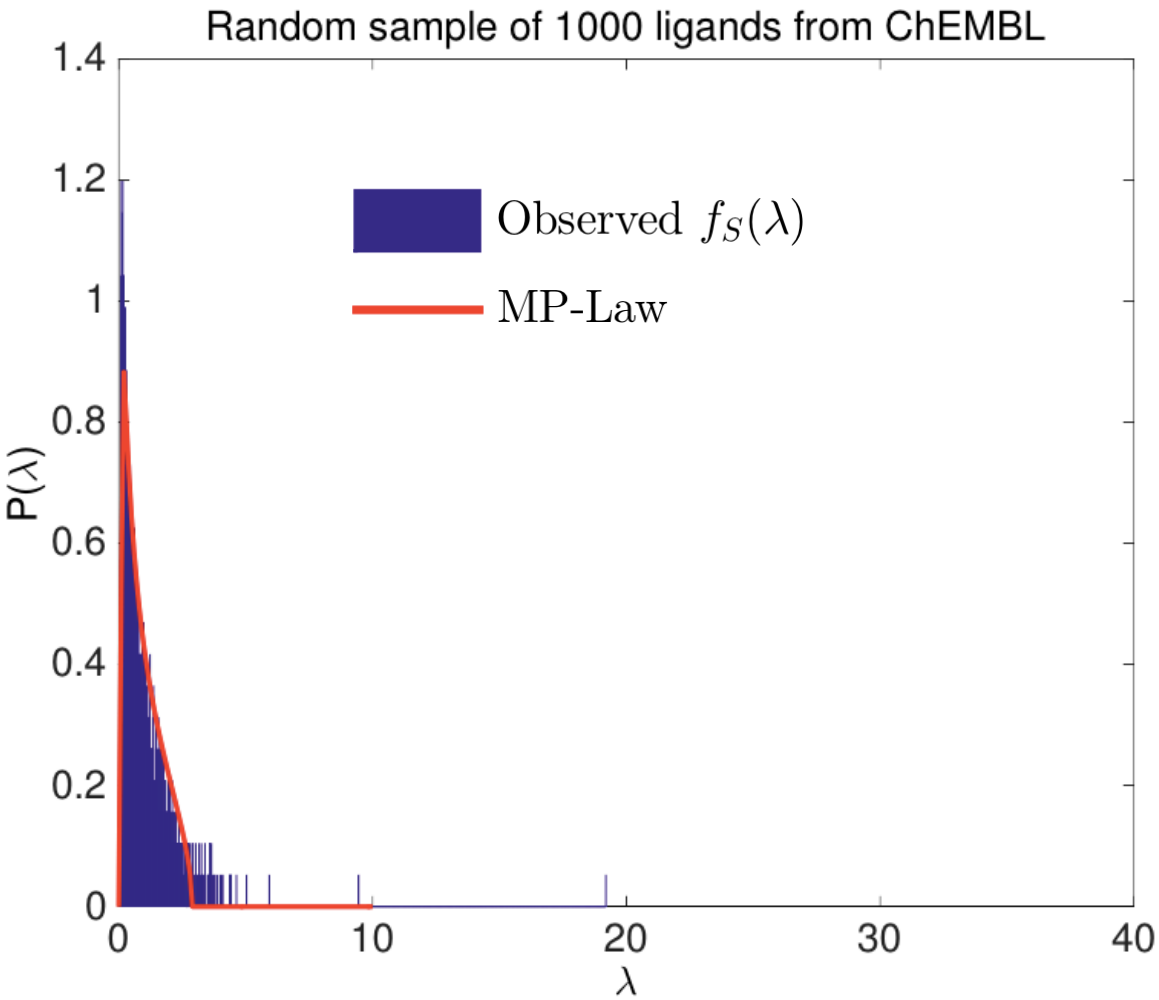
$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}$$

Random matrix theory

Eigenvalues of sample covariance matrix



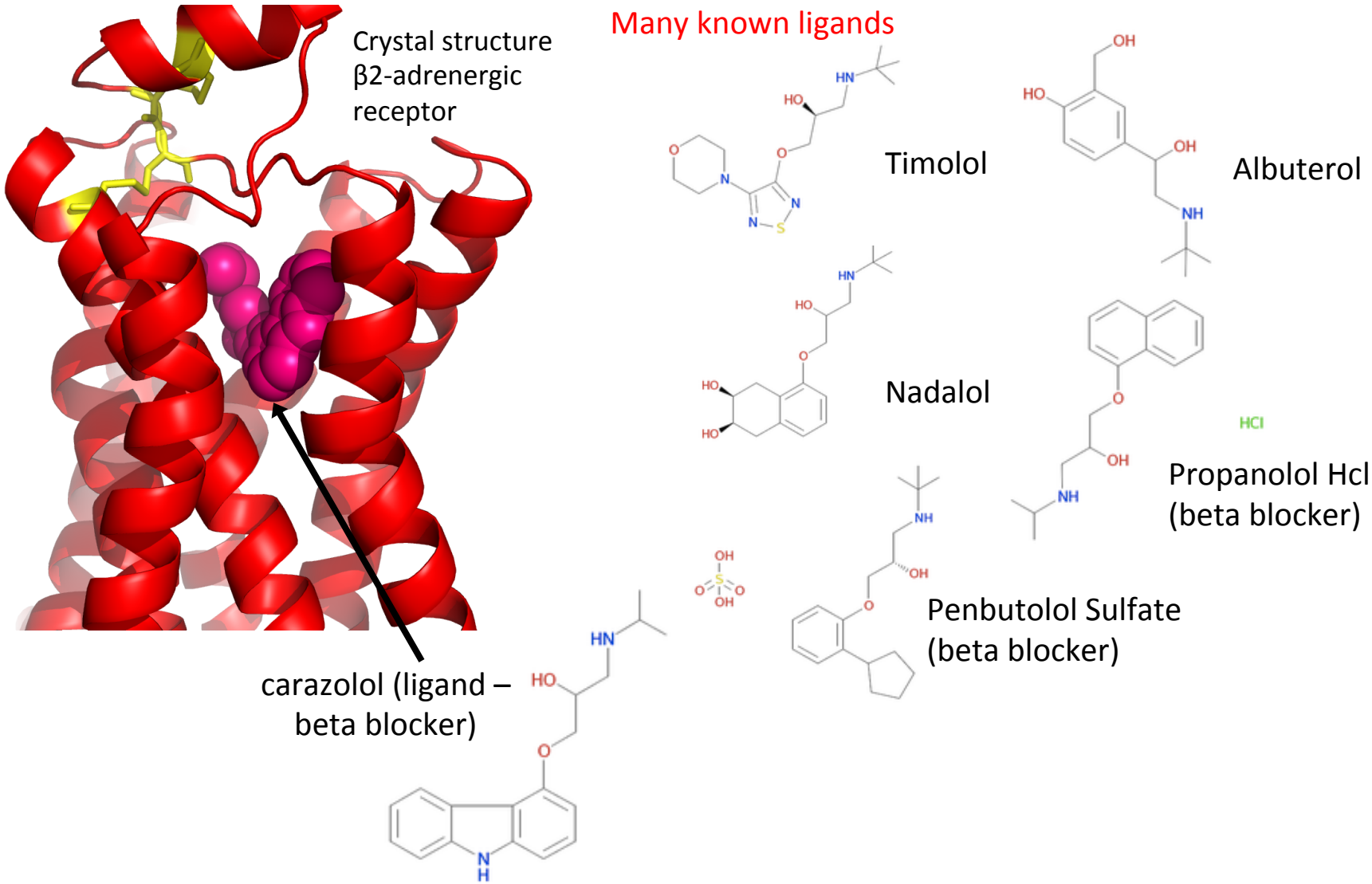
Random matrix theory:
The Marcenko-Pastur law describes the eigenvalue distribution of the sample covariance matrix of Gaussian IID noise.



These molecules are independent samples and follow the Marcenko-Pastur distribution

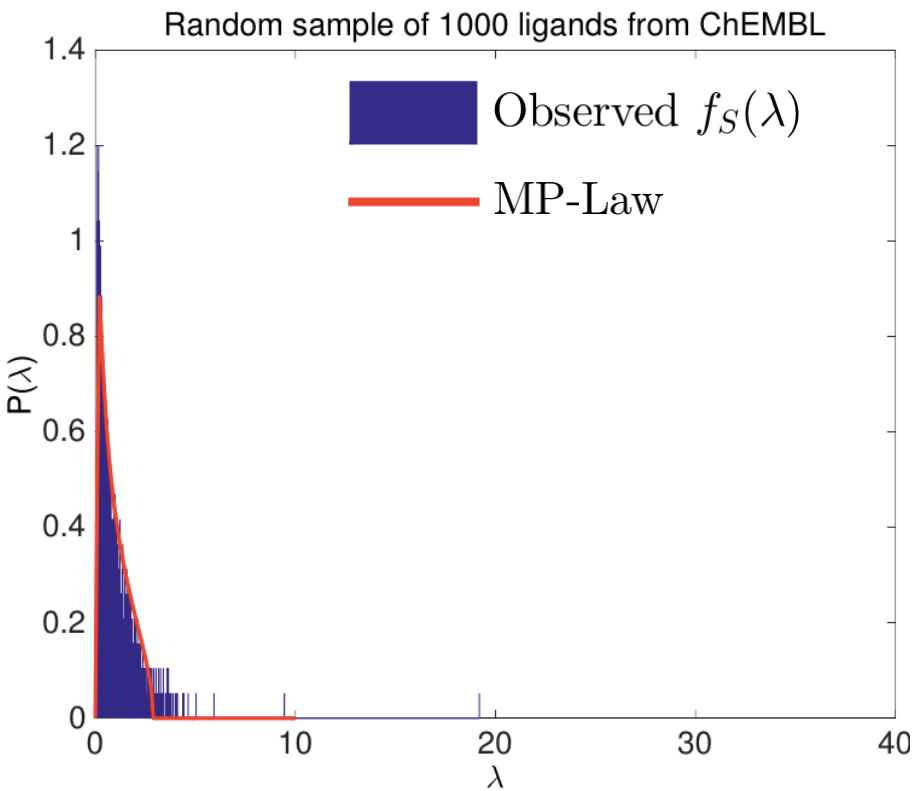
$$f_{MP}(\lambda) = \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{2\pi\lambda}$$

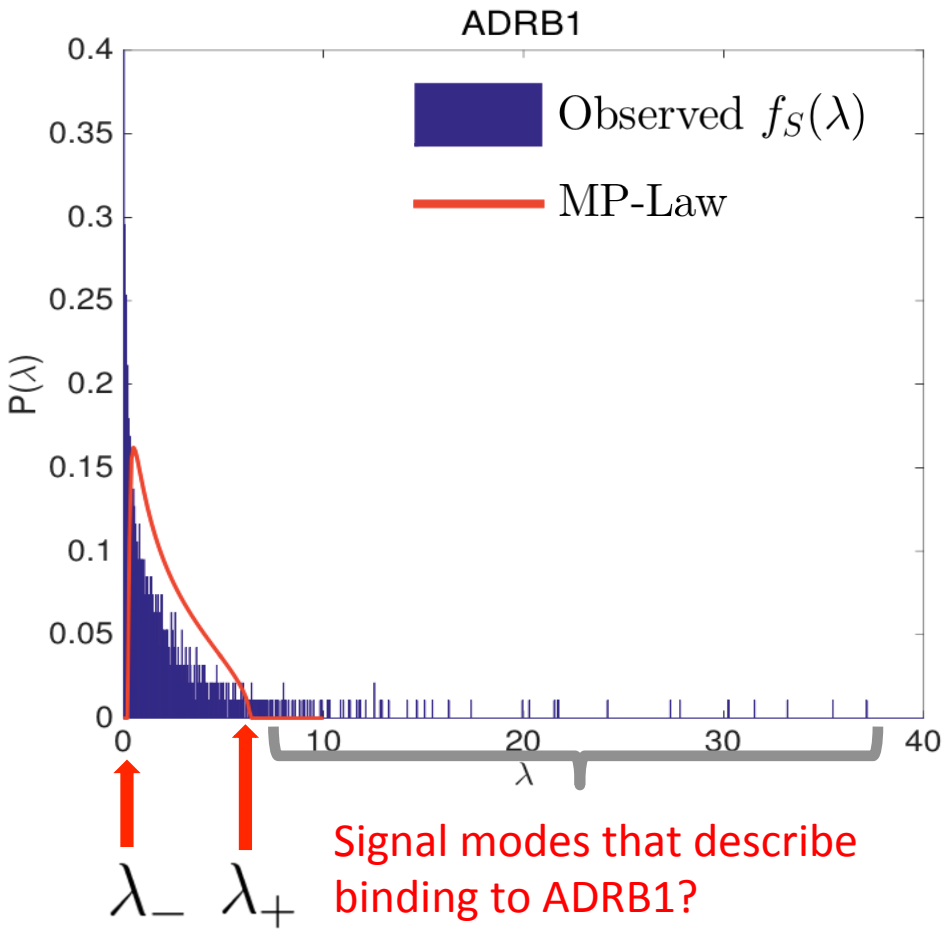
Lee *et al*, PNAS 2016



What about the covariance matrix of ligands that all bind the same protein receptor?

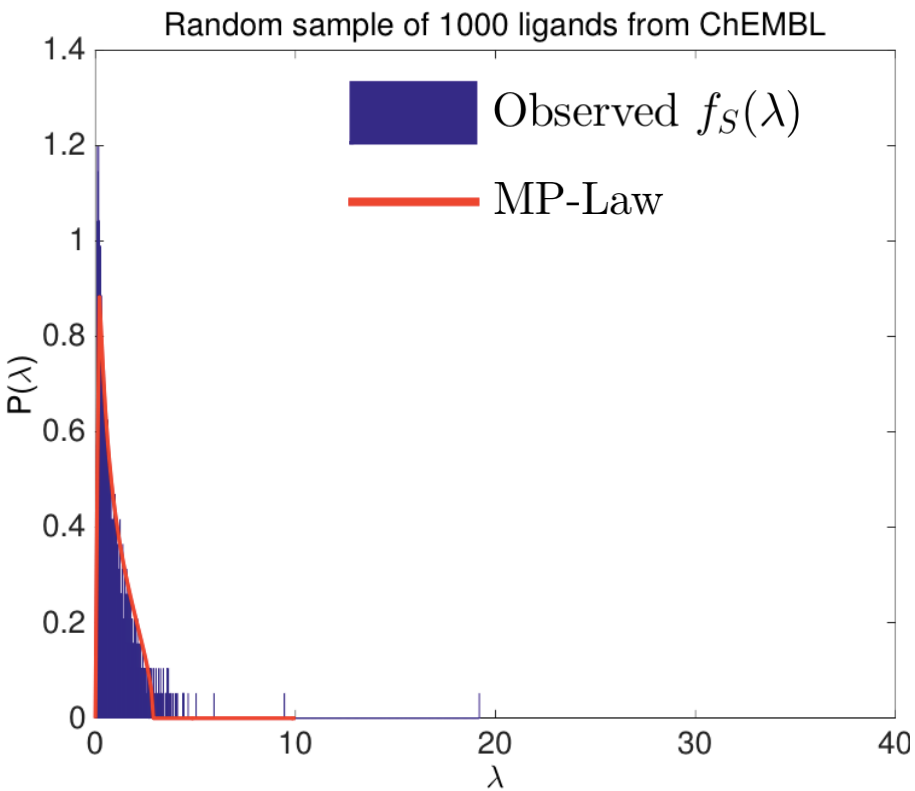
Compare ligands that bind ADRB1 with random molecules

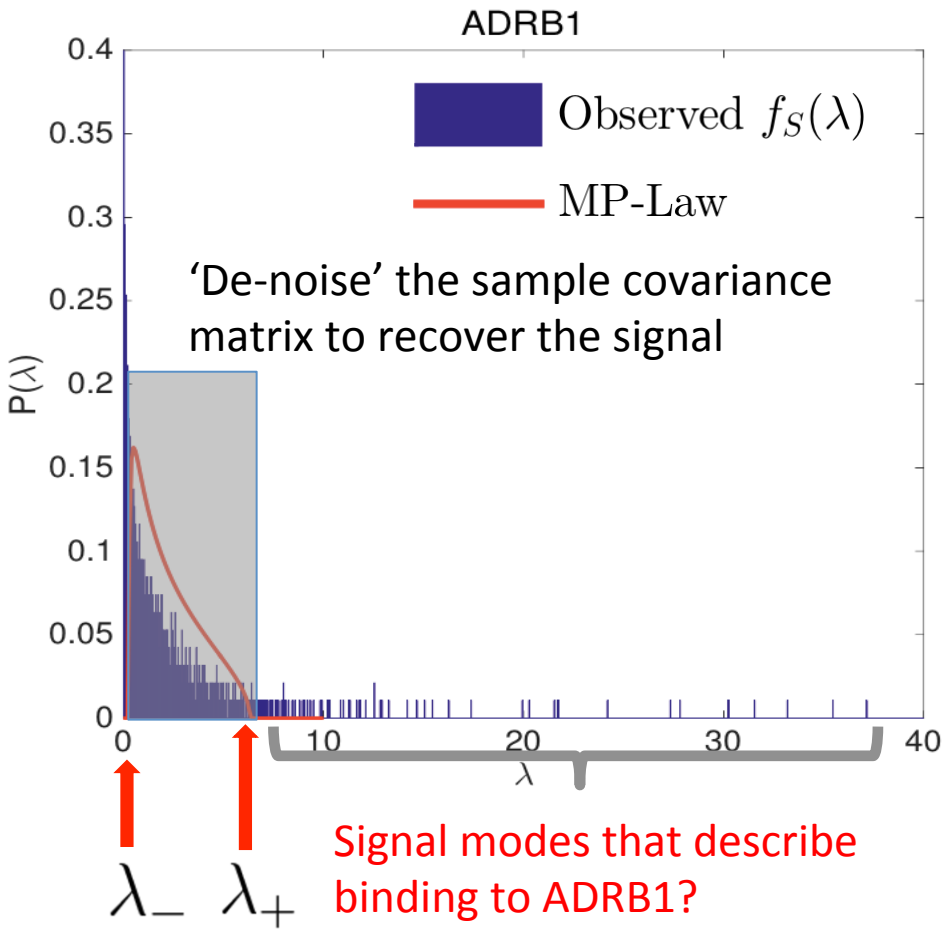




Lee *et al*, PNAS 2016

Compare ligands that bind ADRB1 with random molecules





Lee *et al*, PNAS 2016

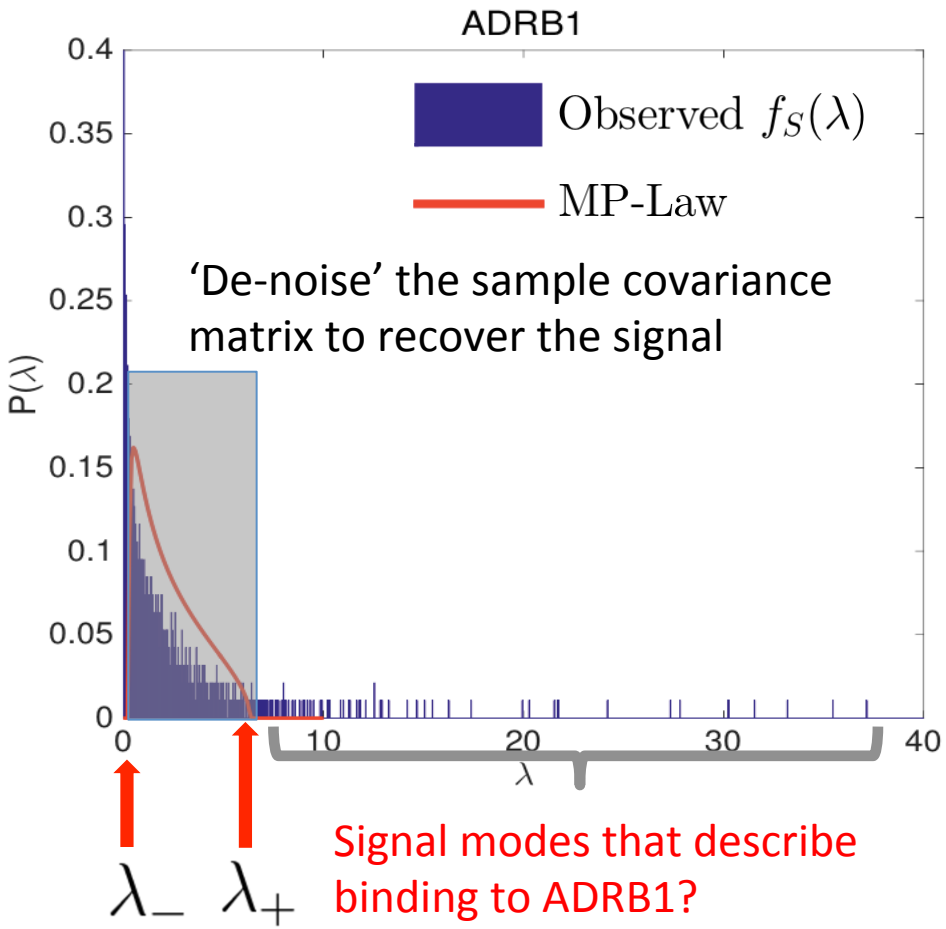
Different truncations yield different models:

$$\mathbf{u}_p = \sum_{i=1}^{m} (\mathbf{v}_i \cdot \mathbf{u})\mathbf{v}_i$$

Here we truncate the spectrum to the space spanned by the largest *m* modes, and ask whether a novel ligand **u** lies in this subspace.

