

Coevolution Methods for Predicting Structure from Large Numbers of Genetic Sequences

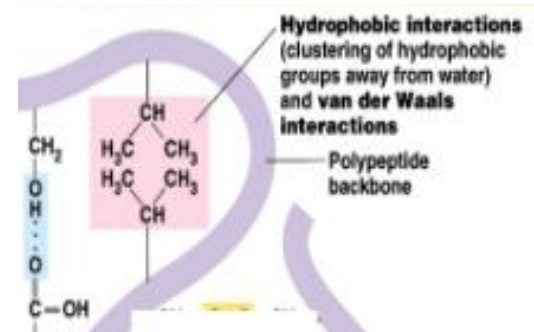
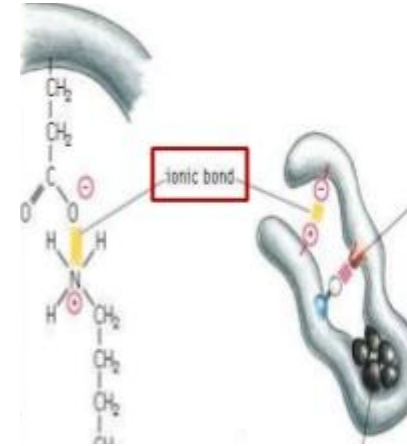
Jason Wang

Kaitlyn Lagattuta

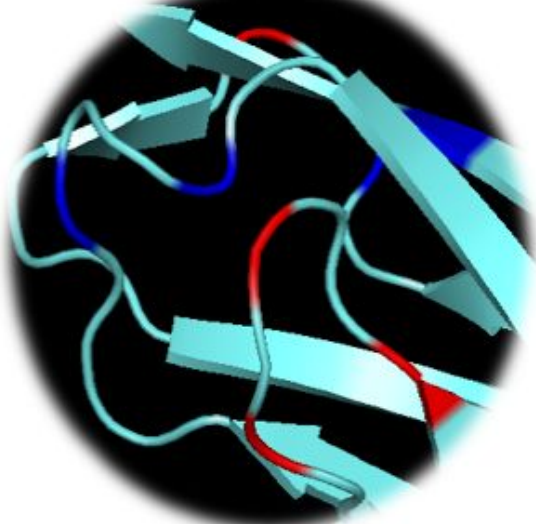
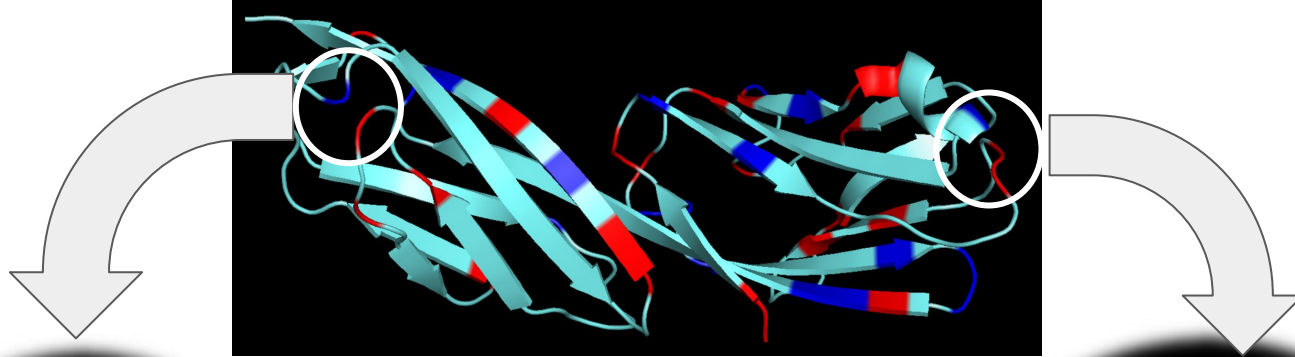
Phillip DiGiacomo

Why Would Protein Residues Coevolve?

- Random mutations occur in protein residues over time. The *ones that survive* create variations in a protein family.
- In order for a protein to accomplish its **function**, it must preserve **structure** relevant to that function.
 - ie pocket shape for ligand binding
- **Interactions between residues** play a large role in determining secondary and tertiary structure of the protein
 - The protein will fold in such a way that negatively and positively charged residues are in contact, hydrophobic residues are in contact in the inside of the protein, etc.
- Residue mutations over time are more likely to persist if they **preserve** the **proximity** of interacting residues-- otherwise the **structure** and **function** of the protein would have failed.



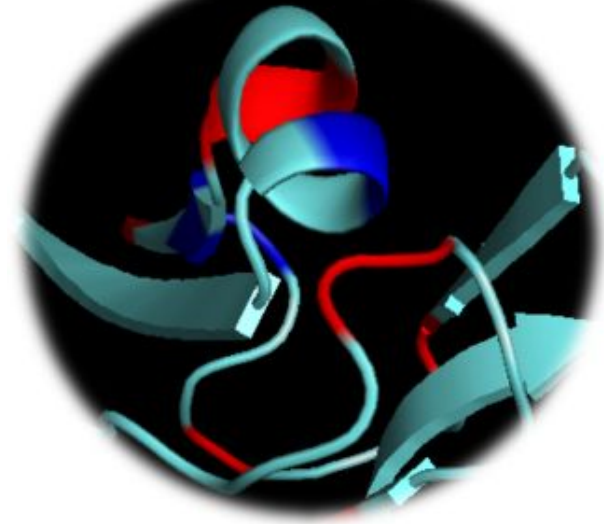
human CD4
primary receptor onto which the HIV virus docks as it enters the T cell (Pymol)



Aspartate (-) at 153 and Arginine (+) at 134

Variations of this protein that *persisted through evolution* most likely **maintained** the shape that occurs when these residues are in **contact**.

These residues are likely to have **co-evolved**.

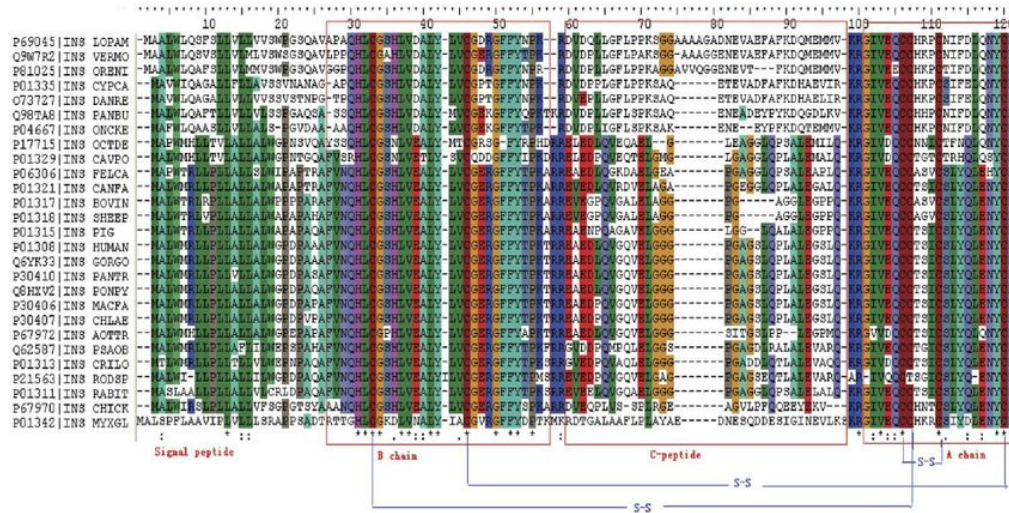


Aspartate (-) at 63 and Lysine (+) at 21

Analytical Approach: Multiple Sequence Alignment (MSA)

- Align sequences of proteins within a given family to identify how protein sequences have changed over time
- Proteins assumed to have an evolutionary relationship by which they share a lineage and are descended from a common ancestor
- Identify differences between nonhomologous sequences to note how they have evolved over time

MSA of insulin protein sequences



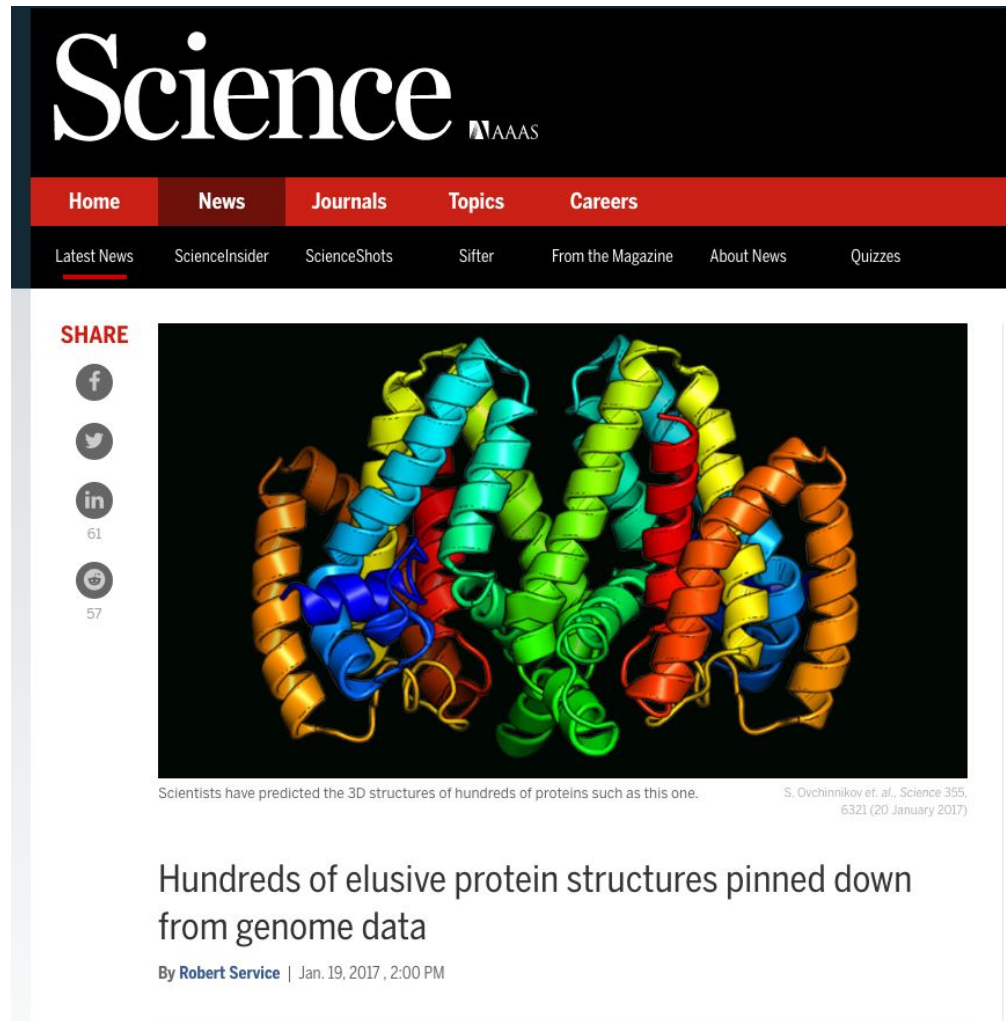
https://www.researchgate.net/figure/234090396_fig1_Figure-1-Multiple-sequence-alignments-of-insulin-protein-sequences-The-species-and

Use of coevolution methods is on the **rise**

Why?

1. The decreasing cost of sequencing

2. The development of improved computational methods



Science AAAS

Home News Journals Topics Careers

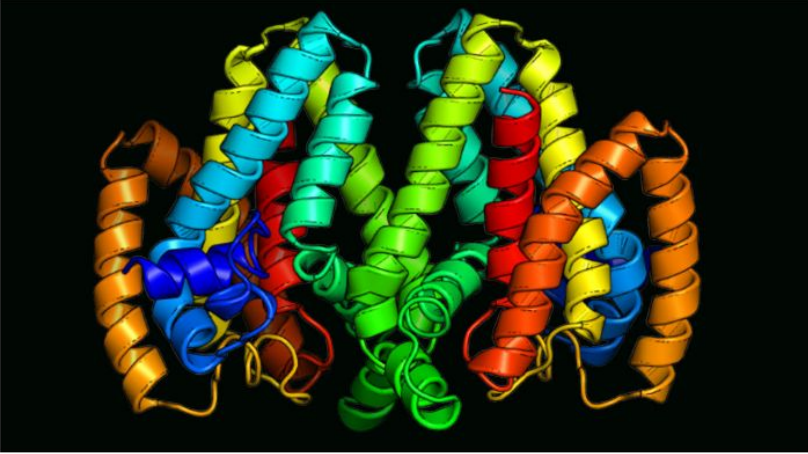
Latest News SciencInsider ScienceShots Sifter From the Magazine About News Quizzes

