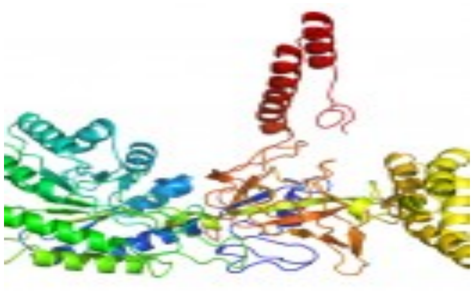


Template Free Protein Structure Modeling

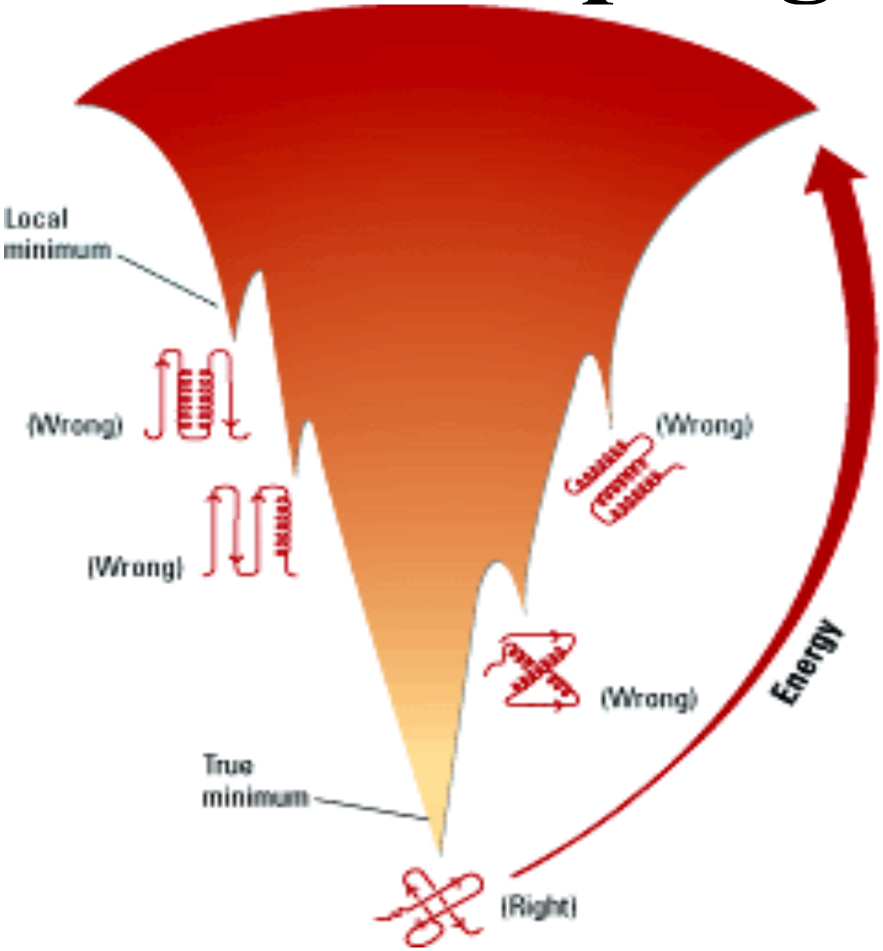
Jianlin Cheng, PhD

Professor
Department of EECS
Informatics Institute
University of Missouri, Columbia
2019

Outline

- Traditional template-free (ab initio) modeling
- Distance-based ab initio modeling empowered by deep learning

Protein Energy Landscape & Free Sampling

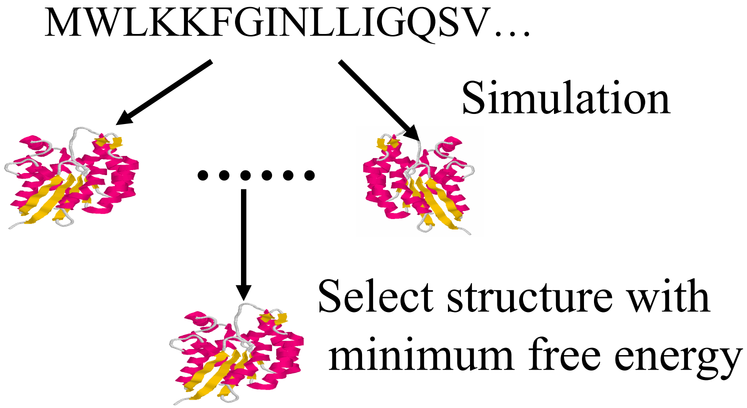


http://pubs.acs.org/subscribe/archive/mdd/v03/i09/html/willis.html

Two Approaches for 3D Structure Prediction

- **Ab Initio Structure Prediction**

Physical force field – protein folding
Contact map - reconstruction



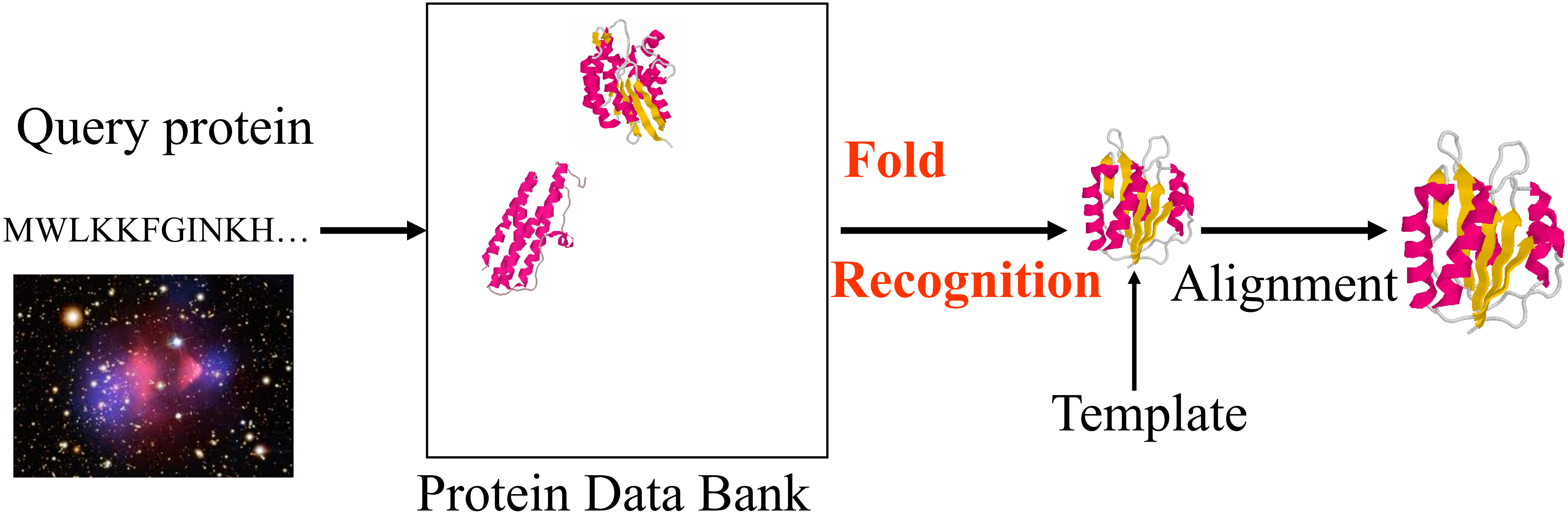
MWLKKFGINLLIGQSV…

Simulation

……

Select structure with
minimum free energy

- **Template-Based Structure Prediction**



Query protein

MWLKKFGINKH…

Protein Data Bank

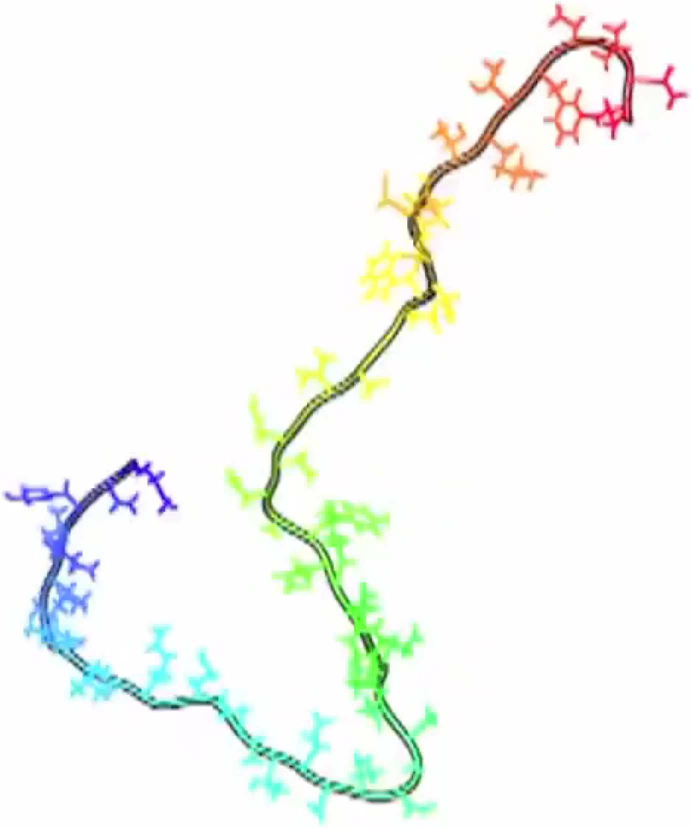
Fold

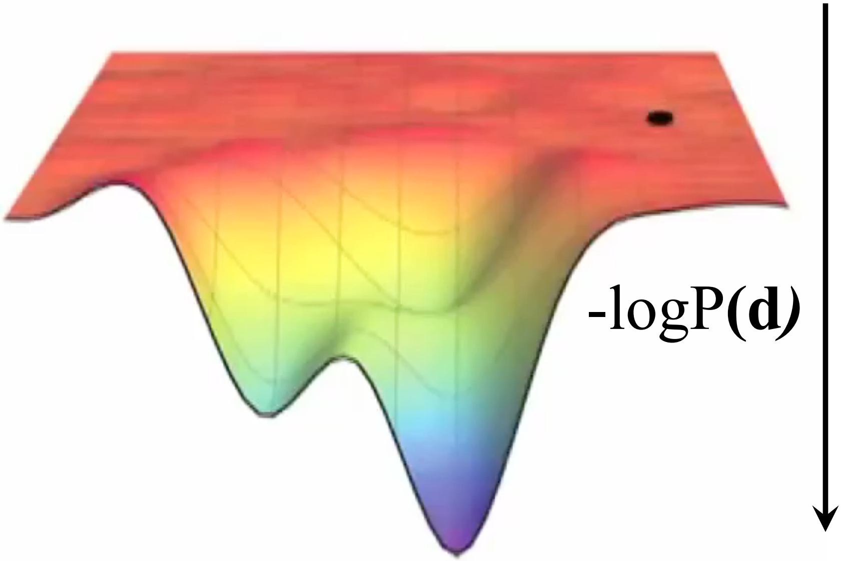
Recognition

Alignment

Template

Demo of Our Protein Structure Prediction Software (FUSION)





Funnel-shaped landscape

Part I. Traditional Ab Initio Modeling Methods

Energy Functions

- T. Lazaridis, M. Karplus. Effective energy functions for protein structure prediction. Current Opinion in Structural Biology. 2000
- A. Liwo, C. Czaplewski, S. Oldiej, H.A. Scheraga. Computational techniques for efficient conformational sampling of proteins. 2008
- K. Simons et al. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. JMB. 1997. (Rosetta – a case study) -- reading assignment due Feb. 26

Protein Energy Function

- The native state of a protein is the state of lowest free energy under physiological conditions
- This state corresponds to the lowest basin of the effective energy surface.
- The term ‘effective energy’ refers to the free energy of the system (protein plus solvent)

Two Kinds of Energy Functions

- Physical effective energy function (PEEF): fundamental analysis of forces between particles
- Statistical effective energy function: data derived from known protein structures (e.g., statistics concerning pair contacts and surface area burial)

Statistical Effective Energy Function (SEEF)

- Less sensitive to small displacements
- Because of their statistical nature, they can, in principle, include all known and unrecognized, physical effects.
- Works better for protein structure prediction

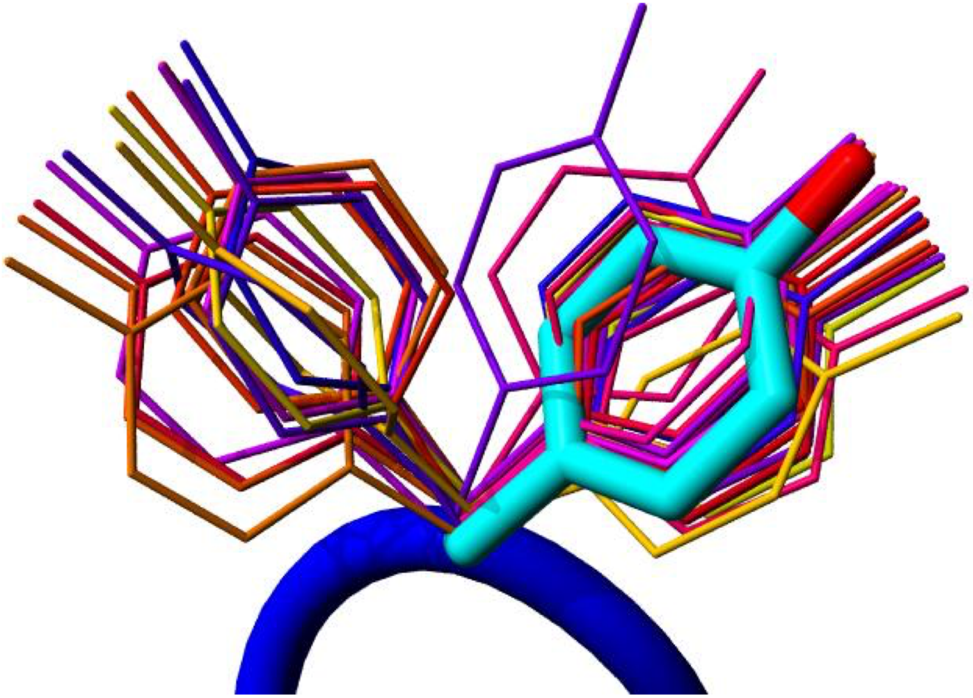
SEEF

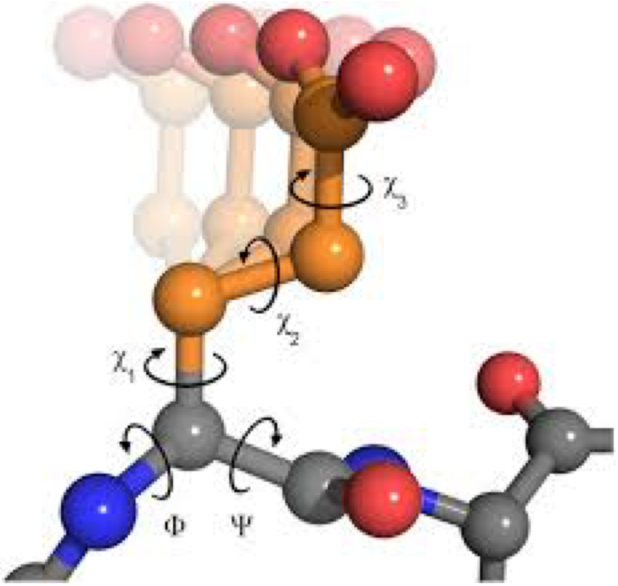
- Employ a reduced representation of the protein: a single interaction center at Ca or Cb for each residue.
- Basic idea: log (P_{ab} / P_a * P_b). P_{ab}: is the observed probability that residues a and b are in contact. P_a is frequency of a and P_b is the frequency of b
- Energy = -log (P_{ab} / P_a * P_b)
- More info: use secondary structure, solvent accessibility, distance as conditions.

Energy Terms

- Pairwise contact potentials
- Hydrogen bonds
- Torsion angle
- Burial energy (solvation energy)
- Sidechain orientation coupling, rotamer energy

Rotamer Energy





http://dunbrack.fccc.edu/scwrl4/

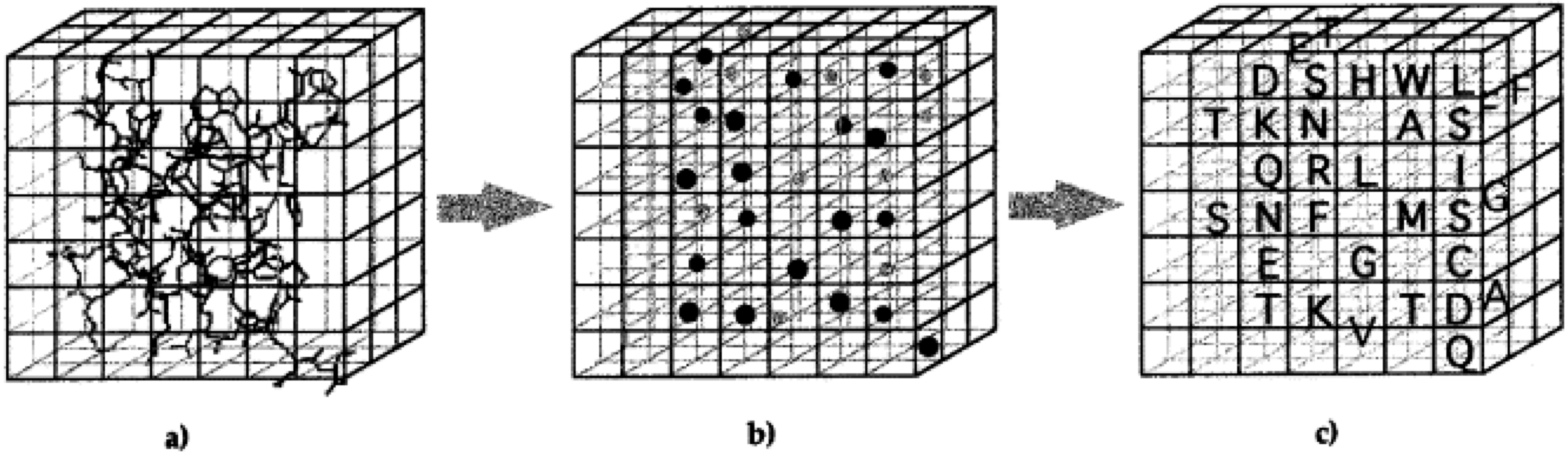
Physical / Statistical Effective Energy Function (PEEF)

- CHARMM implementation (http://www.charmm.org)
- AMBER implementation (http://ambermd.org)
- Dfire energy: http://sparks-lab.org/tools-dfire.html (program)
- RW energy: http://zhanglab.ccmb.med.umich.edu/RW/ (program available)

Benchmark

- Can a function select a native structure from a large pool of decoys?
- Can a function be used effectively in conformation sampling to generate a high proportion of near-native conformations?

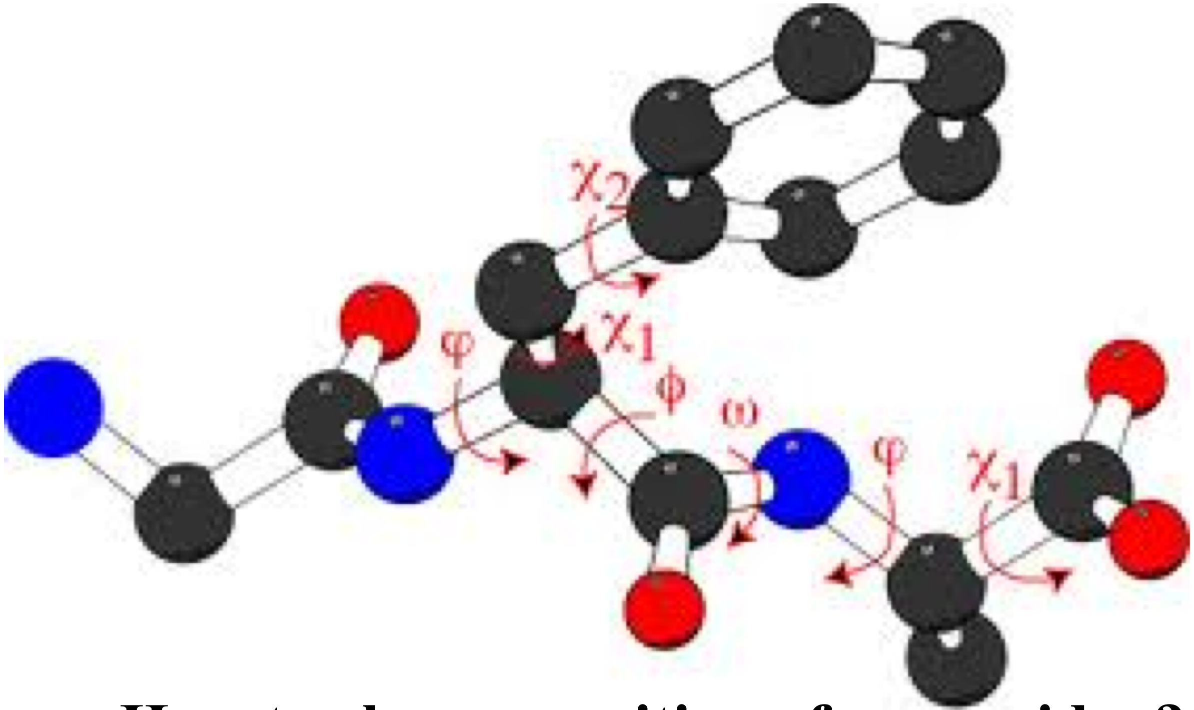
Representation for Conformation Sampling



How to change position of one residue?

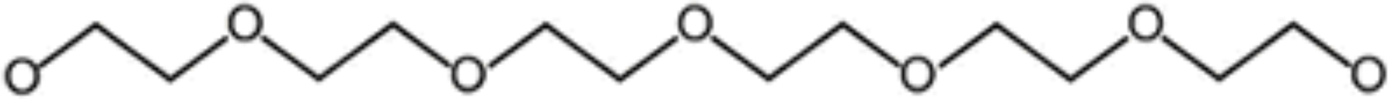
ITASSER: http://zhanglab.ccmb.med.umich.edu/I-TASSER/

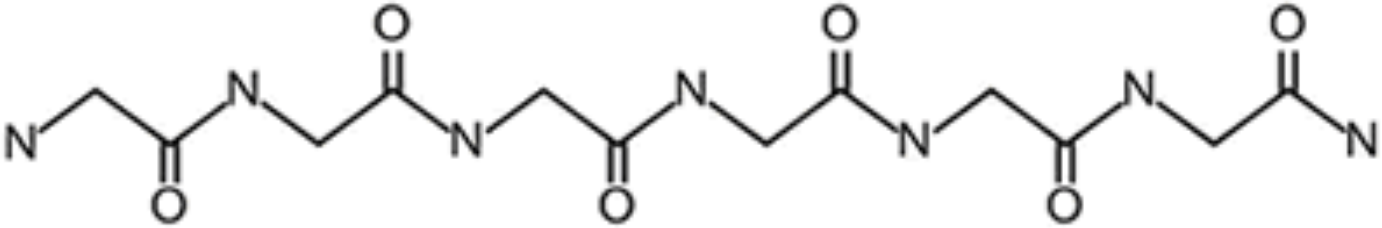
Torsion Angles



How to change position of one residue?