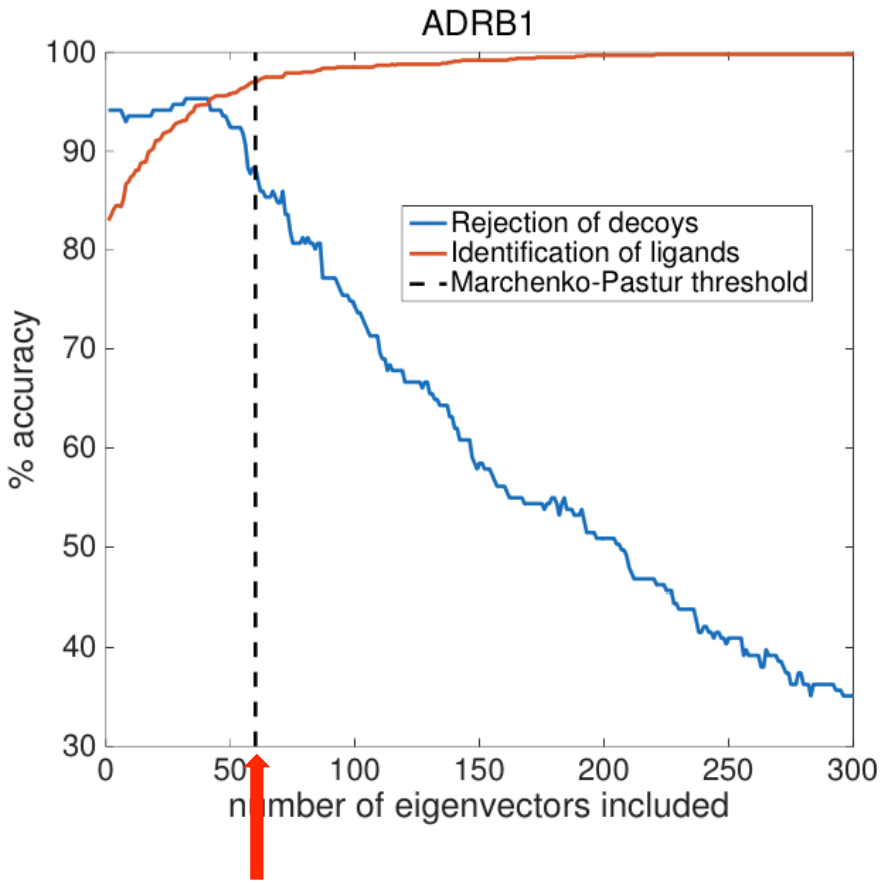
$$\left\| \mathbf{u} - \mathbf{u}_p \right\| < \epsilon$$

This provides a model that predicts whether novel ligands will bind to the ADRB1 receptor or not.

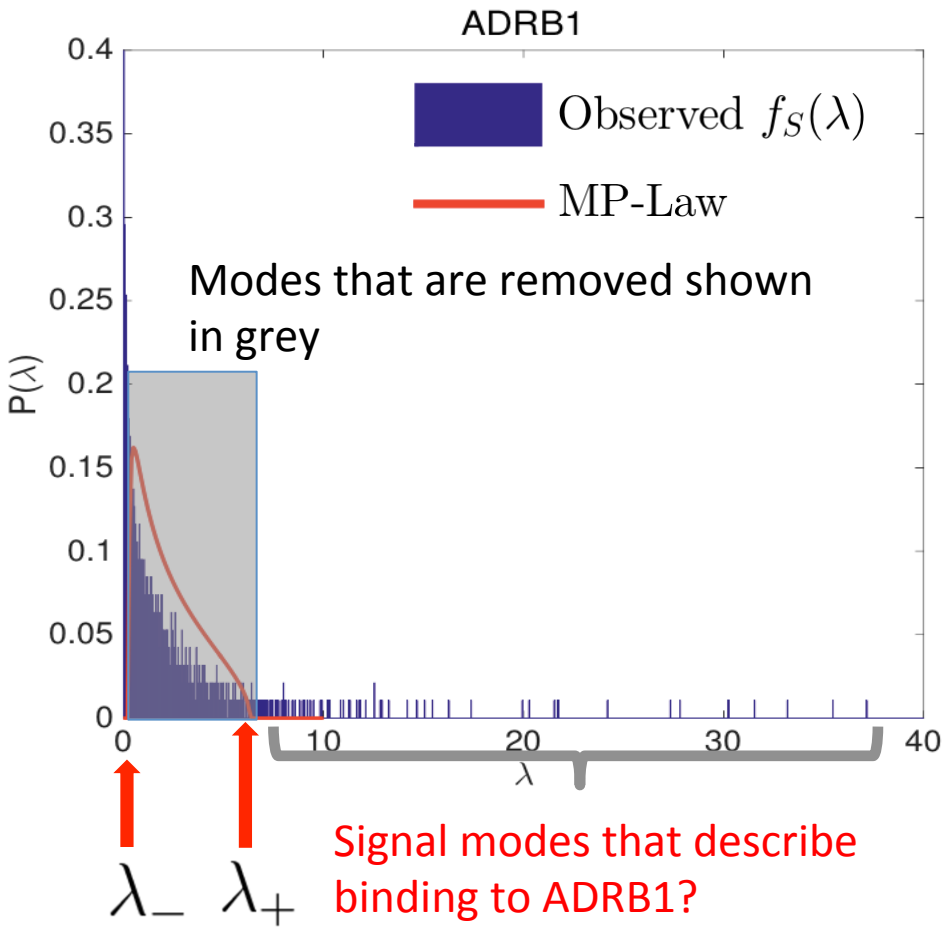


Lee *et al*, PNAS 2016

Test the performance of different truncations



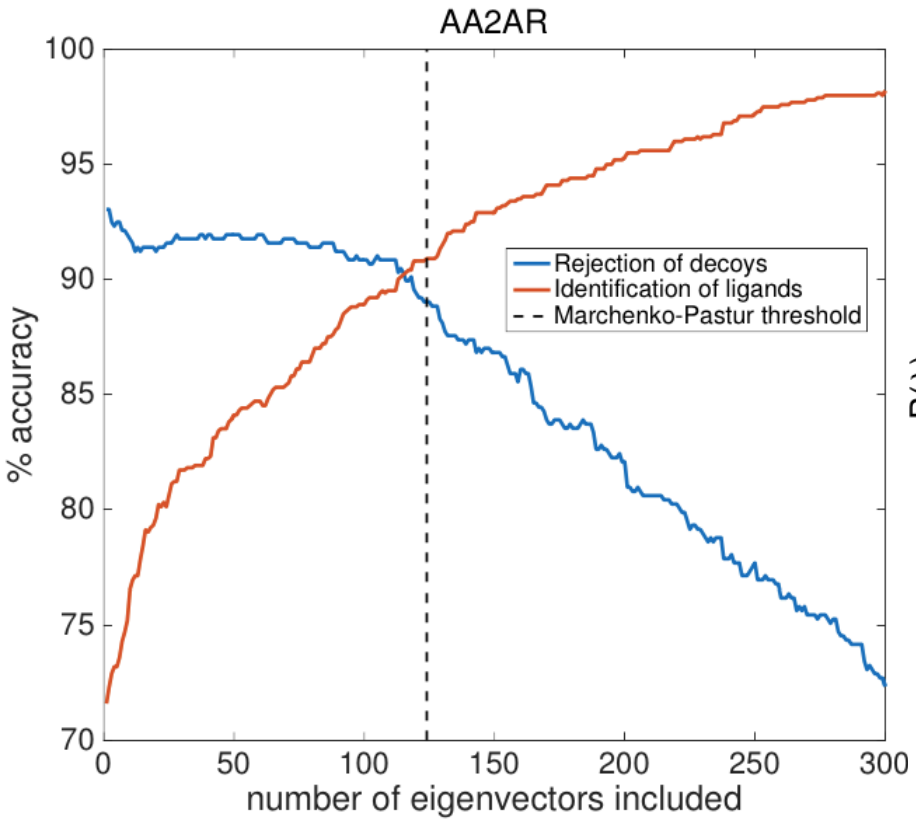
Marcenko-Pastur threshold

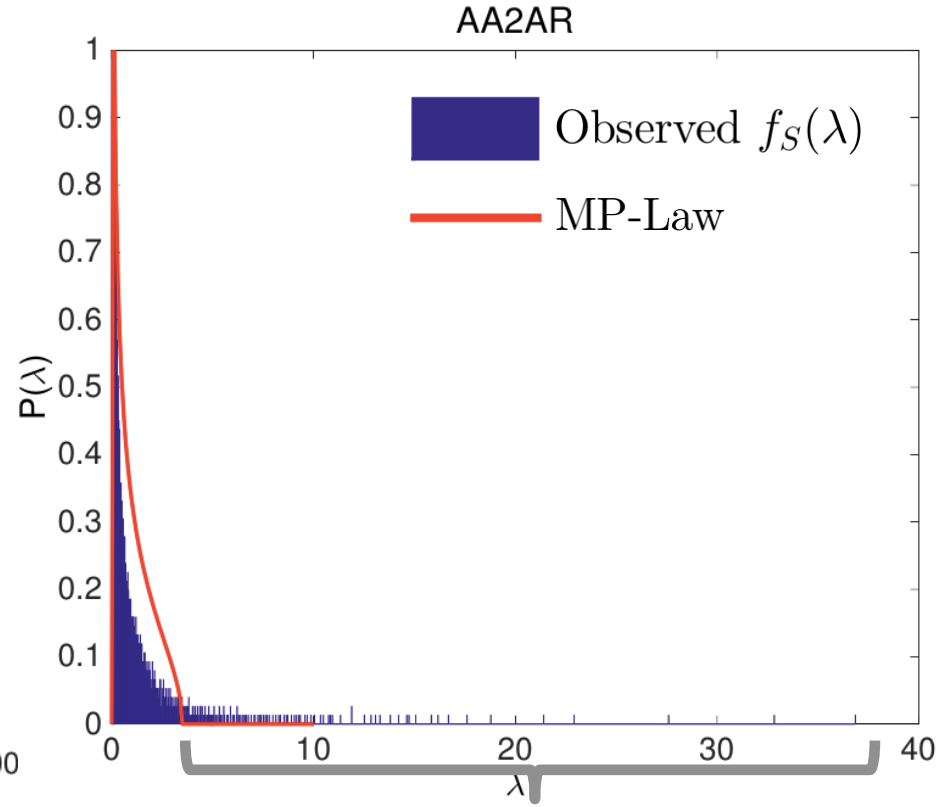


Signal modes that describe binding to ADRB1?

Lee *et al*, PNAS 2016

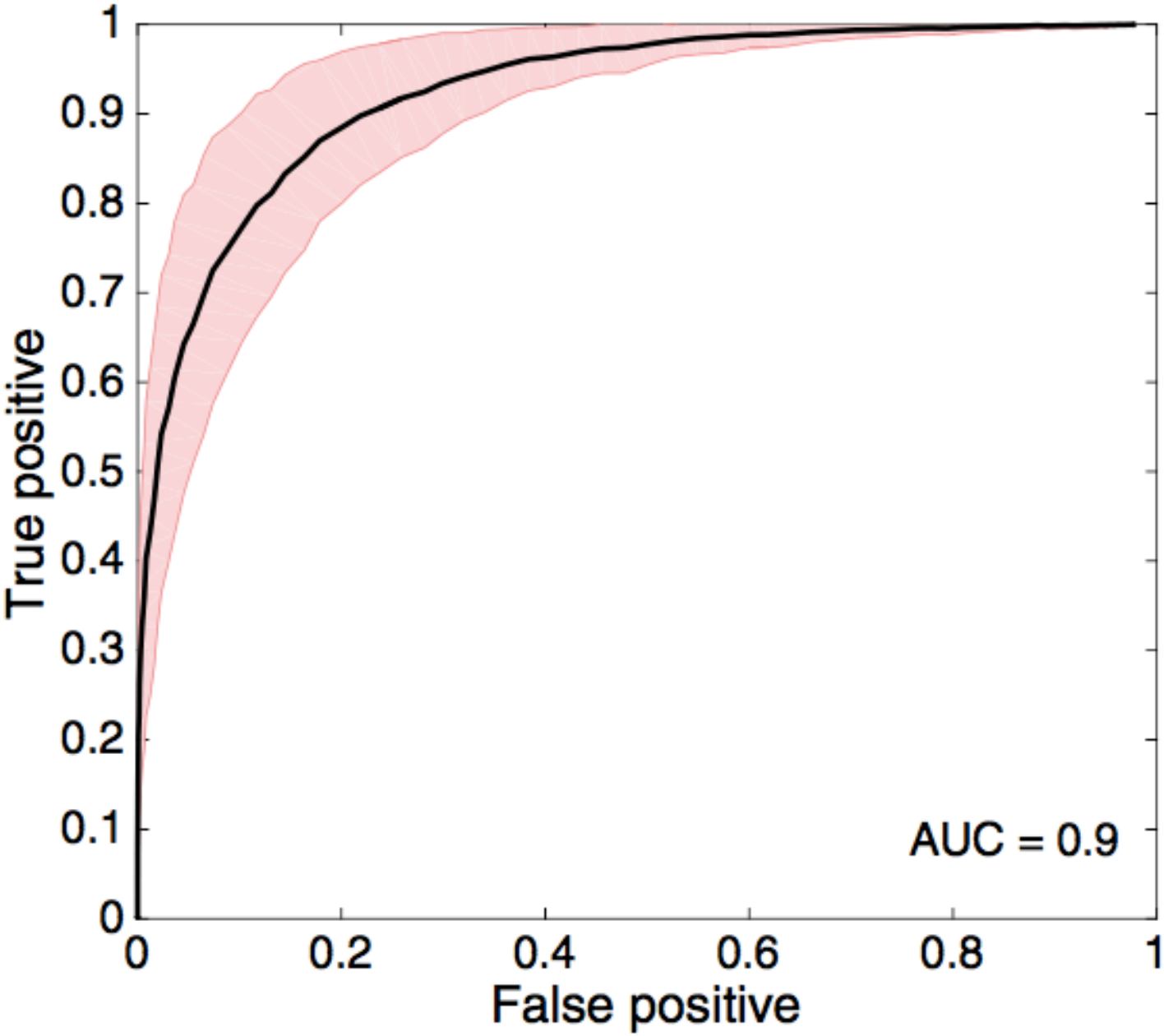
Another example – adenosine receptor





Signal modes that describe binding to AA2AR?

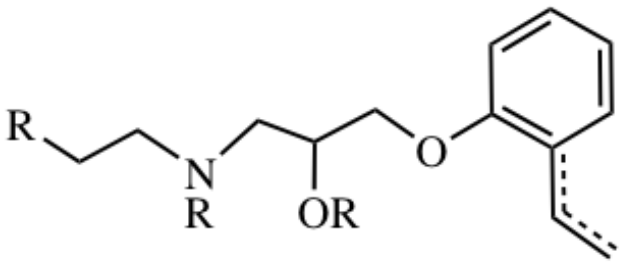
Lee *et al*, PNAS 2016



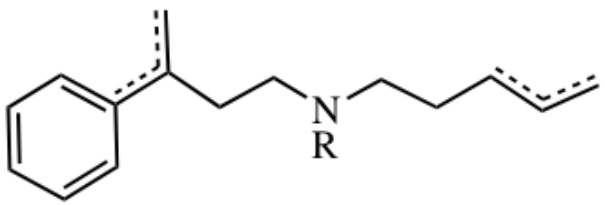
Lee *et al*, PNAS 2016

Attribution - what is the model learning?

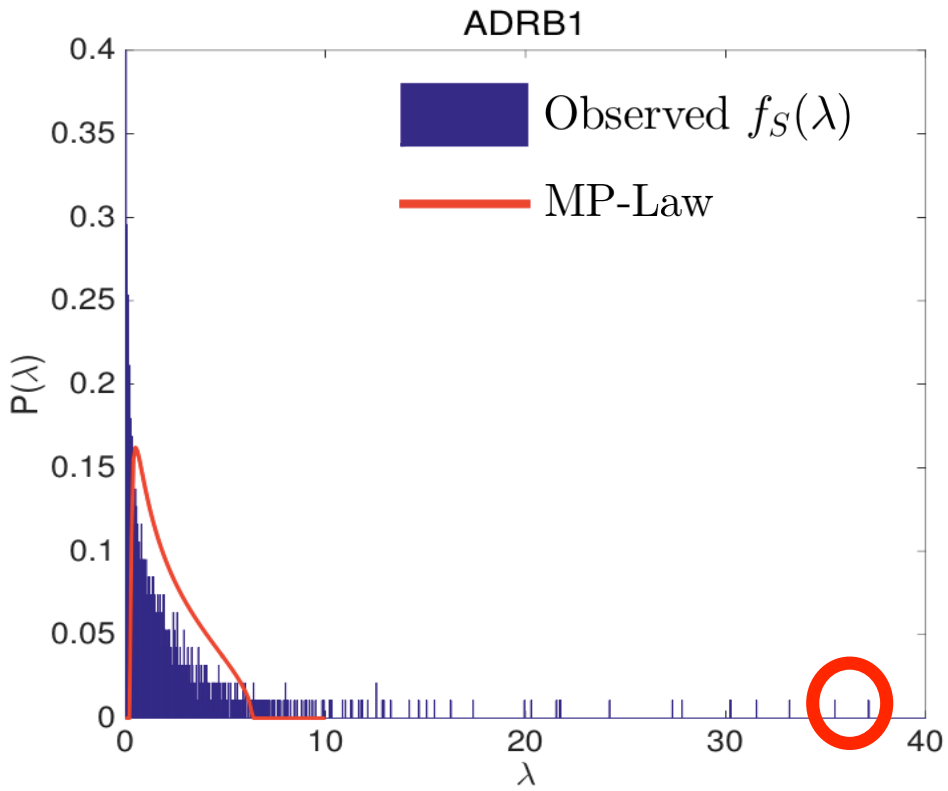
Examine largest two modes of the covariance matrix - infer the structural motif that an eigenvector represents by the common substructure among those ligands that lie closest to that eigenvector:



First eigenvector



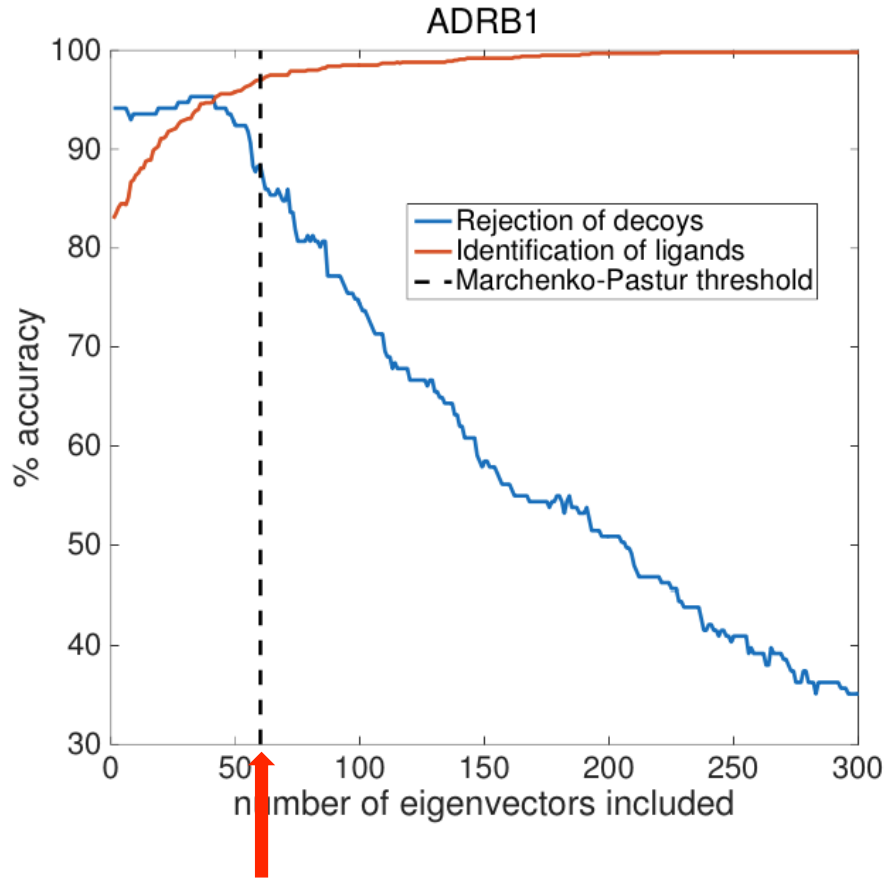
Second eigenvector



Lee *et al*, PNAS 2016

SUMMARY

We showed that removing the noise present in this dataset due to finite sampling greatly improved the ability of our model to make accurate predictions.