SHARE

f 61

61

57

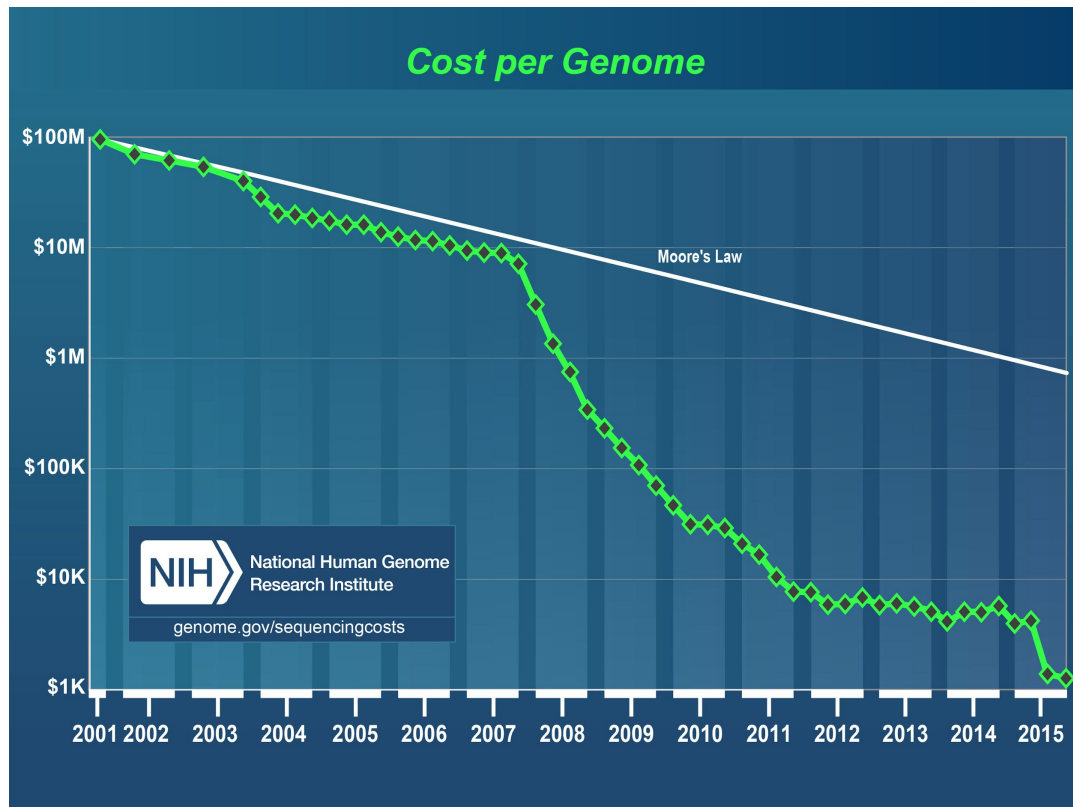


Scientists have predicted the 3D structures of hundreds of proteins such as this one. S. Ovchinnikov et al., *Science* 355, 6321 (20 January 2017)

Hundreds of elusive protein structures pinned down from genome data

By **Robert Service** | Jan. 19, 2017, 2:00 PM

Abundance of sequencing data



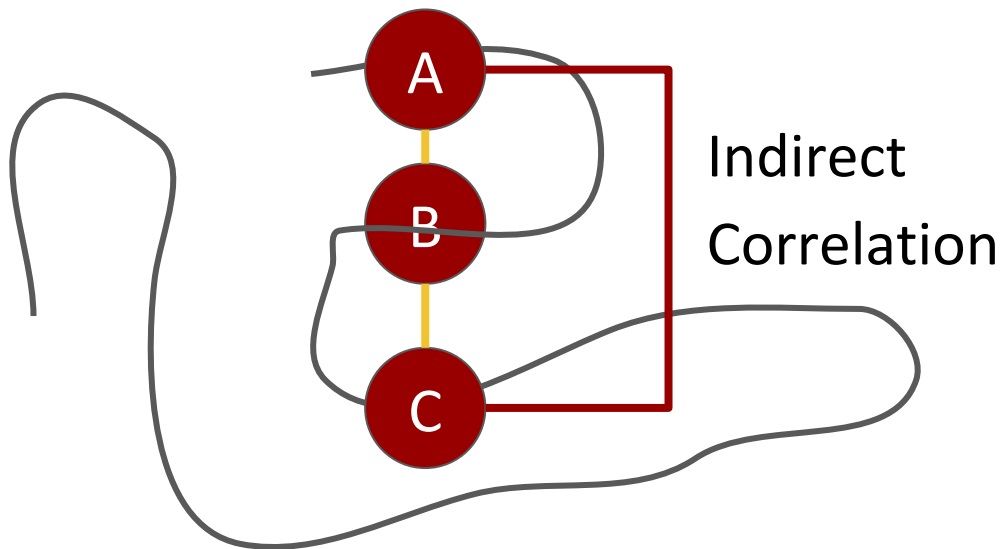
Cost of sequencing is falling faster than Moore's Law

\$1K to sequence human genome

Improved computational **methods**

In particular, methods that have solved the **problem of transitivity!**

New methods find the minimal set of contacts that best explain the sequence data



“Three Dimensional Structures of Membrane Proteins from Genomic Sequencing”

Hopf et al. *Cell* 149 (2012)

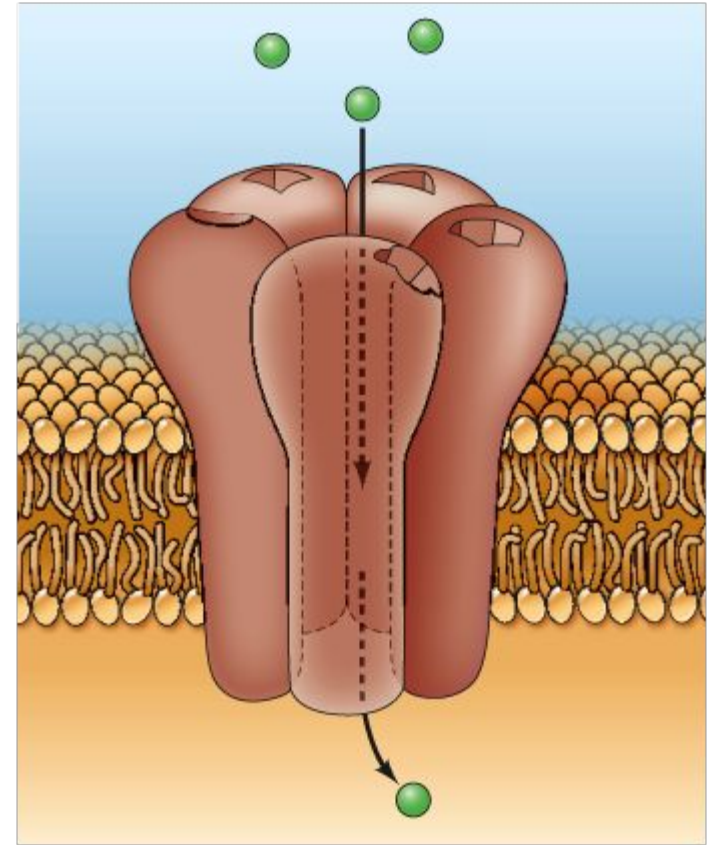
Transmembrane Proteins

Span entirety of a biological membrane

Function as gateways for transport of substances across the membrane

~50% of all drug targets contain a membrane domain

GABA Receptor



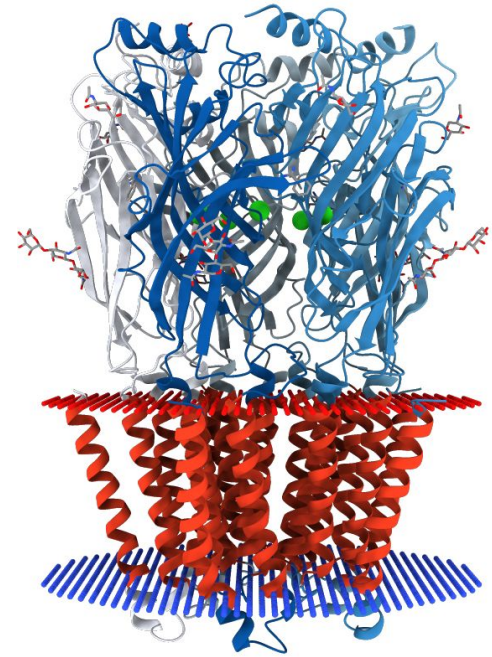
Knowledge of **3D structure** enables discovery

Facilitate characterization of molecular mechanisms

Accelerate development of drugs

Better understand protein function

GABA Receptor



But 3D structures of most transmembrane proteins remain **unknown**

Protein Structure Prediction: 1D sequence → 3D structure

Template-based modeling covers at “max 10% of all human transmembrane proteins”

Homologous protein
with known 3D structure



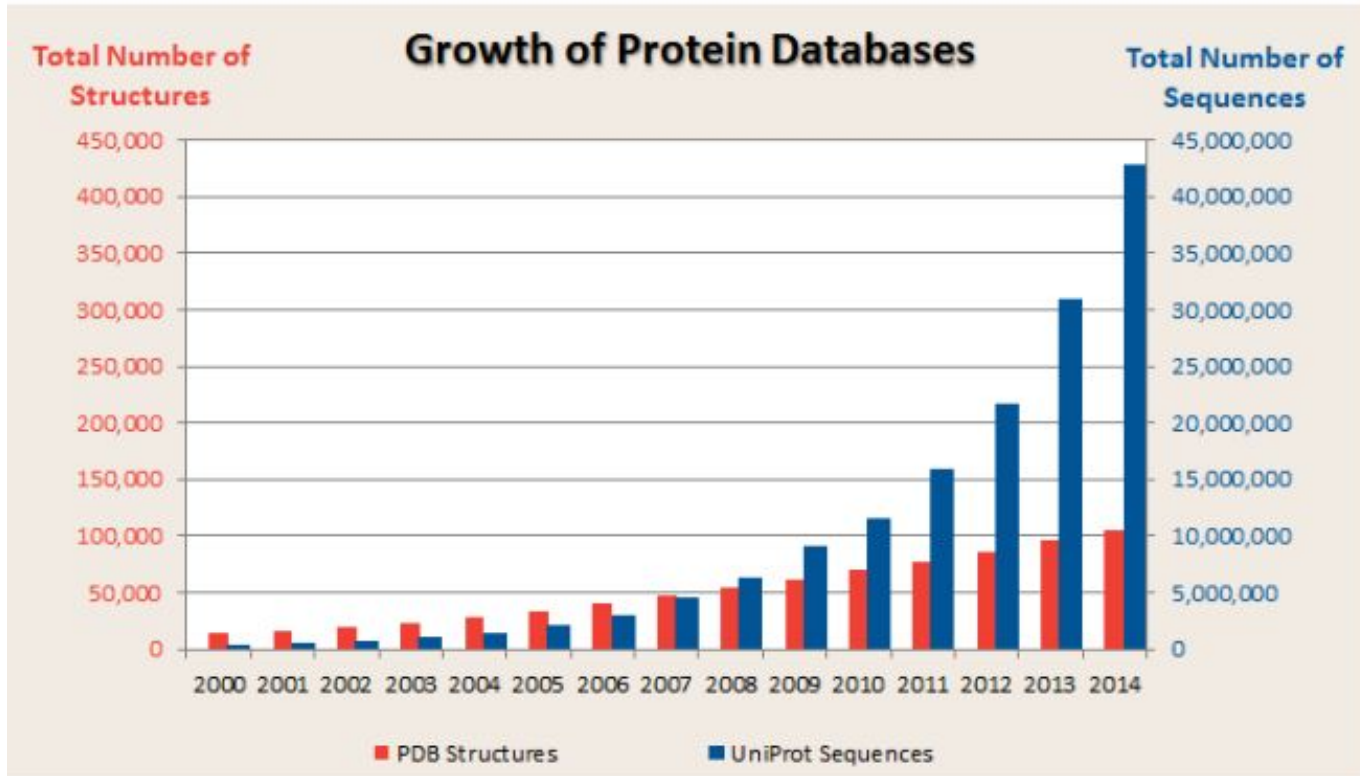
Existing **ab initio methods** are inaccurate and computationally expensive

ALAKYMKRDTENVNDKLRGL...

Protein Sequence

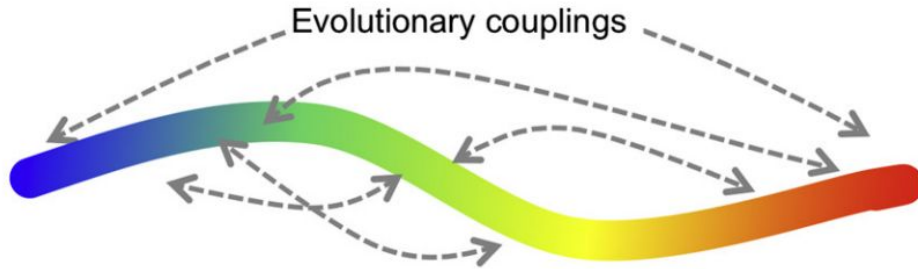


Can we extract further information from **sequence data**?