Vector Space

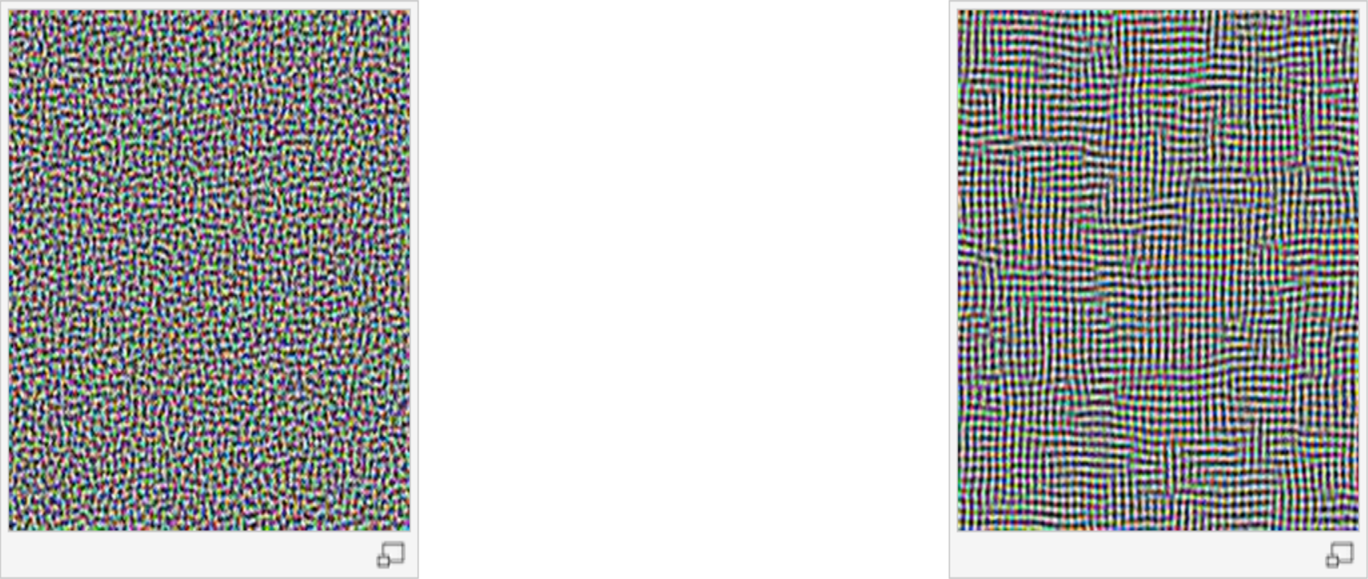




Simulated Annealing

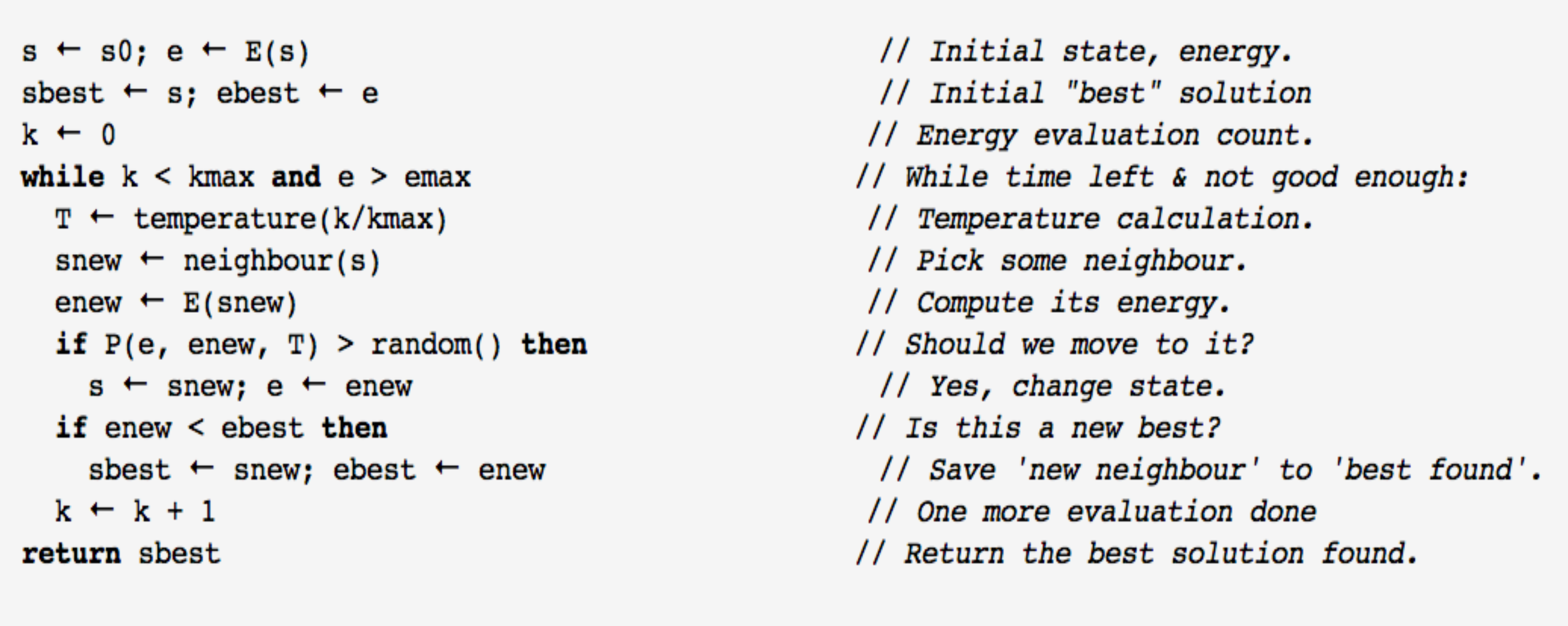
- Accept a move based on a probability related to temperature, e.g., P ~ e^ (-ΔE / T)
- Temperature (T) controls the degree of exploration. Higher temperature, more exploration? Why?
- Temperature decreases as the sampling process progresses (from iteration to iteration): cooling schedule

An Example



Example illustrating the effect of cooling schedule on the performance of simulated annealing. The problem is to rearrange the pixels of an image so as to minimize a certain potential energy function, which causes similar colours to attract at short range and repel at a slightly larger distance. The elementary moves swap two adjacent pixels. These images were obtained with a fast cooling schedule (left) and a slow cooling schedule (right), producing results similar to amorphous and crystalline solids, respectively.

Pseudo Code



```
s ← s0; e ← E(s)                                    // Initial state, energy.
sbest ← s; ebest ← e                                // Initial "best" solution
k ← 0                                              // Energy evaluation count.
while k < kmax and e > emax                       // While time left & not good enough:
  T ← temperature(k/kmax)                          // Temperature calculation.
  snew ← neighbour(s)                              // Pick some neighbour.
  enew ← E(snew)                                   // Compute its energy.
  if P(e, enew, T) > random() then                // Should we move to it?
    s ← snew; e ← enew                              // Yes, change state.
  if enew < ebest then                            // Is this a new best?
    sbest ← snew; ebest ← enew                      // Save 'new neighbour' to 'best found'.
  k ← k + 1                                        // One more evaluation done
return sbest                                      // Return the best solution found.
```

A TFM Example: Rosetta

- K. Simons, C. Kooperberg, E. Huang, D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. JMB, 1997.

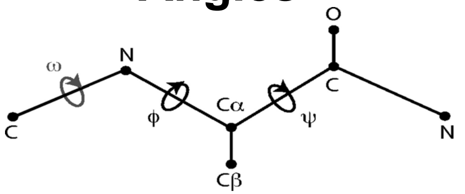
Rosetta: https://www.rosettacommons.org

Basic Idea

- Short sequence segments are restricted to the local structures adopted by the most closely related sequences in the PDB
- Use the observed local conformations of similar local sequences to reduce sampling space

Fragment Assembly (e.g. Rosetta)

Angles



Coordinates

(X, Y, Z)

SDDEVYQYIVSQ**V**KQYGIEPAELLSRK**Y**GDK AKYHLSQ

Randomly pick 9 residues

Reduce search space!

9-Residue Fragment DB

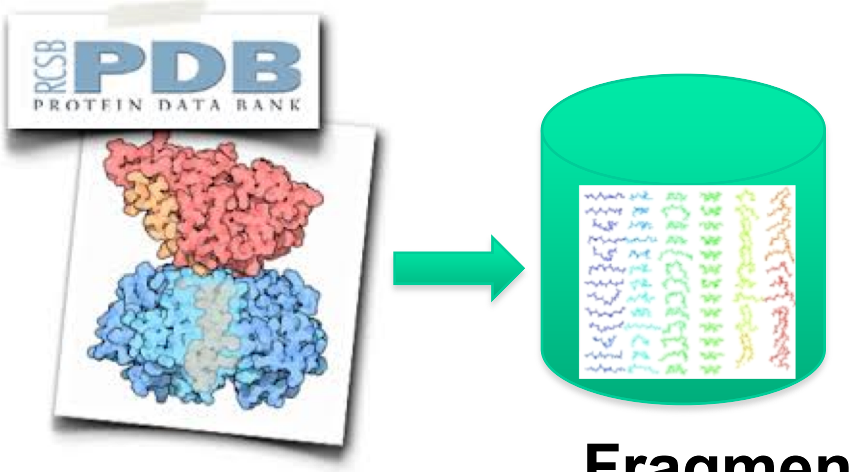
Fragment	**Angles**
SDDEQYQRK	(130,-120, …)
….	
….	

Find a similar fragment
Replace angles

Two ways of obtaining fragments

- Database-based approach: https://www.rosettacommons.org
- Model-based approach: http://sysbio.rnet.missouri.edu/FRAGSION/

Shortcomings of Fragment Assembly Approach Based on Database Search



~80,000 proteins

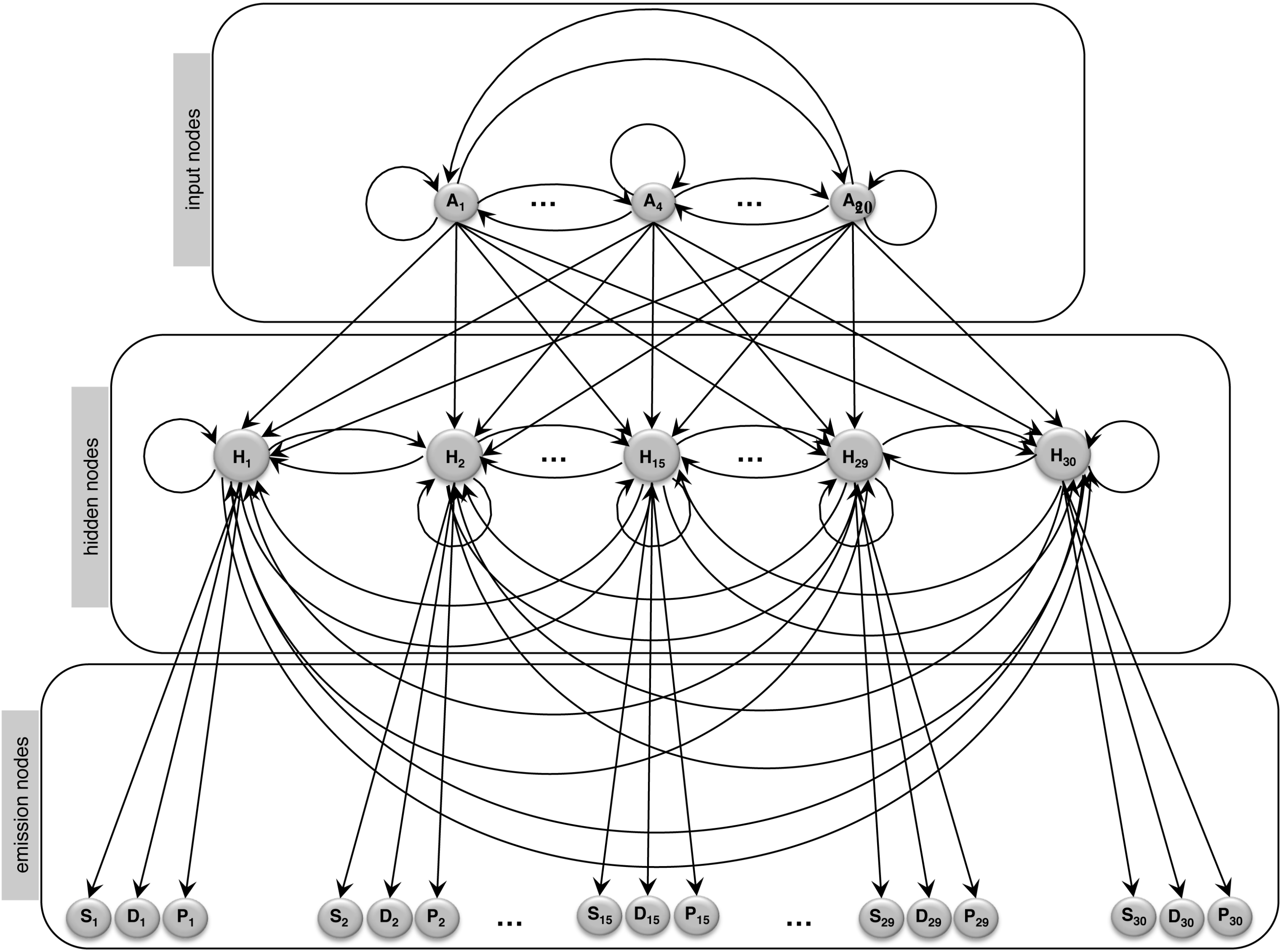
Fragment Structure Database

- Incomplete coverage
- Computationally expensive
- Restricted to small proteins

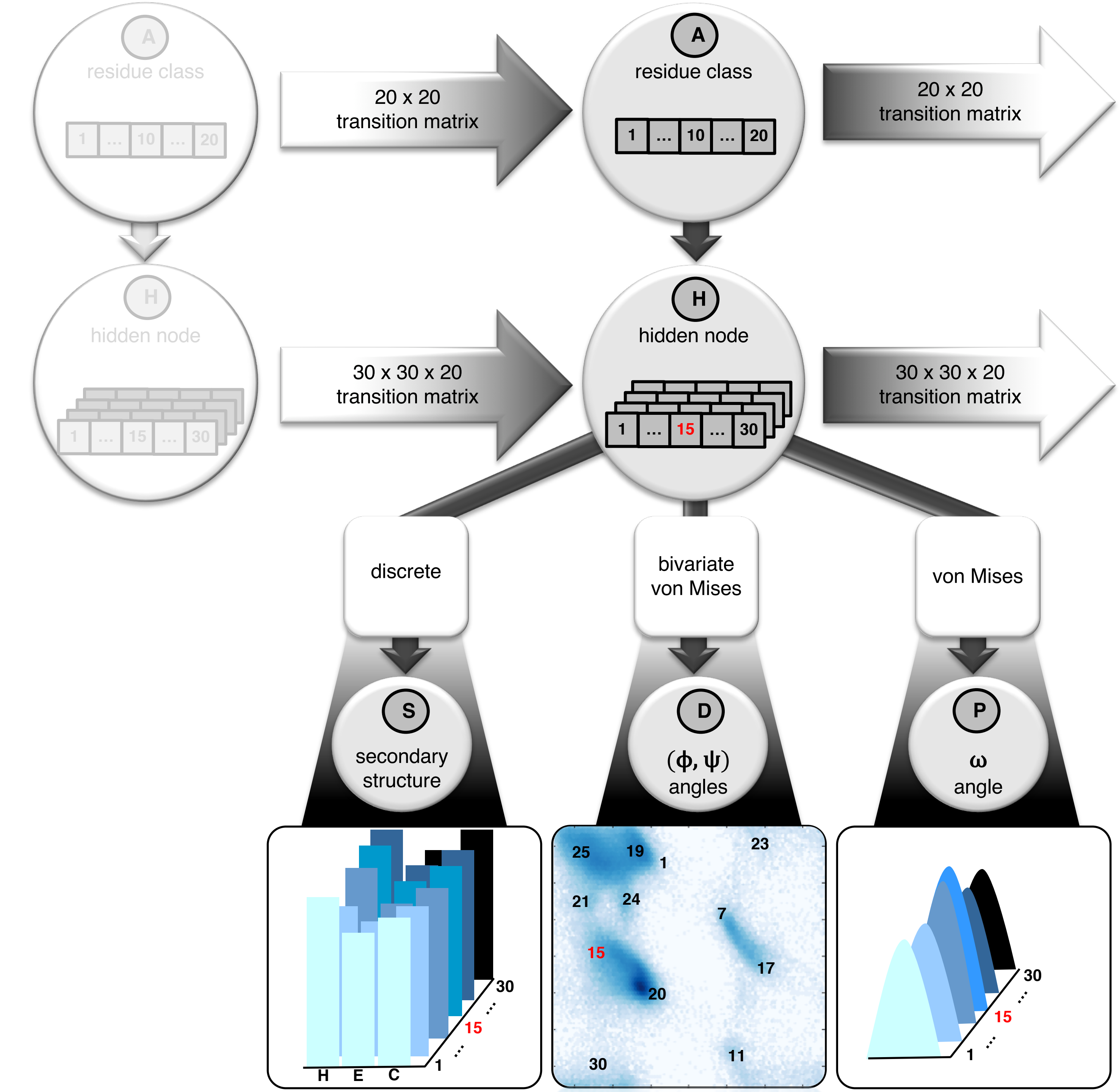
IOHMM (Input-Output Hidden Markov Model)

to model protein conformational space

Bhattacharya & Cheng, Bioinformatics, 2016
Bhattacharya & Cheng, Scientific Reports, 2015



$$P(\mathbf{S},\mathbf{D},\mathbf{P}) = \sum_{\mathbf{H}} P(\mathrm{H}_{\mathit{1}} \mid \mathrm{A}_{\mathit{1}})\ P(\mathrm{S}_{\mathit{1}} \mid \mathrm{H}_{\mathit{1}})\ P(\mathrm{D}_{\mathit{1}} \mid \mathrm{H}_{\mathit{1}})\ P(\mathrm{P}_{\mathit{1}} \mid \mathrm{H}_{\mathit{1}}) \prod_{\mathrm{i}=2}^{n} P(\mathrm{H}_i \mid \mathrm{H}_{i\text{-}1}, \mathrm{A}_i)\ P(\mathrm{S}_i \mid \mathrm{H}_i)\ P(\mathrm{D}_i \mid \mathrm{H}_i)\ P(\mathrm{P}_i \mid \mathrm{H}_i)$$



A

residue class

20 x 20
transition matrix

H

hidden node

30 x 30 x 20
transition matrix

discrete

bivariate
von Mises

von Mises

S

secondary
structure

D

(ϕ, ψ)
angles

P

ω
angle

Bhattacharya **et al.** 2015

Parameter Learning

using EM algorithm

- **1,740 experimentally solved proteins**
- **270,350 observations**
- **Training using stochastic EM algorithm**

Bhattacharya **et al.** 2015; Van **et al.** 2005; Paluszewski **et al.** 2010

Selecting optimal model

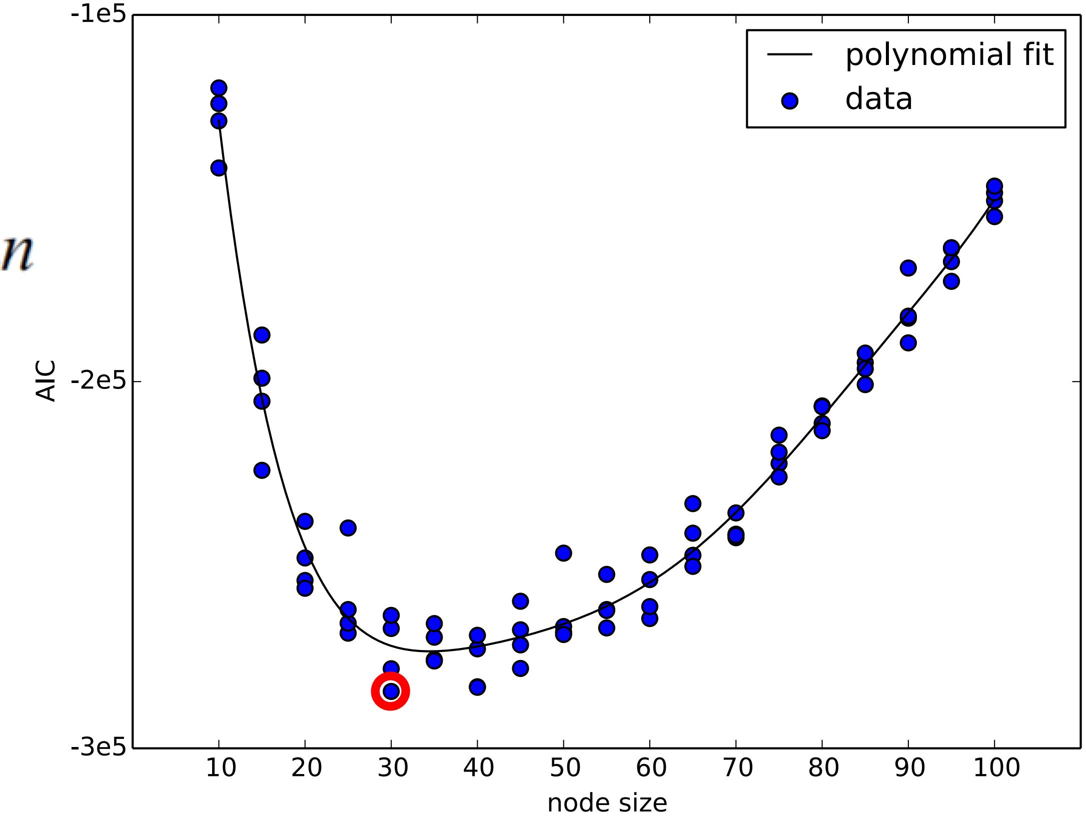
using information theory

$$AIC(n) = -2\log L\ (d|m) + 2n$$

$L(d|m)$: likelihood

d : data

n : parameters



30 hidden nodes
7,812 parameters

Bhattacharya **et al.** 2015; Burnham **et al.** 2002

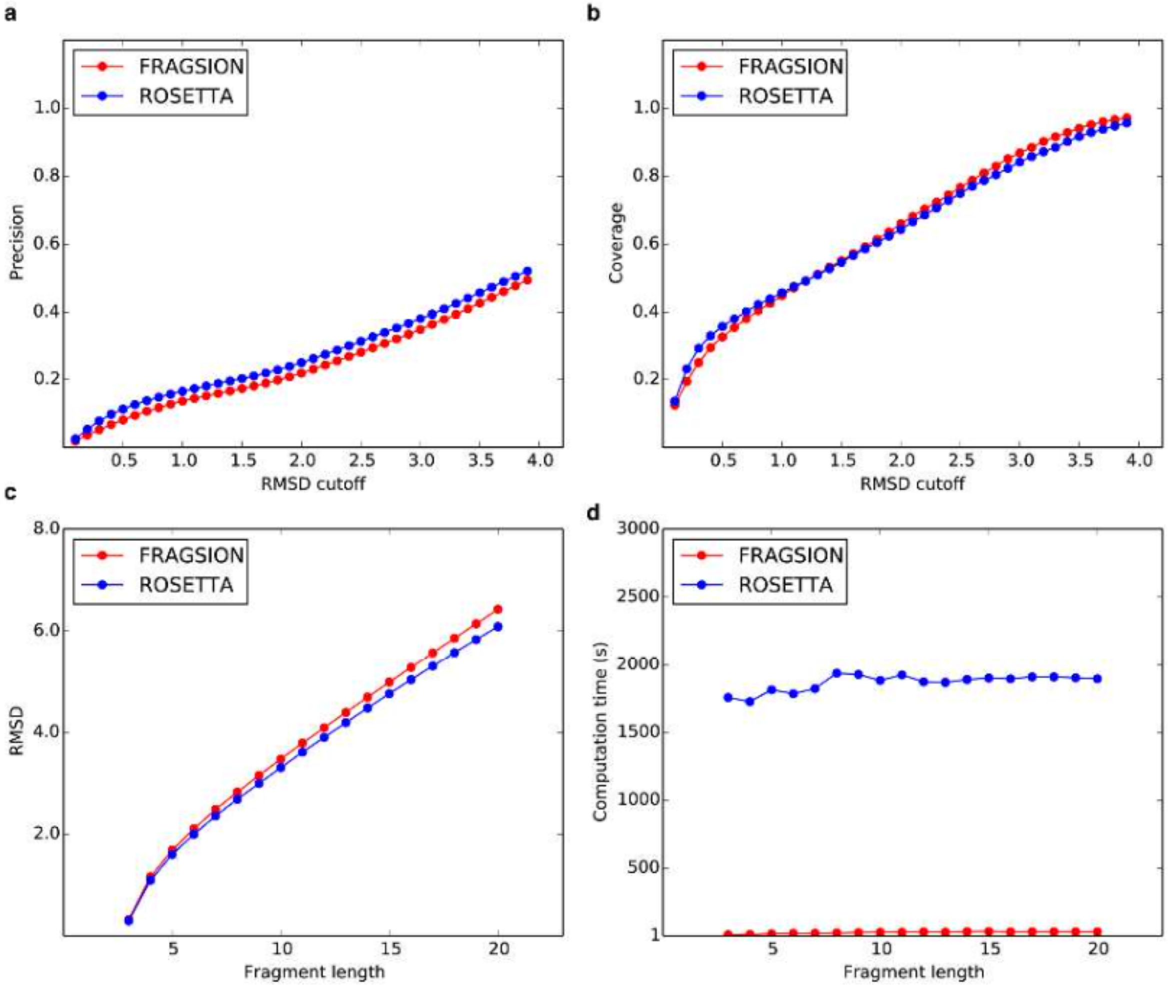


Fig. 1. Comparison between FRAGSION and ROSETTA. Precision (a), coverage (b) at various RMSD cutoffs and RMSD (c), computation time (d) at different fragment lengths averaged over the dataset generated by FRAGSION (red) and ROSETTA (blue).