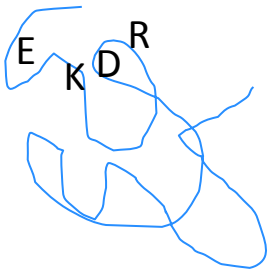


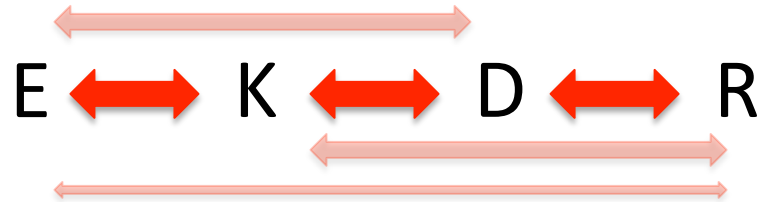
Marcenko-Pastur threshold

But what about the effects of finite sampling in protein sequence data?

Before, I argued that transitive correlations obscure the signal

Wild Type Protein





	K	D	E	R
K	X	I	I	C
D	I	X	C	I
E	I	C	X	C
R	C	I	C	X

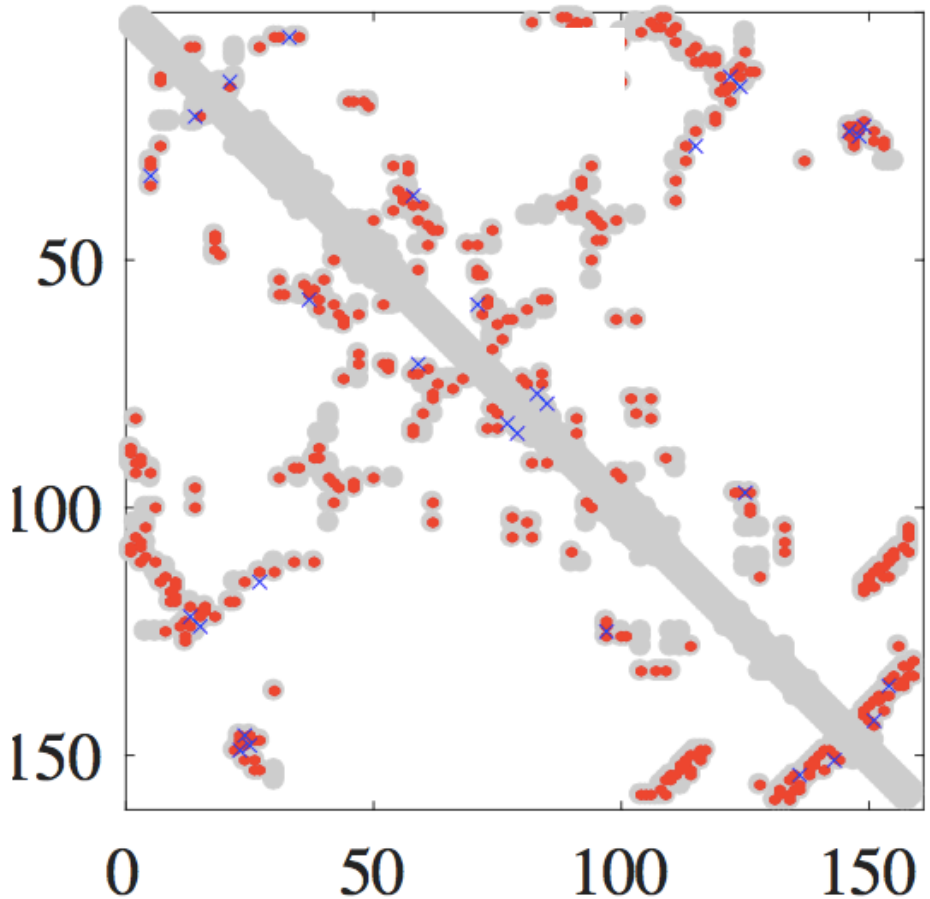
$$P(A_1,\ldots,A_L) = \frac{1}{Z}\exp\left\{-\sum_i h_i A_i - \sum_{(i,j)} e_{ij}\left(A_i, A_j\right)\right\}$$

naïve mean field approximation, assume small couplings

$$e_{ij}\left(A_i, A_j\right) = \left(C^{-1}\right)_{ij}\left(A_i, A_j\right)$$

Is this really true?

The residue covariance matrix: $C_R = X^T * X$

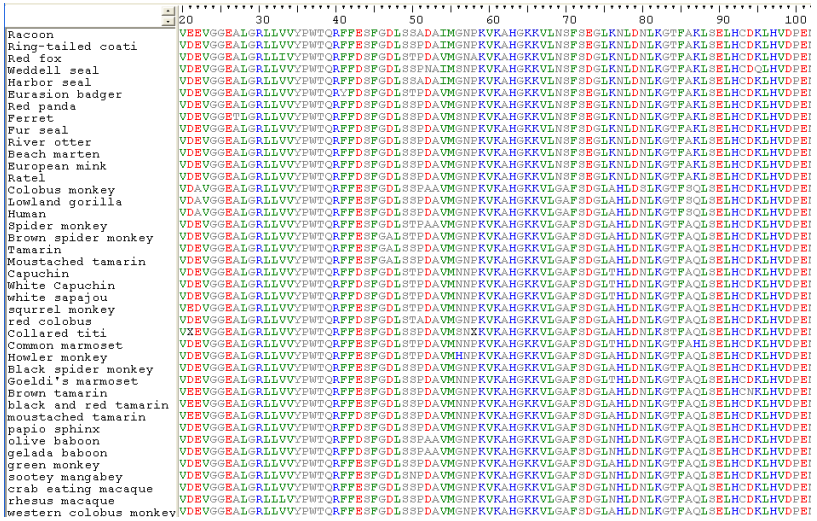


Simulation using a contact map – here for DHFR

What else could cause the signal to get obscured?

What else could cause the signal to be obscured?

Species



Our data matrix X

This data contains ‘nuisance’ correlations because the samples are related to each other, they are not iid. We call this relatedness structure the ‘phylogeny’

What else could cause the signal to be obscured?

Species

```
                                  20        30        40        50        60        70        80        90       100
Racoon                            VEEVGGEALGRLLVVYPWTQRFFESFGDLSSADAIMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Ring-tailed coati                 VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAIMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Red fox                           VDEVGGEALGRLLIVYPWTQRFFDSFGDLSTPDAVMGNAKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Weddell seal                      VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPNAIMSNPKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDQLHVDPEI
Harbor seal                       VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSADAIMGNPKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Eurasion badger                   VDEVGGEALGRLLVVYPWTQRYFDSFGDLSTPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Red panda                         VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Ferret                            VDEVGGETLGRLLVVYPWTQRFFDSFGDLSSPDAVMSNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Fur seal                          VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLNSFSDGLKNLDNLKGTFAKLSELHCDKLHVDPEI
River otter                       VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Beach marten                      VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPEI
European mink                     VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Ratel                             VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLNSFSEGLKNLDNLKGTFAKLSELHCDKLHVDPEI
Colobus monkey                    VDAVGGEALGRLLVVYPWTQRFFESFGDLSSPAAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFSQLSELHCDKLHVDPEI
Lowland gorilla                   VDAVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFSQLSELHCDKLHVDPEI
Human                             VDAVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFSQLSELHCDKLHVDPEI
Spider monkey                     VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPAAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
Brown spider monkey               VDEVGGEALGRLLVVYPWTQRFFESFGALSTPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
Tamarin                           VDEVGGEALGRLLVVYPWTQRFFESFGALSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
Moustached tamarin                VDEVGGEALGRLLVVYPWTQRFFESFGALSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
Capuchin                          VDEVGGEALGRLLVVYPWTQRFFDSFGDLSTPDAVMNNPKVKAHGKKVLGAFSDGLTHLDNLKGTFAQLSELHCDKLHVDPEI
White Capuchin                    VDEVGGEALGRLLVVYPWTQRFFDSFGDLSTPDAVMNNPKVKAHGKKVLGAFSDGLTHLDNLKGTFAQLSELHCDKLHVDPEI
white sapajou                     VDEVGGEALGRLLVVYPWTQRFFDSFGDLSTPDAVMNNPKVKAHGKKVLGAFSDGLTHLDNLKGTFAQLSELHCDKLHVDPEI
squrrel monkey                    VEDVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMNNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
red colobus                       VDEVGGEALGRLLVVYPWTQRFFDSFGDLSTADAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
Collared titi                     VXEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMSNXKVKAHGKKVLGAFSDGLAHLDNLKSTFAQLSELHCDKLHVDPEI
Common marmoset                   VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMNNPKVKAHGKKVLGAFSDGLTHLDNLKGTFAHLSELHCDKLHVDPEI
Howler monkey                     VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMHNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
Black spider monkey               VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
Goeldi's marmoset                 VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMNNPKVKAHGKKVLGAFSDGLTHLDNLKGTFAQLSELHCDKLHVDPEI
Brown tamarin                     VEEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCNKLHVDPEI
black and red tamarin             VEEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMNNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
moustached tamarin                VEEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMNNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
papio sphinx                      VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPEI
olive baboon                      VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPAAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPEI
gelada baboon                     VDEVGGEALGRLLVVYPWTQRFFDSFGDLSSPAAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPEI
green monkey                      VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
sootey mangabey                   VDEVGGEALGRLLVVYPWTQRFFESFGDLSNPDAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPEI
crab eating macaque               VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPEI
rhesus macaque                    VDEVGGEALGRLLLVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLNHLDNLKGTFAQLSELHCDKLHVDPEI
western colobus monkey            VDEVGGEALGRLLVVYPWTQRFFESFGDLSSPDAVMGNPKVKAHGKKVLGAFSDGLAHLDNLKGTFAQLSELHCDKLHVDPEI
```

Our data matrix X

This data contains 'nuisance' correlations because the samples are related to each other, they are not iid. We call this relatedness structure the 'phylogeny'

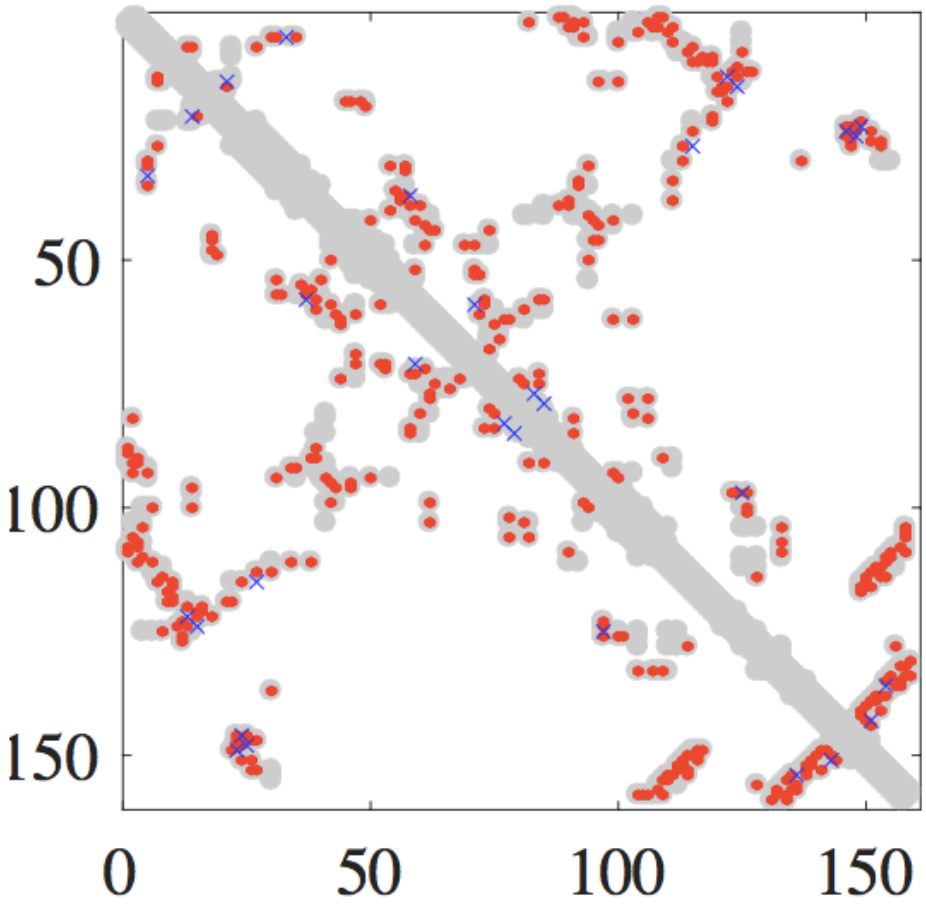
The sequence covariance matrix – $C_S = X*X^T$ used to compute phylogeny

The residue covariance matrix – $C_R = X^T*X$ constraints from structure/function

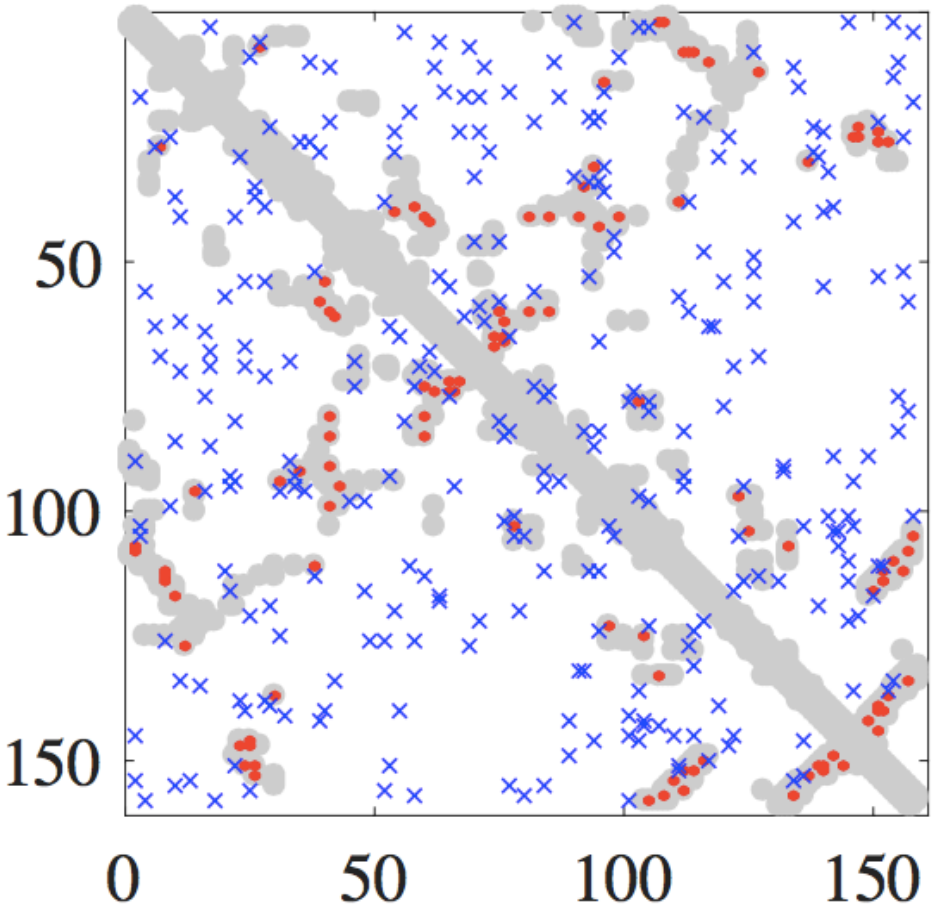
These matrices contain the same information – in particular they have the same eigenvalues (up to the number of zero eigenvalues).

These two sources of covariance mix and confound each other.

Repeat the simulation, but now include phylogeny



Without phylogeny – contact prediction is quite accurate just using $C_R = X^T*X$



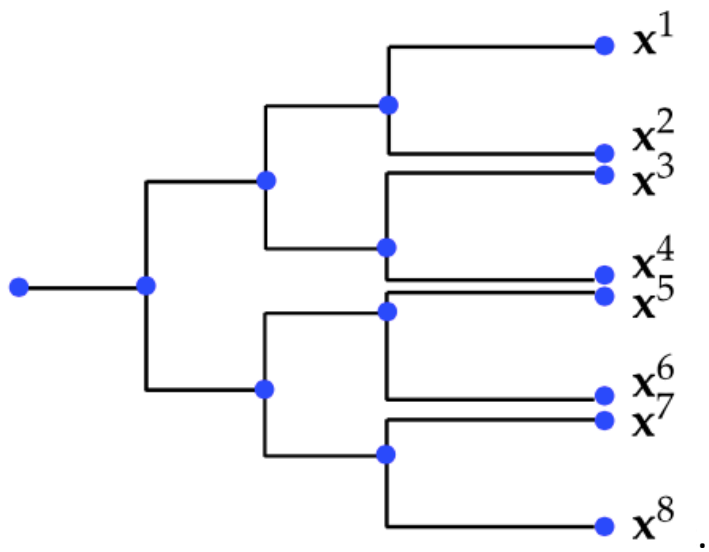
With phylogeny – contact prediction is much less accurate using $C_R = X^T*X$

We need to know what covariance is caused by the phylogeny

Consider a perfect 'tree phylogeny'.

A founder sequence of length p residues are independently evolved for m mutation steps.

Sequences are duplicated and the process repeats through b such branching events to yield n sequences.