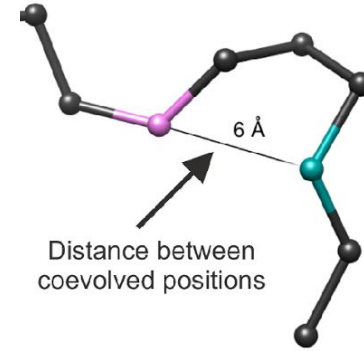


Yes, via **Coevolution!**

Key idea: use coevolving residue pairs as distance constraints for improved ab initio folding



Pairwise distance constraints



Assumption: coevolving residue pairs are in direct contact

EVfold_membrane

Improved de novo prediction of 3D structures of transmembrane proteins using information from evolutionary constraints

Unlike previous approaches, uses neither fragments nor homologous 3D structures

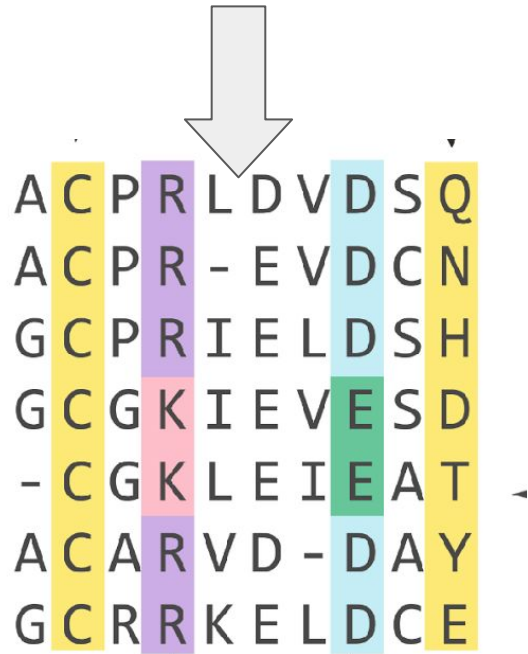
Input: 1D protein sequence



Output: predicted 3D structure (all atom-coordinates)

Build MSA

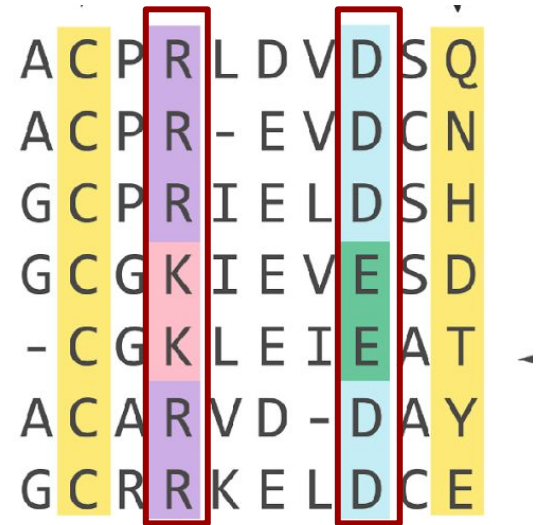
Query Sequence:
ACPRLDVDSQ...



Entropy Maximization

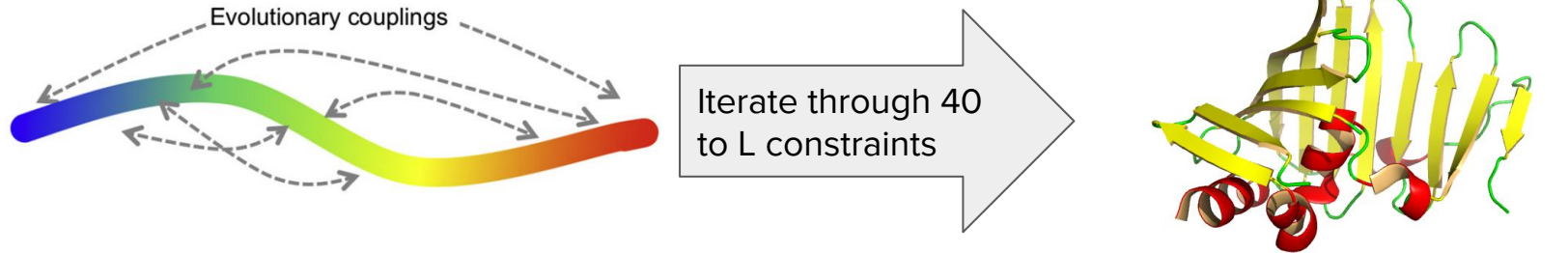
Find residue pairs that best explain the data

Filter based on predicted secondary structure and membrane topology



Ab Initio Folding with Evolutionary Distance Constraints

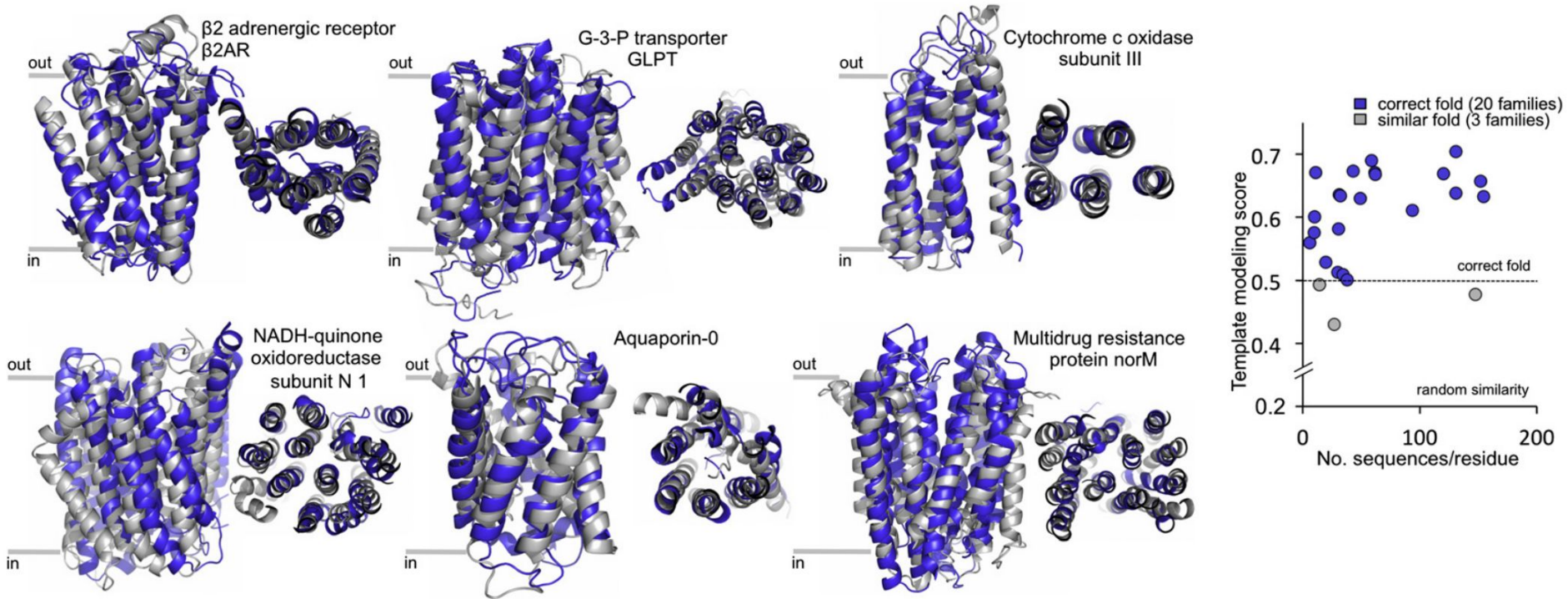
Fold fully extended polypeptide chain ab initio using resulting set of evolutionary distance constraints



Choose top model based on 1) lipid accessibility of residues, 2) quality of secondary structure formation

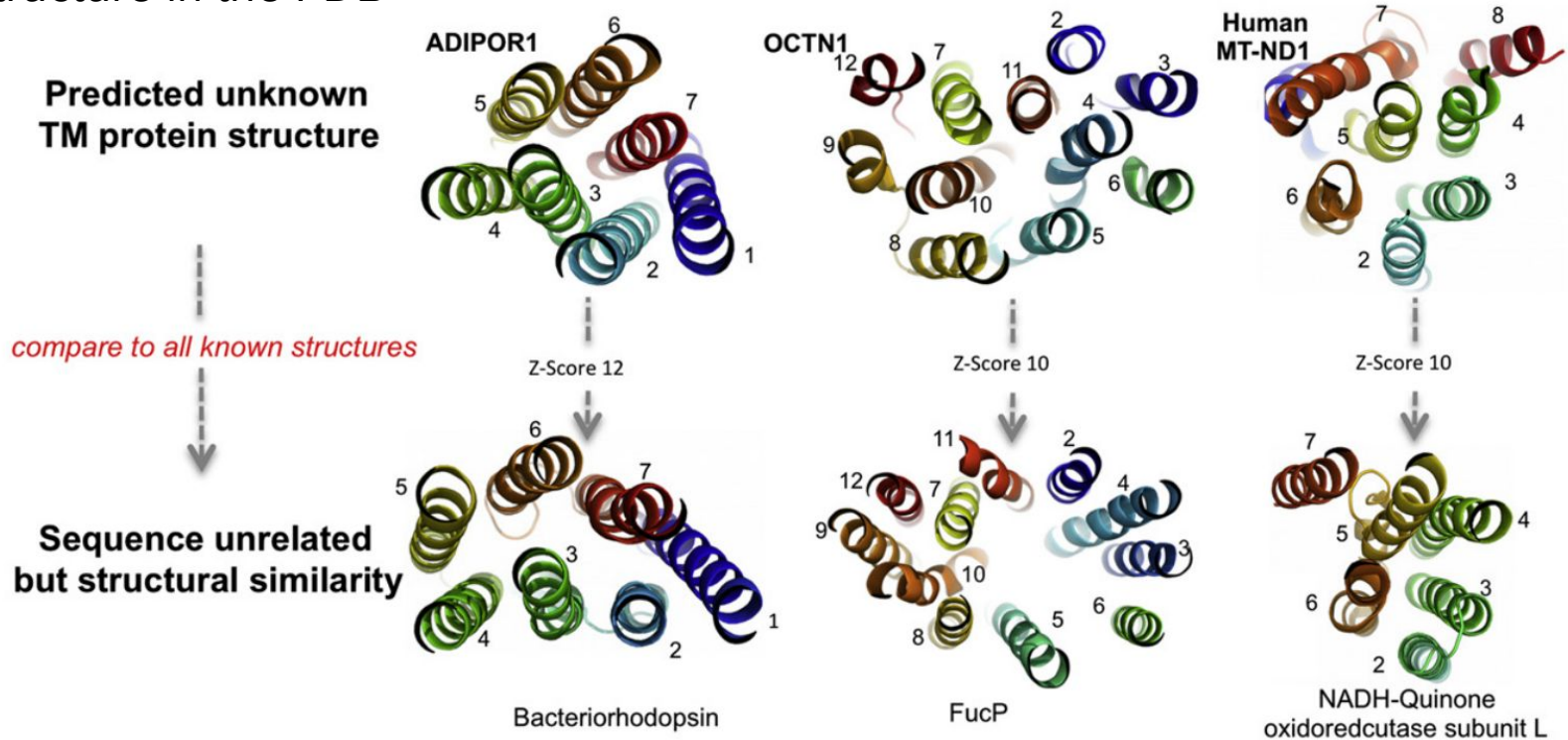
Evaluation on **known** 3D structures

Use template modeling score (TM-score) to compare predicted with experimental structure



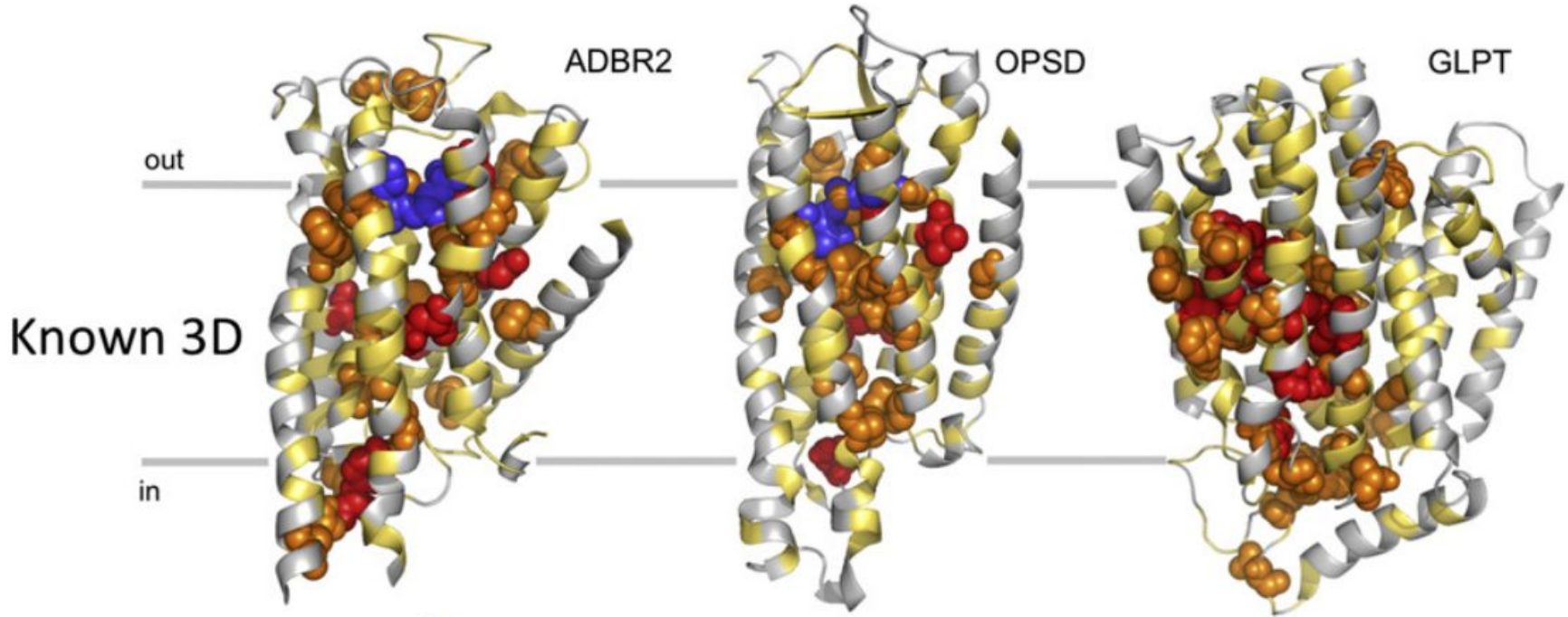
Evaluation on **unknown** 3D structures

Use DALI Z-score to compare predicted model to most structurally similar known 3D structure in the PDB



What do **evolutionary constraints** really represent?