Function of IOHMM Model of Protein Conformation

- Sample the conformation of a (sub) sequence of any size
- Software: Fragsion: http://sysbio.rnet.missouri.edu/FRAGSION/

Protein Folding Video

- https://www.youtube.com/watch?v=HBONCqN9U4k

Scoring Functions of Selecting Local Conformations

- Knowledge-based potential functions
- Bayesian scoring function

$$P(\mathit{structure} \mid \mathit{sequence}) = P(\mathit{structure}) \times \frac{P(\mathit{sequence} \mid \mathit{structure})}{P(\mathit{sequence})}$$

One native assumption is P(structure) = 1 / # of structures.

P(a structure)

- 0 for configurations with overlaps between atoms
- Proportional to exp(-radius of gyration^2) for all other configurations.
- Independent of secondary structure elements

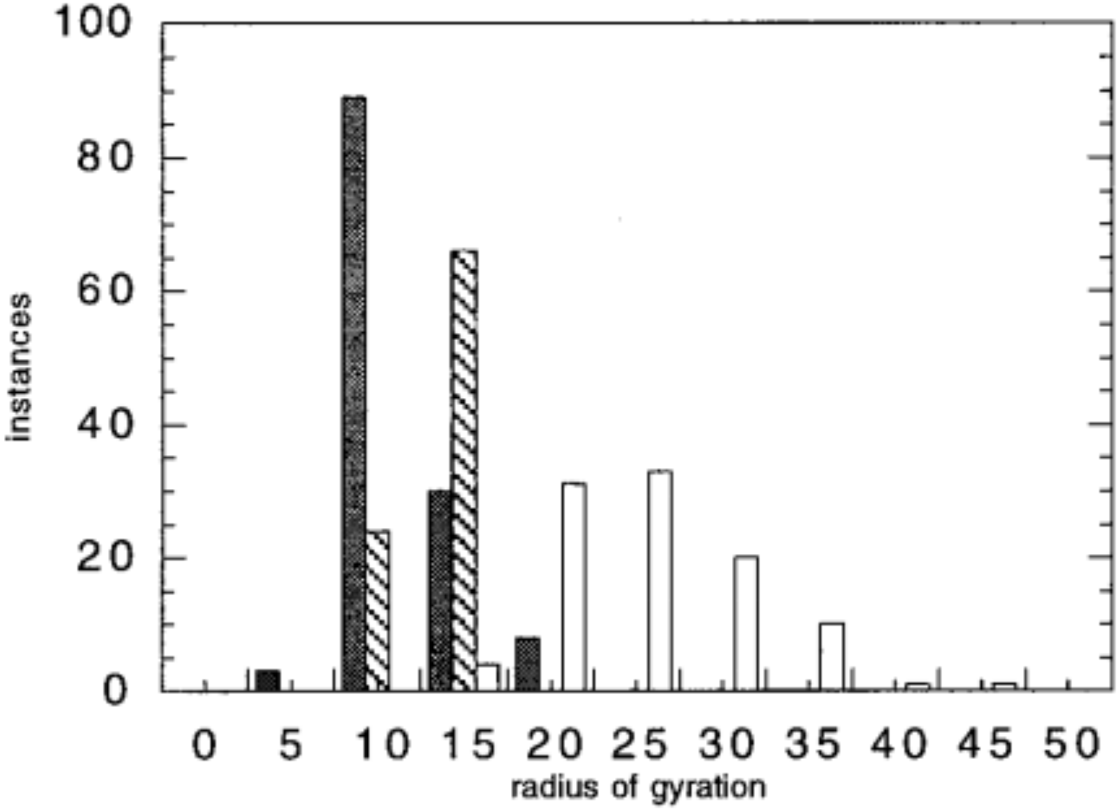
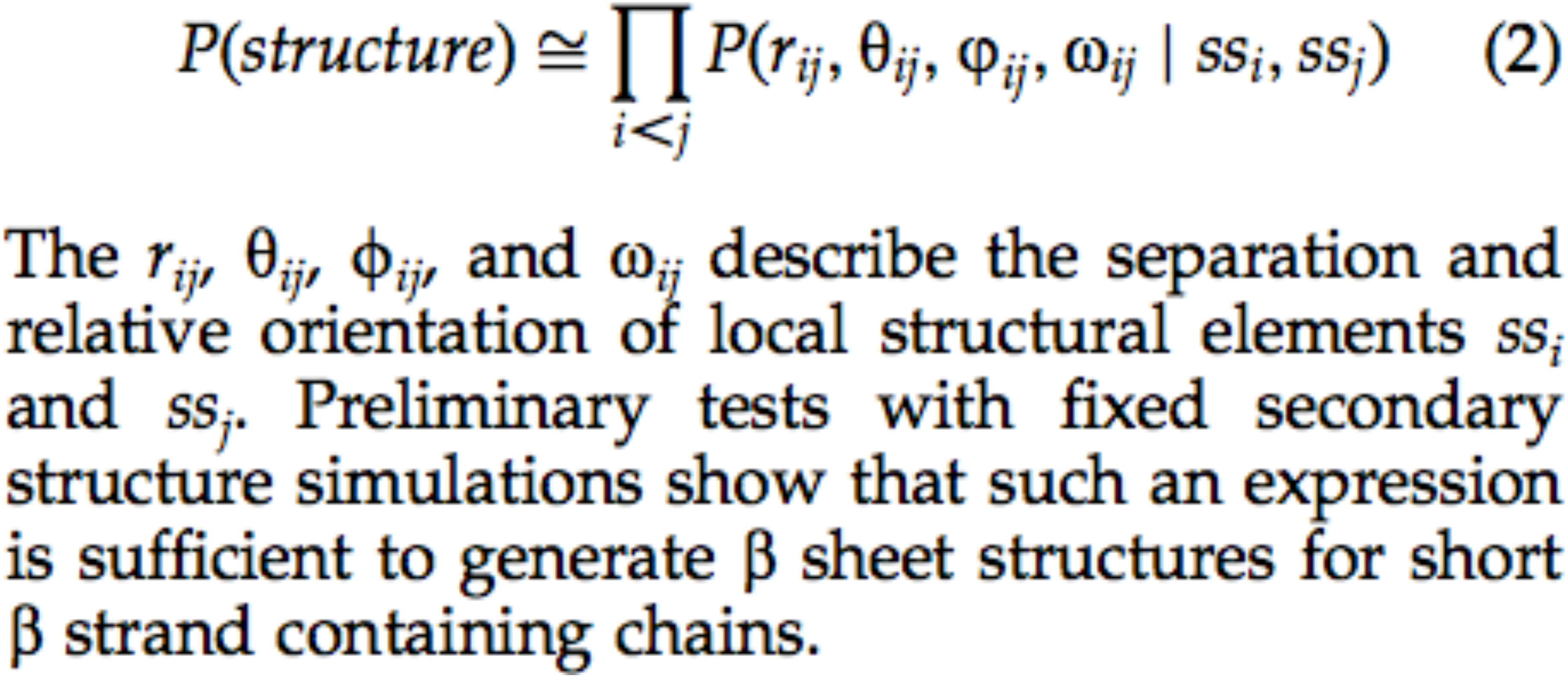


Figure 1. Comparison of the radii of gyrations of simulated and native structures. 100 structures were generated for chains of 100 residues by splicing together protein fragments as described in Methods using either no scoring function (open bars), or the square of the radius of the gyration as the scoring function (hatched bars). Histograms were computed using 5 Å bins. The distribution of radii of gyrations for the small (50 to 150 residue) proteins in the pdbselect 25 set is shown for comparison (filled bars).

Considering Beta-Sheet Pairing



$$P(structure) \cong \prod_{i<j} P(r_{ij}, \theta_{ij}, \varphi_{ij}, \omega_{ij} \mid ss_i, ss_j) \quad (2)$$

The r_{ij}, θ_{ij}, ϕ_{ij}, and ω_{ij} describe the separation and relative orientation of local structural elements ss_i and ss_j. Preliminary tests with fixed secondary structure simulations show that such an expression is sufficient to generate β sheet structures for short β strand containing chains.

Scoring – P(Sequence | Structure)

$$P(aa_1, aa_2, \ldots, aa_n \mid structure) \cong \prod_i P(aa_i \mid E_i)$$

$$\times \prod_{i<j} \frac{P(aa_i, aa_j \mid r_{ij}, E_i, E_j)}{P(aa_i \mid r_{ij}, E_i, E_j) P(aa_j \mid r_{ij}, E_i, E_j)} \qquad (8)$$

E_i can represent a variety of features of the local structural environment around residue i.

Implementation

- Second term: for pairs separated for more than 10 residues along the chain
- Buried environment: >16 other Cb atoms within 10 Angstrom of the Cb atom of the residue; otherwise, exposed

Negative Log of Interaction Probability Function

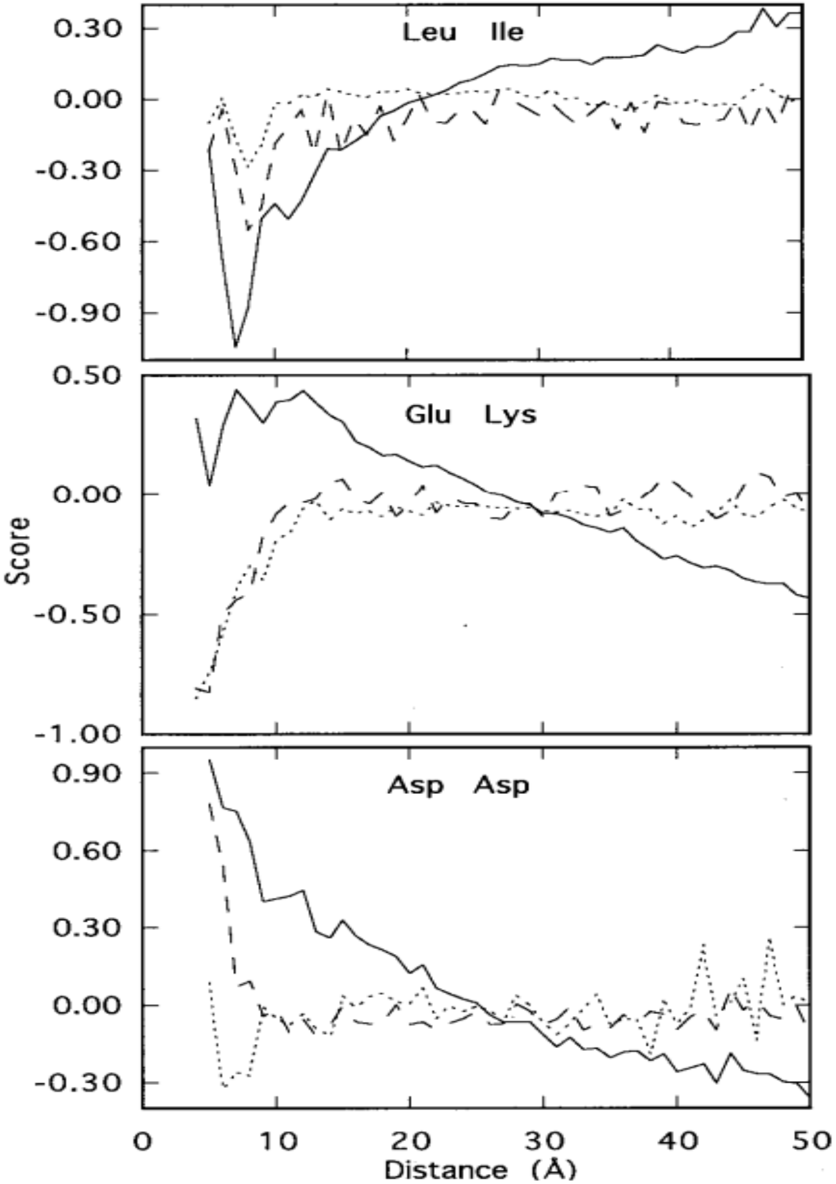


Figure 4. Comparison of the negative logarithms of equation (5) and the residue pair specific second term in equation (8) for sequence separations greater than ten. Residues with greater than 16 neighbors were considered buried. Continuous lines, equation (5); dotted lines, equation (8) both residues buried; broken line, equation (8) both residues exposed.

Structure Generation

- Initialization:

$$P(structure \mid sequence) \cong e^{-radius\ of\ gyration^2} \times \prod_{i<j} \frac{P(r_{ij} \mid aa_i, aa_j)}{P(r_{ij})} \quad (6)$$

Splicing together fragments of proteins of known structure with similar local sequences and evaluating them initially using equation.

Simulated Annealing

- Low scoring conformations with distributions of residues similar to those of known proteins are resampled by simulated annealing in conjunction with a simple move set that involves replacing the torsion angles of a segment of the chain with the torsion angles of a different protein fragment with a related amino acid sequence.
- The simulated conformation is evaluated by (8)

Methods

- Structures are represented using a simplified model consisting of heavy atoms of the main-chain and the C_b atom of the side chain.
- All bond lengths and angles are held constant according to the ideal geometry of alanine (Engh & Huber 91); the only remaining variables are the backbone torsional angles.

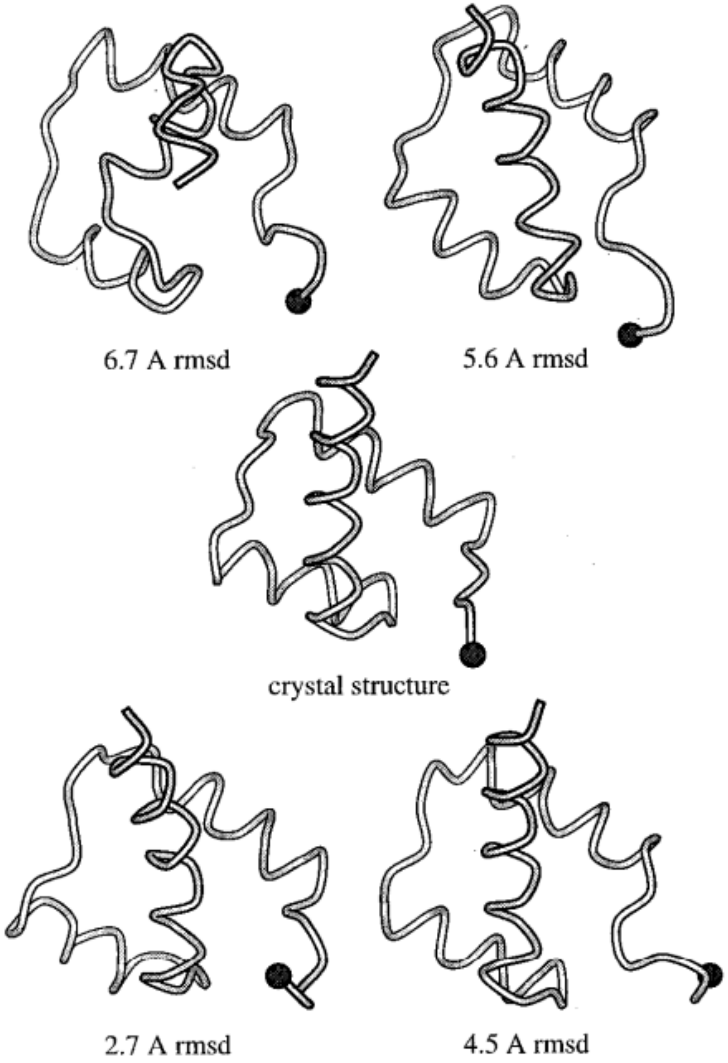
Fragment Databases

- Nimers / trimers (sequences) and their conformations extracted from known structures in the database
- Identify sequence neighbors: simple amino acid frequency matching score.

Simulation

- The starting configuration in all simulations was the fully extended chain.
- A move consists of substituting the torsional angles of a randomly chosen neighbor at a randomly chosen position for those of the current configuration.
- Moves which bring two atoms within 2.5 Angstrom are immediately rejected; other moves are evaluated according to the Metropolis criterion using the scoring equation.
- Simulated annealing was carried out by reducing the temperature from 2500 to 10 linearly over the course of 10,000 cycles (attempted moves).

Simulated Structure Examples



6.7 A rmsd

5.6 A rmsd

crystal structure

2.7 A rmsd

4.5 A rmsd

Figure 5. Simulated homeodomain structures with different rms deviations from the native structure. The N termini are displayed as black spheres.

Table 1. Folding simulation results

	<7 Å rmsd	<6 Å rmsd	<5 Å rmsd	<4 Å rmsd	Lowest rmsd	*Q*
A. *Unconstrained simulations*						
Homeodomain						
dist_env filter + msa (100)	65	47	31	17	2.75	–1.7
dist_env filter – msa	63	45	31	16	2.75	–1.8
No filter	63	48	38	8	2.75	–1.5
Random sequence	31	11	1	0	4.89	–0.2
Random fragments	16	4	1	0	4.73	–0.6
Random all	6	2	0	0	5.82	0
Calbindin						
dist_env filter + msa (64)	31	17	2	0	4.70	–1.7
dist_env filter – msa	24	14	1	0	4.70	–1.9
No filter	17	3	2	0	4.86	–1.4
Random sequence	3	0	0	0	6.18	–0.2
Random fragments	6	1	0	0	5.71	–0.4
Random all	0	0	0	0	7.63	0
Protein A						
dist_env filter	96	95	93	41	3.29	–2.3
No filter	86	85	77	41	3.16	–2.0
Random sequence	33	25	8	1	3.52	–0.2
Random fragments	48	32	9	1	3.97	–0.6
Random all	32	14	1	0	4.58	0
Cro repressor						
dist_env filter + msa (4)	39	18	8	0	4.20	–1.7
dist_env filter – msa	35	20	10	0	4.20	–1.9
No filter	24	11	4	0	4.26	–1.5
Random sequence	7	1	0	0	5.95	–0.3
Random fragments	5	0	0	0	6.14	–0.7
Random all	0	0	0	0	7.26	0
Protein G						
dist_env filter + msa (5)	3	0	0	0	6.33	–1.5
dist_env filter – msa	2	0	0	0	6.33	–1.5
No filter	1	0	0	0	6.89	–1.2
Random sequence	0	0	0	0	8.43	–0.4
Random fragments	0	0	0	0	7.80	–0.6
Random all	0	0	0	0	8.35	0

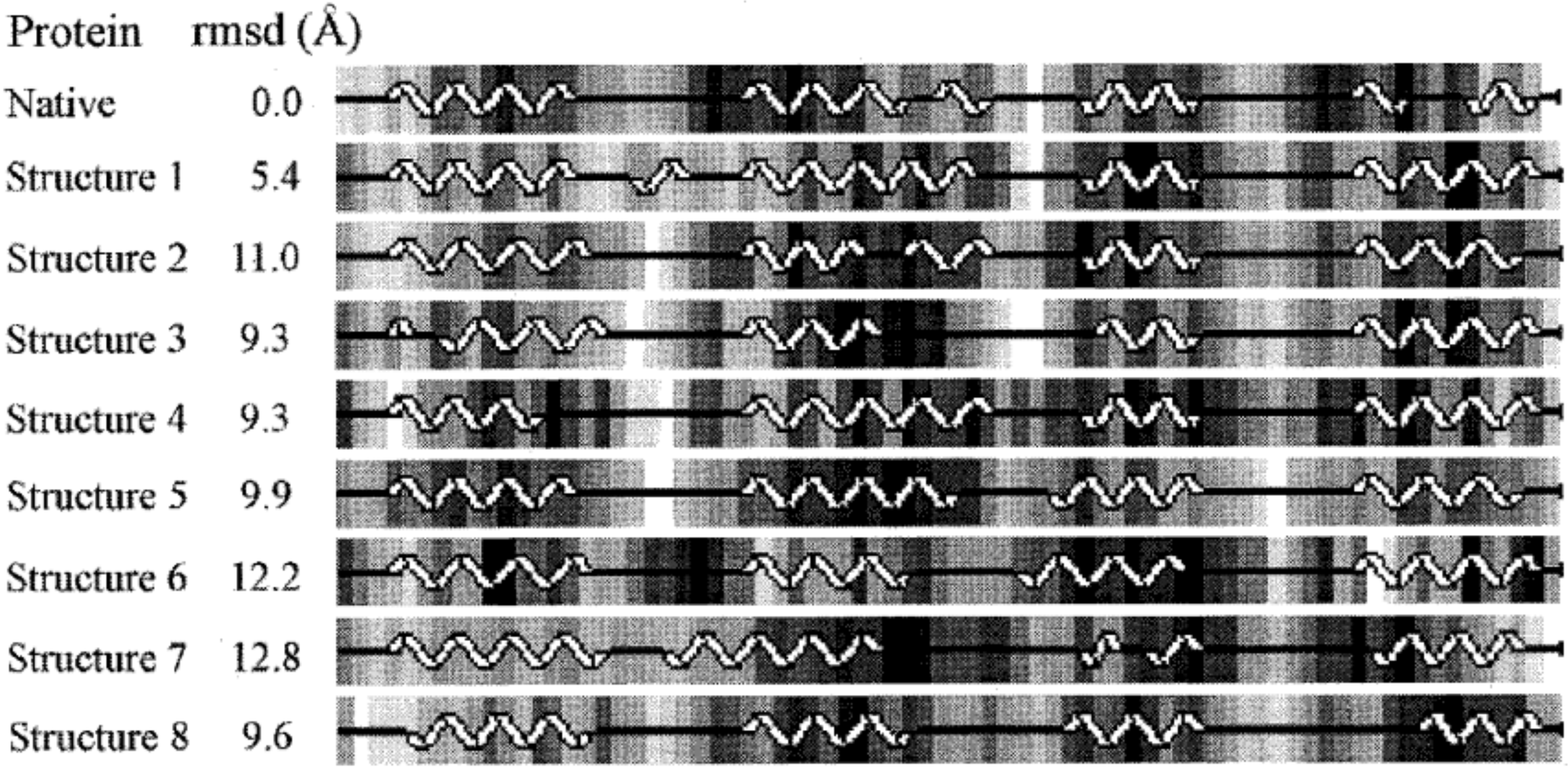


Figure 6. Solvent accessibility and secondary structure of a number of simulated non-native calbindin structures as depicted by PROCHECK (Laskowski *et al.*, 1993). The structures were randomly drawn from the simulated structure set prior to filtering. The rmsd to the native structure is shown in the second column; the rmsd between all pairs of structures is greater than 5 Å. White, solvent accessible; black, buried.