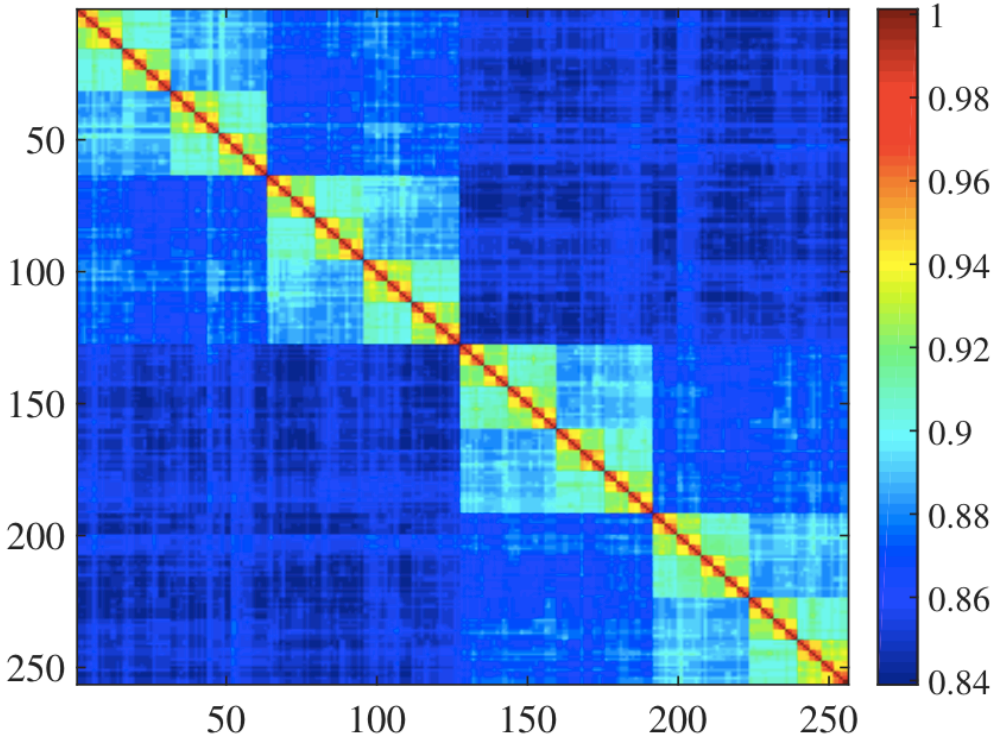


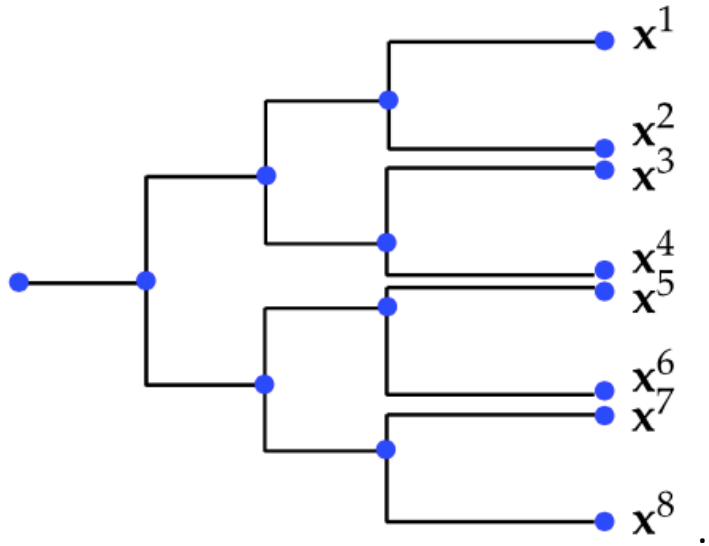
We need to know what covariance is caused by the phylogeny

Consider a perfect 'tree phylogeny'.

A founder sequence of length p residues are independently evolved for m mutation steps.

Sequences are duplicated and the process repeats through b such branching events to yield n sequences.





Sequence covariance matrix is highly structured..... like batch effects.

What effect does this have on the eigenvalues?

We need to know what covariance is caused by the phylogeny

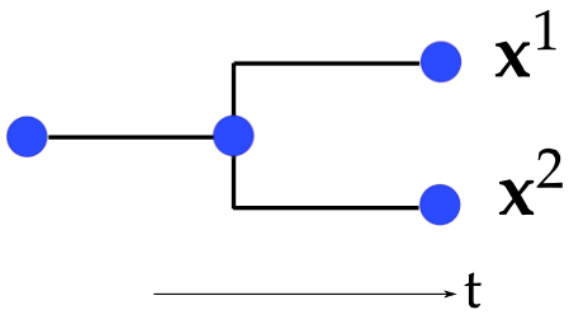
We define:

$$\alpha(m) = \exp\left[-\frac{qm}{(q-1)p}\right]$$

m mutations per branch, sequences of length p, alphabet of size q. Start with q = 2.

The elements of the true covariance matrix are given by: $(\Sigma_S)_{\mathrm{ij}} = \alpha^{b_{\mathrm{ij}}}$ where b_{ij} is the number of branching events between sequences i and j.

Consider a single branching event:



$$\Sigma_S = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}$$

$$\lambda_\pm = 1 \pm \alpha$$

We need to know what covariance is caused by the phylogeny

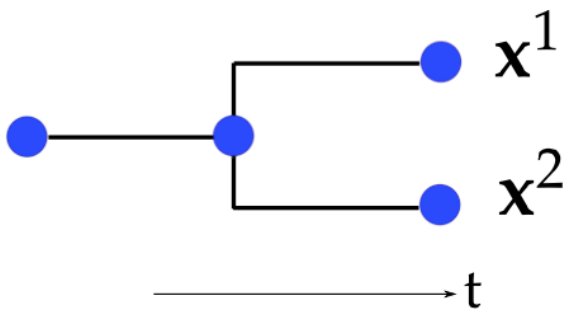
We define:

$$\alpha(m) = \exp\left[-\frac{qm}{(q-1)p}\right]$$

m mutations per branch, sequences of length p, alphabet of size q. Start with q = 2.

The elements of the true covariance matrix are given by: $(\Sigma_S)_{\mathrm{ij}} = \alpha^{b_{\mathrm{ij}}}$ where b_{ij} is the number of branching events between sequences i and j.

Consider a single branching event:



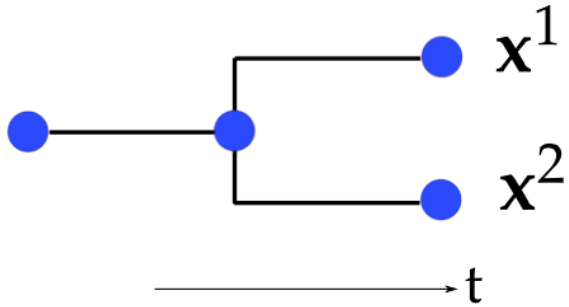
$$\Sigma_S = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}$$

$$\lambda_\pm = 1 \pm \alpha$$

What is the spectrum of the sample covariance matrix?

Calculating the effects of finite sampling

Recall, for a single branching event:



$$\Sigma_S = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}$$

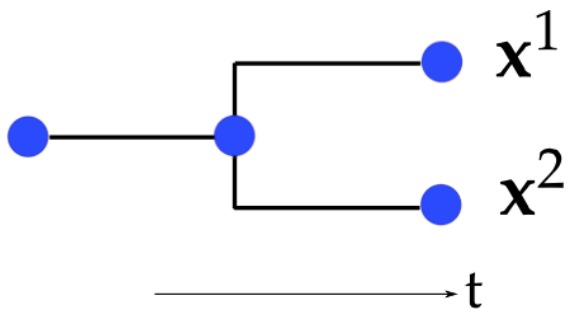
$$\lambda_\pm = 1 \pm \alpha$$

To compute the expected eigenvalue distribution of the sample covariance matrix, we use the Stieltjes transform:

$$G(z, c) = \int_{-\infty}^{+\infty} \frac{\mathrm{d}F(\lambda)}{z - \lambda}$$

Calculating the effects of finite sampling

Recall, for a single branching event:



$$\Sigma_S = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}$$

$$\lambda_\pm = 1 \pm \alpha$$

To compute the expected eigenvalue distribution of the sample covariance matrix, we use the Stieltjes transform:

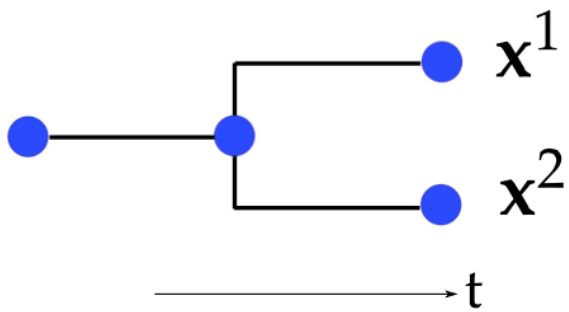
$$G(z,c) = \int_{-\infty}^{+\infty} \frac{\mathrm{d}F(\lambda)}{z-\lambda}$$

In 1967 Marcenko and Pastur used a pde-based approach to write the Stieltjes transform in terms of the eigenvalue distribution of the true covariance matrix $T(\lambda)$

$$G(z,c) = -1\left(z - c\int_{-\infty}^{\infty} \frac{\lambda \mathrm{d}T(\lambda)}{1+\lambda G(z,c)}\right)^{-1}$$

Calculating the effects of finite sampling

Recall, for a single branching event:



$$\Sigma_S = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}$$

$$\lambda_\pm = 1 \pm \alpha$$

To compute the expected eigenvalue distribution of the sample covariance matrix, we use the Stieltjes transform:

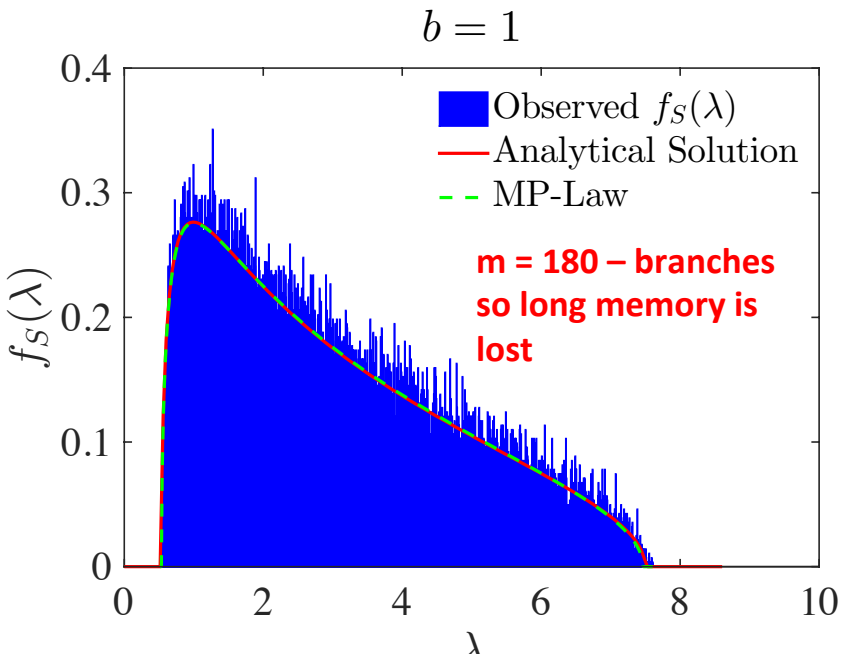
$$G(z, c) = \int_{-\infty}^{+\infty} \frac{\mathrm{d}F(\lambda)}{z - \lambda}$$

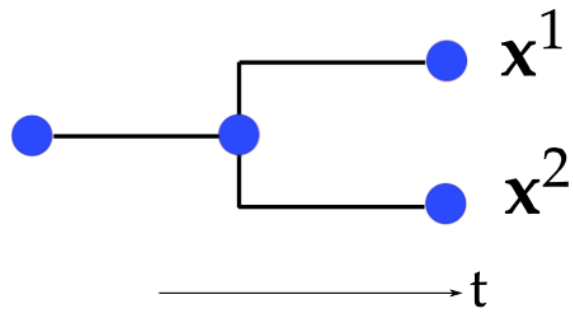
In 1967 Marcenko and Pastur used a pde-based approach to write the Stieltjes transform in terms of the eigenvalue distribution of the true covariance matrix $T(\lambda)$. In this case, the eigenvalues are discrete so the inverse Stieltjes transform, which describes the finite sampling effects, reduces to finding the positive imaginary root of this polynomial:

$$z - \frac{c}{2}\left(\frac{1+\alpha}{1+(1+\alpha)G}\right) - \frac{c}{2}\left(\frac{1-\alpha}{1+(1-\alpha)G}\right) = -\frac{1}{G}$$

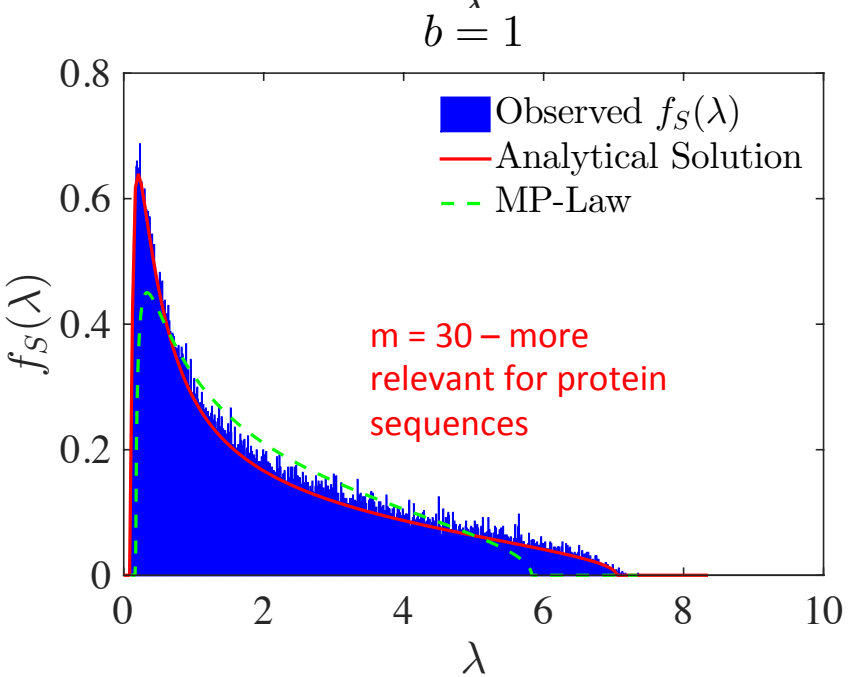
Red lines computed using this approach:

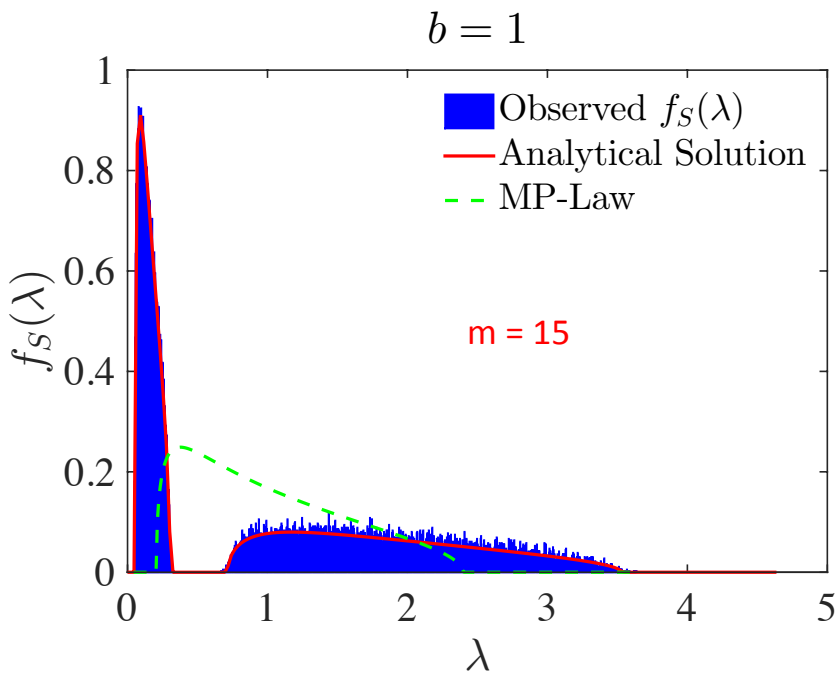
Simulations of the simplest trees produce non-MP spectra

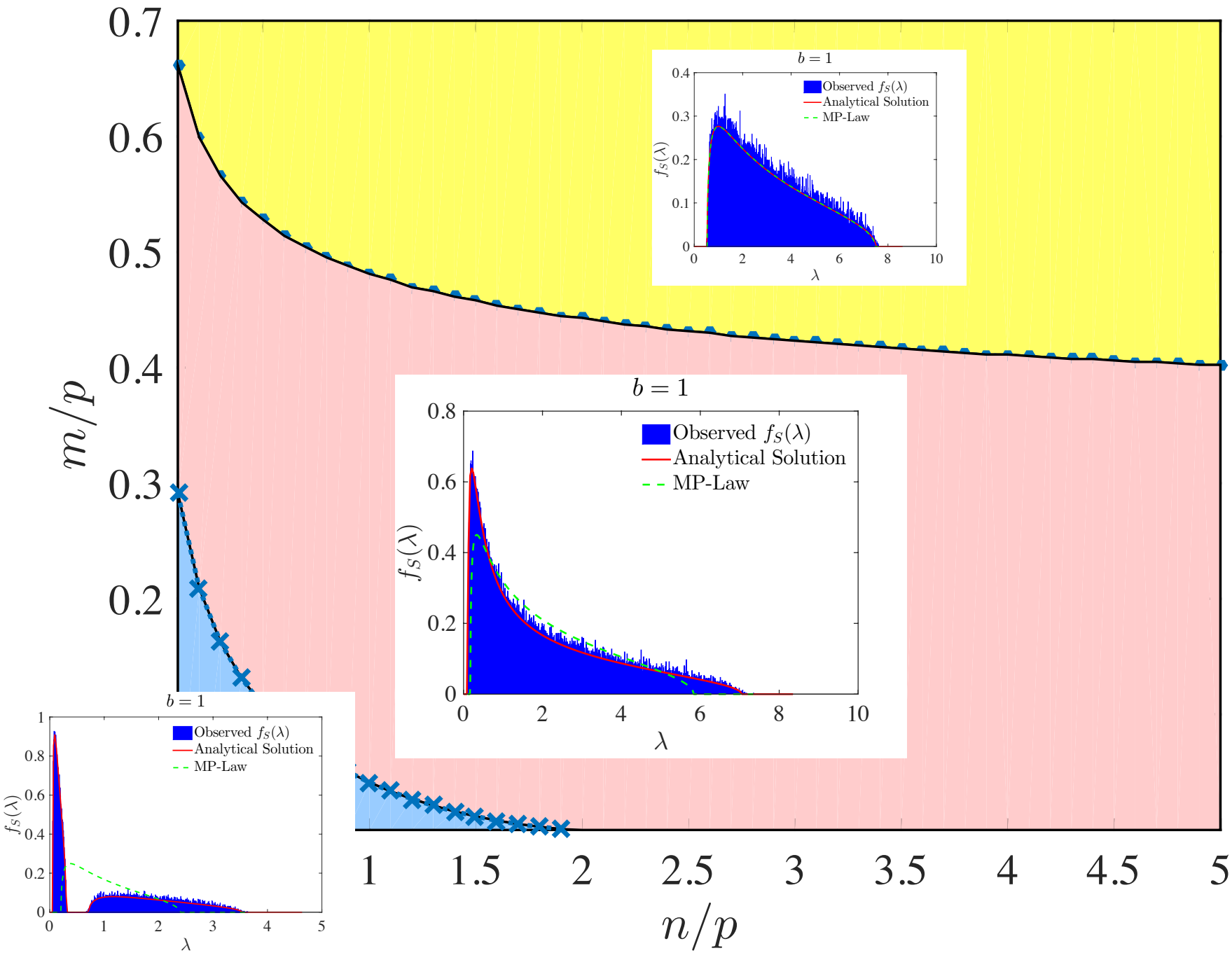




Consider sequences of length p = 300 in a model with a single duplication event, with branch length m.

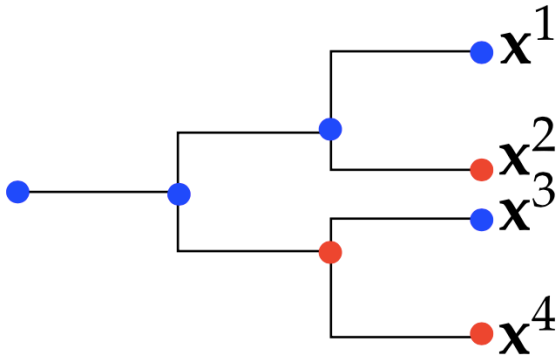





0.7
0.6
0.5
0.4
0.3
0.2
m/p
1
1.5
2
2.5
3
3.5
4
4.5
5
n/p
$b = 1$
Observed $f_S(\lambda)$
Analytical Solution
MP-Law
$f_S(\lambda)$
λ
$b = 1$
Observed $f_S(\lambda)$
Analytical Solution
MP-Law
$f_S(\lambda)$
λ
$b = 1$
Observed $f_S(\lambda)$
Analytical Solution
MP-Law
$f_S(\lambda)$
λ

What about deeper trees?