Is the strength of evolutionary coupling on a residue an indication of its **functional importance** within the protein?



Strengths of EVfold_membrane

Takes ~1-2 min of computing time per model on a single CPU (household laptop) for protein of average size; reduces the conformational search space

Can be applied to large protein sizes up to 14 helices; previous methods were limited to 4-7 helices

Input **requires neither homologous 3D structures nor database fragments**

Substantially **improves prediction accuracy**

Study **Limitations**

Ranking of predicted models is **poorly explained**

The evaluation of predictions for proteins of unknown 3D structure uses the **most structurally similar protein** in the entire PDB

Validation was done on proteins with >1000 known sequences per family and high coverage (sequence/residue ratio); **unlikely to be scalable** for all proteins

Fundamentally **limited by the accuracy of the statistical method** used to identify residue pairs in direct contact

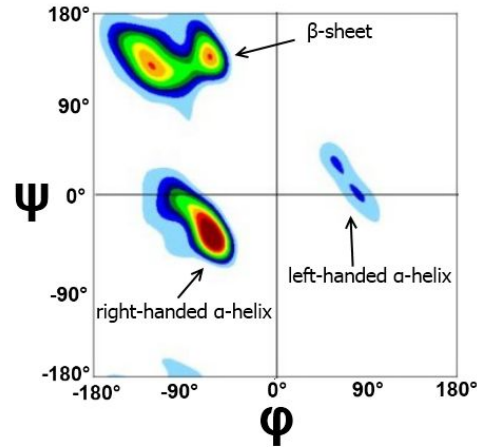
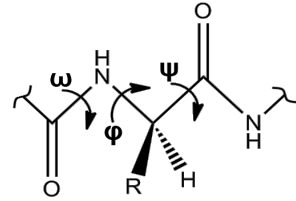
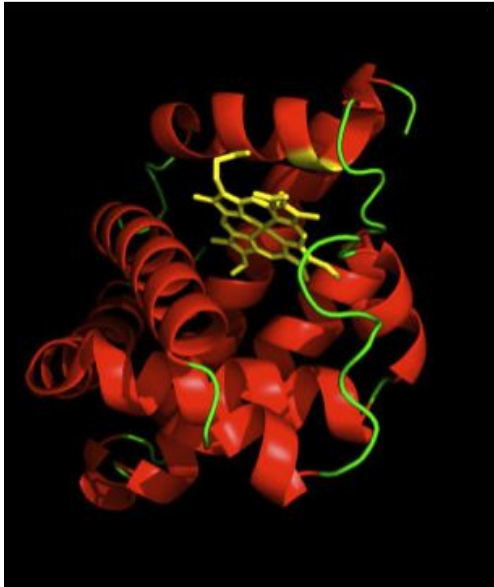
“Improved Contact Predictions Using the Recognition of Protein Like Contact Patterns”

Deep learning meets statistical inference contact prediction

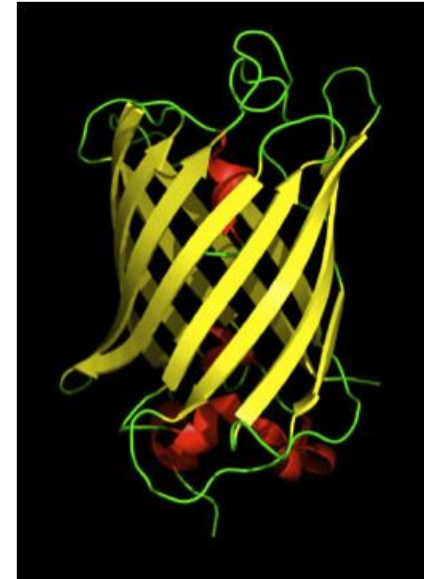
Skwark et al. *PLoS Computational Biology* 10.11 (2014)

A Priori knowledge: Common Contact Patterns

α -Helices



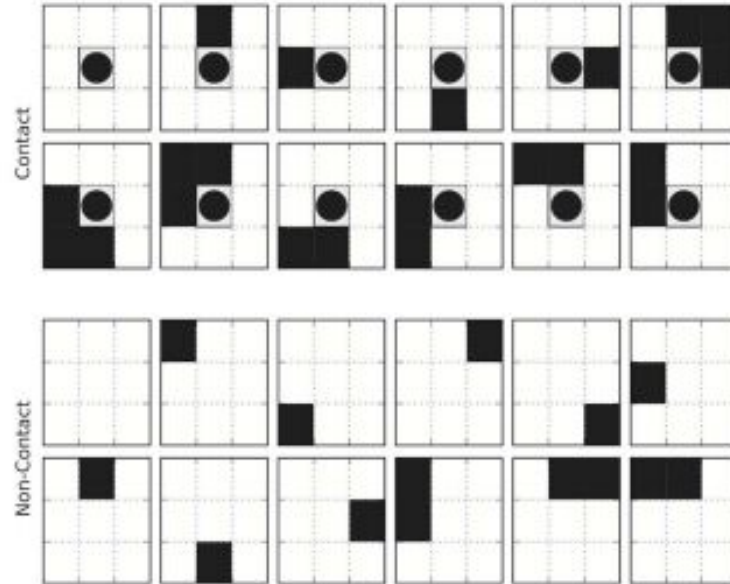
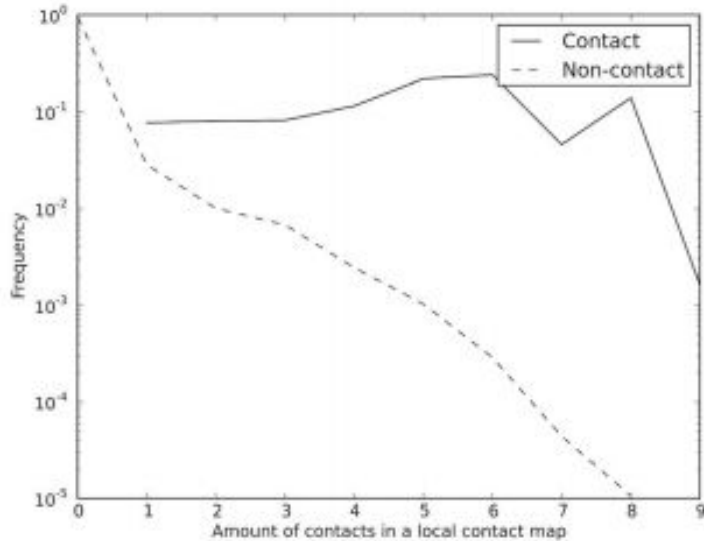
β -Sheets



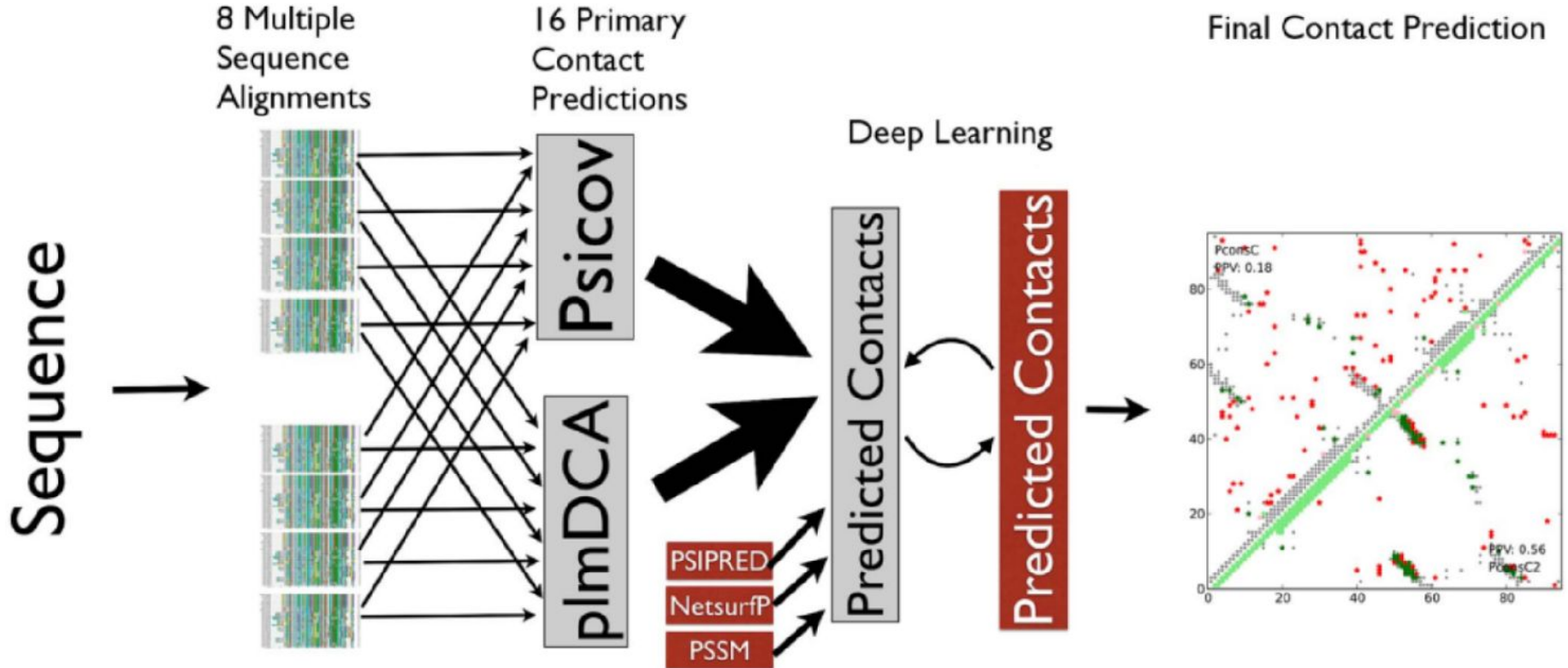
A Priori Knowledge Can Guide Prediction

- Technology in this field attempts to predict a series of residues in contact that describe overall shape of protein
- We know that proteins tend to take on certain structures, and these structures correspond to contact patterns
- Idea: pay attention to *surrounding contacts*, and if they indicate a commonly observed contact pattern
- It's a pretty safe bet to generalize to the pattern!

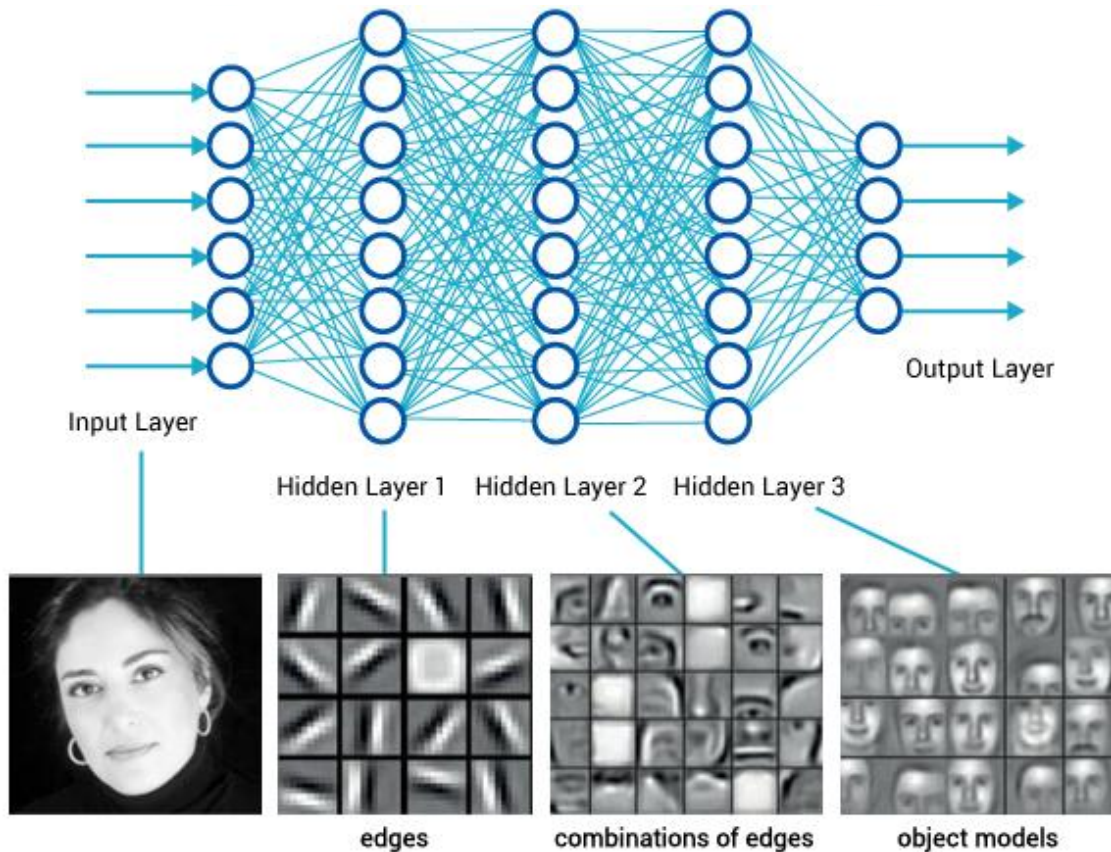
Strength: successive layer considers nearby contacts in receptive field