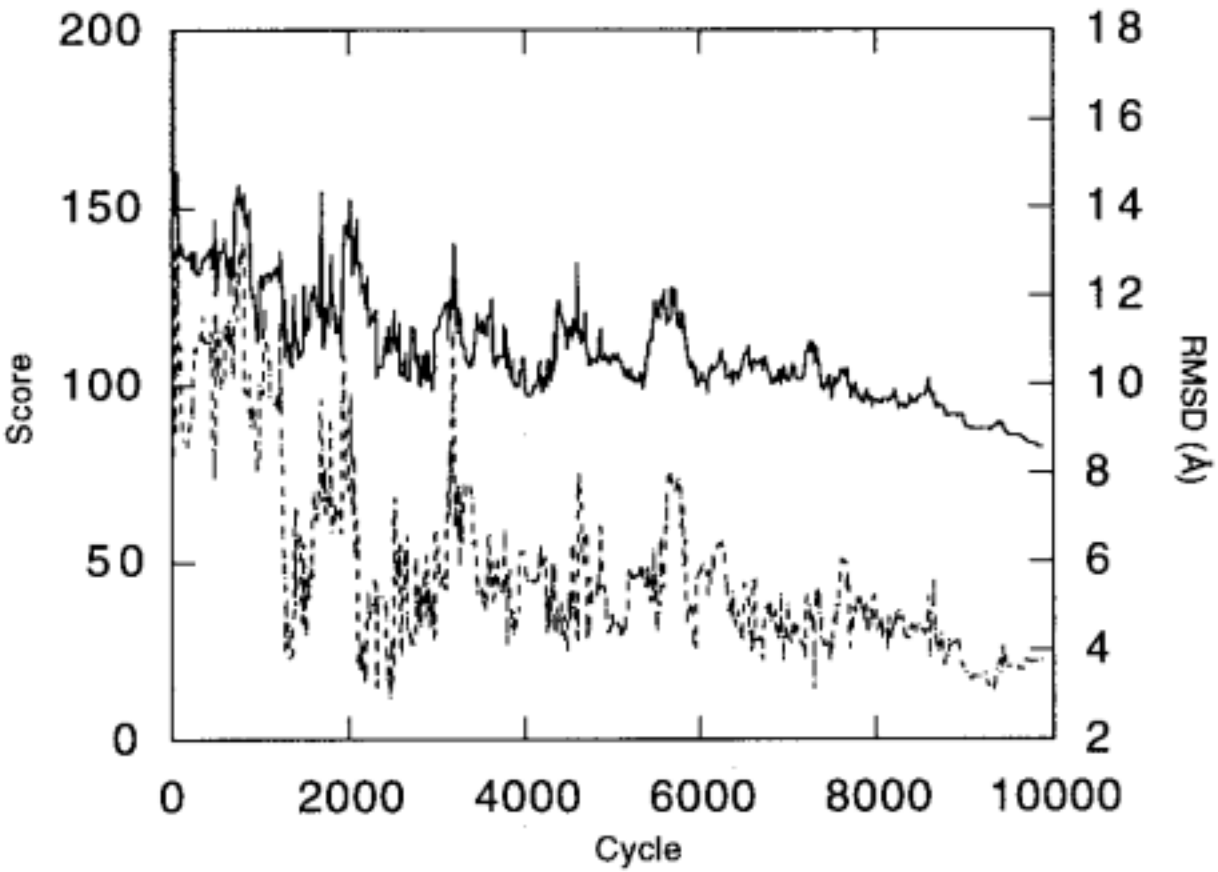


Figure 7. Progression of a homeodomain folding simulation. Continuous line, score; broken line, rmsd from the native structure. A cycle is an attempted replacement of the current torsion angles of a segment of the structure with the torsion angles of a fragment from the protein database with similar local sequence.

Table 2. Origins of fragments contributing to final simulated structures

	Structure I	Structure II
Residue	(2.7 Å rmsd, 2.1 Å dme)	(3.0 Å rmsd 2.1 Å dme)
1	Methyltransferase (1hmy)	Endonuclease III (1abk)
2	Creatinase (1chm)	Endonuclease III (1abk)
3	Cytochrome *c* (1ccr)	Endonuclease III (1abk)
4	Cytochrome *c* (1ccr)	Recoverin (1rec)
5	Cytochrome *c* (1ccr)	Recoverin (1rec)
6	Barley seed protein (1bw4)	Recoverin (1rec)
7	Hydrolase inhibitor (1hle)	3-isopropyl malate DH (1hex)
8	Ribose binding protein (2dri)	3-isopropyl malate DH (1hex)
9	HIN recombinase (1hcr)	Proteinase inhibitor (1cew)
10	HIN recombinase (1hcr)	Proteinase inhibitor (1cew)
11	HIN recombinase (1hcr)	Proteinase inhibitor (1cew)
12	Aspartate aminotransferase (1ars)	Histidine binding protein (1hsl)
13	Apolipoprotein-E3 (1lpe)	Cutinase (1cus)
14	Apolipoprotein-E3 (1lpe)	Leghemoglobin (1gdm)
15	Apolipoprotein-E3 (1lpe)	Leghemoglobin (1gdm)
16	Glutathione transferase (1gst)	Leghemoglobin (1gdm)
17	Glutathione transferase (1gst)	Uteroglobin (1utg)
18	Acyl transferase (3cla)	Uteroglobin (1utg)
19	Interleukin-10 (1ilk)	Uteroglobin (1utg)
20	Thermolysin (8tln)	Alpha-parvalbumin (1rtp)
21	Immunoglobin FC (1fc2)	Adenovirus fiber protein (1knb)
22	Immunoglobin FC (1fc2)	Adenovirus fiber protein (1knb)
23	Immunoglobin FC (1fc2)	Adenovirus fiber protein (1knb)
24	Dihydrofolate reductase (3dfr)	Alpha-parvalbumin (1rtp)
25	Dihydrofolate reductase (3dfr)	Phosphotransferase (1npk)

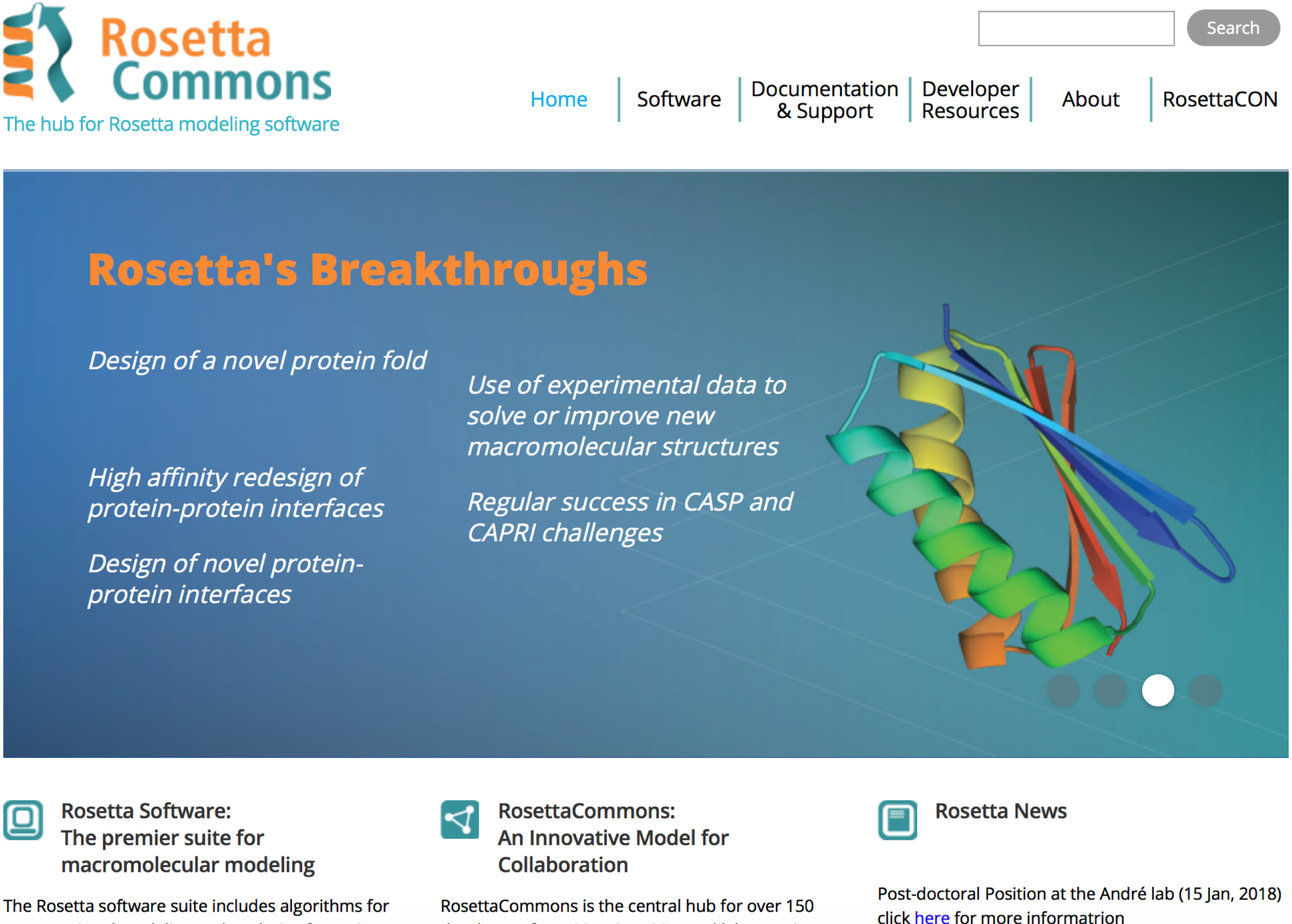
The proteins from which the final torsion angles of two simulated homeodomain structures originate are indicated for residues 1 to 25 of both structures.

Table 3. *Z*-scores for native-like conformations with different scoring functions

	1FC2A	1HDD	2CRO	4ICB	Average
Surface	−0.52	−0.23	−0.38	−0.48	−0.40
HF	−0.46	−0.68	−0.04	−0.69	−0.47
Contact(HL)	−0.41	−0.19	0.08	−0.38	−0.23
Contact(MJ)	−0.30	−0.13	0.08	−0.59	−0.24
Shell	−0.41	−0.48	−0.55	−1.05	−0.63
Shelltop	−0.39	−0.37	−0.42	−1.02	−0.55
Histogram	0.00	−0.04	−0.70	−0.48	−0.31
VdW(HL4)	−0.36	−0.69	−0.39	−1.31	−0.69
Shellm	−0.43	−0.54	−0.66	−0.59	−0.56
Shelltopm	−0.38	−0.56	−0.64	−0.89	−0.62
Eq(8)	−0.32	−0.69	−1.12	−0.87	−0.75
Eq(8) + msa	−0.32	−0.79	−1.08	−1.29	−0.87

The cutoff below which conformations were taken to be native-like was 4 Å rmsd for protein A and the homeodomain, and 5 Å rmsd for calbindin and cro repressor. The *Z*-scores (the number of standard deviations separating the scores of the native-like conformations from the ensemble average) were calculated over ensembles of 500 conformations for each protein generated using the "no filter" condition of Table 1.

Rosetta Software



Rosetta's Breakthroughs

Design of a novel protein fold

High affinity redesign of protein-protein interfaces

Design of novel protein-protein interfaces

Use of experimental data to solve or improve new macromolecular structures

Regular success in CASP and CAPRI challenges

Rosetta Software: The premier suite for macromolecular modeling

The Rosetta software suite includes algorithms for

RosettaCommons: An Innovative Model for Collaboration

RosettaCommons is the central hub for over 150

Rosetta News

Post-doctoral Position at the André lab (15 Jan, 2018)
click here for more informatrion

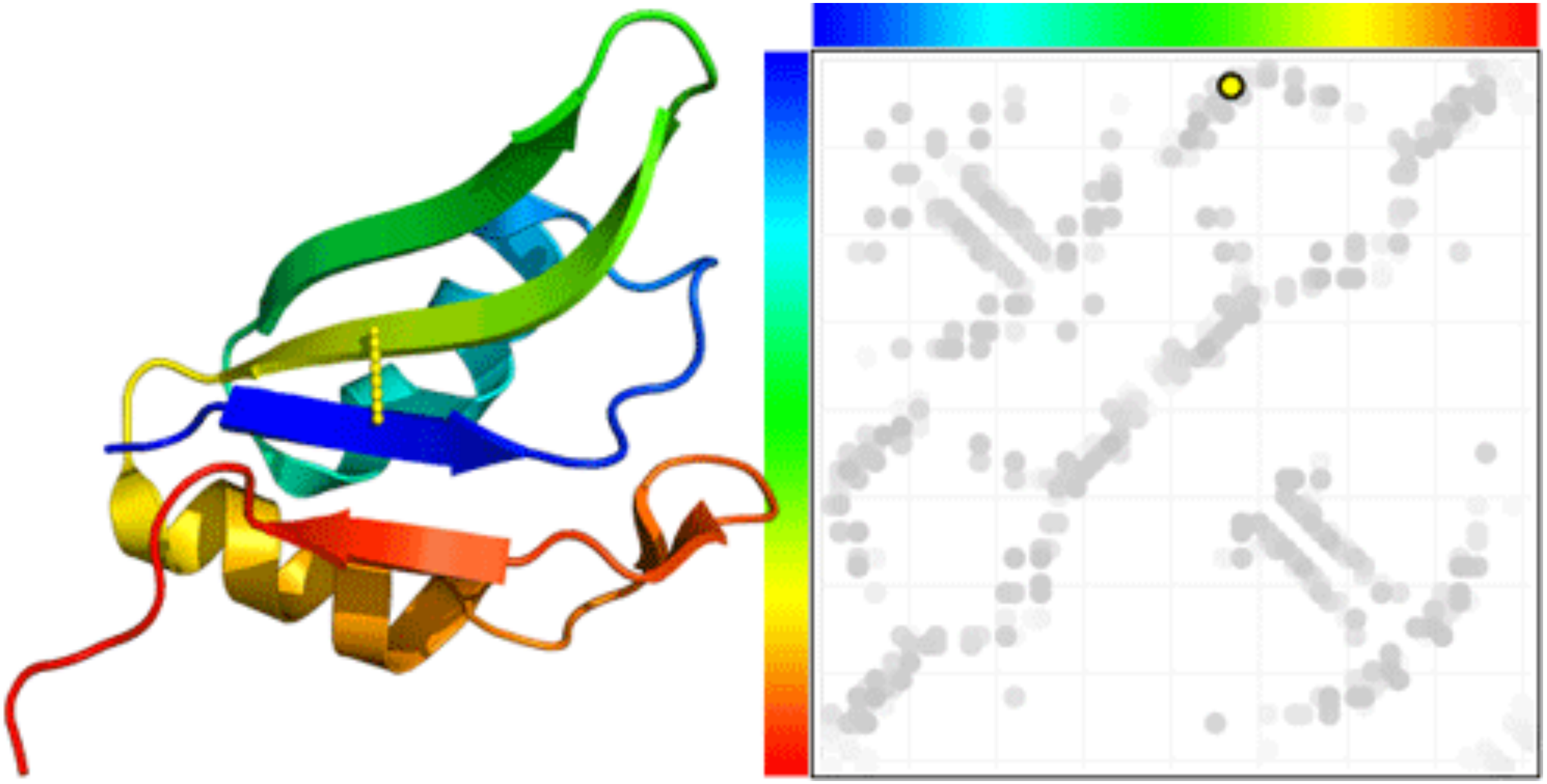
Part II. Distance-Based Ab Initio Modeling Empowered by Deep Learning

Limitations of Fragment-Assembly

- Work better on small, simple topology
- Low accuracy (0.2 – 0.3 GDT-TS score)
- Huge bottleneck (30% proteins)

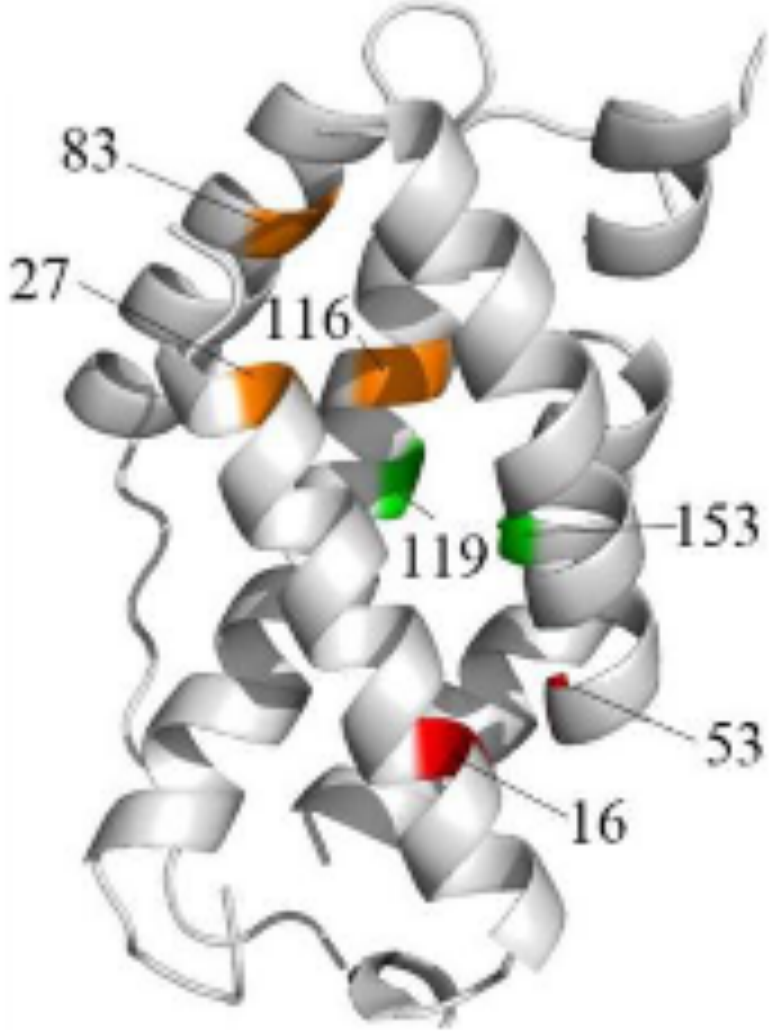
Walk in darkness without much clue!

Protein Contact Map



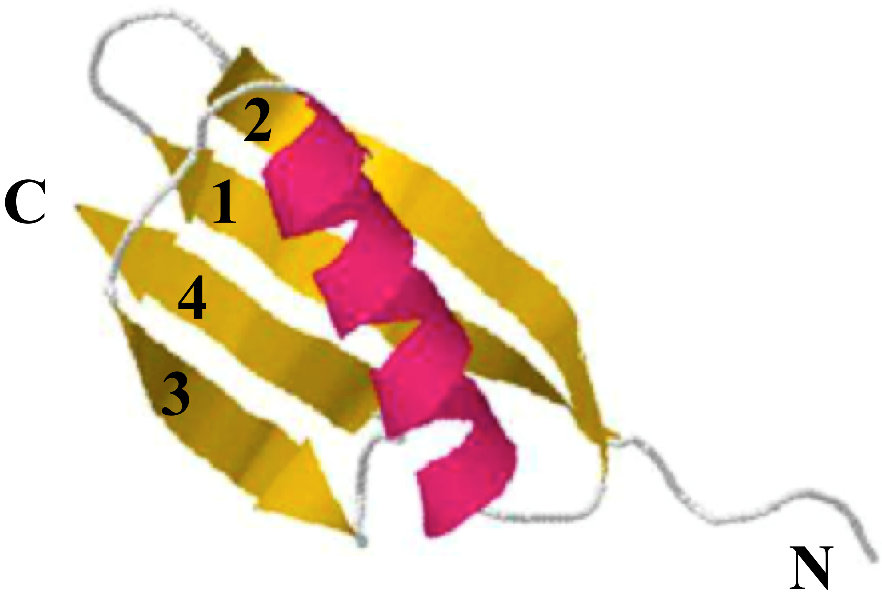
http://gremlin.bakerlab.org/gremlin_faq.php

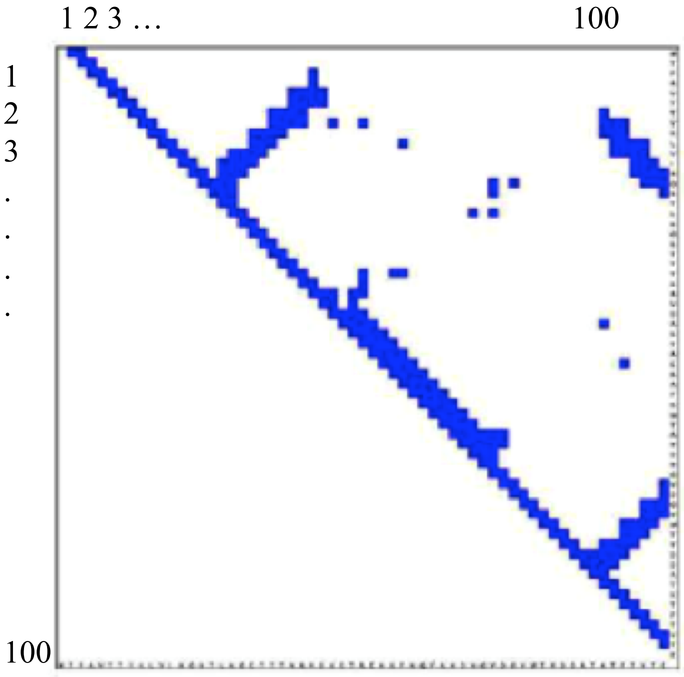
Residue-Residue Contact Prediction: A Binary Classification