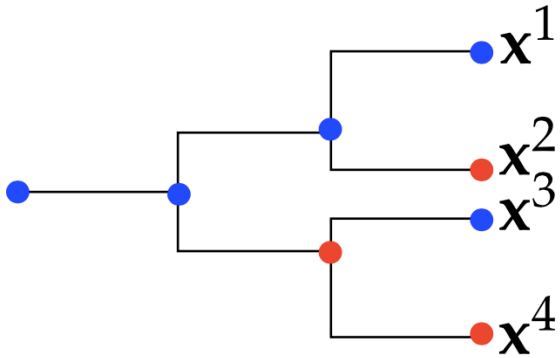
Move to deeper trees:



$$\Sigma_S = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^2 \\ \alpha & 1 & \alpha^2 & \alpha^2 \\ \alpha^2 & \alpha^2 & 1 & \alpha \\ \alpha^2 & \alpha^2 & \alpha & 1 \end{pmatrix}$$

What about deeper trees?

Move to deeper trees:



$$\Sigma_S = \begin{pmatrix} 1 & \alpha & \alpha^2 & \alpha^2 \\ \alpha & 1 & \alpha^2 & \alpha^2 \\ \alpha^2 & \alpha^2 & 1 & \alpha \\ \alpha^2 & \alpha^2 & \alpha & 1 \end{pmatrix}$$

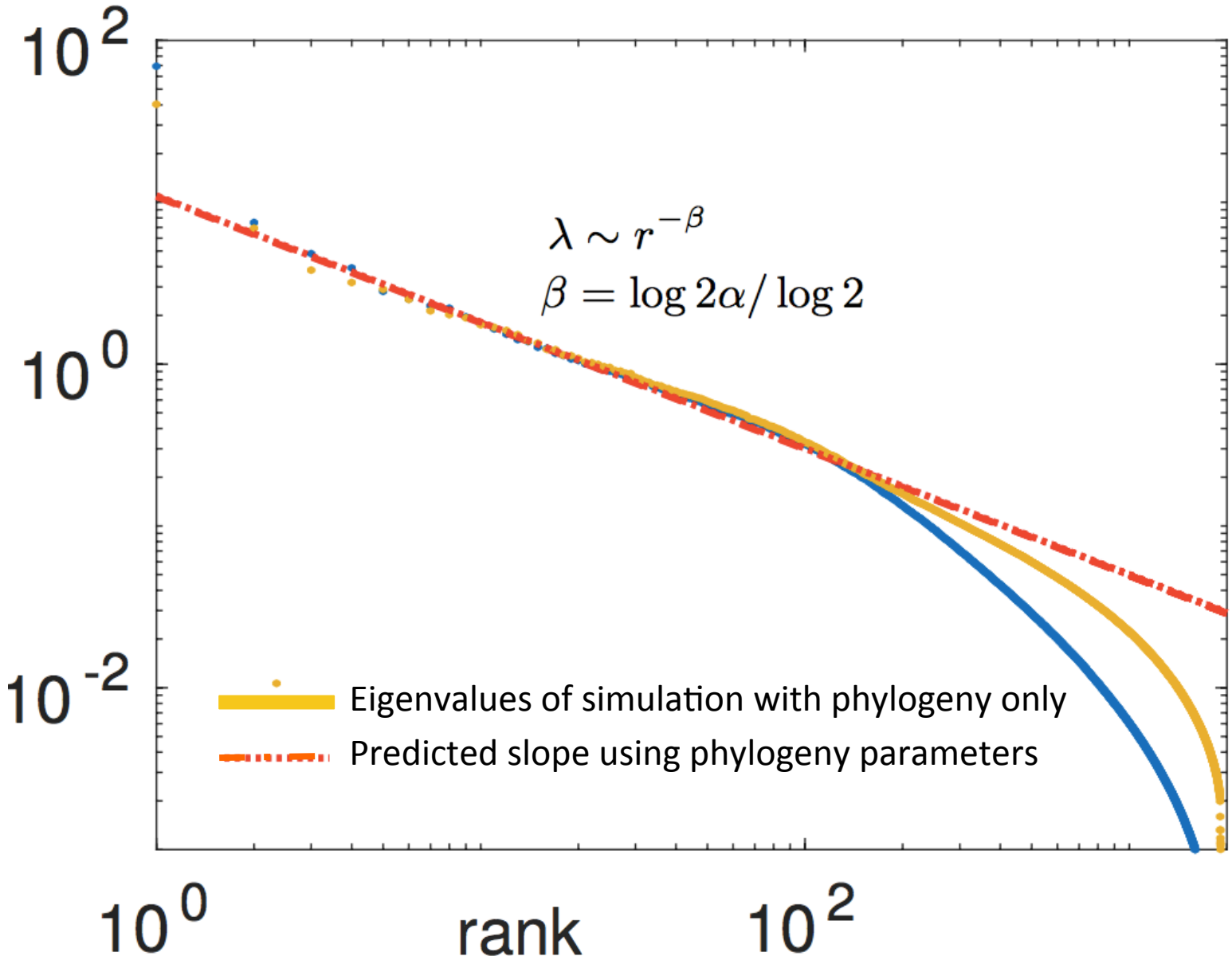
Eigenvalues of the true covariance matrix (more generally):

$$\lambda = \begin{cases} 1 + \sum_{\mathrm{i}=1}^{\mathrm{j}-1} 2^{\mathrm{i}-1}\alpha^{\mathrm{i}} - 2^{\mathrm{j}-1}\alpha^{\mathrm{j}} & \forall \mathrm{j} < b \\ 1 + \sum_{\mathrm{i}=1}^{b} 2^{\mathrm{i}-1}\alpha^{\mathrm{i}} \end{cases}$$

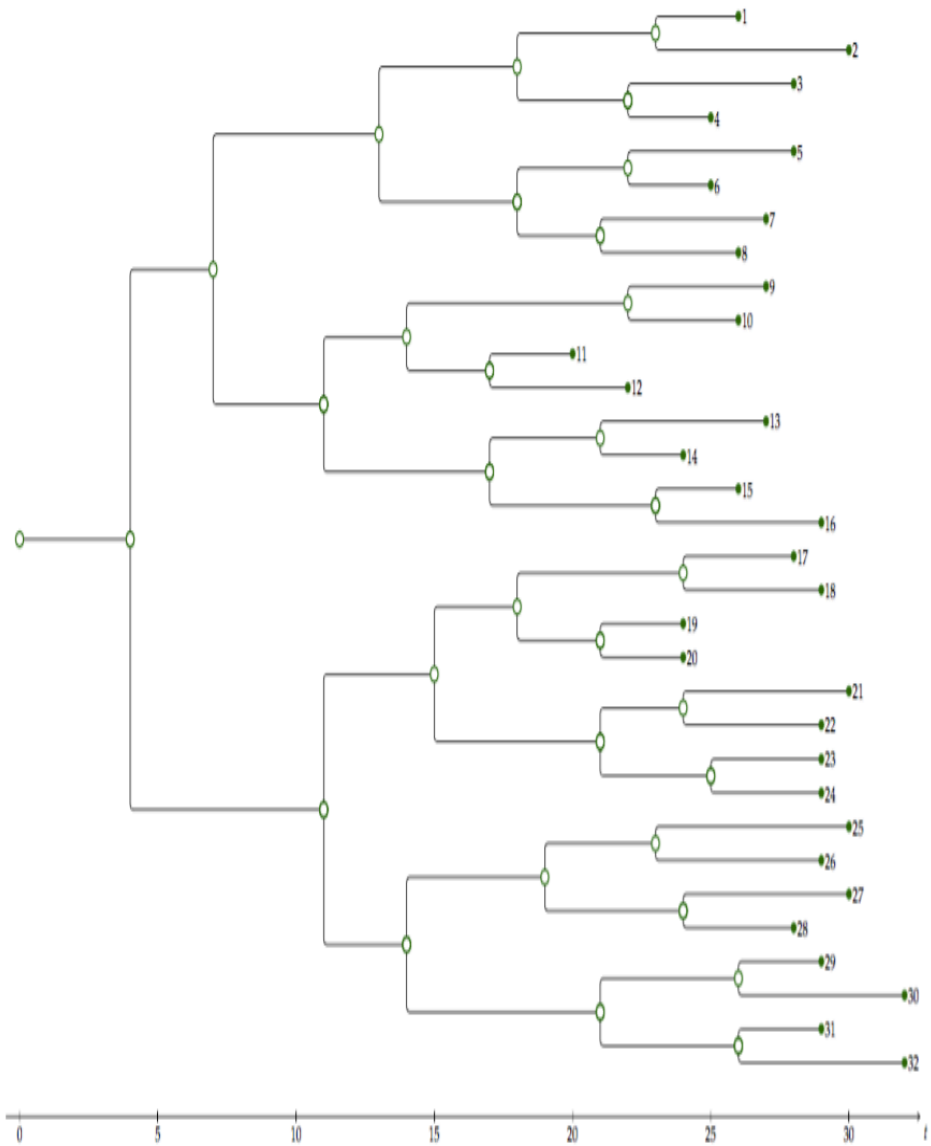
where: $P(\lambda = \lambda_i) = p_i = 2^{-i}$

This results in a power law, if *r* is the rank then: $\lambda \sim r^{-\beta}$

$$\beta = \log 2\alpha / \log 2$$


10^2
10^0
10^{-2}
$\lambda \sim r^{-\beta}$
$\beta = \log 2\alpha / \log 2$
Eigenvalues of simulation with phylogeny only
Predicted slope using phylogeny parameters
10^0
rank
10^2

“Real trees” are Inhomogeneous



Mathematics is more complicated.

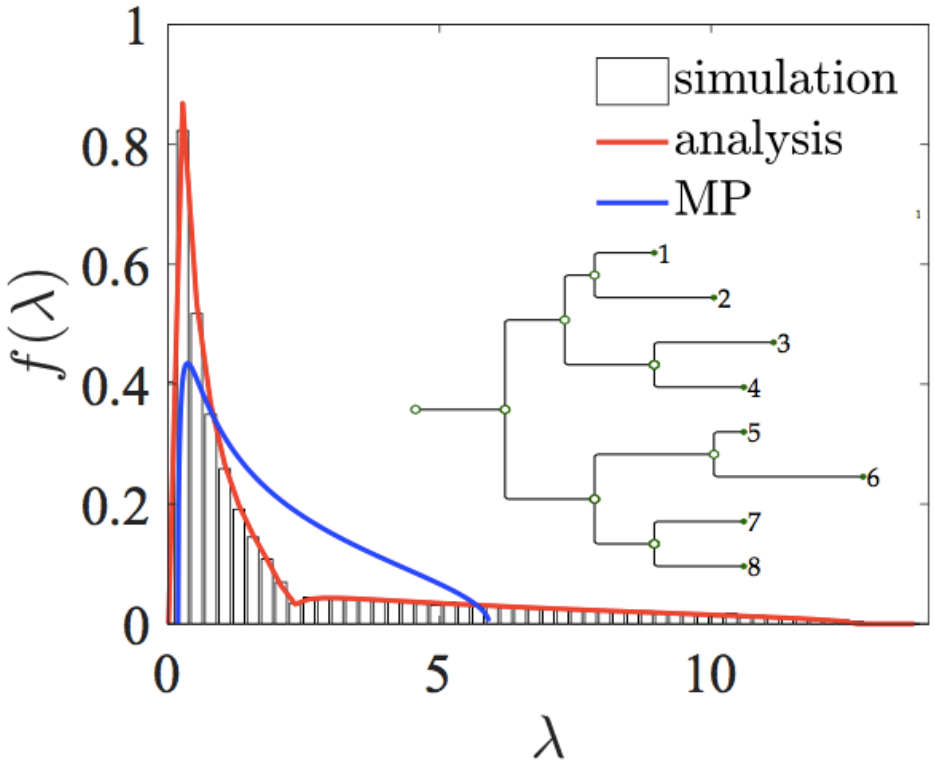
But we can prove that if the branch lengths are drawn from a distribution with mean *m* and bounded higher moments, then the empirical spectral distribution is exactly the same as for a homogeneous tree with branch length *m*.

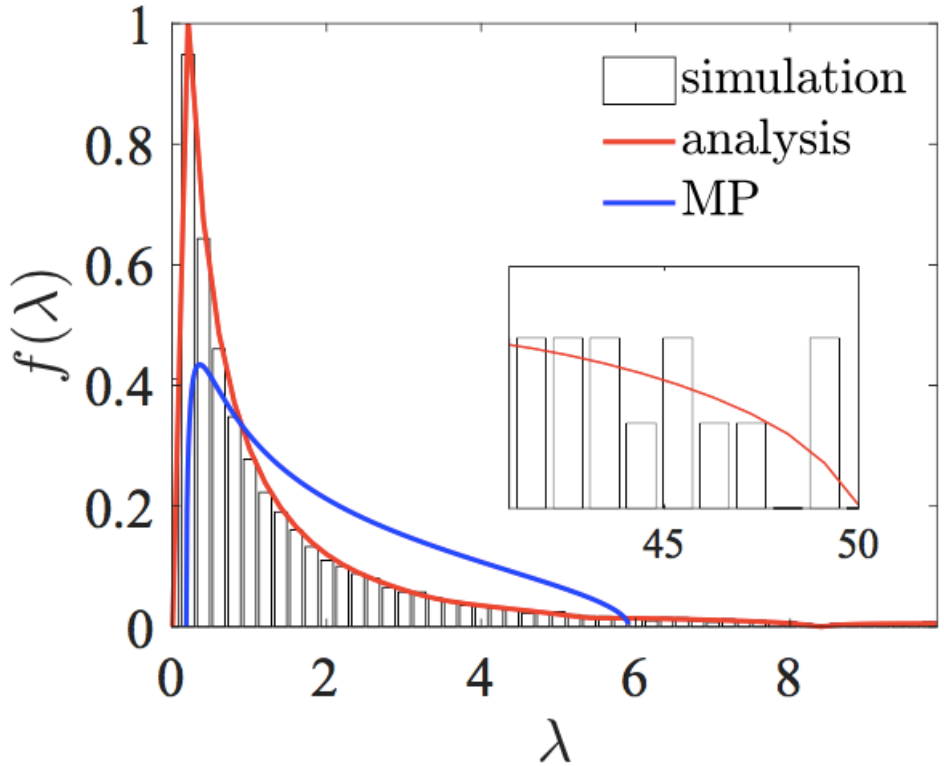
Key parameters:

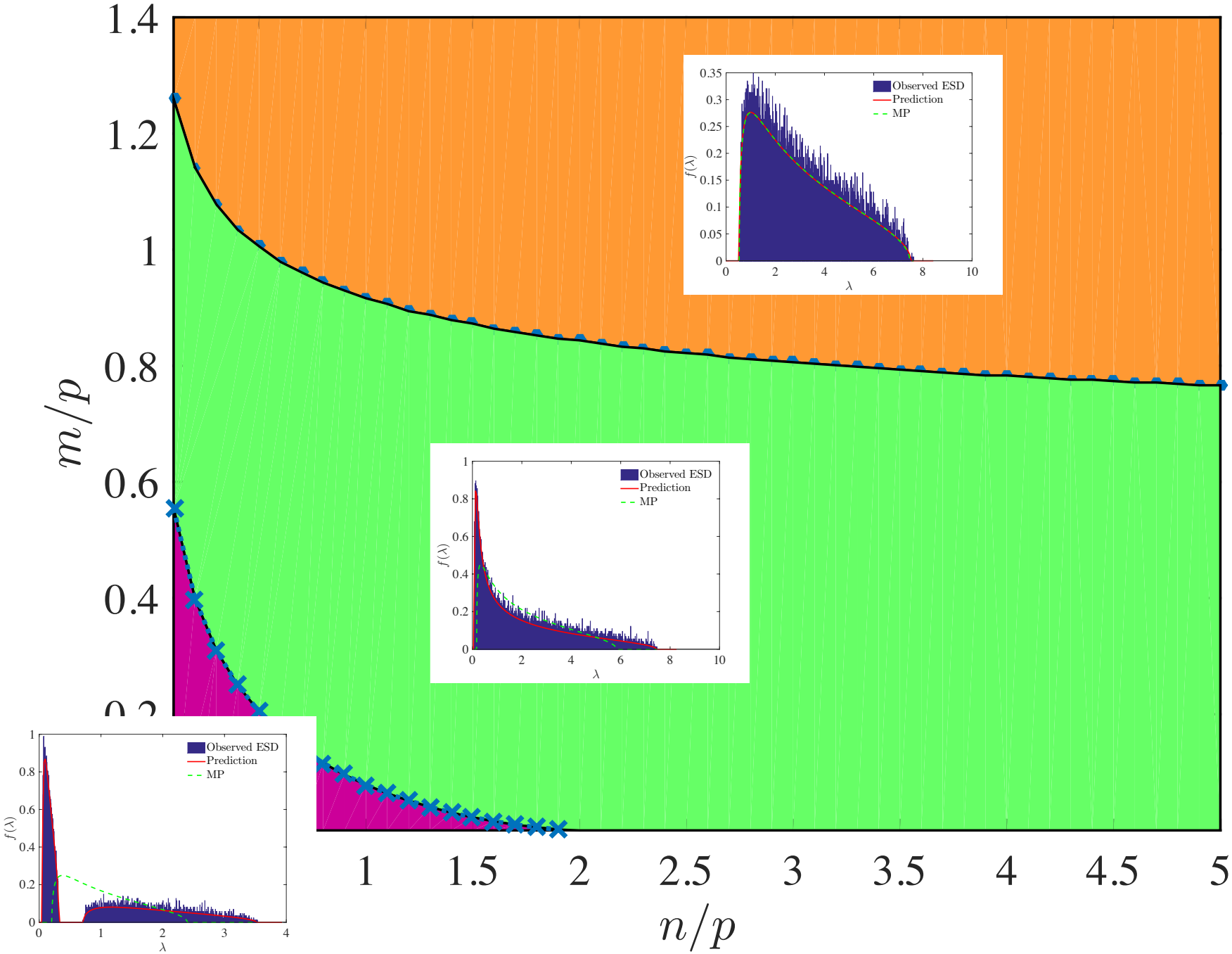
b: number of branching events

m: average mutations per branch

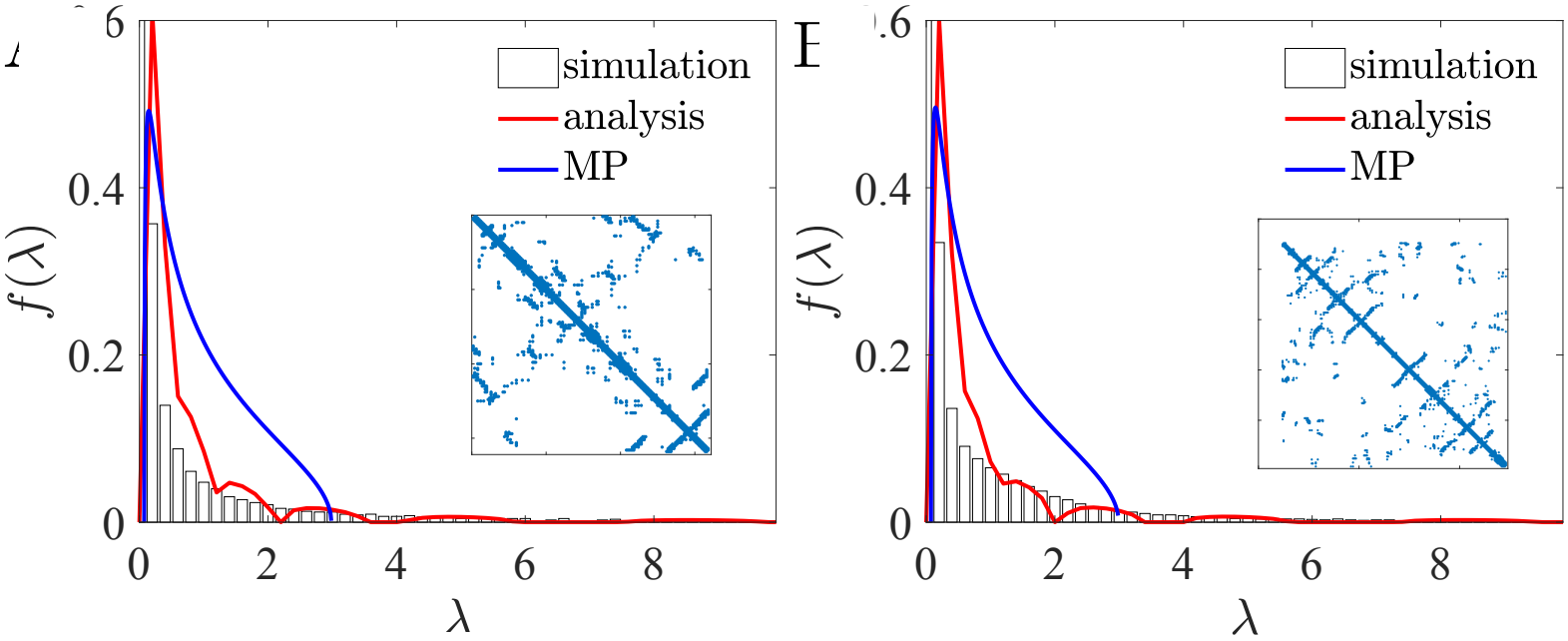
Theory also works for more realistic trees