PonsC2 Pipeline



Why deep learning? Finds general abstractions in data



Five additional features for input data

1. Contact prediction from PSICOV and plmDCA
2. Amino acid separation
3. Predicted secondary structure
4. Position specific score for amino acids
5. Predicted relative surface accessibility

1. PSICOV & pIMDCA: contact predictions from sequence correlation

BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 2 2012, pages 184–190
doi:10.1093/bioinformatics/btr638

Sequence analysis

Advance Access publication November 17, 2011

PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments

David T. Jones^{1,*}, Daniel W. A. Buchan¹, Domenico Cozzetto¹ and Massimiliano Pontil²

2012:

- sparse inverse covariance estimation
- corrections for phylogenetic and entropic correlation noise



Fast pseudolikelihood maximization for direct-coupling analysis of protein structure from many homologous amino-acid sequences

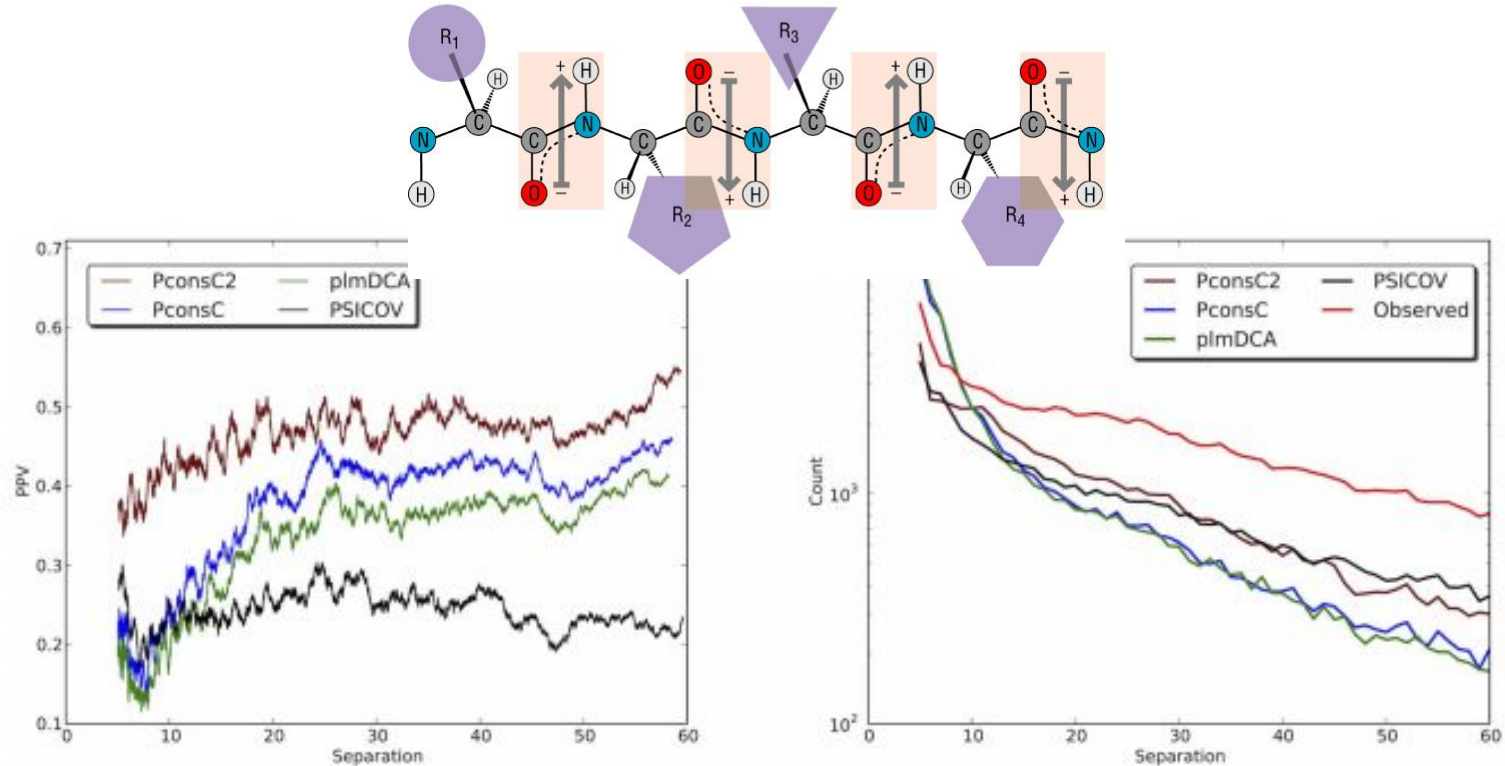
Magnus Ekeberg^{a,b,*}, Tuomo Hartonen^{c,d,1}, Erik Aurell^{b,c,e}



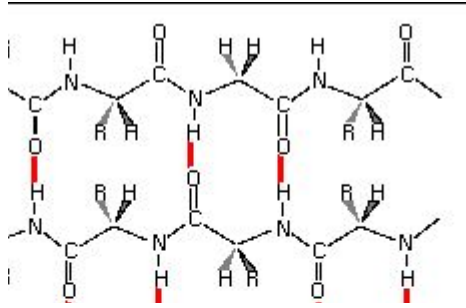
2014:

- correlation can arise from firsthand interaction but can also be network-propagated via intermediate sites
- separates direct from indirect interactions

2. Amino acid separation: improvement at all distances

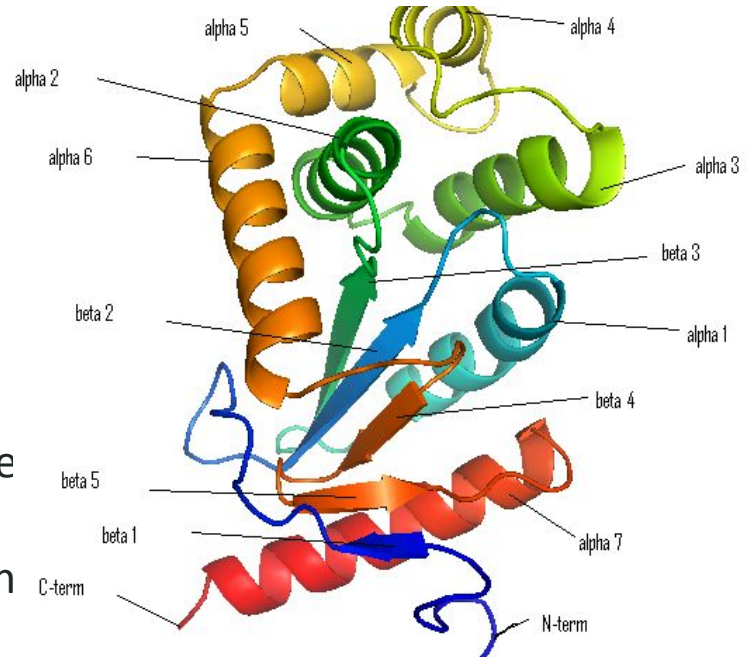


3. Secondary structure: improvement largest for β -sheets



- Contacts between β -sheets primarily mediated through backbone H bonds
- Side chains not under co-evolutionary pressure

- Least improvement for α proteins
- Performance highest for mixed α/β proteins
- General limitation in the field: different technologies catering to specific types of proteins.
- Does this imply over-fitting, or risk in generalization



4. Position specific score matrix incorporates likely evolutionary changes in family

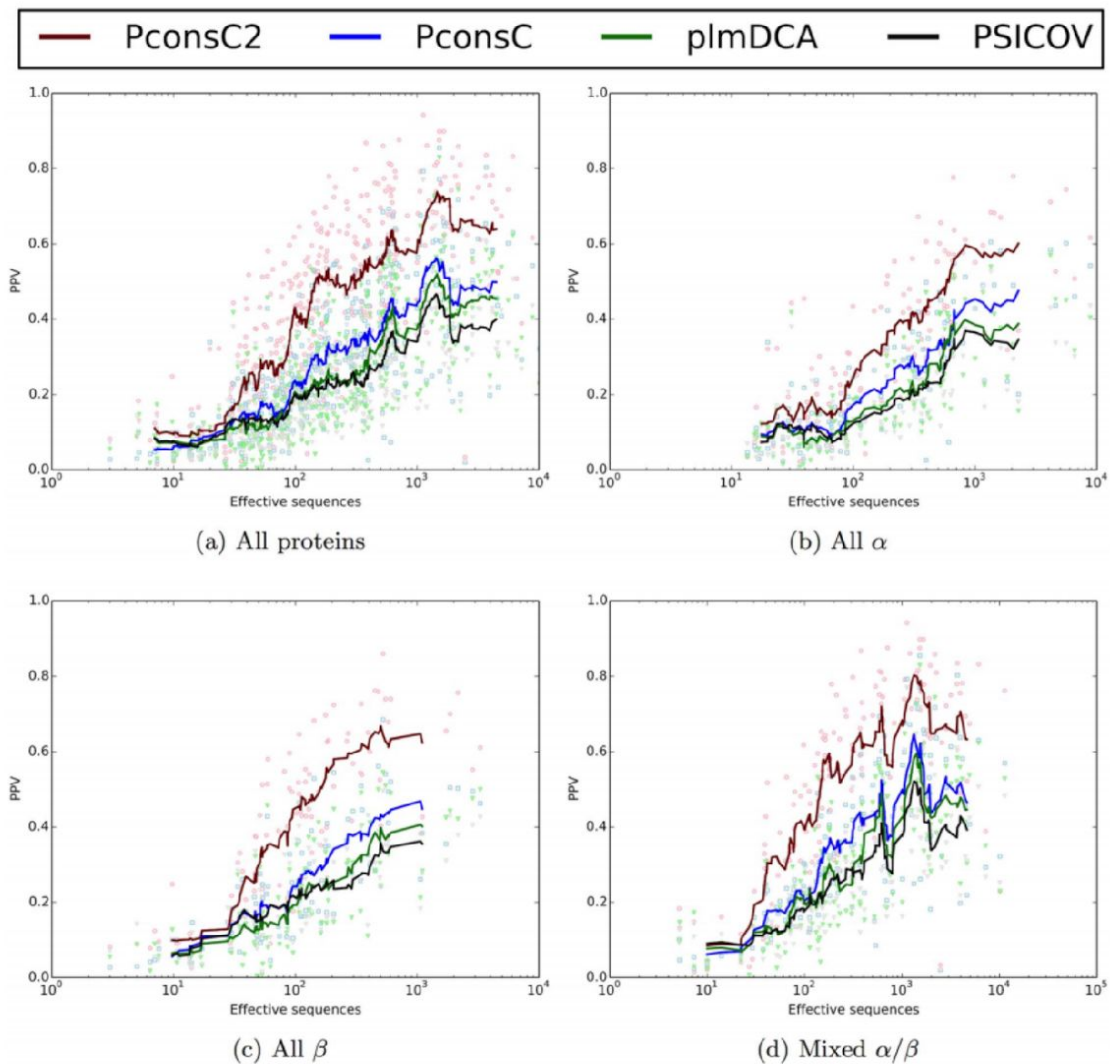
Amino acid	1	2	3	4	5	6	7
N	0.541	-0.061	-0.061	-0.061	-0.061	-0.061	-0.061
T	-0.061	0.240	0.240	-0.061	-0.061	-0.061	-0.061
E	-0.061	-0.061	0.240	-0.061	0.416	0.240	-0.061
G	-0.061	-0.061	-0.061	0.416	0.240	-0.061	-0.061
W	-0.061	-0.061	-0.061	-0.061	-0.061	0.240	0.240
I	-0.061	0.416	-0.061	-0.061	-0.061	-0.061	0.240
H	-0.061	-0.061	-0.061	-0.061	-0.061	-0.061	-0.061
R	-0.061	-0.061	-0.061	0.240	-0.061	-0.061	-0.061
A	-0.061	-0.061	0.240	-0.061	-0.061	-0.061	-0.061
C	-0.061	-0.061	-0.061	-0.061	-0.061	0.240	0.240
...

The matrix assigns positive scores to residues that appear more often than expected by chance and negative scores to residues that appear less often than expected by chance.

Strength: significant prediction improvement in small protein families

- “vast majority of protein families may never reach the thousands of members that are needed for successful predictions”

- close homologs don't provide as much co-variation as distantly related proteins– number of *efficient* sequences takes redundancy into account



Limitation: predicting loop contacts

Table 5. PPV values at $L=1$ for contacts between different secondary structure classes.