Eickholt et al., 2011

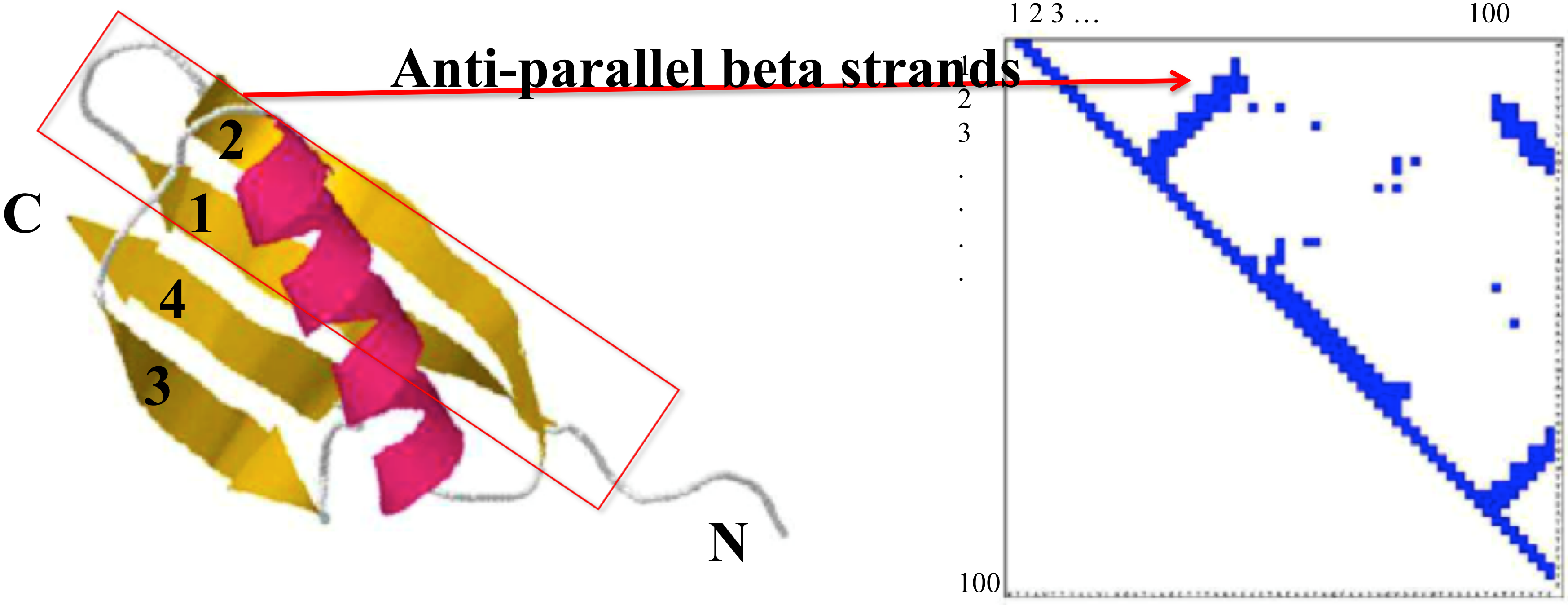
Residue-Residue Contact Prediction





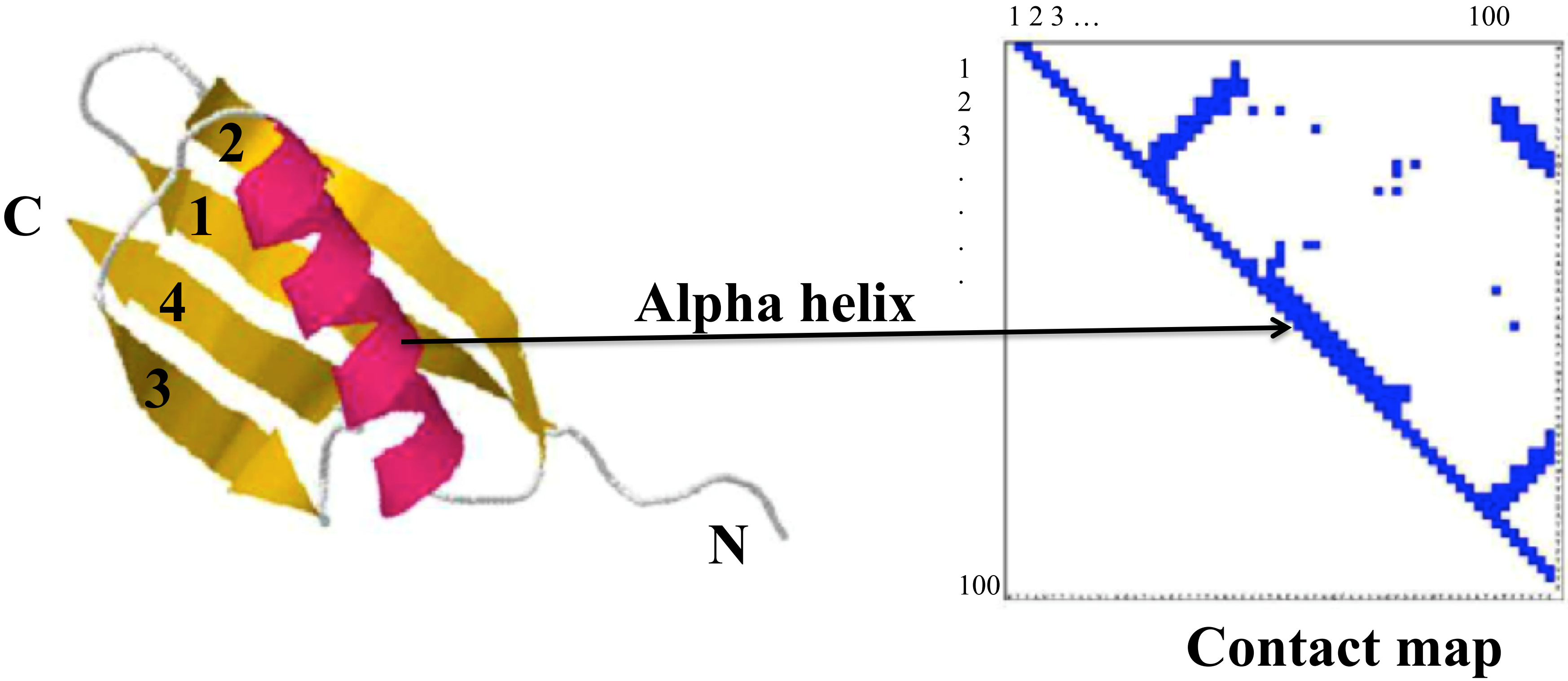
Contact map

Residue-Residue Contact Prediction

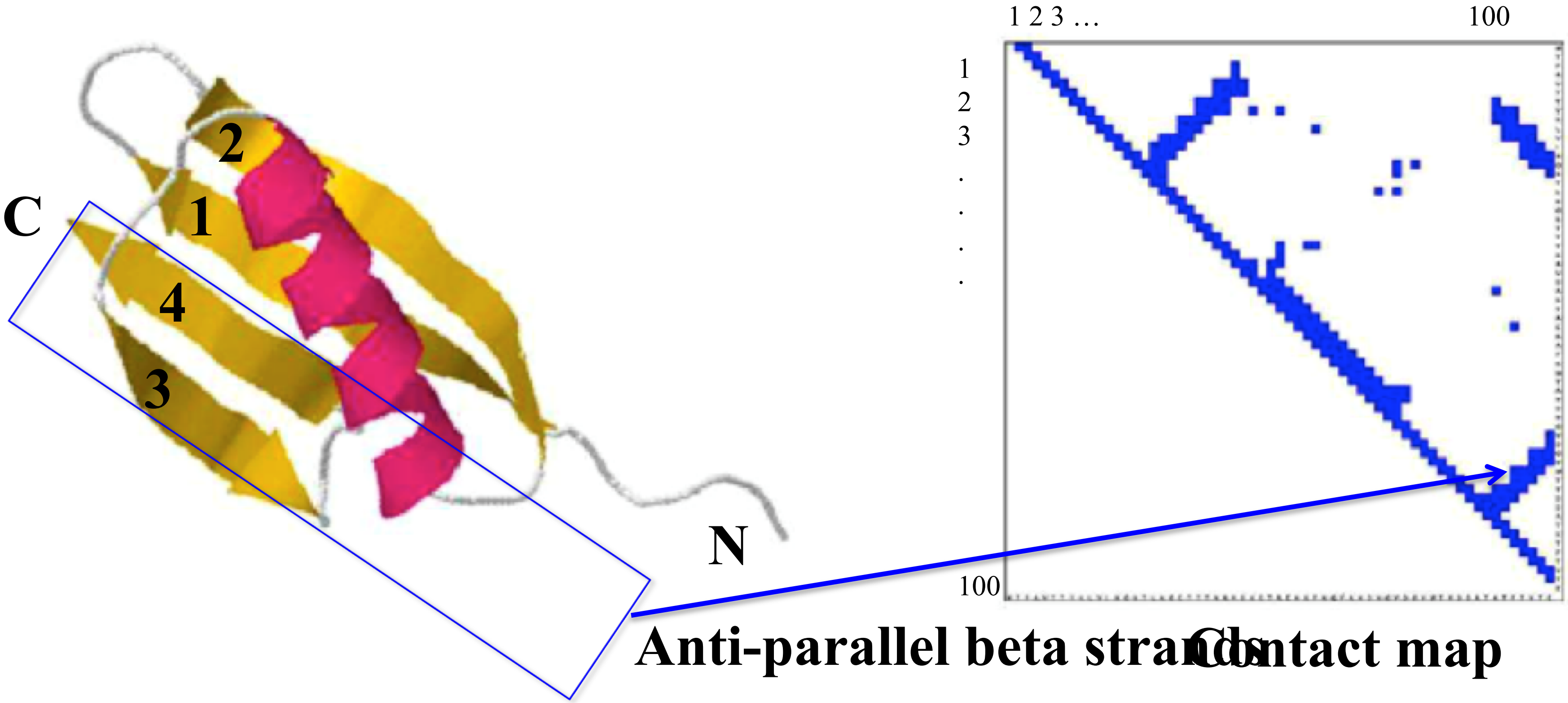


Contact map

Residue-Residue Contact Prediction



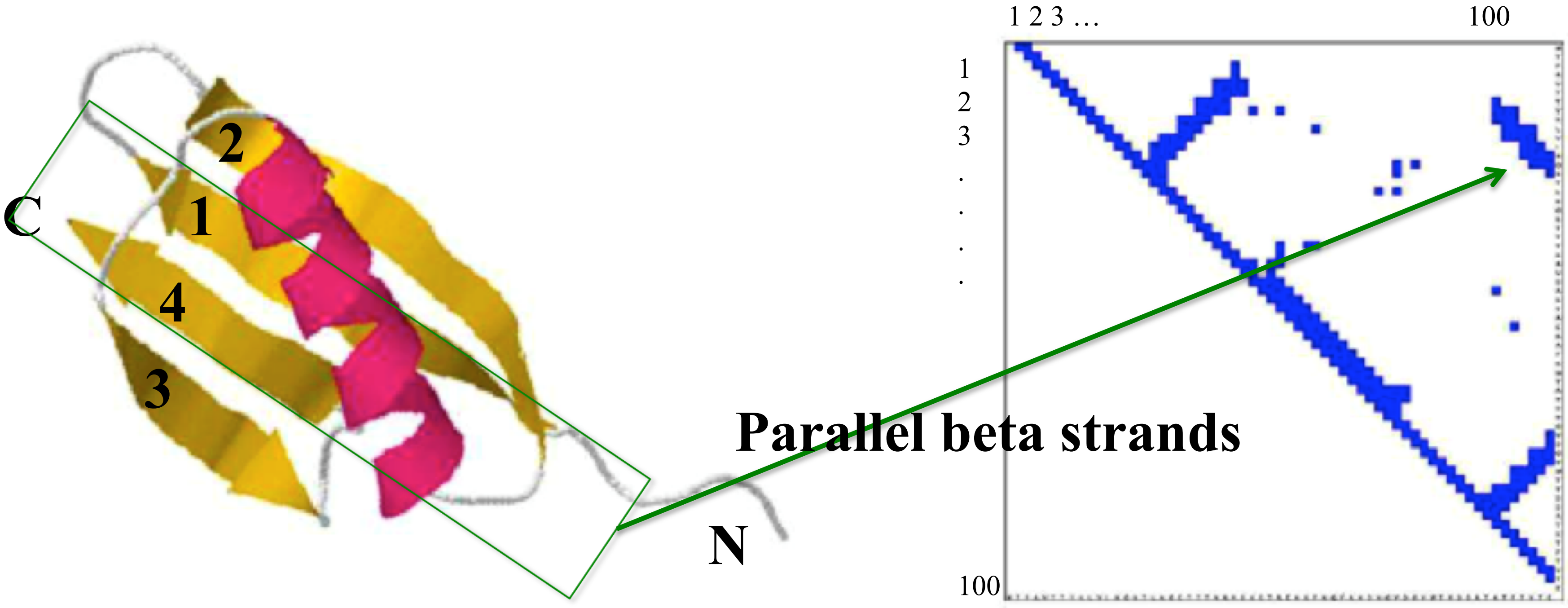
Residue-Residue Contact Prediction



Anti-parallel beta strands

Contact map

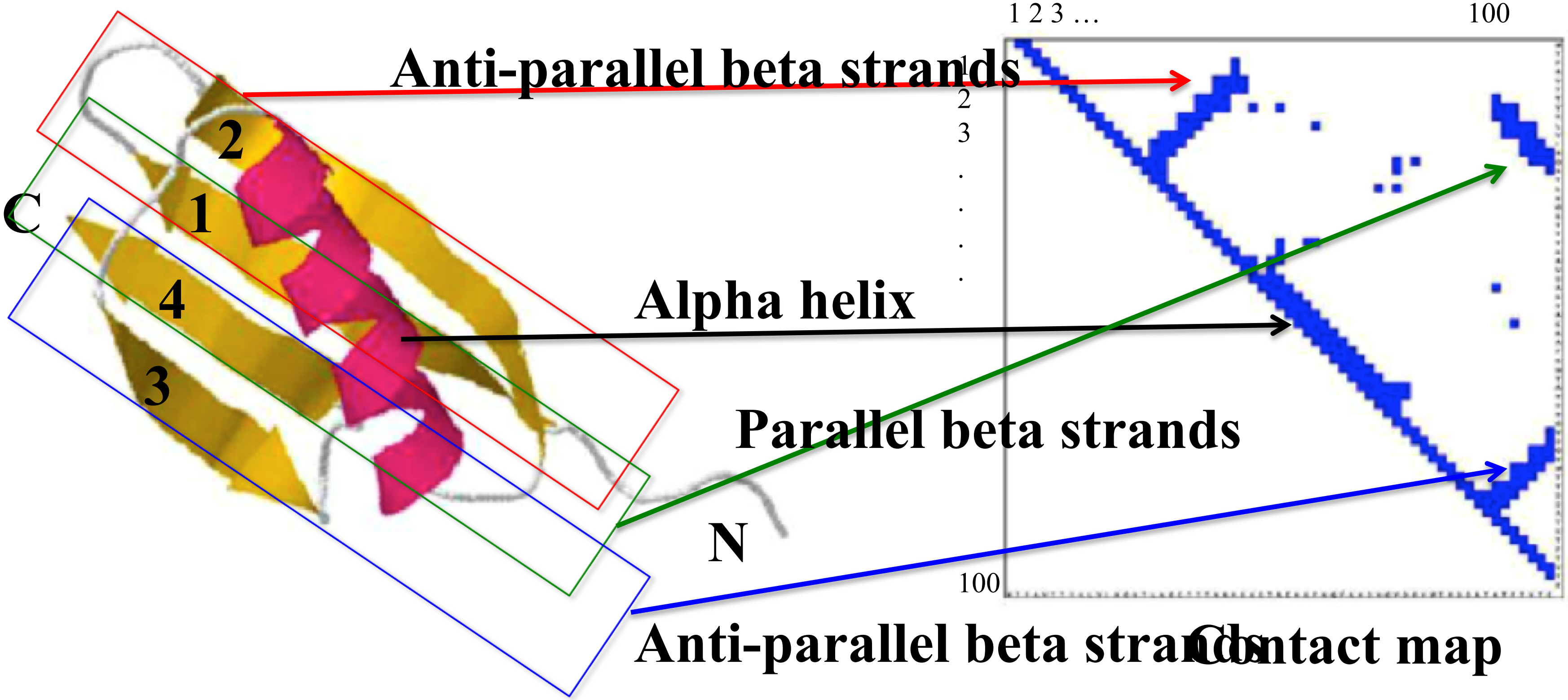
Residue-Residue Contact Prediction



Parallel beta strands

Contact map

Residue-Residue Contact Prediction



Anti-parallel beta strands

Alpha helix

Parallel beta strands

Anti-parallel beta strands

Contact map

SDDEVYQYIVSQVKQYGIEPAELLSRKYGDKAKYHLSQRW

i *j*

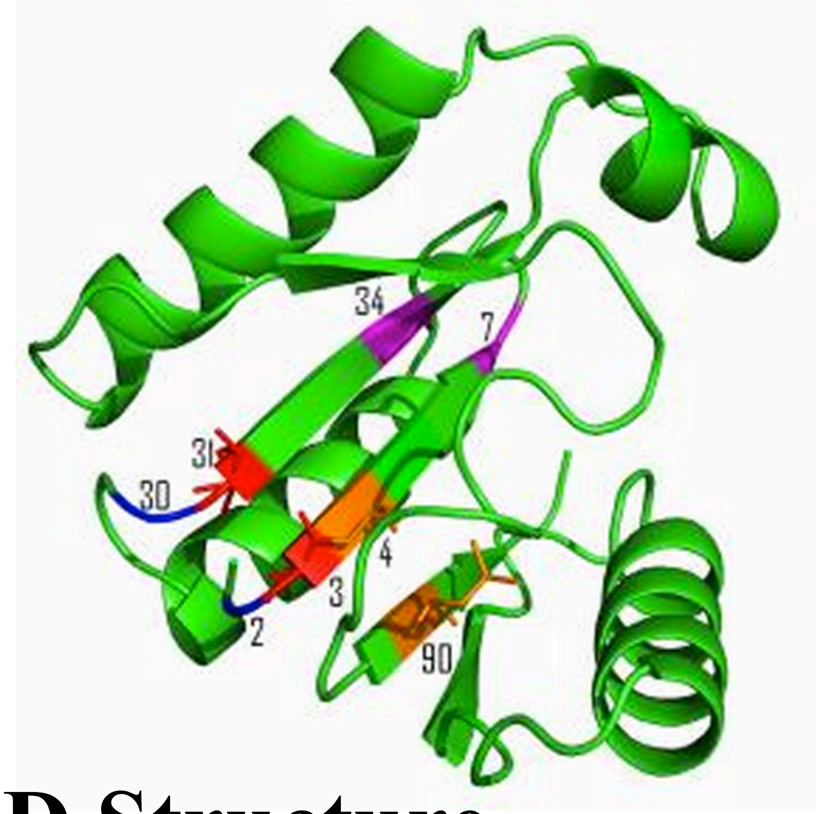
Objective:

Predict if two residues (*i*, *j*) are in contact, i.e. distance(*i*, *j*) < 8 Å, for |*i*-*j*| >= 6

Residue-Residue Contact Prediction

1D Sequence

SDDEVYQYIVSQVKQYGIEPAELLSRKYGDKAKYHLSQRW



3D Structure

Objective:

Predict if two residues (*i*, *j*) are in contact (spatially close), i.e. Distance(i, j) < 8 Angstrom

Cheng, Baldi, 2007; Eickholt, Cheng, 2012

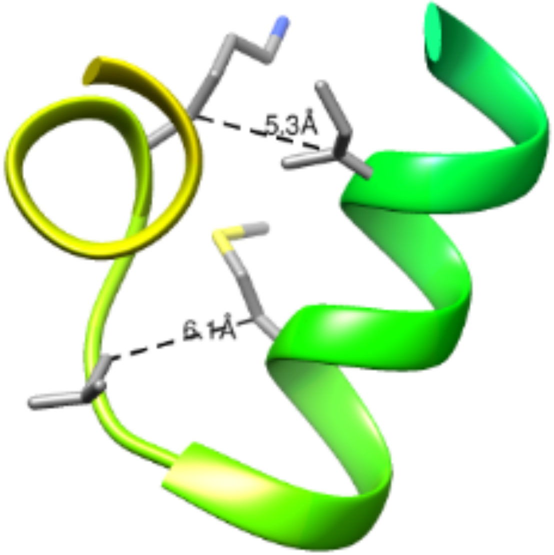
ConEVA Demo

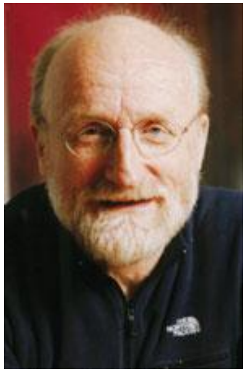
- ConEVA: http://iris.rnet.missouri.edu/cgi-bin/coneva/main_v2.0.cgi
- A protein structure: CASP13 target - T0958

Protein Contact Distance Prediction – A Major Breakthrough in *Ab Initio* Protein Structure Prediction in the Last 20 Years

- **Contact prediction (1994)**
- **Contact prediction until 2010 (little attention)**
- **Co-evolution and deep learning (2011 and 2012 in CASP10) – <u>two major advances</u>**
- **Contact prediction improved *ab initio* structure prediction (CASP11, 2014 and CASP12, 2016)**
- **CASP13 (Google's AlphaFold, MULTICOM, etc)**

Breakthrough I – Residue-Residue Co-evolutionary Analysis





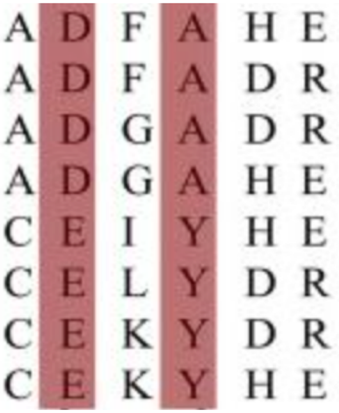
EVFOLD
Dr. Chris Sander at Memorial Sloan Kettering Cancer Center

EVFOLD
Dr. Debora Marks
Harvard Medical School

MetaPSICOV
Dr. David Jones at University College London (UCL)

GREMLIN
Dr. David Baker at University of Washington

Contact Prediction



A	D	F	A	H	E
A	D	F	A	D	R
A	D	G	A	D	R
A	D	G	A	H	E
C	E	I	Y	H	E
C	E	L	Y	D	R
C	E	K	Y	D	R
C	E	K	Y	H	E

FreeContact
Dr. Burkhard Rost at Technische Universität München (TUM)

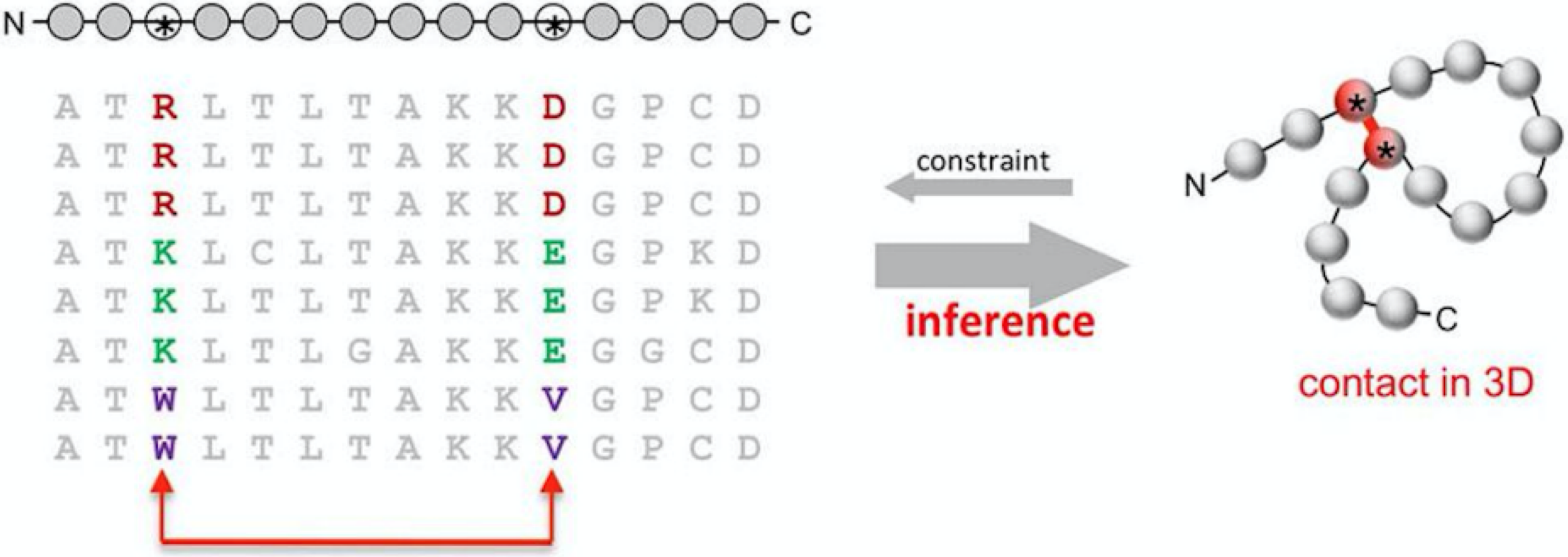
CCMpred
Dr. Johannes Söding at University of Munich

CMAppro
Dr. Pierre Baldi
UC Irvine

Distill
Dr. Gianluca Pollastri
U. College. Dublin

DNcon / SVMcon / NNcon
Dr. Jianlin Cheng at University of Missouri Columbia

Direct Co-Evolutionary Coupling Analysis



```
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T R L T L T A K K D G P C D
A T K L C L T A K K E G P K D
A T K L T L T A K K E G P K D
A T K L T L G A K K E G G C D
A T W L T L T A K K V G P C D
A T W L T L T A K K V G P C D
```