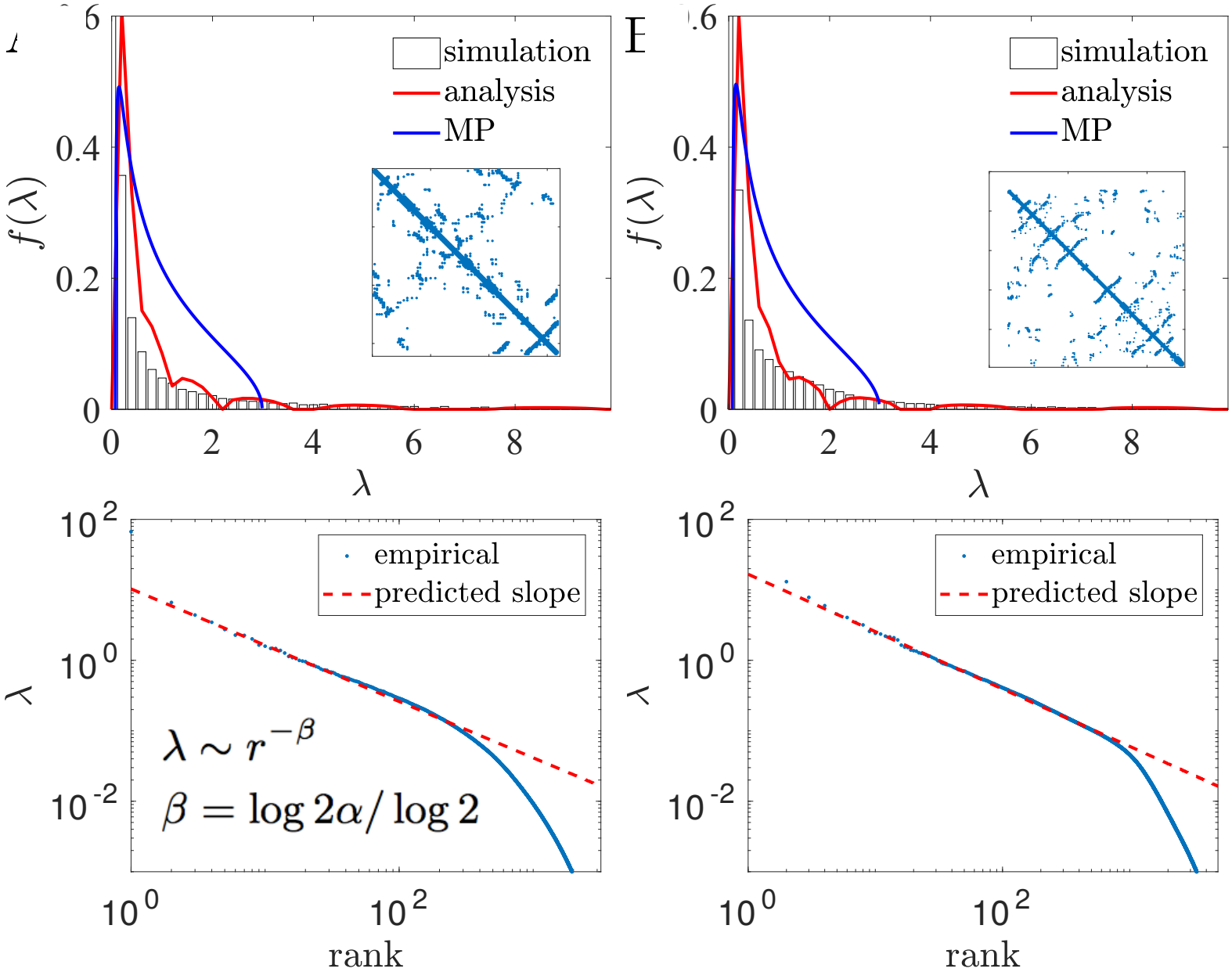



1.4
1.2
1
0.8
0.6
0.4
0.2
m/p
1
1.5
2
2.5
3
3.5
4
4.5
5
n/p
Observed ESD
Prediction
MP
$f(\lambda)$
λ

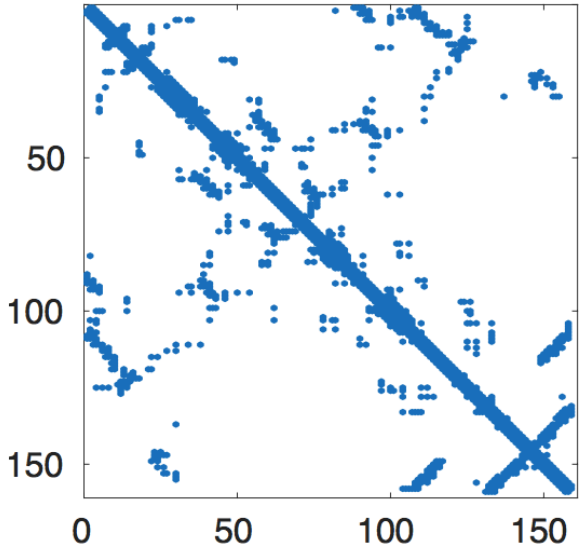
What about simulations with phylogeny and interactions?

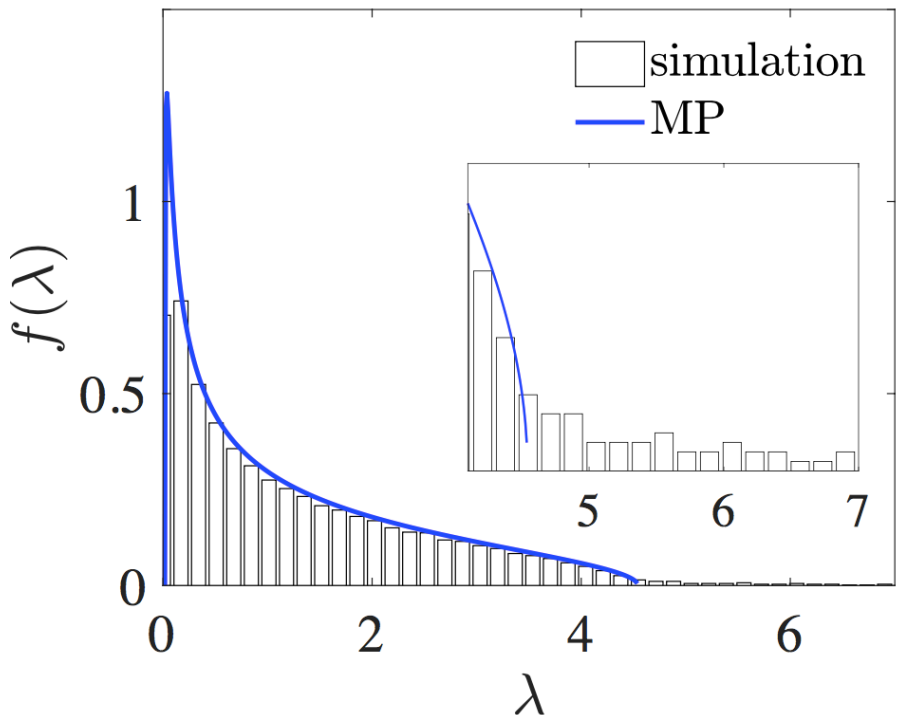


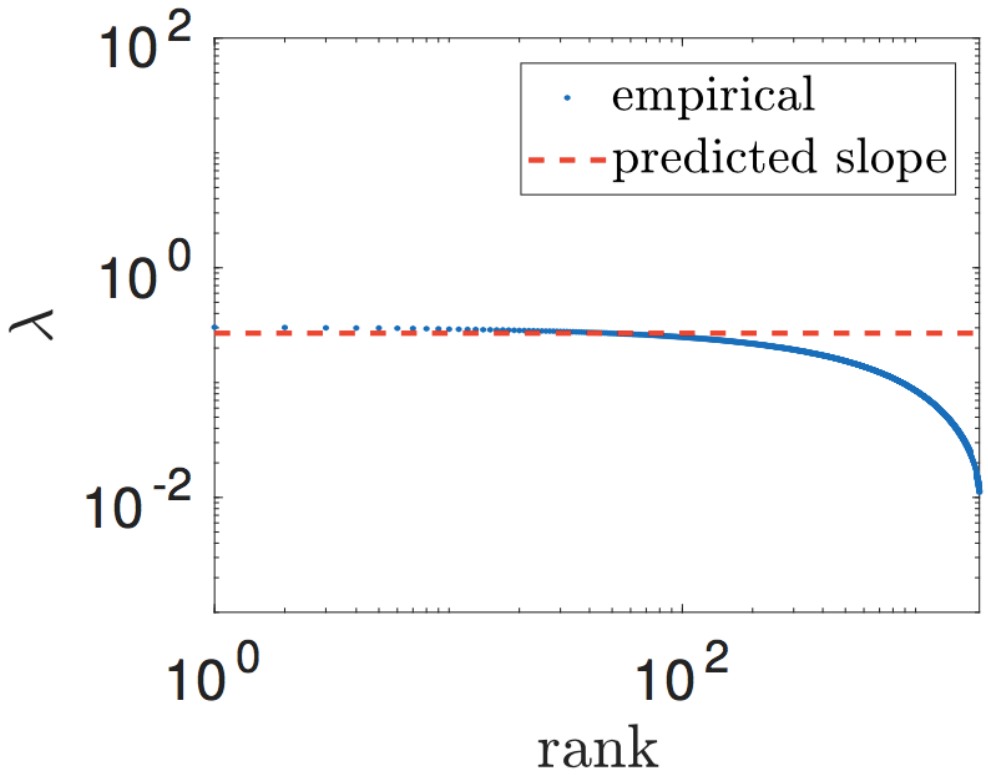
Simulations with phylogeny and interactions follow the power law

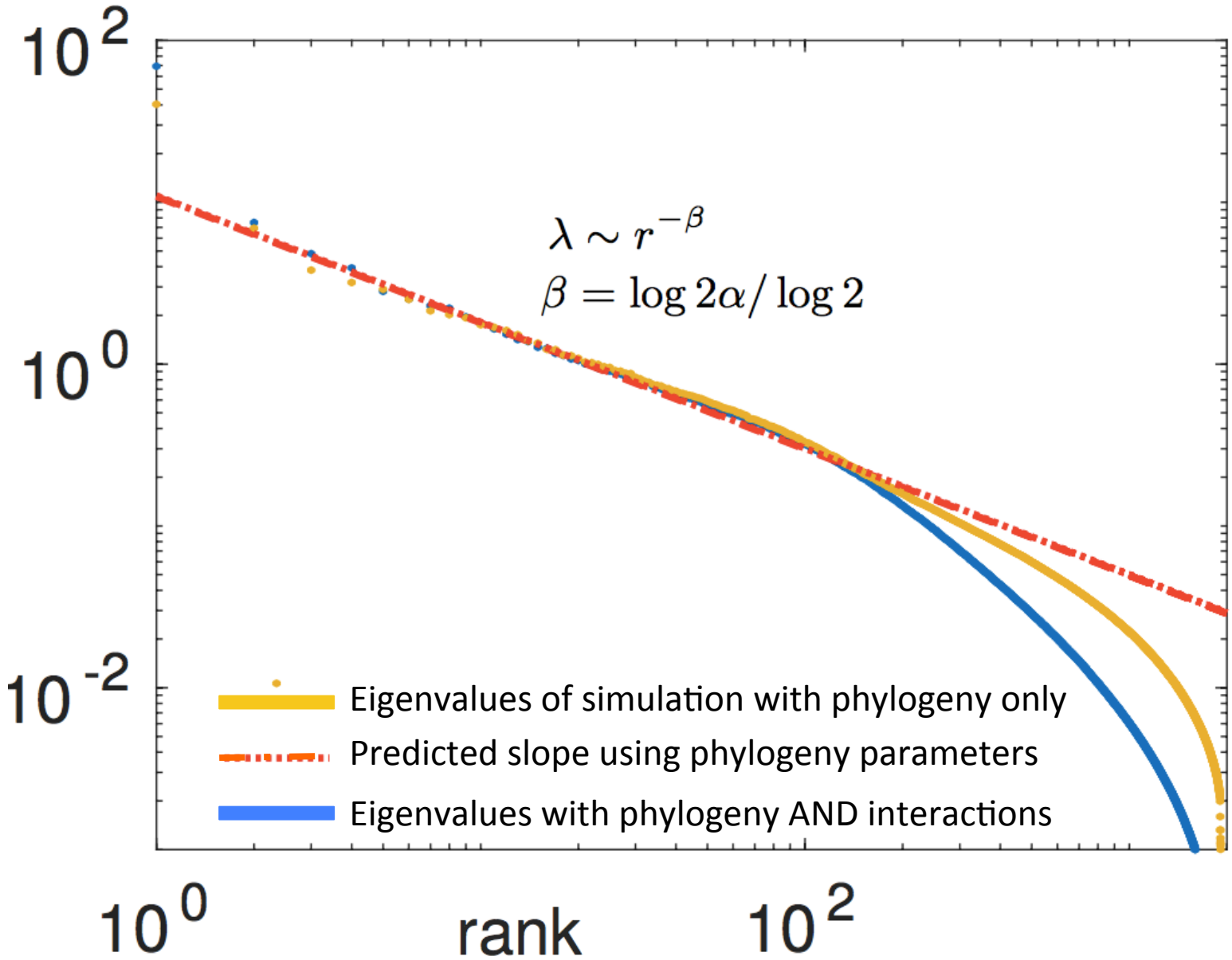


But interactions alone have a very different signature



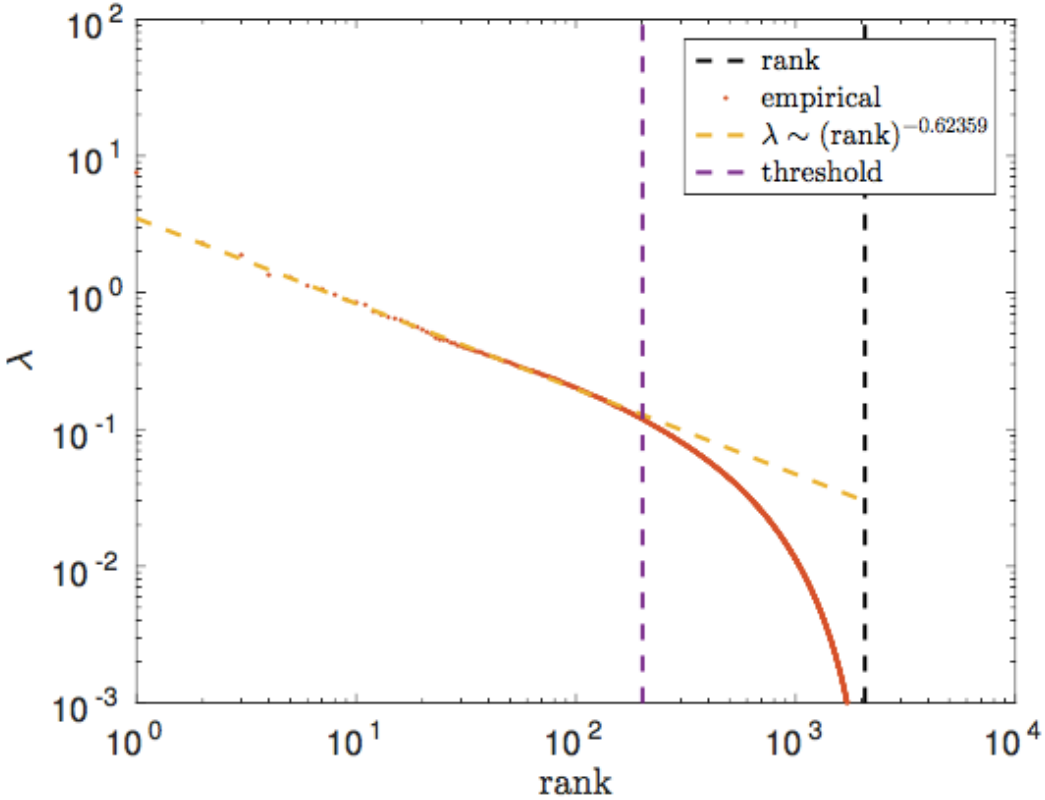




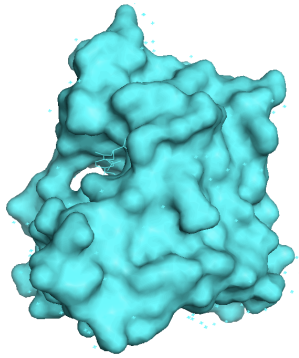


This suggests a strategy to 'clean' the phylogenetic corruption from the sequence covariance matrices……. what about real proteins???

What about real sequence data – can we use this?