Structural category	Real	PSICOV	plmDCA	PconsC	PconsC2
$\alpha - \alpha$	11%	0.23 (15%)	0.22 (17%)	0.28 (16%)	0.40 (15%)
$\alpha - \beta$	8%	0.31 (9%)	0.46 (7%)	0.48 (8%)	0.54 (7%)
$\alpha - \text{loop}$	16%	0.18 (21%)	0.22 (21%)	0.27 (21%)	0.37 (13%)
$\beta - \beta$	22%	0.41 (13%)	0.51 (12%)	0.52 (13%)	0.57 (34%)
$\beta - \text{loop}$	22%	0.20 (23%)	0.22 (21%)	0.26 (22%)	0.46 (16%)
loop -loop	21%	0.19 (18%)	0.17 (22%)	0.24 (19%)	0.30 (16%)
ALL		0.24	0.26	0.31	0.46

Limitation: predicting loop contacts

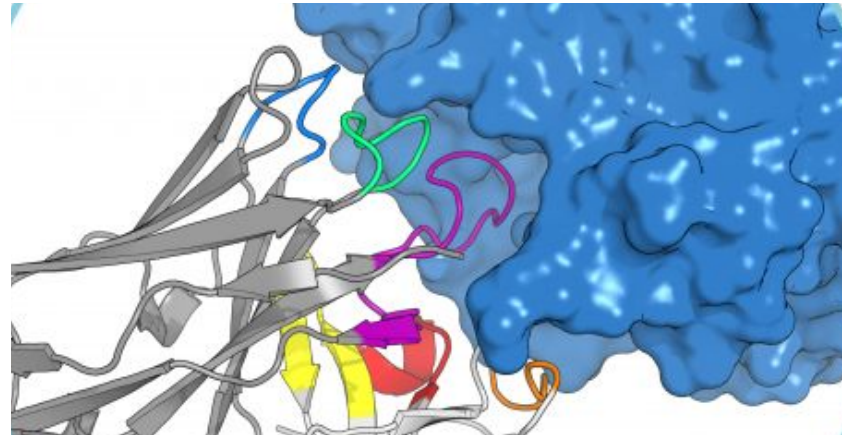
- **Variation in loops:**

- Less *a priori* knowledge of observed contact patterns

- **Ligand binding research:**

- “In many cases, a protein’s function depends on the ability of its loops to adopt different conformations” -Oxford Protein Informatics Group

- Could we use ligands that drug binds as further input data?



PconsC2 performs better:

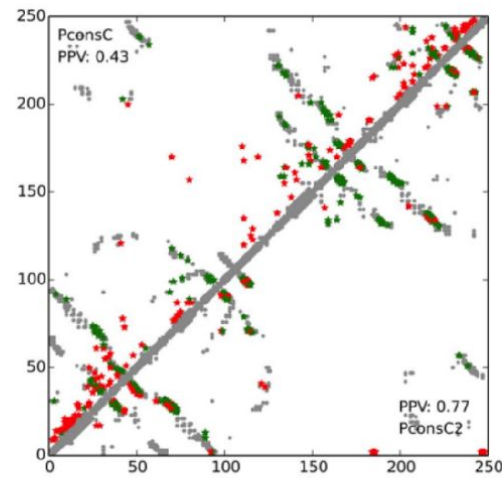
(a) Predominantly β conformation. PconsC2 filtered out spurious predictions at termini from PconsC

(b) limited overlap between individual predictions– PconsC2 reconciles conflicts

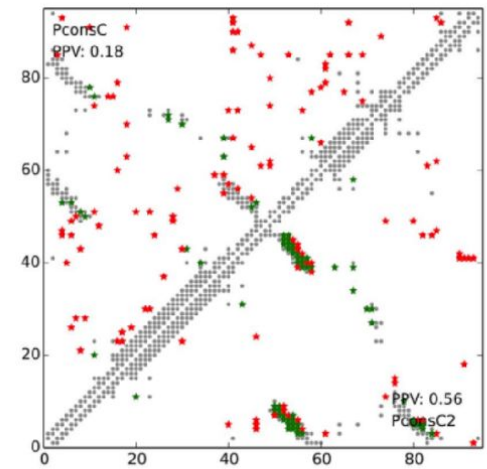
PconsC2 performs worse:

(c) protein contains few secondary structures

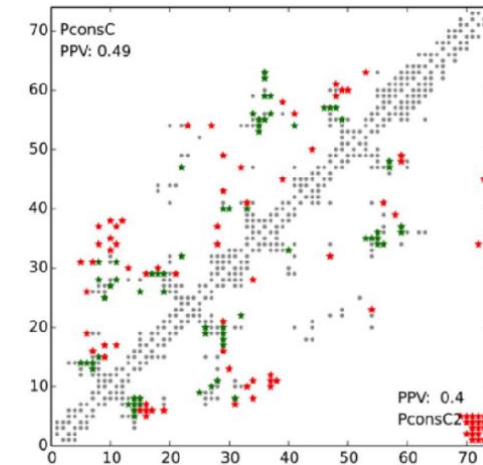
(d) PconsC2 invents contacts between secondary structures



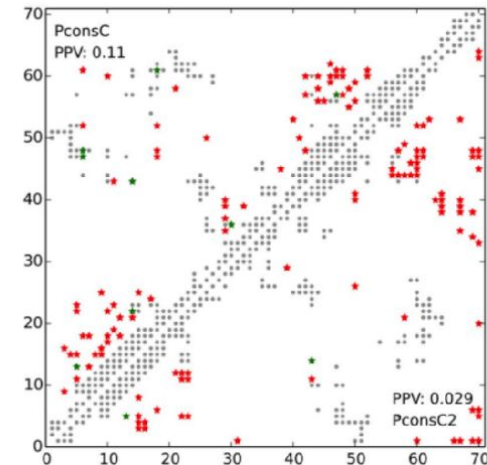
(a) liz4A



(b) 3znuG



(c) 1wigA



(d) 1imxA

“Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model”

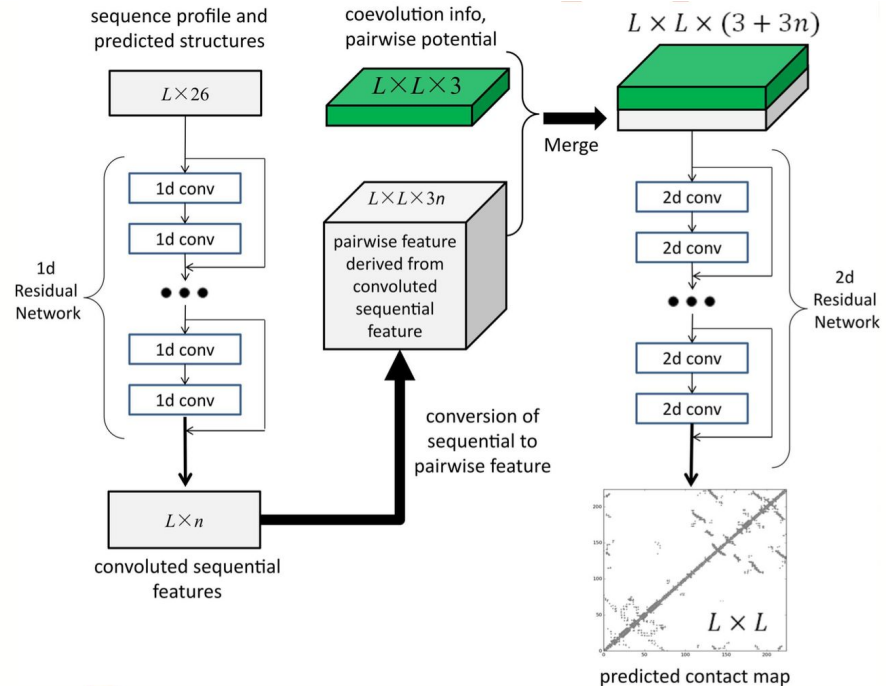
Wang et al. *PLOS Computational Biology* 13.1 (2017)

General Idea

- Improve upon MSA co-evolution techniques for protein contact prediction by using deep learning techniques
- Combine various types and large quantities of input data
 - Protein sequence profile
 - Predicted 3-state secondary structure and solvent accessibility
 - Direct co-evolutionary information, generated by CCMpred
 - Mutual information and pairwise potential

Residual Neural Network Method

- Novel deep learning model formed by two deep residual neural networks
- Integrates evolutionary coupling (EC) and sequence conservation information
- Approaches problem of protein contact map prediction like pixel-level image labeling



Training and Testing

- Trained on subset of PDB25 proteins with solved structures
- Redundancy removal strategy
- Tested on publicly available CASP11 and CAMEO targets, as well as many membrane proteins
- Tested against currently available DCA methods and supervised machine learning methods
 - PSICOV, Evfold, CCMpred, plmDCA, Gremlin, and MetaPSICOV

Results

Analyzed performance of the method in a number of contexts:

1. General performance against previous methods
2. As a function of # of sequence homologs
3. Contact-assisted protein folding
4. Difficult case studies

1. General performance against prior methods

Method	Short				Medium				Long			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
EVfold	0.17	0.13	0.11	0.09	0.23	0.19	0.13	0.10	0.25	0.22	0.17	0.13
PSICOV	0.20	0.15	0.11	0.08	0.24	0.19	0.13	0.09	0.25	0.23	0.18	0.13
CCMpred	0.22	0.16	0.11	0.09	0.27	0.22	0.14	0.10	0.30	0.26	0.20	0.15
pImDCA	0.23	0.18	0.12	0.09	0.27	0.22	0.14	0.10	0.30	0.26	0.20	0.15
Gremlin	0.21	0.17	0.11	0.08	0.27	0.22	0.14	0.10	0.31	0.26	0.20	0.15
MetaPSICOV	0.56	0.47	0.31	0.20	0.53	0.45	0.32	0.22	0.47	0.42	0.33	0.25
Our method	0.67	0.57	0.37	0.23	0.69	0.61	0.42	0.28	0.69	0.65	0.55	0.42

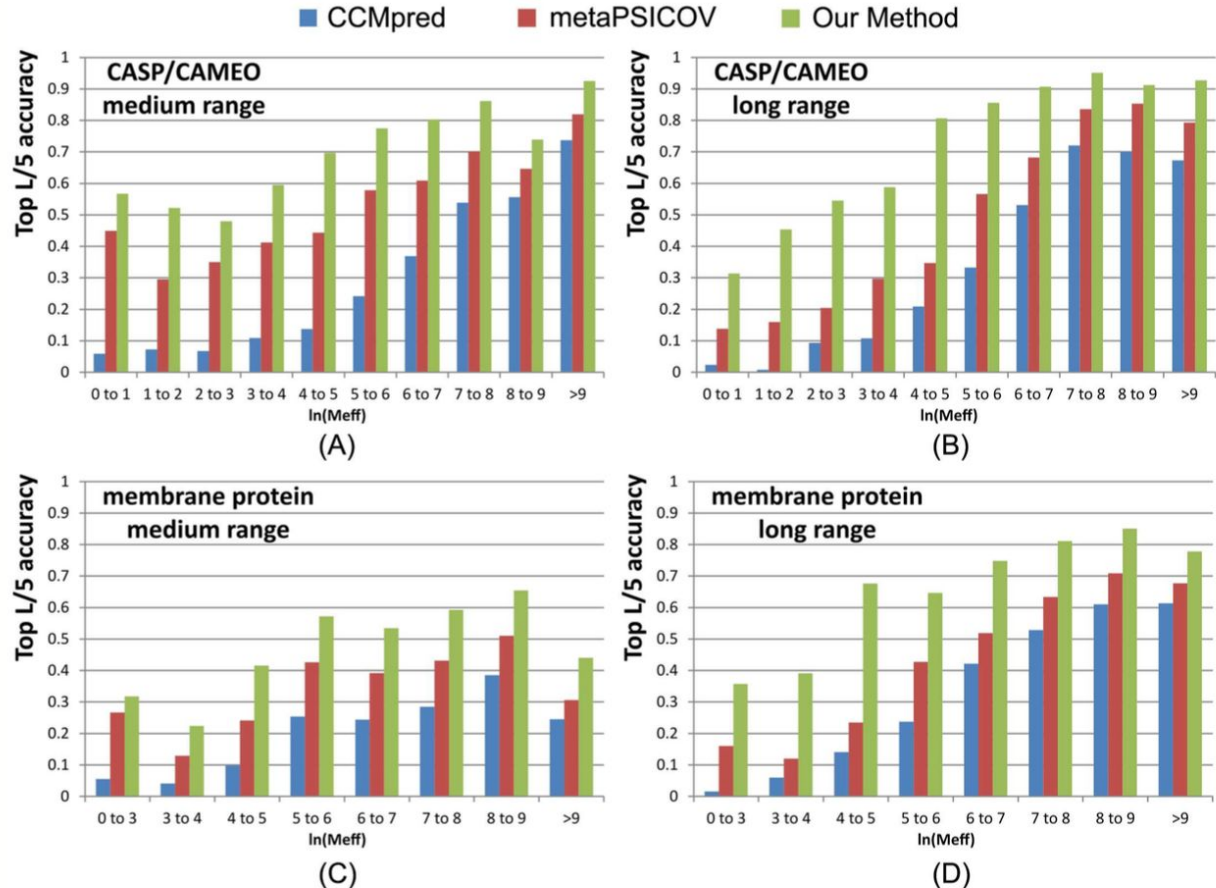
Contact prediction accuracy on 76 past CAMEO hard targets

Contact prediction accuracy on 398 membrane proteins

Method	Short				Medium				Long			
	L/10	L/5	L/2	L	L/10	L/5	L/2	L	L/10	L/5	L/2	L
EVfold	0.16	0.13	0.09	0.07	0.28	0.22	0.13	0.09	0.44	0.37	0.26	0.18
PSICOV	0.22	0.16	0.10	0.07	0.29	0.21	0.13	0.09	0.42	0.34	0.23	0.16
CCMpred	0.27	0.19	0.11	0.08	0.36	0.26	0.15	0.10	0.52	0.45	0.31	0.21
pImDCA	0.26	0.18	0.11	0.08	0.35	0.25	0.14	0.09	0.51	0.42	0.29	0.20
Gremlin	0.27	0.19	0.11	0.07	0.37	0.26	0.15	0.10	0.52	0.45	0.32	0.21
MetaPSICOV	0.45	0.35	0.22	0.14	0.49	0.40	0.27	0.18	0.61	0.55	0.42	0.30
Our method	0.60	0.46	0.27	0.16	0.66	0.53	0.33	0.22	0.78	0.73	0.62	0.47