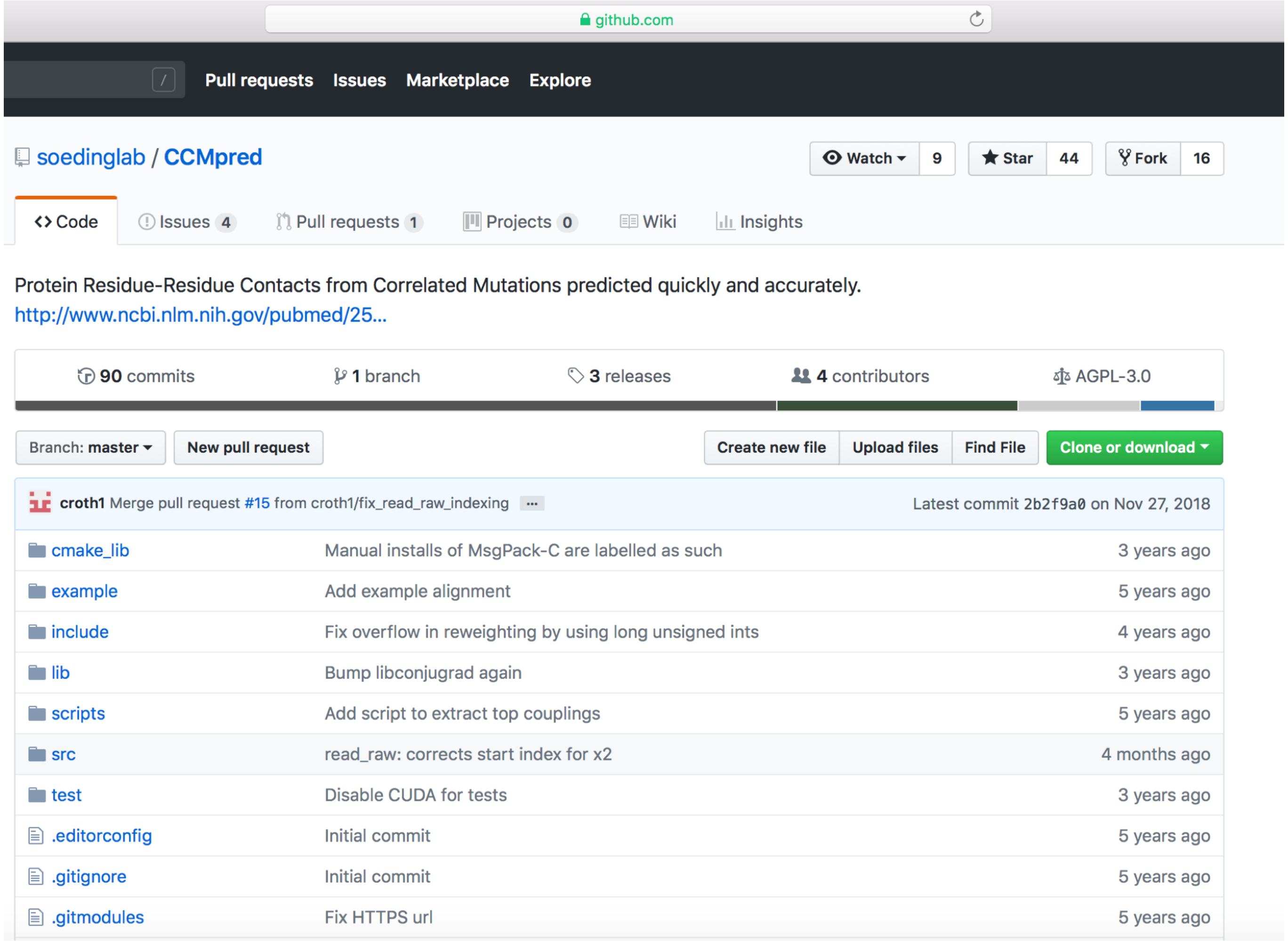
Calculate direct correlation caused by co-evolution (Marks et al., 2011)

Co-evolution plus neural networks (Jones et al., 2014; CASP11)

How to Get Multiple Sequence Alignment

- Hhblits – search a sequence against UniRef protein sequence database: https://github.com/soedinglab/hh-suite
- Jackhmmer – search a sequence aginast UniRef protein sequence database: http://hmmer.org

CCMpred



soedinglab / CCMpred

Protein Residue-Residue Contacts from Correlated Mutations predicted quickly and accurately.
http://www.ncbi.nlm.nih.gov/pubmed/25...

90 commits · **1** branch · **3** releases · **4** contributors · AGPL-3.0

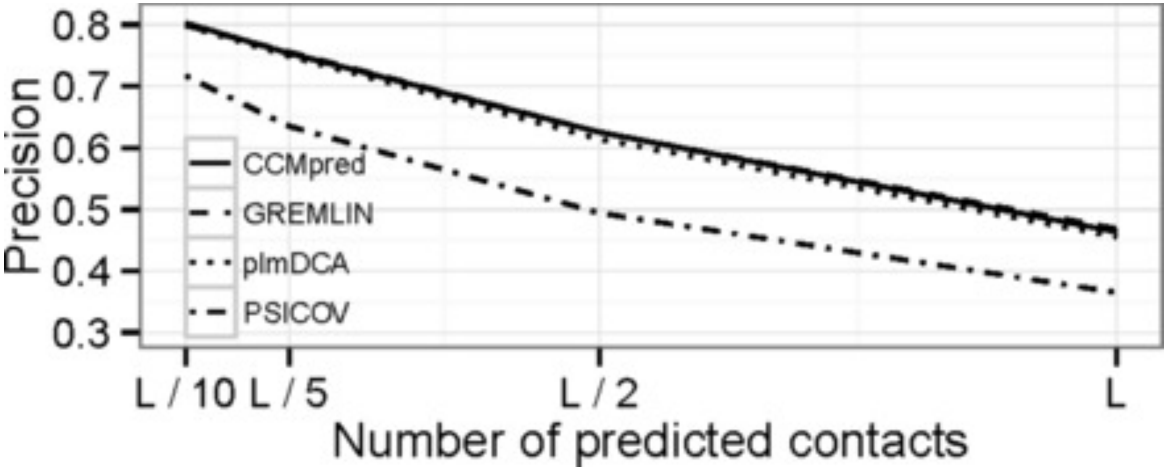
croth1 Merge pull request #15 from croth1/fix_read_raw_indexing ... Latest commit 2b2f9a0 on Nov 27, 2018

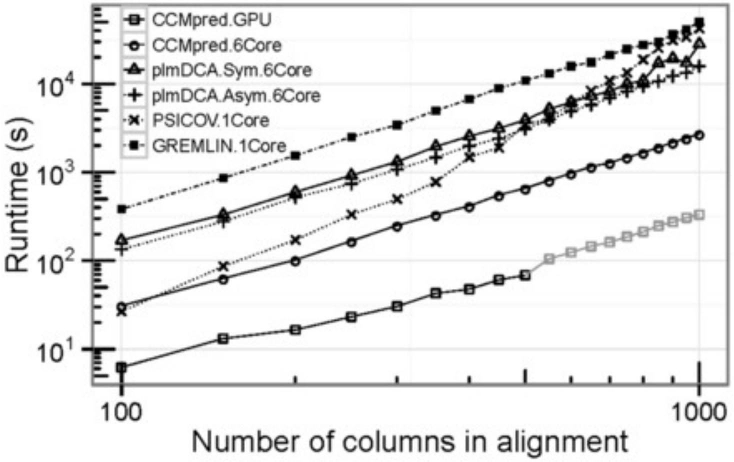
cmake_lib	Manual installs of MsgPack-C are labelled as such	3 years ago
example	Add example alignment	5 years ago
include	Fix overflow in reweighting by using long unsigned ints	4 years ago
lib	Bump libconjugrad again	3 years ago
scripts	Add script to extract top couplings	5 years ago
src	read_raw: corrects start index for x2	4 months ago
test	Disable CUDA for tests	3 years ago
.editorconfig	Initial commit	5 years ago
.gitignore	Initial commit	5 years ago
.gitmodules	Fix HTTPS url	5 years ago

https://github.com/soedinglab/CCMpred

How to generate co-evolutionary scores

CCMPred



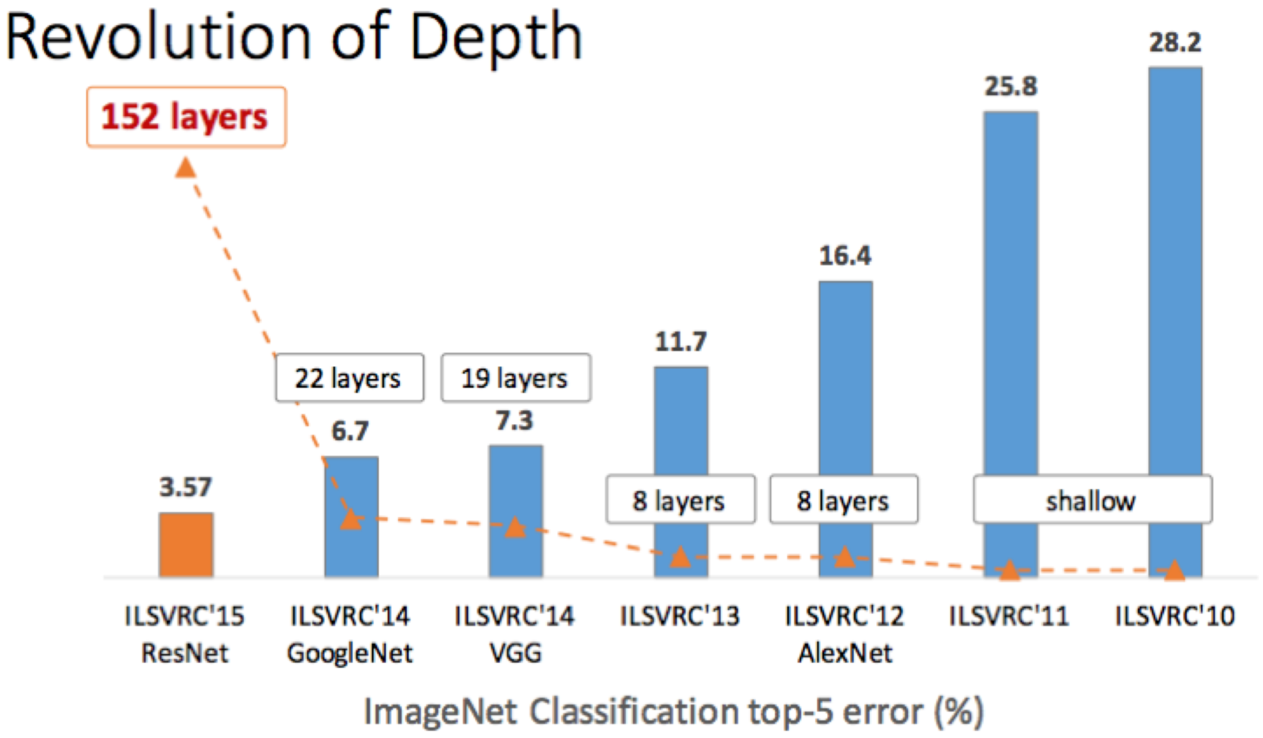


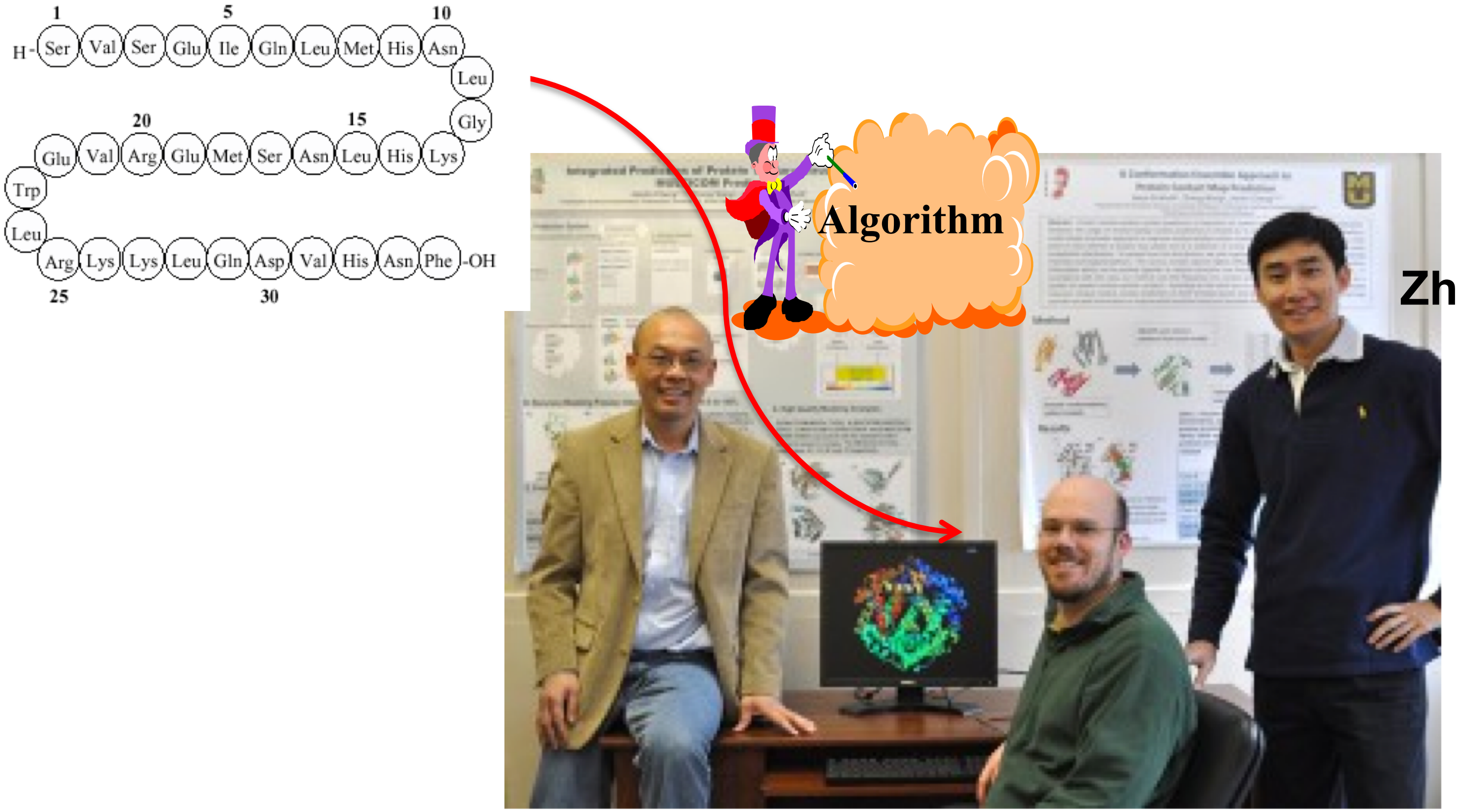
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4201158/

Breakthrough II

- Deep Learning for Contact Prediction (DNCON1) (Eickholt, Cheng, 2012)
- No. 1 in CASP10, 2012
- One of the first deep learning methods for bioinformatics

Deep Learning Revolution

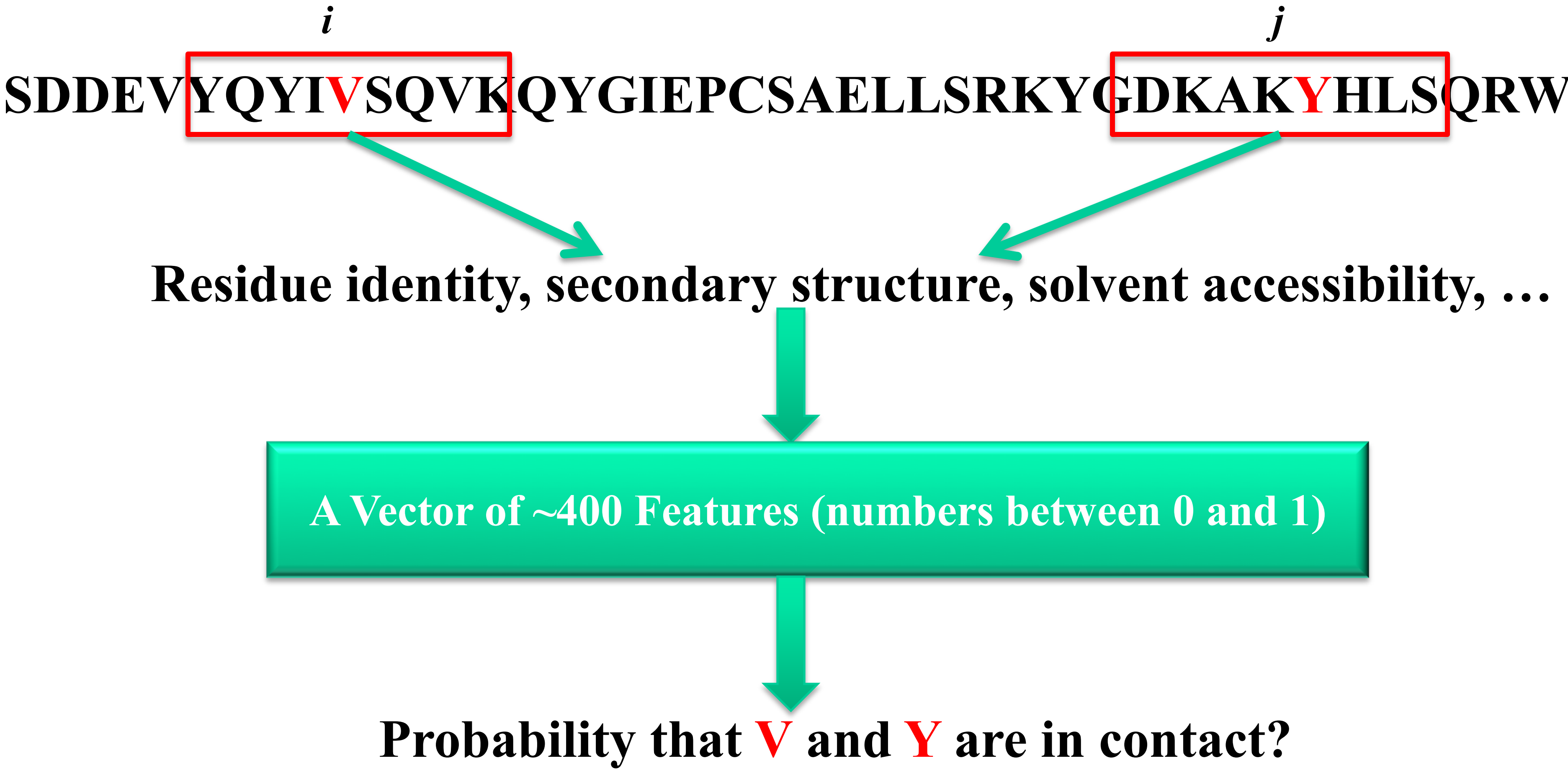



1
5
10
H-Ser Val Ser Glu Ile Gln Leu Met His Asn
Leu
Gly
20
15
Glu Val Arg Glu Met Ser Asn Leu His Lys
Trp
Leu
Arg Lys Lys Leu Gln Asp Val His Asn Phe-OH
25
30
Algorithm

Zheng Wang

Jesse Eickholt

A Binary Classification Problem



i *j*

SDDEVYQYIVSQVKQYGIEPCSAELLSRKYGDKAKYHLSQRW

Residue identity, secondary structure, solvent accessibility, …

A Vector of ~400 Features (numbers between 0 and 1)

Probability that V and Y are in contact?

Cheng & Baldi, 2007; Tegge et al., 2009; Eickholt, Cheng, 2012

Input Features

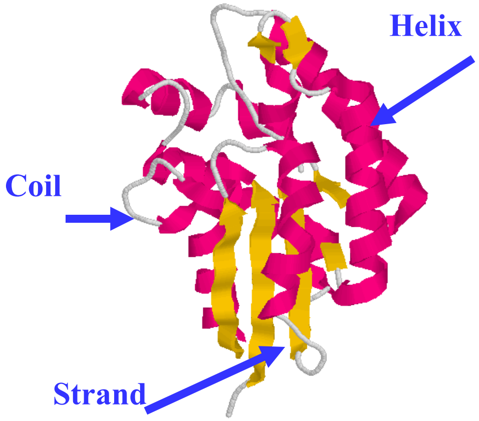
i *j*

SDDEVYQYIVSQVKQYGIEPCSAELLSRKYGDKAKYHLSQRW

20 binary numbers

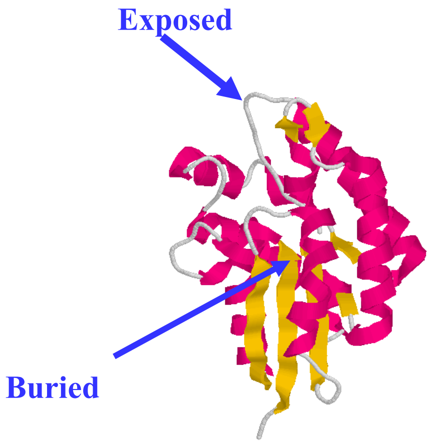
A	10000000000000000000
C	01000000000000000000
D	00100000000000000000
.	
.	
.	
.	
.	
.	
.	
Y	00000000000000000001

3 numbers



Helix	100
Strand	010
Coil	001

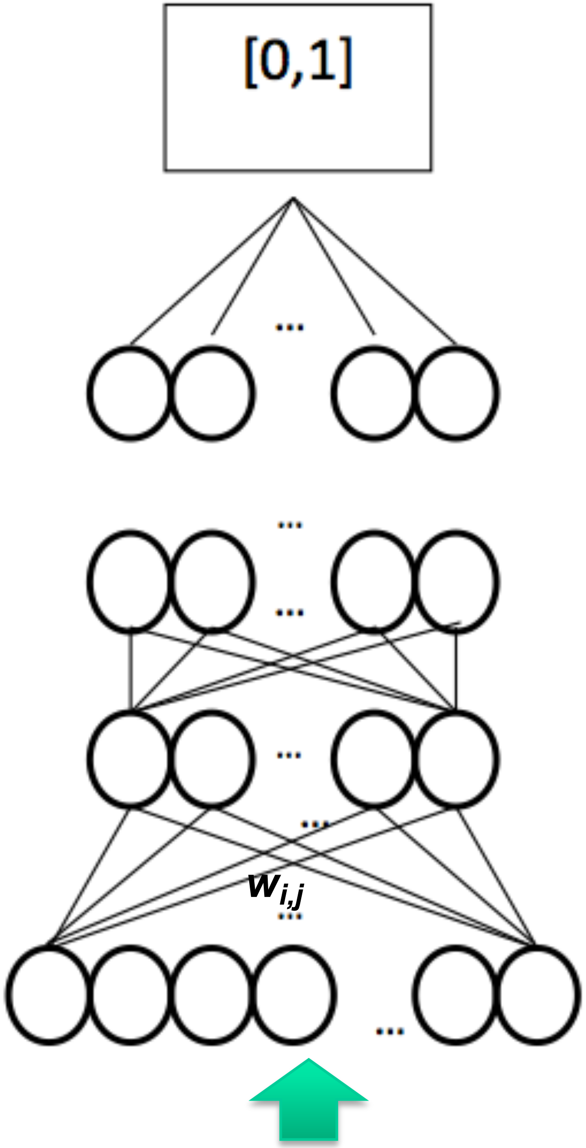
2 numbers



Exposed	10
Buried	01

25 * 18 = 400 features for a pair (*i*, *j*)