DHFR

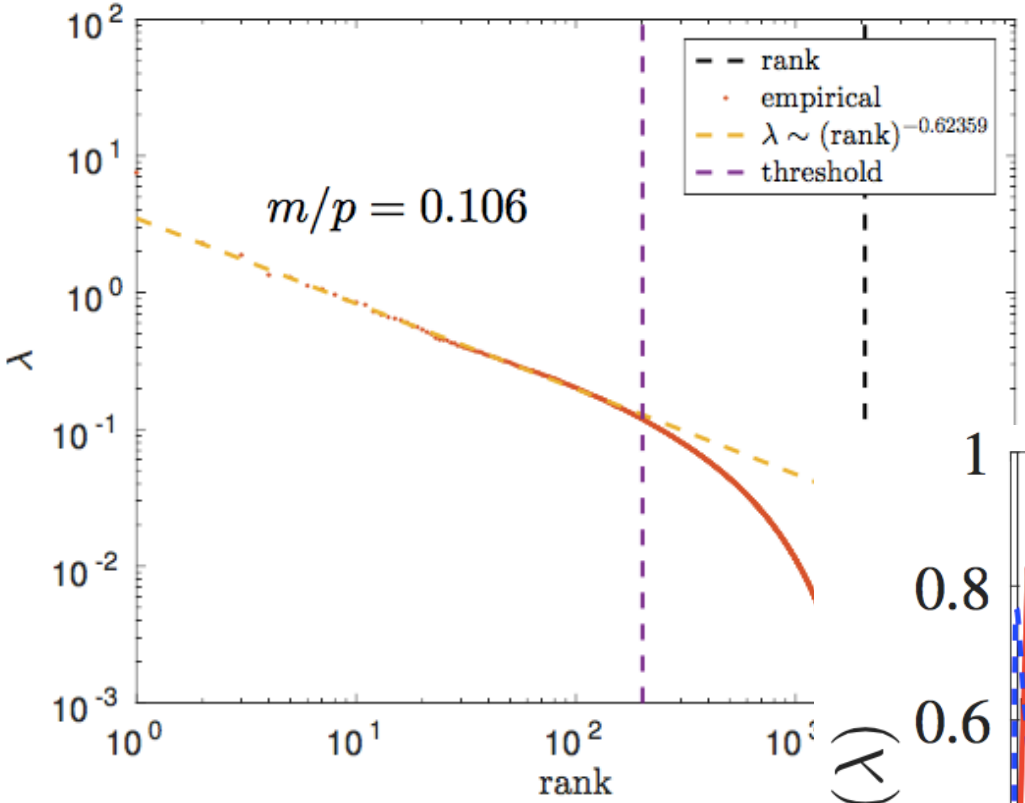


We have the covariance matrix here ~8k sequences. Calculate eigenvalues, fit the power law and extract the parameters:

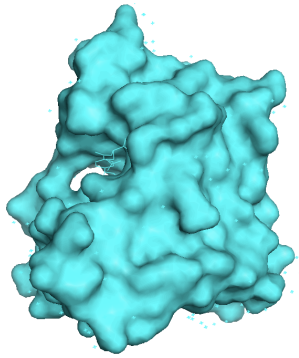
$$\lambda \sim r^{-\beta}$$

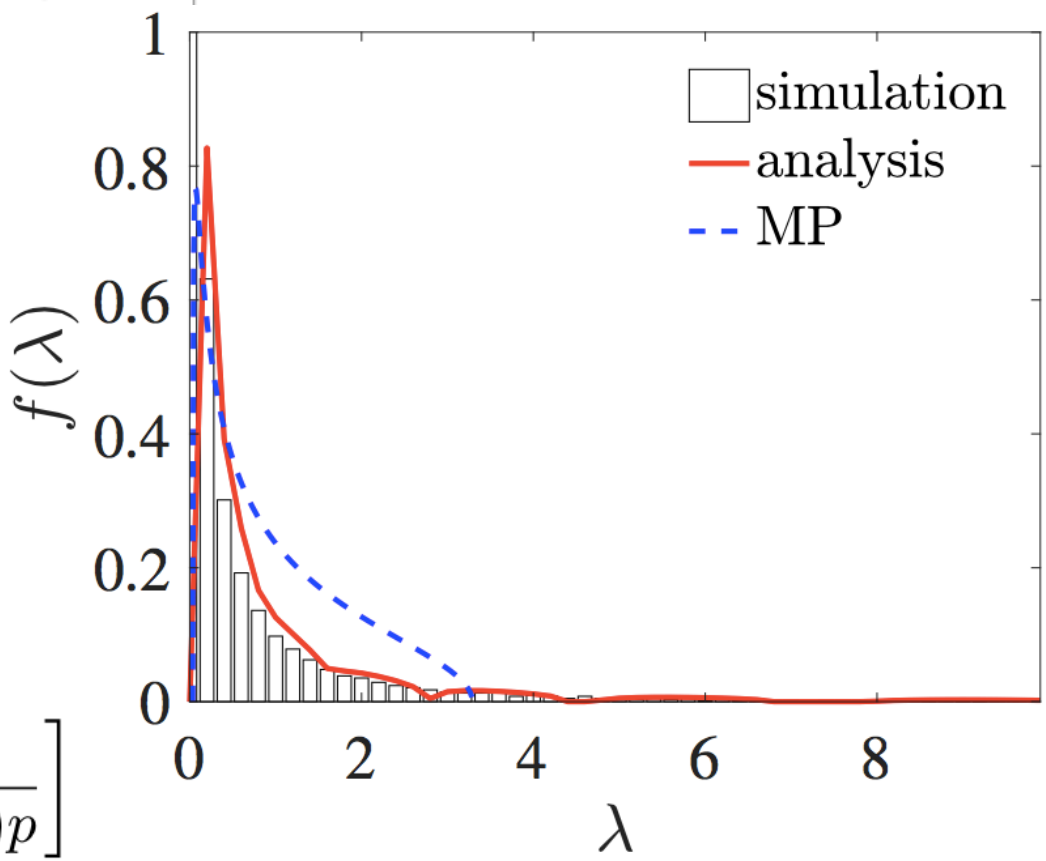
$$\beta = \log 2\alpha / \log 2$$

$$\alpha(m) = \exp\left[-\frac{qm}{(q-1)p}\right]$$



DHFR



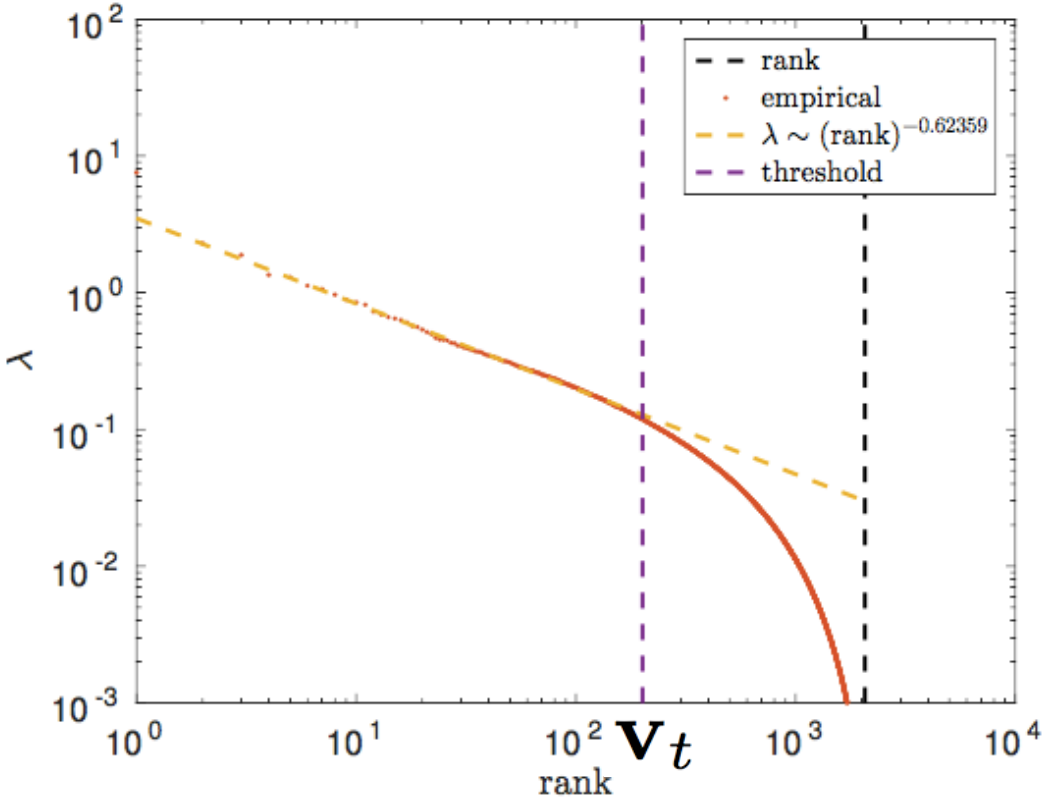


We have the covariance matrix here ~8k sequences. Calculate eigenvalues, fit the power law and extract the parameters:

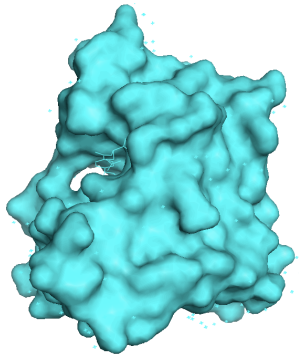
$$\lambda \sim r^{-\beta}$$

$$\beta = \log 2\alpha / \log 2$$

$$\alpha(m) = \exp\left[-\frac{qm}{(q-1)p}\right]$$



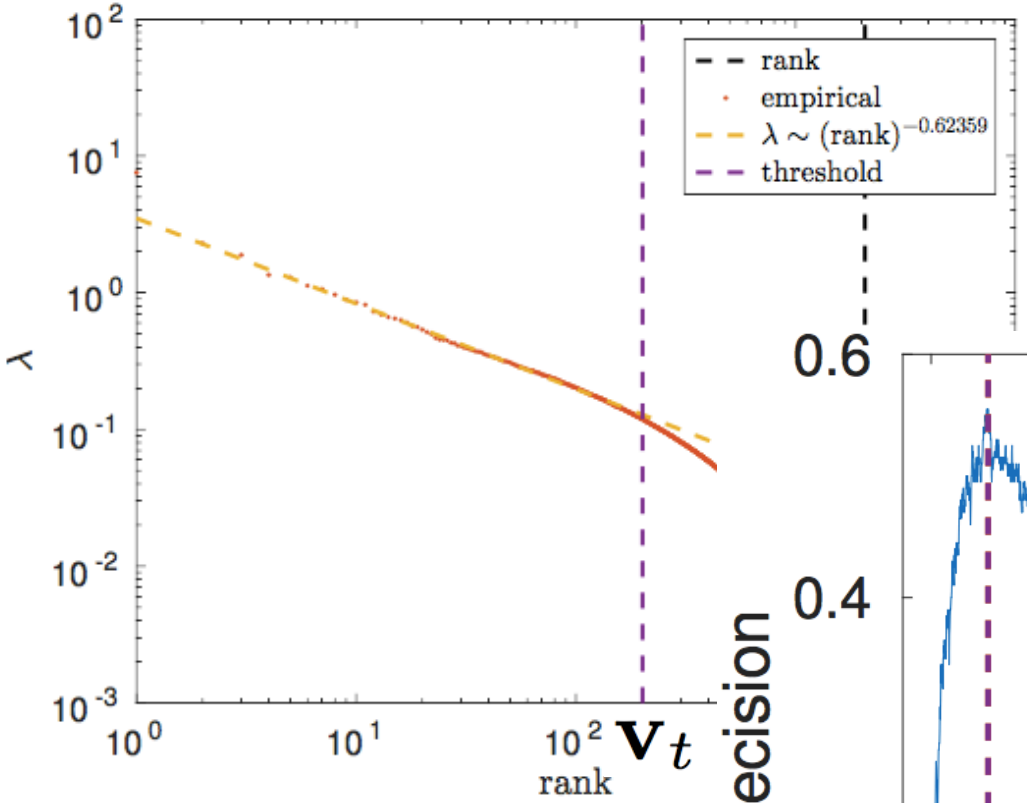
DHFR



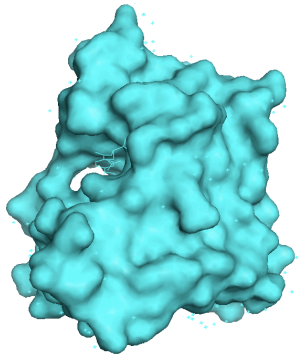
Calculated from the covariance matrix of ~8k sequences.

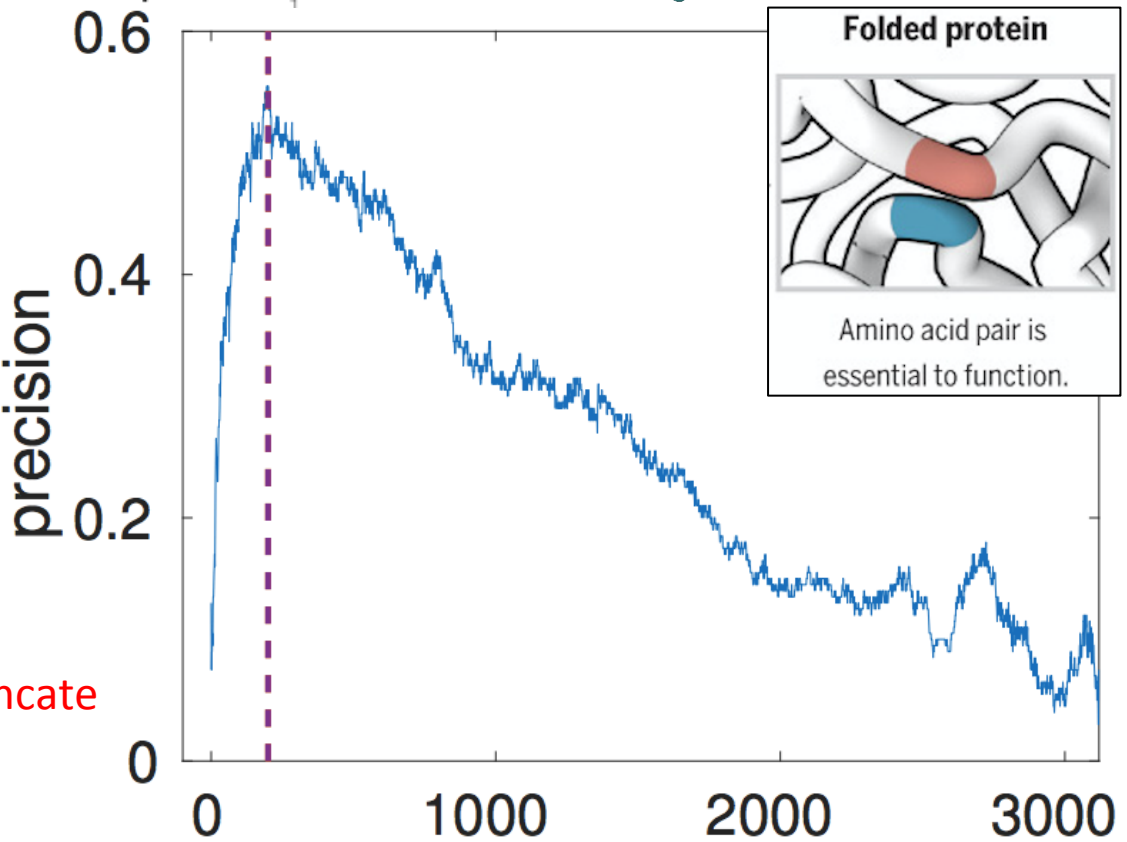
Can we use this understanding to remove the covariance caused by phylogeny?

$$S_p = \mathbf{v}_1\mathbf{v}_1^T + \cdots + \mathbf{v}_t\mathbf{v}_t^T, \quad \lambda_1 \leq \cdots \leq \lambda_t \leq \cdots \leq \lambda_{p(q-1)}$$



DHFR

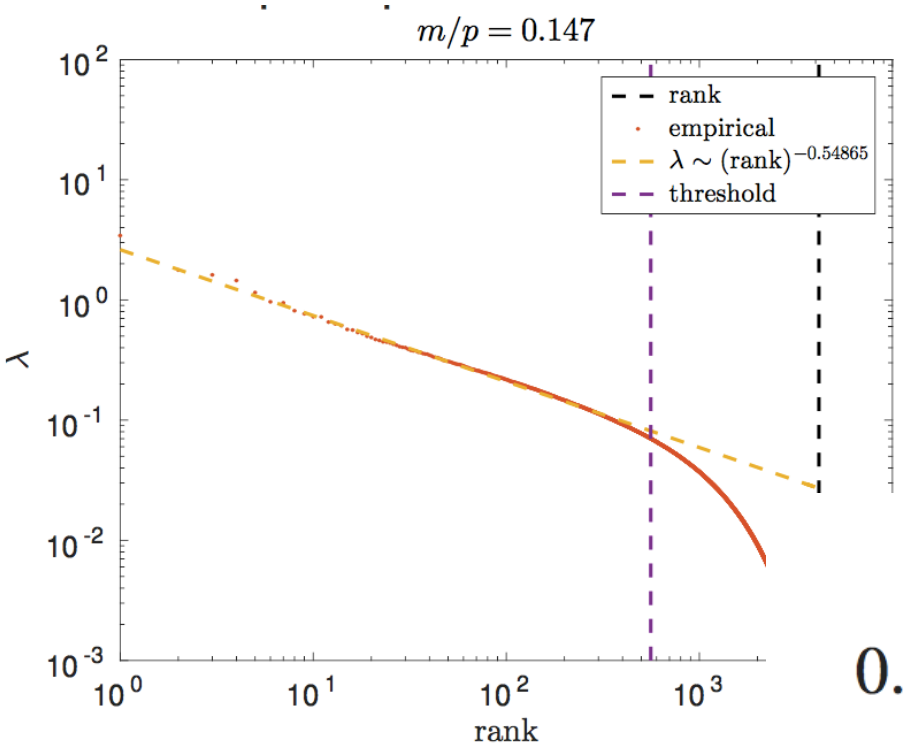




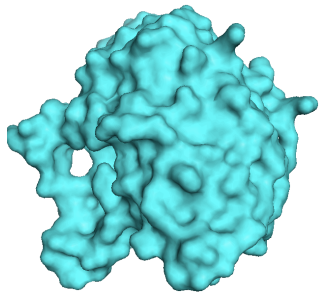
Calculated from the covariance matrix of ~8k sequences.

Using Random Matrix Theory to truncate the largest modes recovers contact prediction from raw covariance.

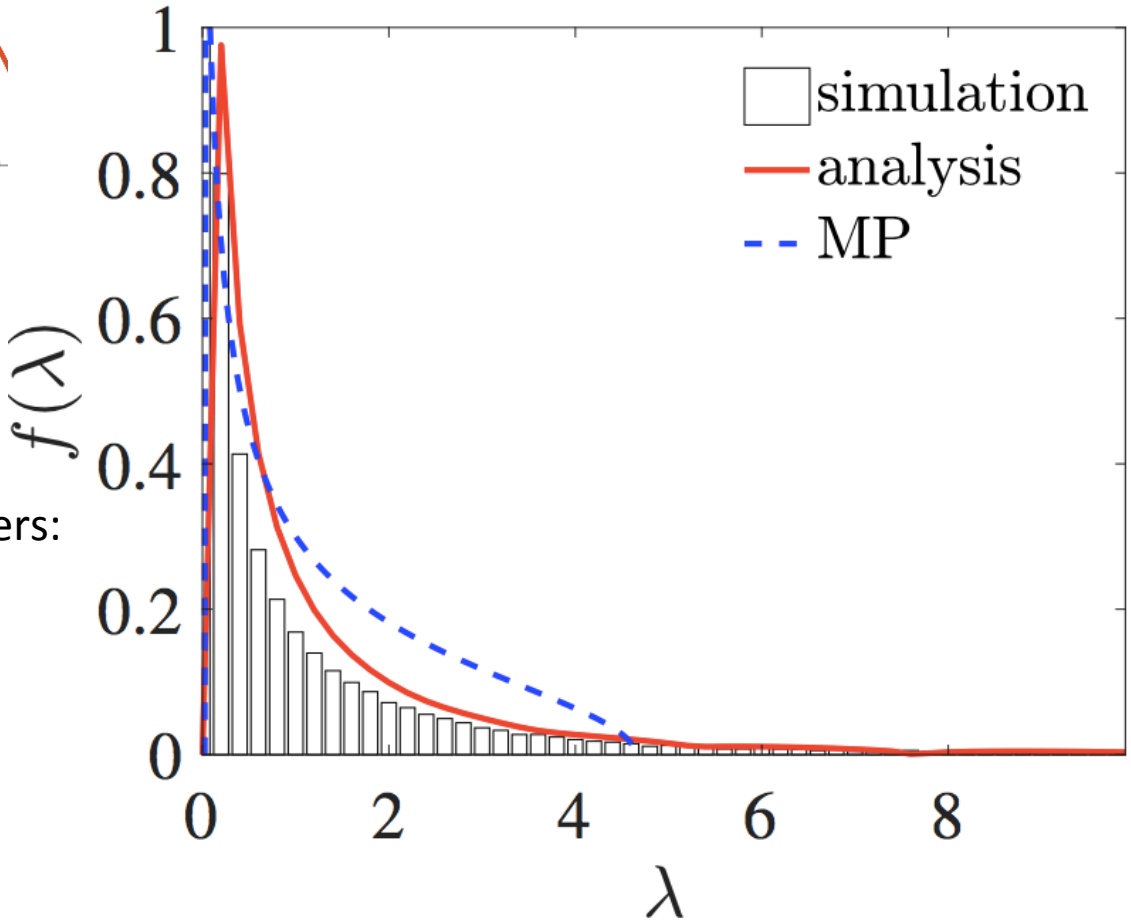
$$S_p = \mathbf{v}_1\mathbf{v}_1^T + \cdots + \mathbf{v}_t\mathbf{v}_t^T, \quad \lambda_1 \leq \cdots \leq \lambda_t \leq \cdots \leq \lambda_{p(q-1)}$$