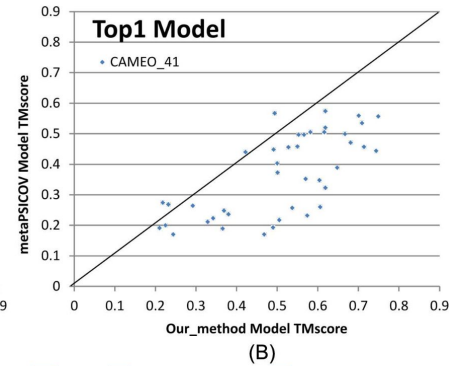
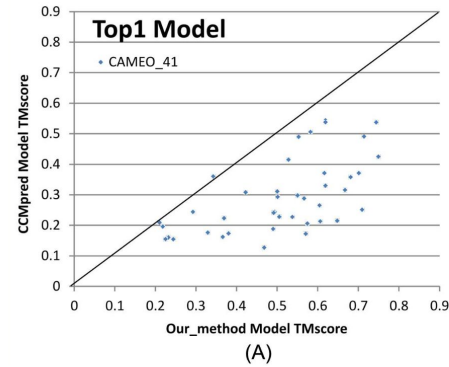
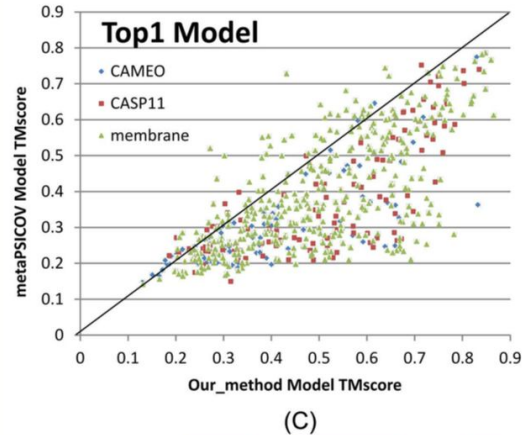
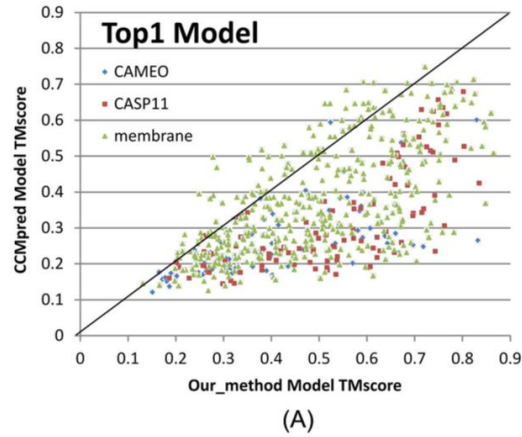
2. Accuracy as Function of # of Sequence Homologs

Accuracy of this method (green), CCMpred (blue) and MetaPSICOV (red) with respect to the amount of homologous information measured by $\ln(\text{Meff})$



3. Contact-assisted protein folding

Quality comparison of contact assisted models generated by this method and (A) CCMpred and (C) MetaPSICOV on CASP11 targets (red), CAMEO targets (blue), and membrane proteins (green)



Quality comparison of contact-assisted models generated by this method and (A) CCMpred and (B) MetaPSICOV on the 41 CAMEO hard targets

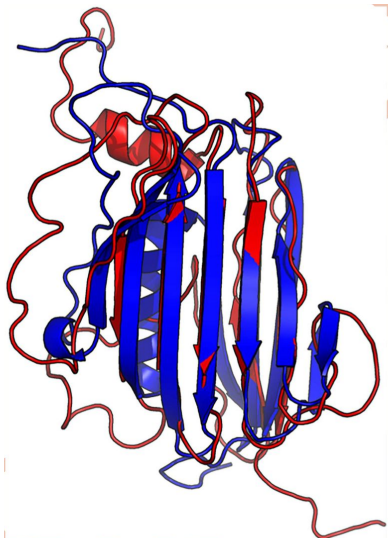
Wang et al. "Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model." PLOS Computational Biology 13.1 (2017)

4. Study of difficult CAMEO targets

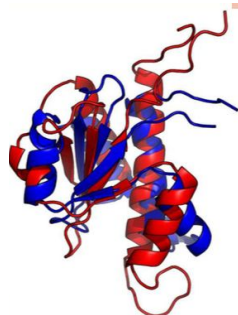
- Successfully folded the hardest targets for structure prediction released by CAMEO
- Identified novel folds in several CAMEO protein targets

Summary of results on 5 CAMEO hard targets with novel folds

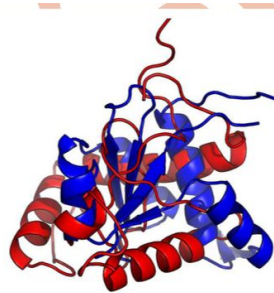
Target	CAMEO ID	Type	Len	Meff	Method	RMSD(A)	TMscore
2nc8A	2016-09-10_00000002_1	β	182	250	Our server	6.5	0.61
					Best of the others	12.18	0.47
5dcjA	2016-09-17_00000018_1	$\alpha+\beta$	125	180	Our server	7.9	0.52
					Best of the others	10.0	0.53
5djeB	2016-09-24_00000052_1	α	140	330	Our server	5.81	0.65
					Best of the others	14.98	0.34
5f5pH	2016-10-15_00000047_1	α	217	65	Our server	4.21	0.71
					Best of the others	>40.0	0.48
5flgB	2016-11-12_00000046_1	α/β	260	113	Our server	7.12	0.61
					Best of the others	16.9	0.25



Predicted model (red) and its native structure (blue) for the CAMEO test protein (PDB ID 2nc8 and chain A)



(A)



(B)



(C)

Predicted models (red) and native structure (blue) for the CAMEO test protein (PDB ID 5dcj and chain A) by (A) this method, (B) CCMpred, and (C) MetaPSICOV

Novel Aspects and Advantages

1. Concatenates two deep residual neural networks
2. Predicts all contacts of a protein simultaneously
3. Deeper architecture
4. Trains all 2D convolution layers simultaneously
5. Learns sequence-structure relationship from thousands of protein families

Additional Interesting Findings

1. Without using membrane proteins in training, this method has comparable accuracy on that set to methods trained with them
2. Identified model parameters most important to accuracy
 - a. Co-evolution strength produced by CCMpred
 - b. Depth of deep model

Study Limitations

1. Not particularly well-explained/well-written
2. Comparison to prior methods might not be right comparison
3. Limited size and scope of training set
 - a. Consists of only about 100 membrane proteins
4. Redundancy between training set and test set
5. Does not consider energy functions or fragment assembly

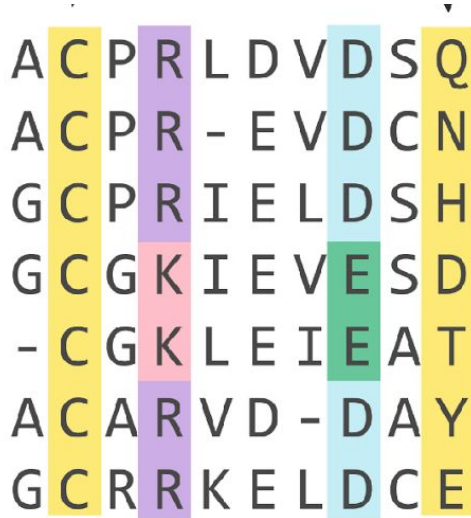
Future Directions and Further Applications

- Introduction of additional convolution layers
- Prediction of protein-protein and protein-RNA interfacial contacts

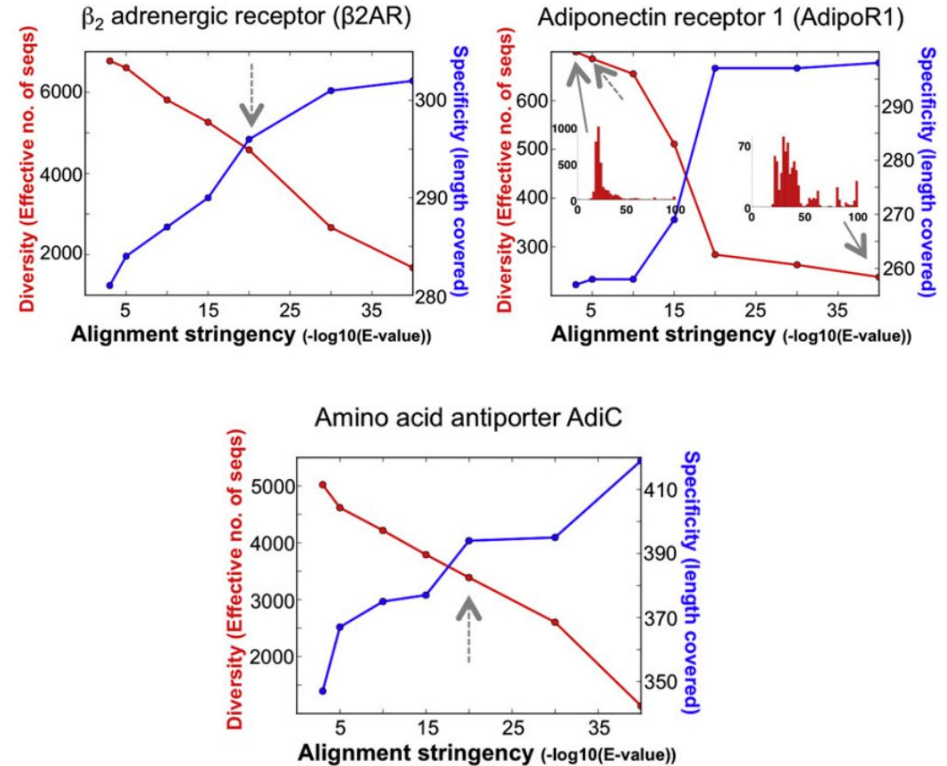
Appendix

Additional Slides

Build MSA

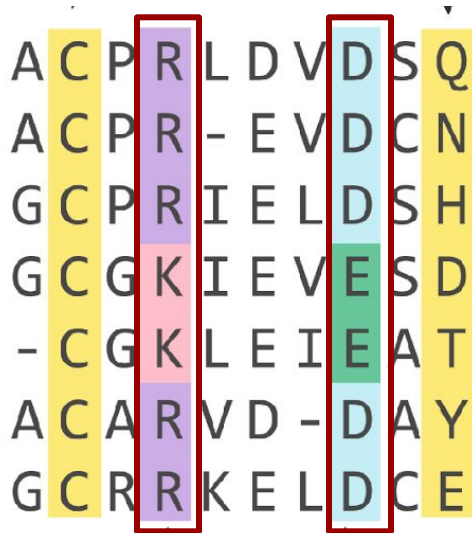


Maximize power of detection



Entropy maximization

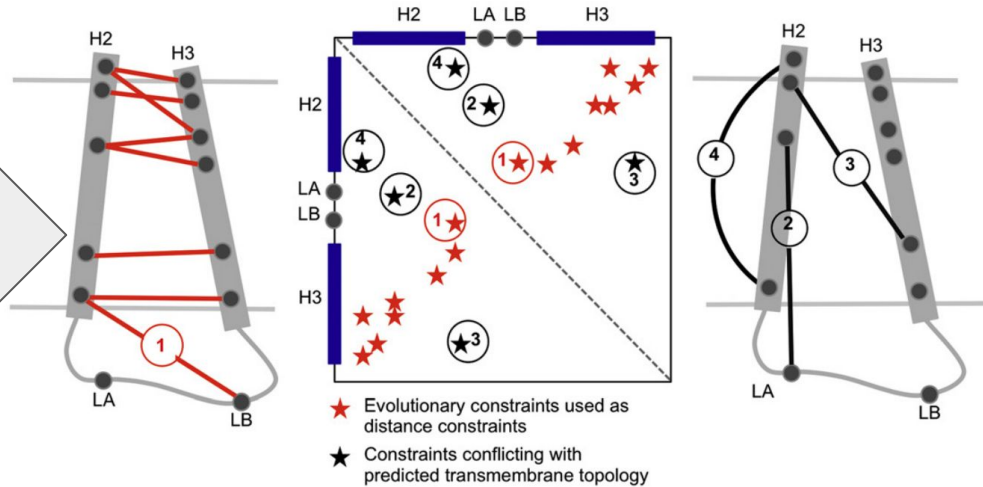
Find EV couplings (residue pairs) that best explain the data