Deep Learning Network Architecture



~350 nodes

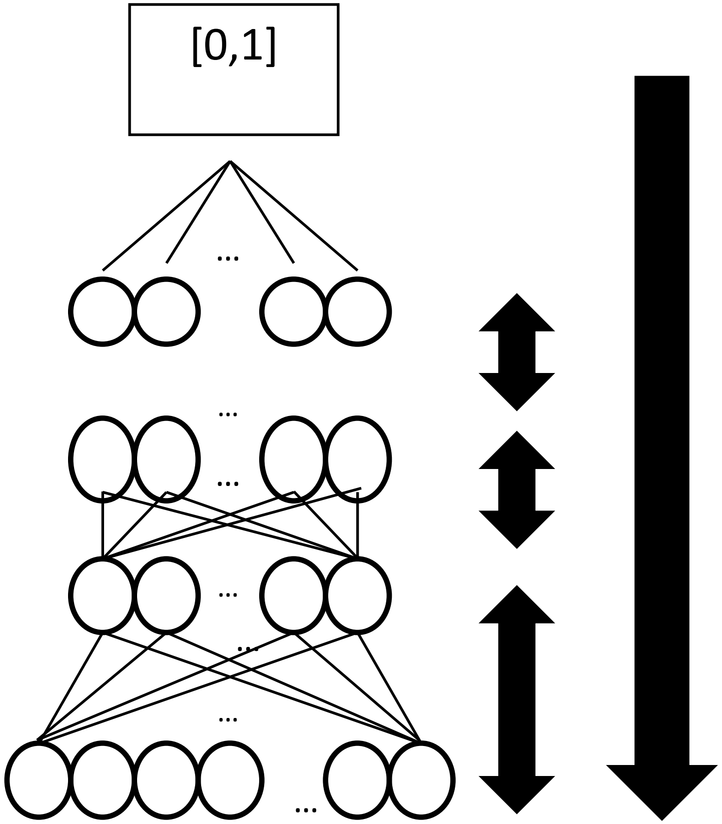
~500 nodes

~500 nodes

~400 input nodes

A Vector of ~400 Features (numbers between 0 and 1)

Training a Deep Network

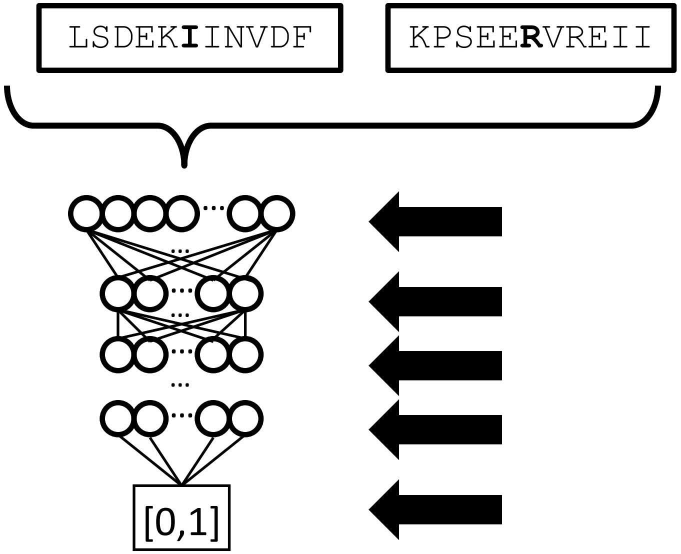


1239 Proteins for Training
Residue Pairs (|i-j| >= 6)

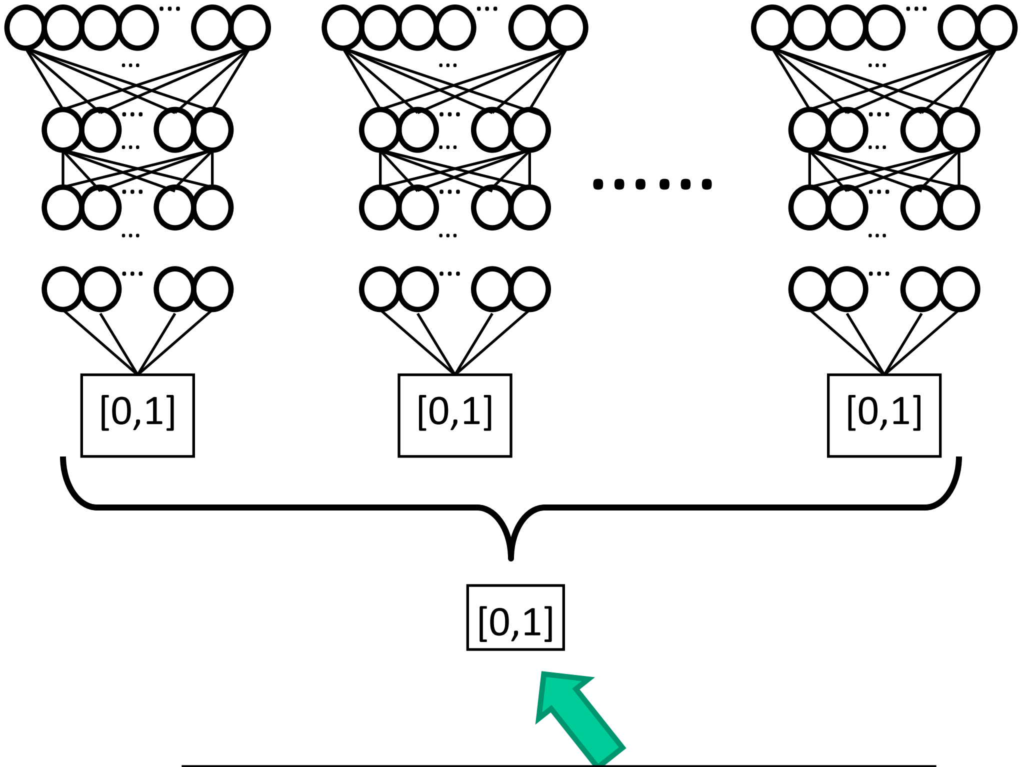
Specific Implementation on GPU

Speed up training by CUDAMat and GPUs

Train DNs with over 1M parameters in about an hour



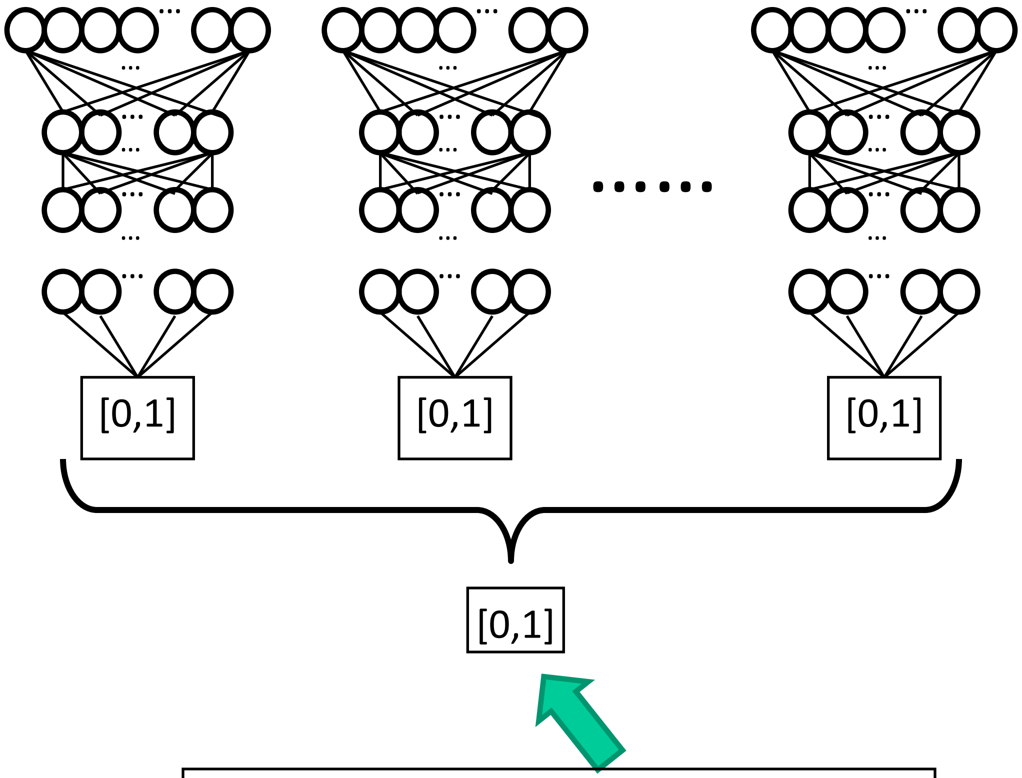
Boosted Ensembles for Contact Prediction



Final output of ensemble is a performance weighted sum of individual DN outputs.

Eickholt and Cheng, *Bioinformatics* (2012)

Boosted Ensembles for Contact Prediction



Final output of ensemble is a performance weighted sum of individual DN outputs.

Eickholt and Cheng, *Bioinformatics* (2012)

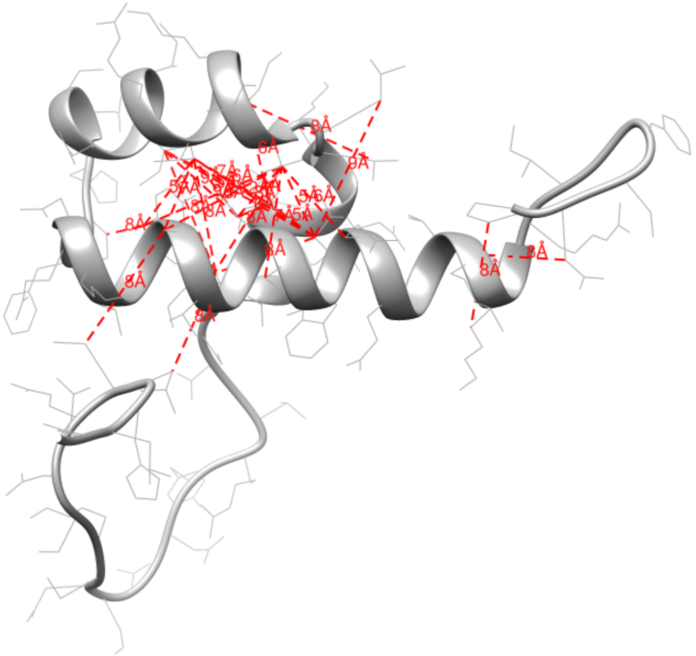
Benchmarking and Evaluation Metrics

Accuracy of top L, L/5, or L/10 predictions for various ranges of sequence separation (medium- and long-range): [TP/(TP+FP)]

Results on Test Data Set (196 Proteins)

Metric	Acc. L/5	Acc. L/5 (one shift)
Short Range (6 <= |i-j| <12)	0.51	0.79
Medium Range (12 <= |i-j| <24)	0.38	0.65
Long Range (|i-j| >= 24)	0.34	0.55

An Example:



Blind Test on CASP10 Targets

Exact match (96 proteins, long-range contacts)

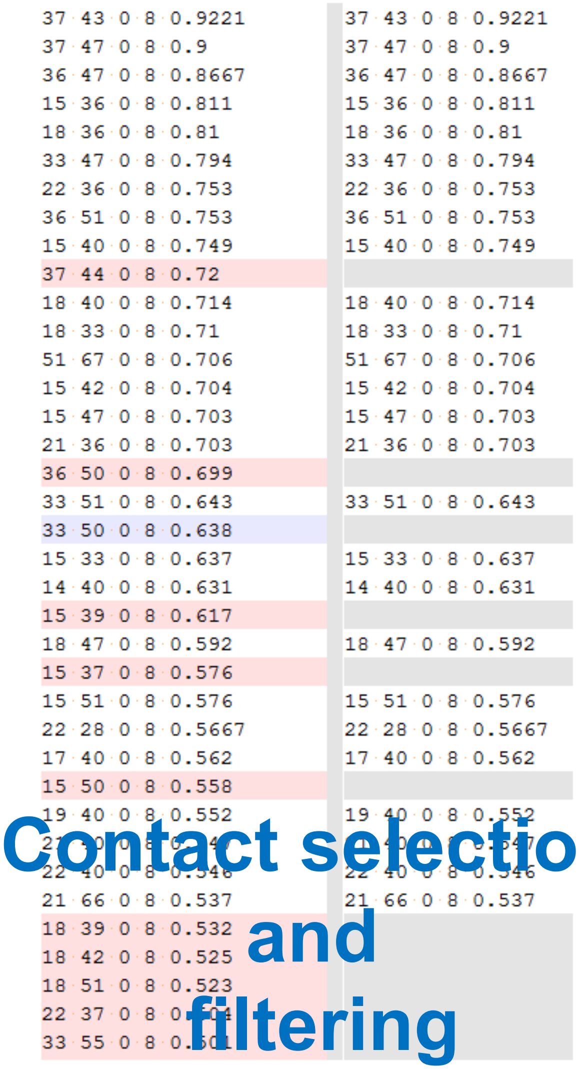
Method	Acc. L/5
DNcon	**0.30**
SVMcon	0.19

→ 9-fold better than random

Inexact match with minor shifts

Method	δ	Acc. L/5
DNcon	**1**	**0.53**
SVMcon	1	0.37
DNcon	**2**	**0.62**
SVMcon	2	0.45

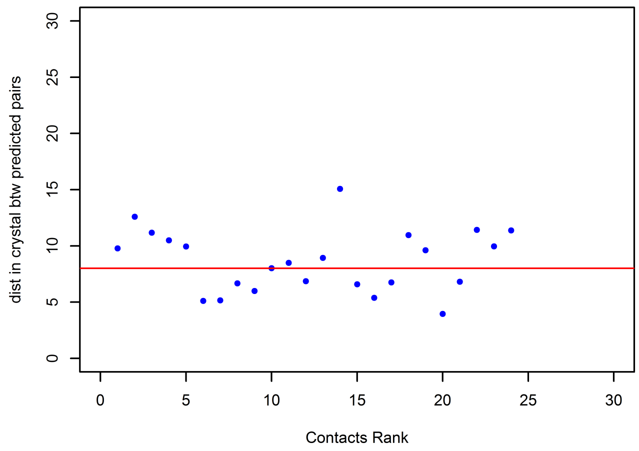
3D Reconstruction from Predicted Contacts (CASP Target T0716)

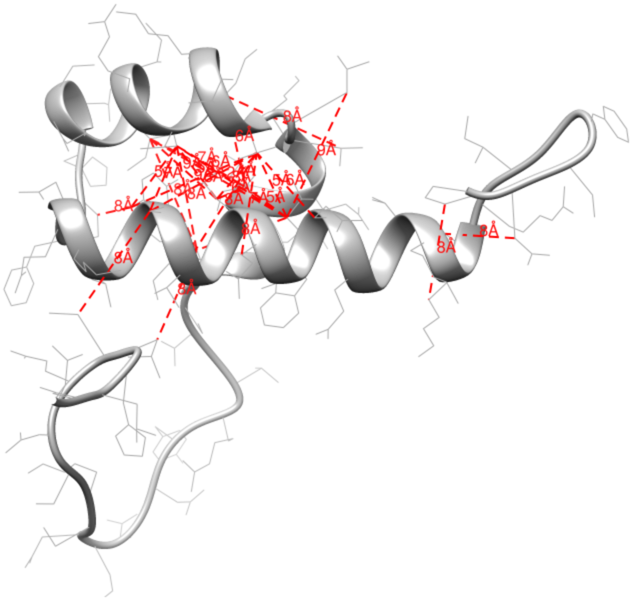


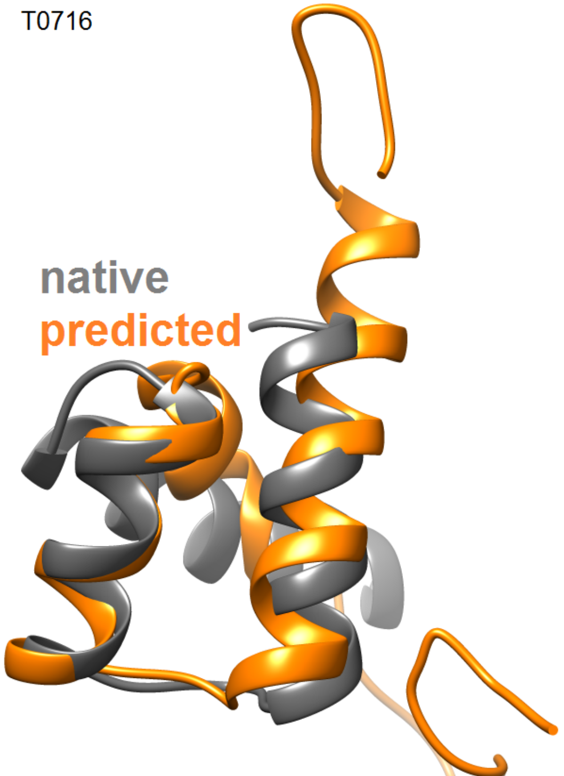
original	top 0.4L (33% SR, 33% MR, 33%SR)
37 43 0 8 0.9221	37 43 0 8 0.9221
37 47 0 8 0.9	37 47 0 8 0.9
36 47 0 8 0.8667	36 47 0 8 0.8667
15 36 0 8 0.811	15 36 0 8 0.811
18 36 0 8 0.81	18 36 0 8 0.81
33 47 0 8 0.794	33 47 0 8 0.794
22 36 0 8 0.753	22 36 0 8 0.753
36 51 0 8 0.753	36 51 0 8 0.753
15 40 0 8 0.749	15 40 0 8 0.749
37 44 0 8 0.72	
18 40 0 8 0.714	18 40 0 8 0.714
18 33 0 8 0.71	18 33 0 8 0.71
51 67 0 8 0.706	51 67 0 8 0.706
15 42 0 8 0.704	15 42 0 8 0.704
15 47 0 8 0.703	15 47 0 8 0.703
21 36 0 8 0.703	21 36 0 8 0.703
36 50 0 8 0.699	
33 51 0 8 0.643	33 51 0 8 0.643
33 50 0 8 0.638	
15 33 0 8 0.637	15 33 0 8 0.637
14 40 0 8 0.631	14 40 0 8 0.631
15 39 0 8 0.617	
18 47 0 8 0.592	18 47 0 8 0.592
15 37 0 8 0.576	
15 51 0 8 0.576	15 51 0 8 0.576
22 28 0 8 0.5667	22 28 0 8 0.5667
17 40 0 8 0.562	17 40 0 8 0.562
15 50 0 8 0.558	
19 40 0 8 0.552	19 40 0 8 0.552
21 40 0 8 0.547	21 40 0 8 0.547
22 40 0 8 0.546	22 40 0 8 0.546
21 66 0 8 0.537	21 66 0 8 0.537
18 39 0 8 0.532	
18 42 0 8 0.525	
18 51 0 8 0.523	
22 37 0 8 0.504	
33 55 0 8 0.501	

Contact selection and filtering

T0716 top 0.4L contacts (30% SR, 30% MR, 30% LR







Target : T0716 (CASP10)

Length : 71

RMSD : 4.3A

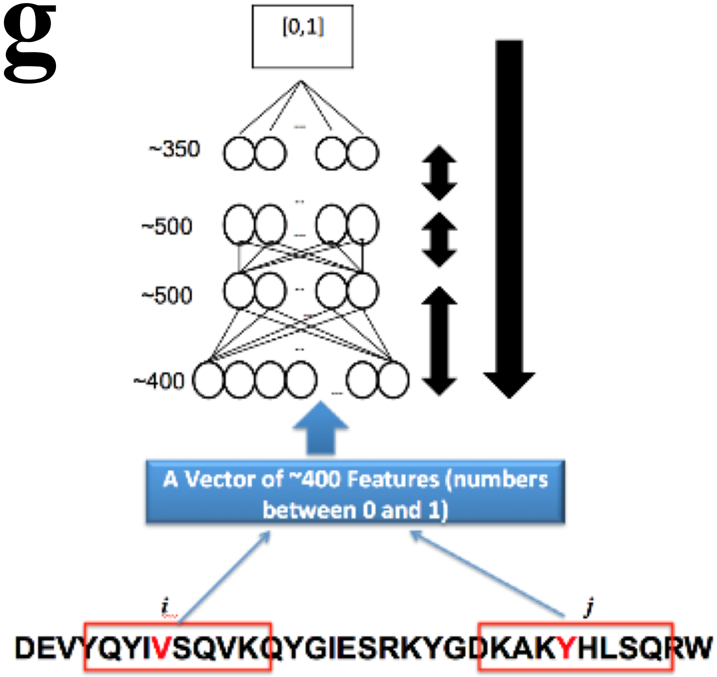
GDT-TS : 0.58

Contacts : DNcon (filtered and selected 0.4L)

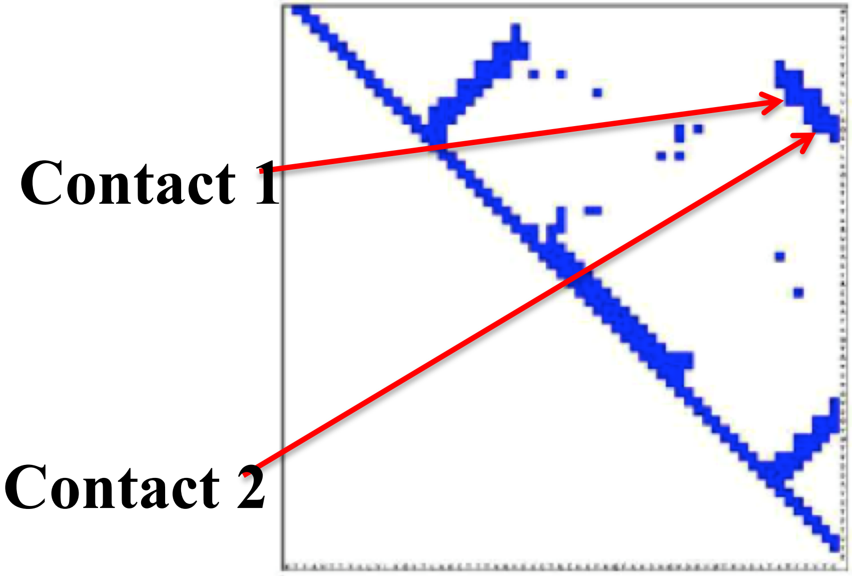
Selection: Best Structure

Deep Learning

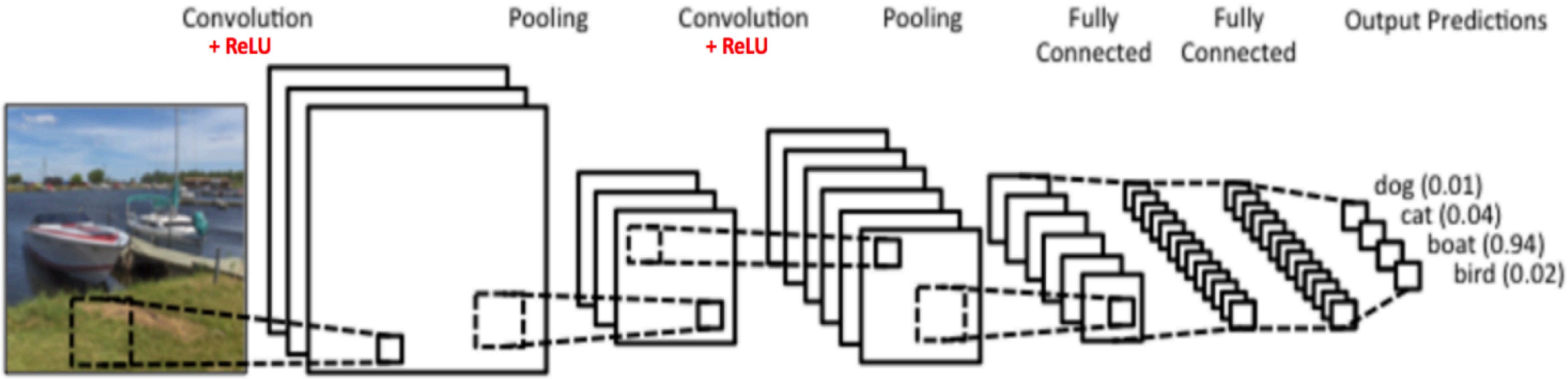
- **Deep Learning (CASP10; Eickholt and Cheng, 2012)**