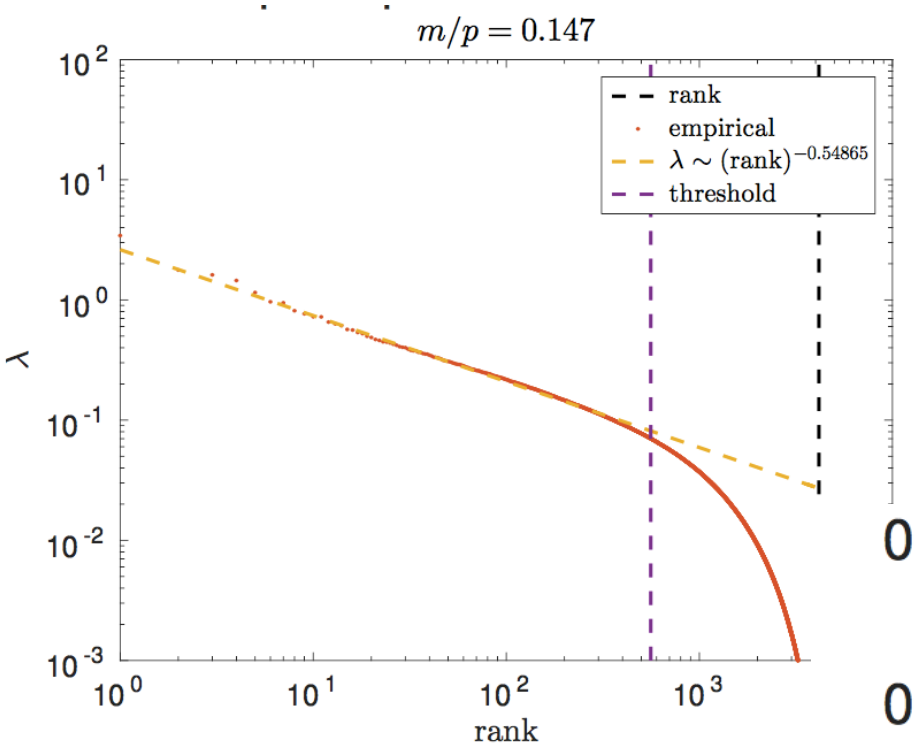


Trypsin

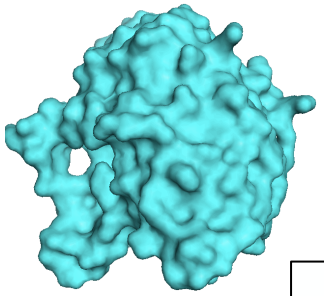


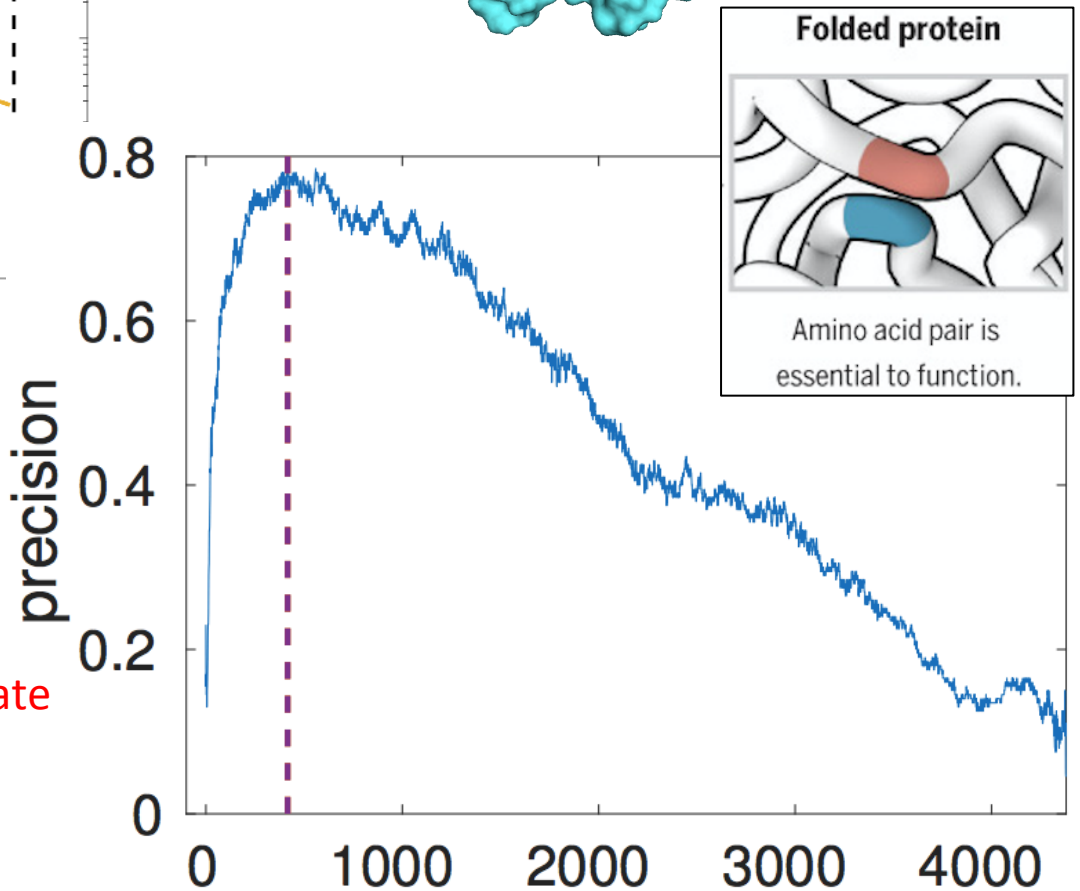
Calculated from the covariance matrix of ~30k sequences. Fit the power law and extract the parameters:





Trypsin



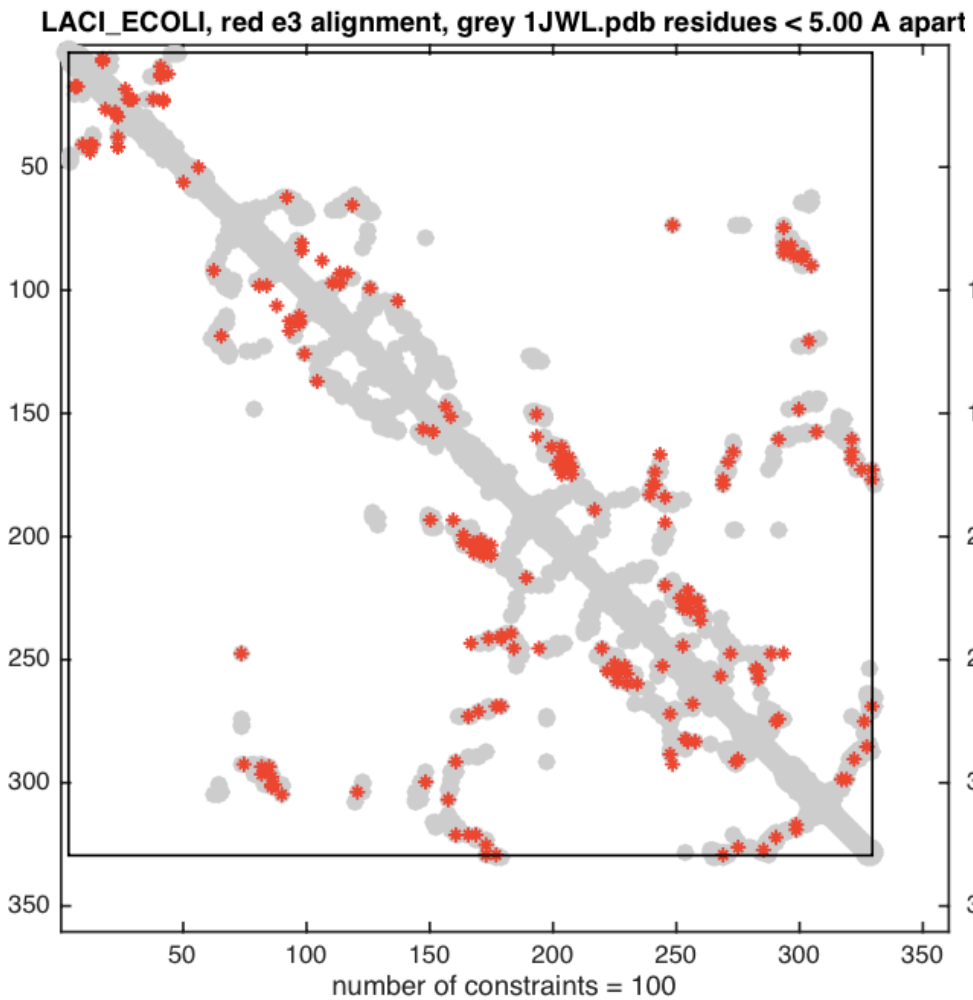


Calculated from the covariance matrix of ~30k sequences.

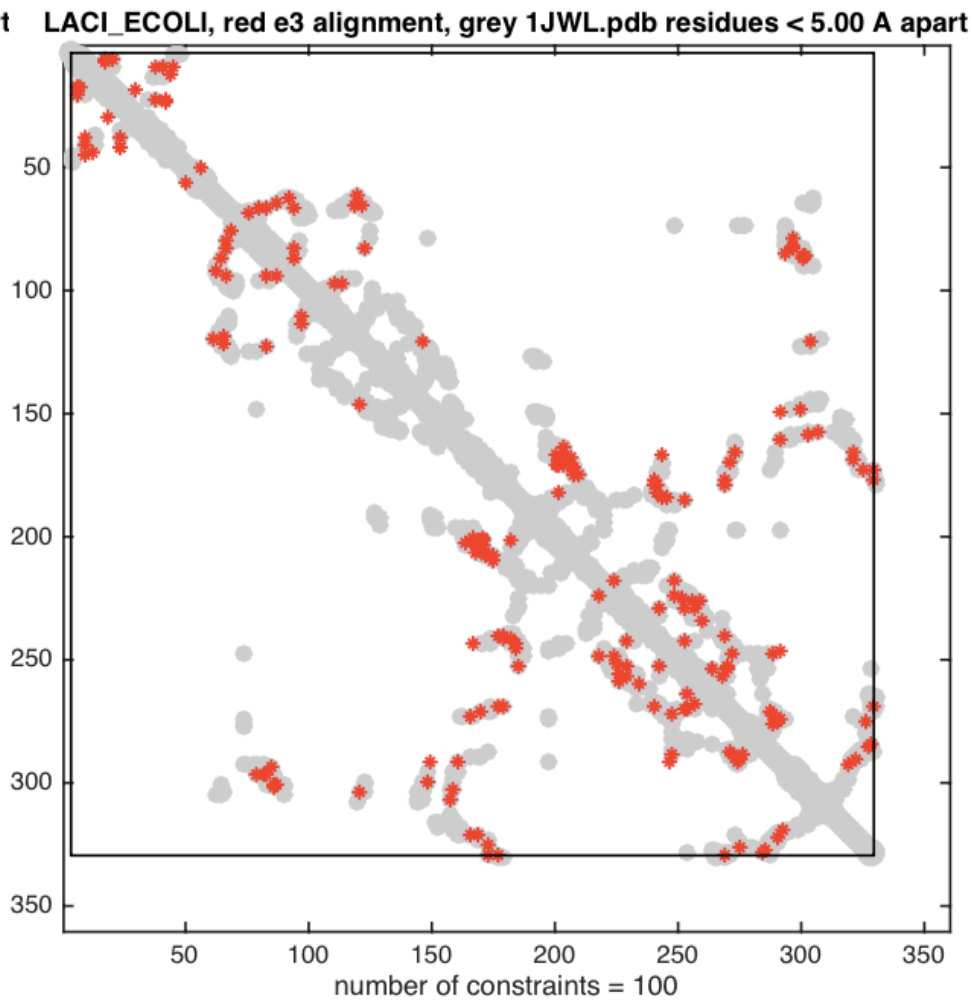
Using Random Matrix Theory to truncate the largest modes recovers contact prediction from raw covariance.

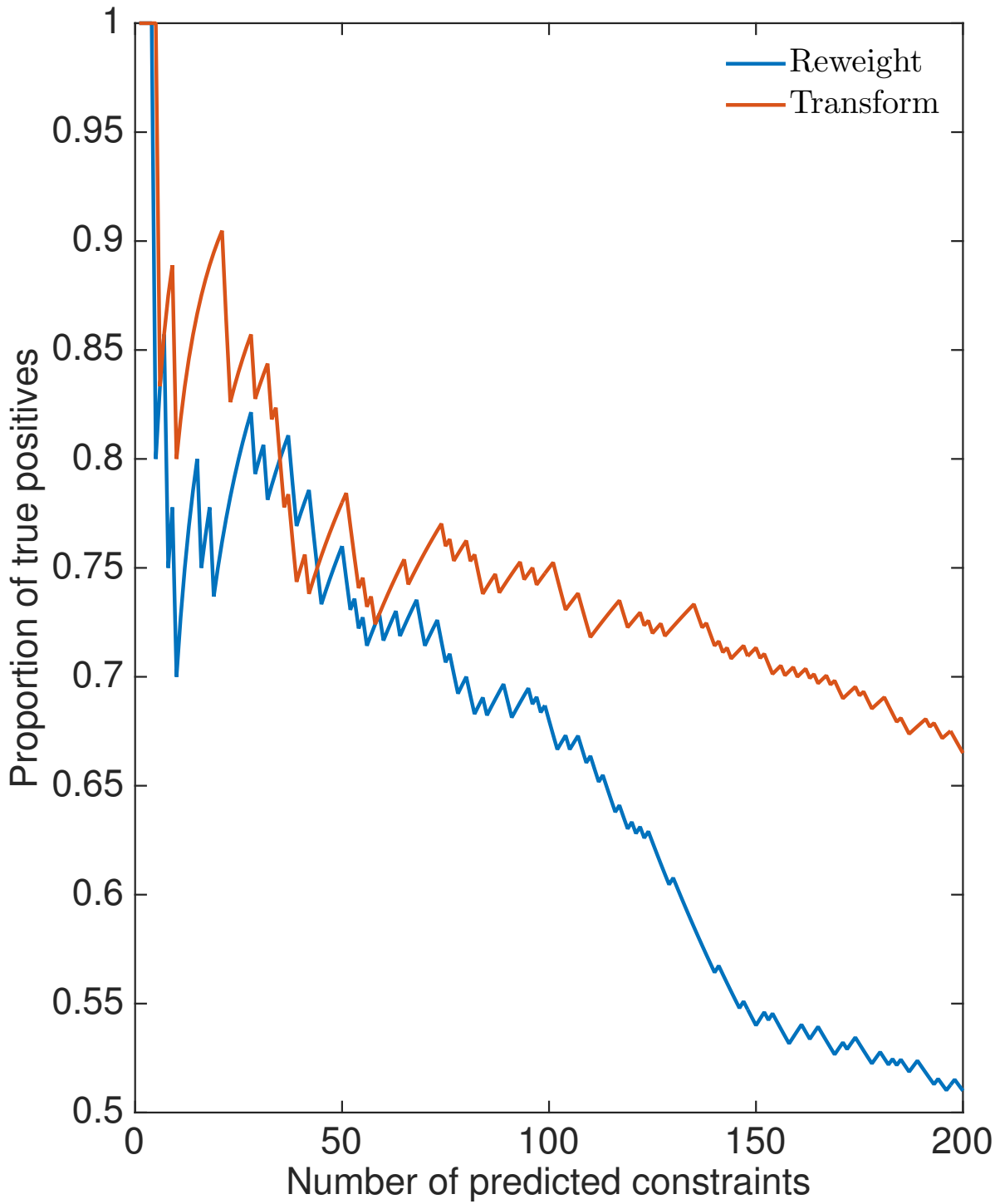
$$S_p = \mathbf{v}_1\mathbf{v}_1^T + \cdots + \mathbf{v}_t\mathbf{v}_t^T, \quad \lambda_1 \leq \cdots \leq \lambda_t \leq \cdots \leq \lambda_{p(q-1)}$$

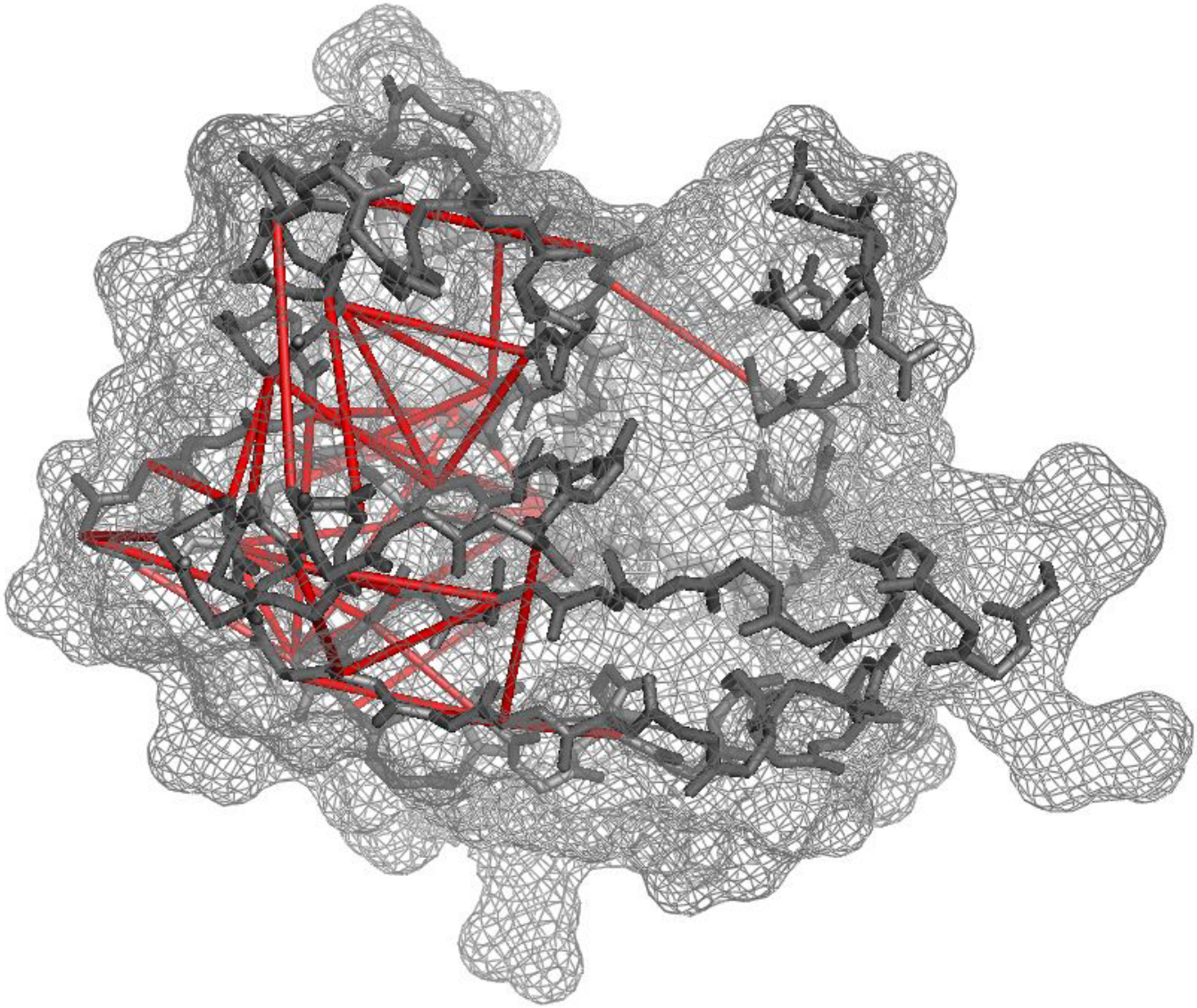
Phylogenetic reweighting



RMT-based linear transformation




Reweight
Transform
Proportion of true positives
Number of predicted constraints
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0
50
100
150
200



SUMMARY

1. Covariance is a very powerful tool for uncovering biological information from large data sets:

   - Protein structure and function
   - Which proteins interact with each other
   - Ligand binding specificity

2. A critically important problem is to understand the limits of these methods, how much data is needed for accurate prediction and so forth.

   - Random Matrix Theory provides a natural framework for solving this problem.