Evolutionary Couplings

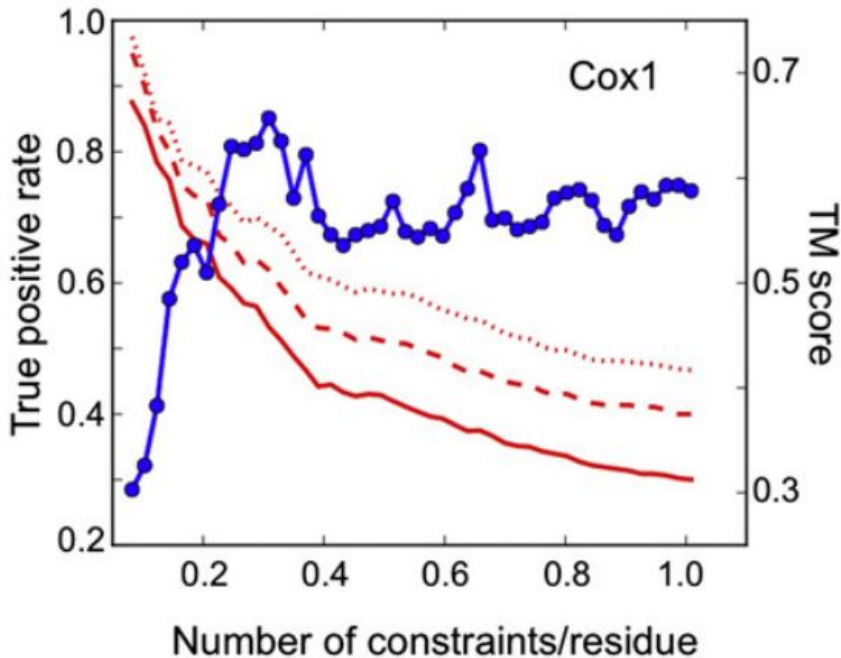
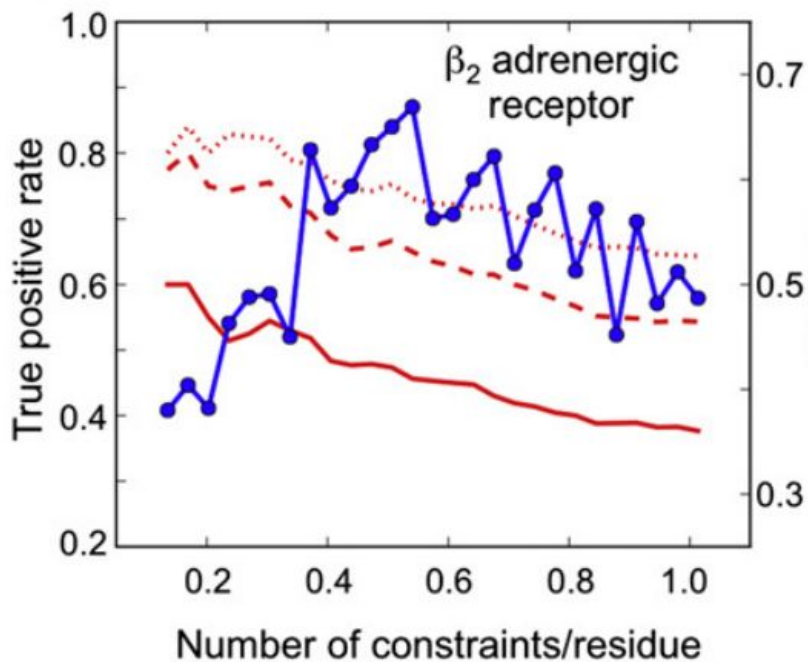
Constraint resolution

Remove couplings that are not consistent with predicted membrane topology + secondary structure



Evaluation on known 3D Structures

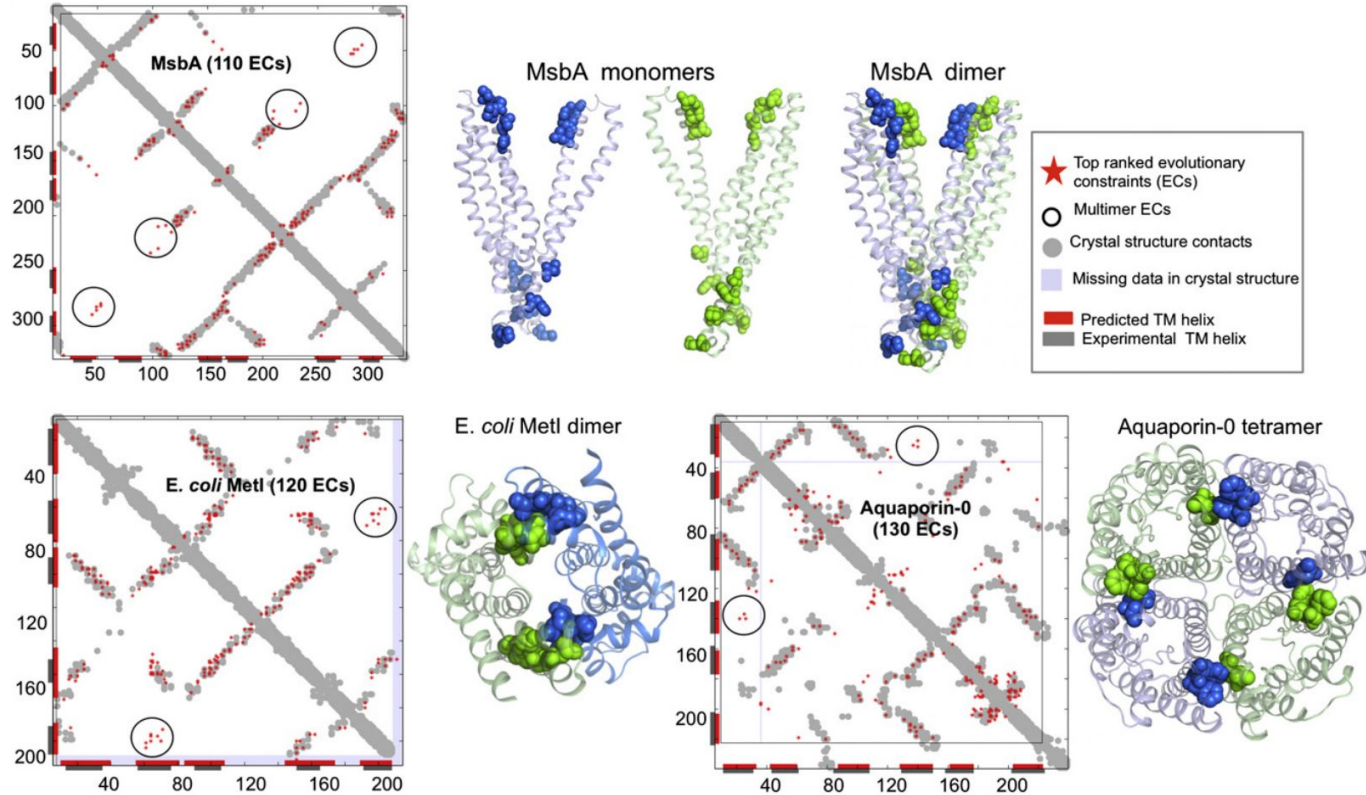
Interesting insight: 3D prediction accuracy is stable even as TPR of evolutionary constraints decreases



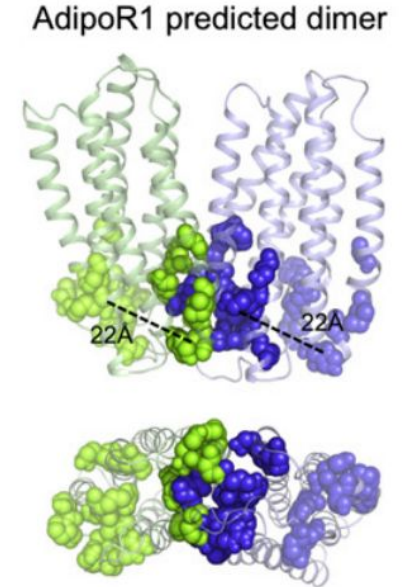
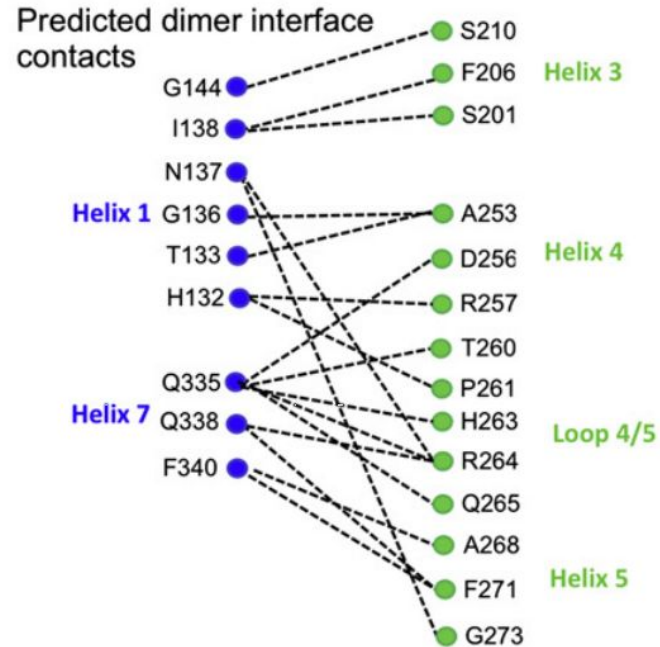
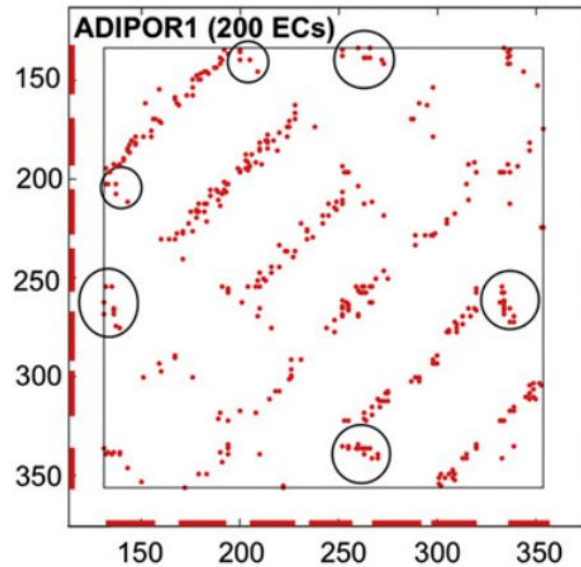
What do evolutionary constraints really represent?

1. Positions of Conformational Change
2. Functional Sites

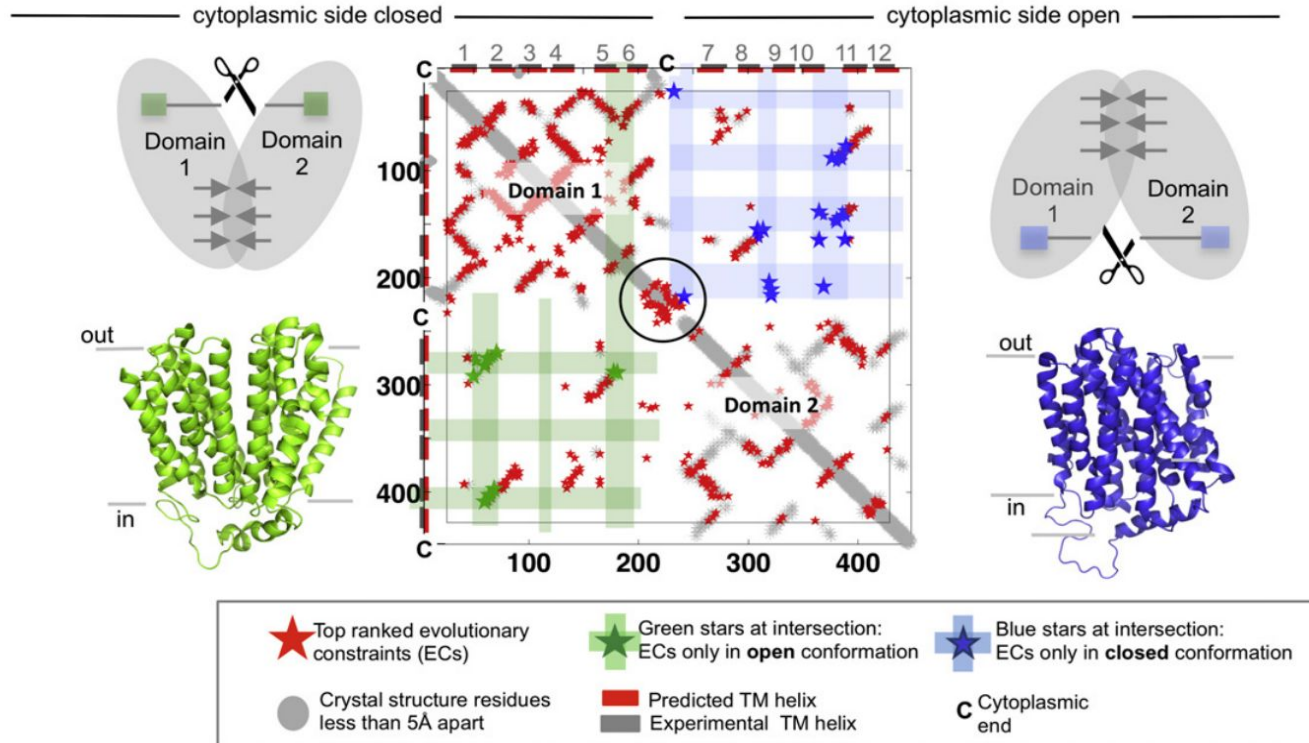
1. Homo-Oligomer Contacts



1. Homo-Oligomer Contacts



Can coevolution predict **more than 1** 3D conformation?



Sequence data far **outpaces** structure data