[0,1]
~350
~500
~500
~400
A Vector of ~400 Features (numbers between 0 and 1)
i
j
DEVYQYIVSQVKQYGIESRKYGDKAKYHLSQRW

- **2D Convolutional Neural Networks (CASP12; Wang et al., 2017; Adhikari et al., 2017)**



Deep Convolutional Neural Network

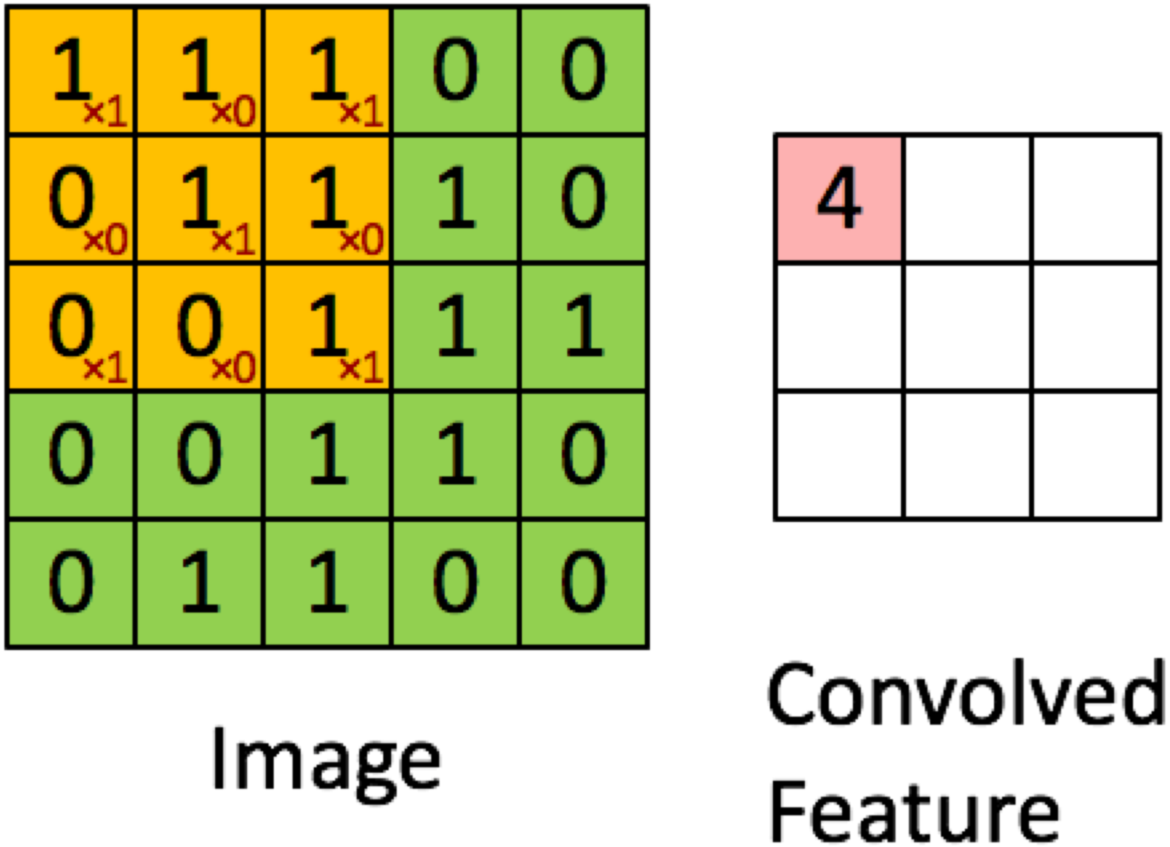


- **Automatic feature extraction without hand crafting**

- **Feature composition from local (low level) to global (high level)**

Google Image

A Convolution Example



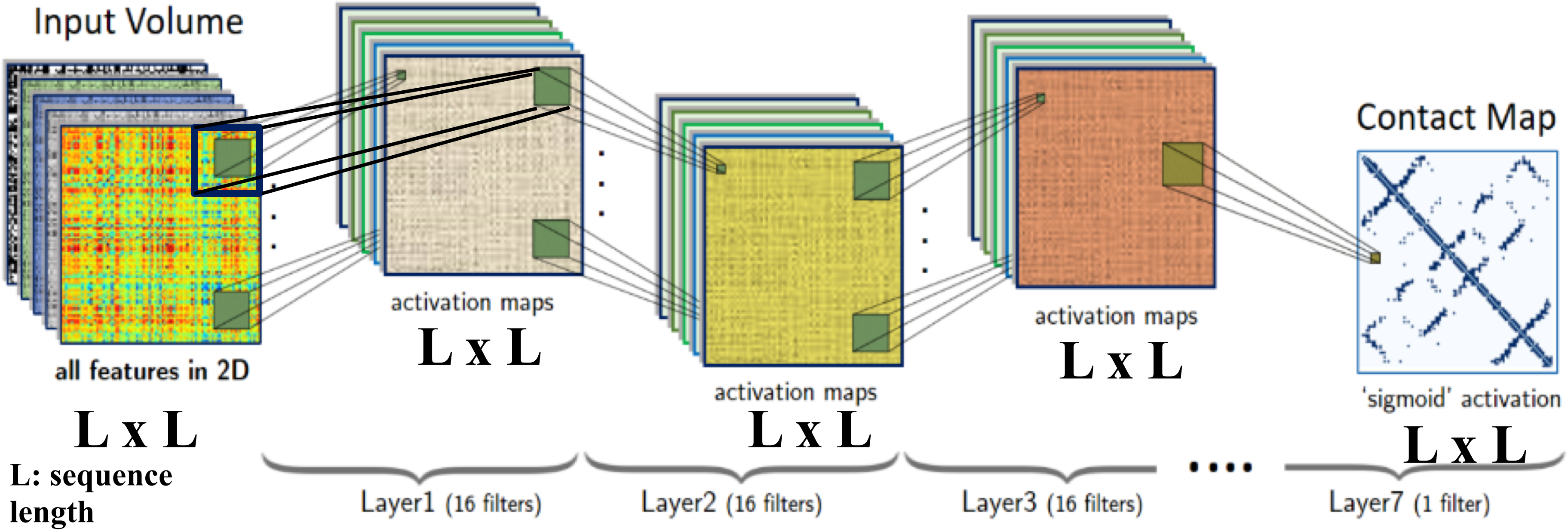
1 ×1	1 ×0	1 ×1	0	0
0 ×0	1 ×1	1 ×0	1	0
0 ×1	0 ×0	1 ×1	1	1
0	0	1	1	0
0	1	1	0	0

4		

Deep Learning lecture, Google

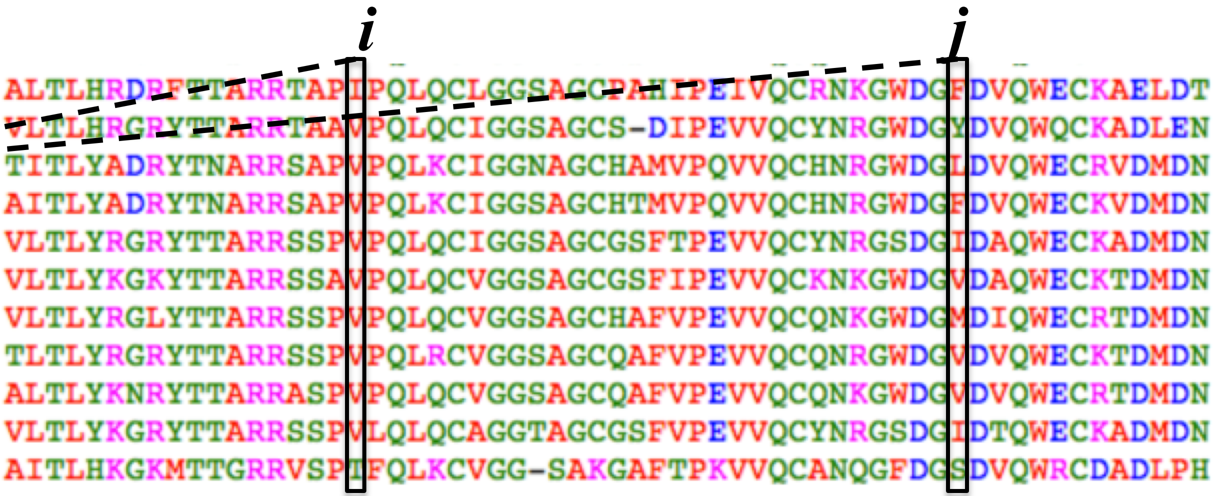
2D Convolutional Neural Network for Contact Prediction (DNCON2)

Adhikari et al., 2017



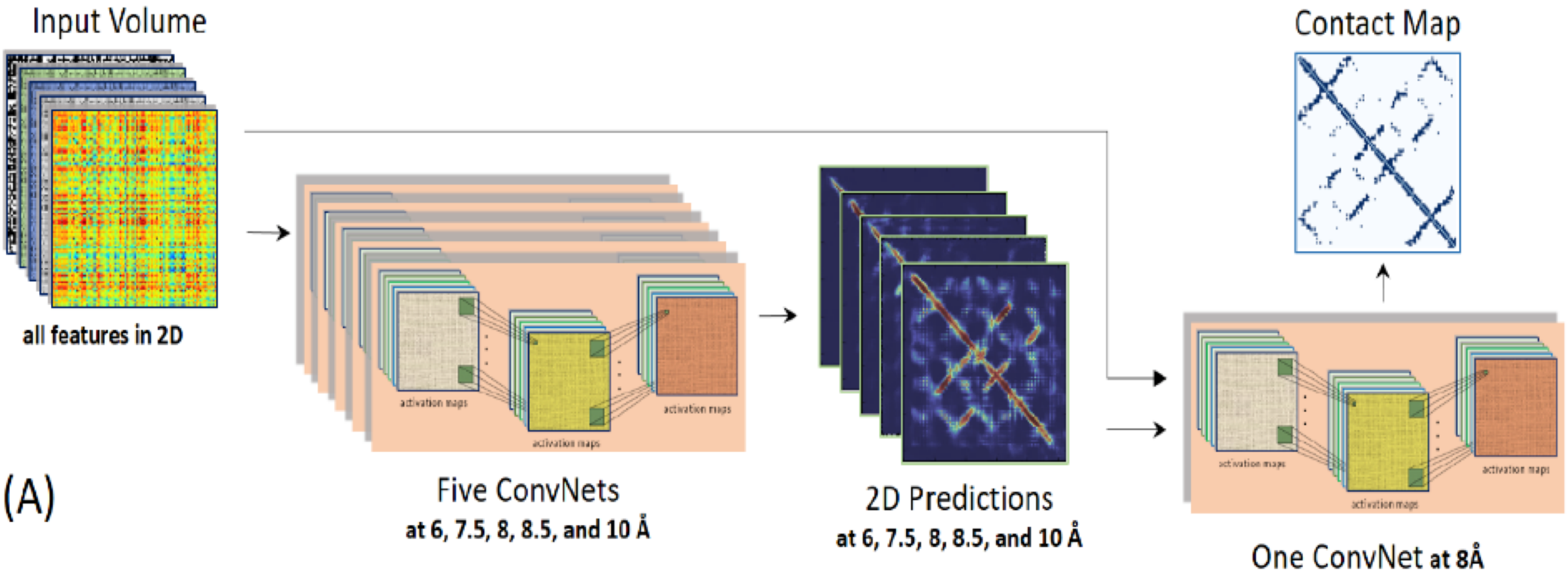
2D Input Matrices

- **Co-evolution**
- Secondary structure
- Solvent accessibility
- Mutual information
- Contact potentials
- …



```
ALTLHRDRFTTARRTAPIPQLQCLGGSAGCPAHIPEIVQCRNKGWDGFDVQWECKAELDT
VLTLHRGRYTTARRTAAVPQLQCIGGSAGCS-DIPEVVQCYNRGWDGYDVQWQCKADLEN
TITLYADRYTNARRSAPVPQLKCIGGNAGCHAMVPQVVQCHNRGWDGLDVQWECRVDMDN
AITLYADRYTNARRSAPVPQLKCIGGSAGCHTMVPQVVQCHNRGWDGFDVQWECKVDMDN
VLTLYRGRYTTARRSSPVPQLQCIGGSAGCGSFTPEVVQCYNRGSDGIDAQWECKADMDN
VLTLYKGKYTTARRSSAVPQLQCVGGSAGCGSFIPEVVQCKNKGWDGVDAQWECKTDMDN
VLTLYRGLYTTARRSSPVPQLQCVGGSAGCHAFVPEVVQCQNKGWDGMDIQWECRTDMDN
TLTLYRGRYTTARRSSPVPQLRCVGGSAGCQAFVPEVVQCQNRGWDGVDVQWECKTDMDN
ALTLYKNRYTTARRASPVPQLQCVGGSAGCQAFVPEVVQCQNKGWDGVDVQWECRTDMDN
VLTLYKGRYTTARRSSPVLQLQCAGGTAGCGSFVPEVVQCYNRGSDGIDTQWECKADMDN
AITLHKGKMTTGRRVSPTFQLKCVGG-SAKGAFTPKVVQCANQGFDGSDVQWRCDADLPH
```

Two-Level Deep Convolutional Neural Networks



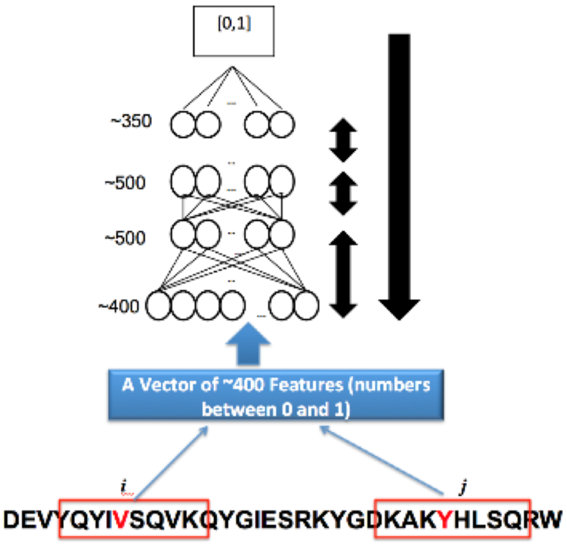
Level 1 **Level 2**

- **Training dataset:** **1426 proteins with known contact maps**
- **Validation dataset:** **196 proteins**
- **Test datasets:** **CASP10, CASP11 and CASP12 datasets**
- **Implementation:** **Keras and TensorFlow**
- **Hardware:** **Tesla K20 Nvidia GPUs**

Adhikari et al., 20

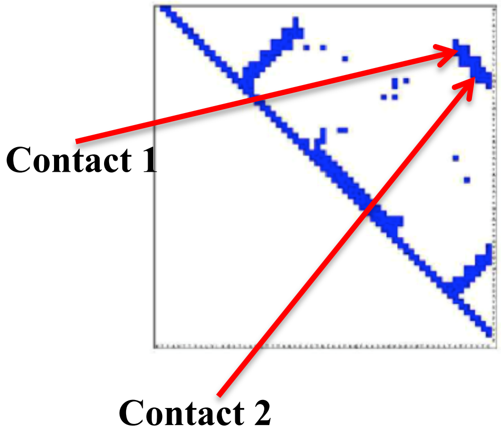
Key advantages:

- Use global information

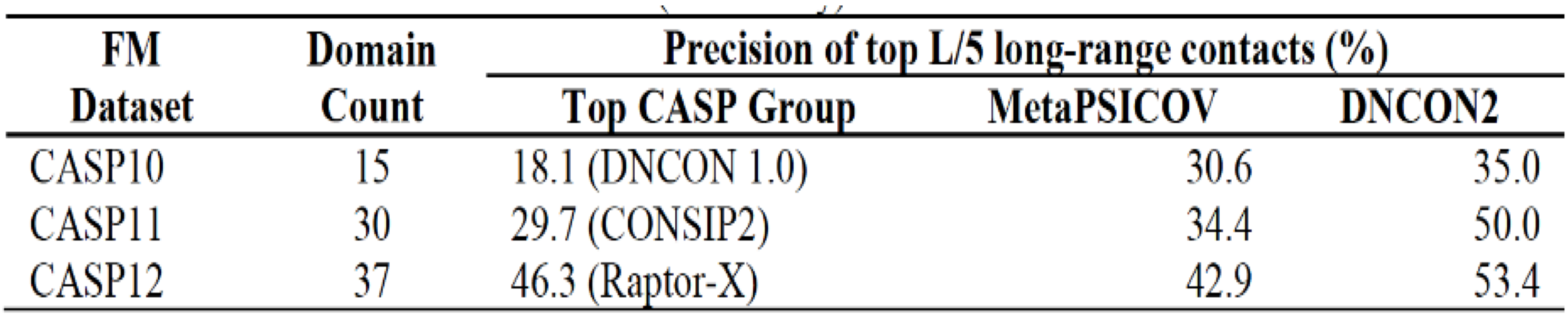

[0,1]
~350
~500
~500
~400
A Vector of ~400 Features (numbers between 0 and 1)
i
j
DEVYQYIVSQVKQYGIESRKYGDKAKYHLSQRW

Local Window

- Capture correlation between contacts (high-level contact patterns / clusters)



Test on CASP Datasets



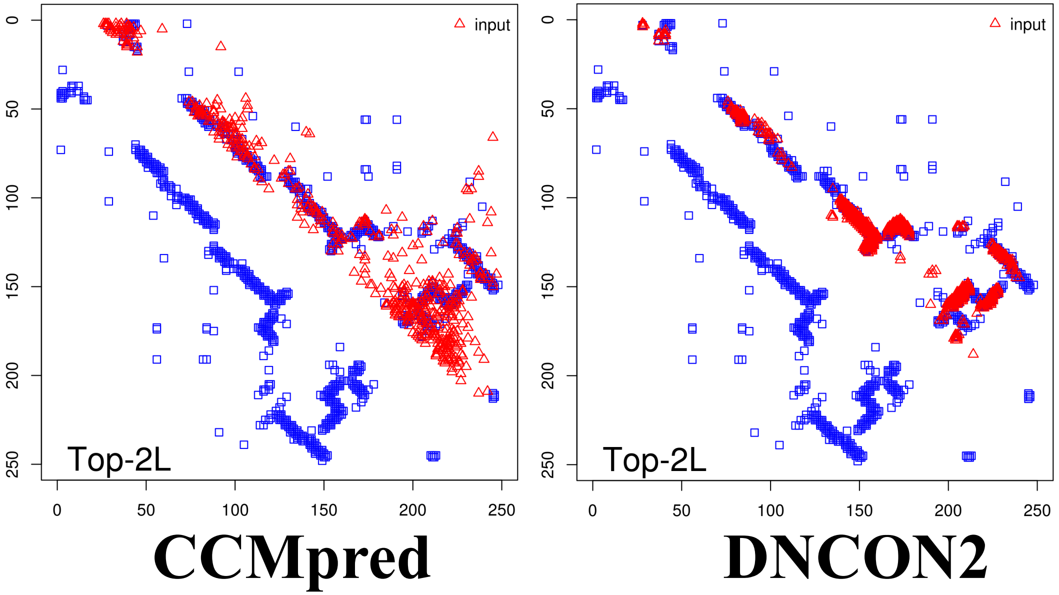
FM	Domain	Precision of top L/5 long-range contacts (%)		
Dataset	Count	Top CASP Group	MetaPSICOV	DNCON2
CASP10	15	18.1 (DNCON 1.0)	30.6	35.0
CASP11	30	29.7 (CONSIP2)	34.4	50.0
CASP12	37	46.3 (Raptor-X)	42.9	53.4

Method	Accuracy of top L/5 contacts on 115 CASP13 domains
DNCON2 (deep learning)	75%
CCMpred (co-evolution)	45%

What are deep learning methods doing that other methods do not?

- **Use more long-range information (CNN, RNN, LSTM, ResNet, …)**
- **One deep model for proteins of variable length**
- **Capture correlations between contacts (clusters), signal reinforcement, chain propagation**
- **Recall missing contacts and remove noise**
- **More powerful in recognizing weak patterns (*deep learning versus shallow learning*)**

Co-Evolution VS Deep Learning: T0953S2 (blue: true; red: predicted)



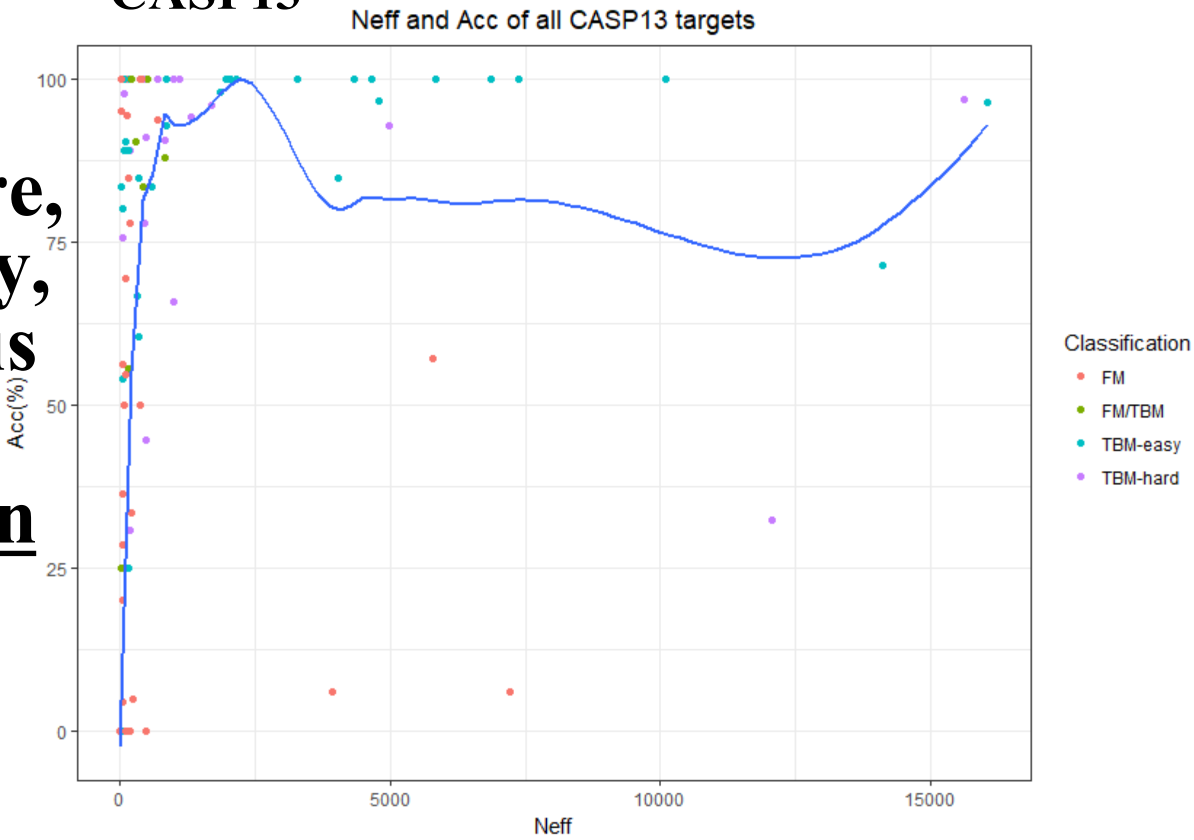
CCMpred **DNCON2**

	Top 5	Top L/5	Top L
CCMpred	60	59	33
DNCON2	100	75	61

When did the deep learning methods perform well or poorly?

- **Key factor: num. of effective sequences (high versus low)**
- **Other features: secondary structure, solvent accessibility, etc (accurate versus inaccurate)**
- **Topology of protein structure (alpha, beta, alpha/beta, alpha+beta, and non-globular)**

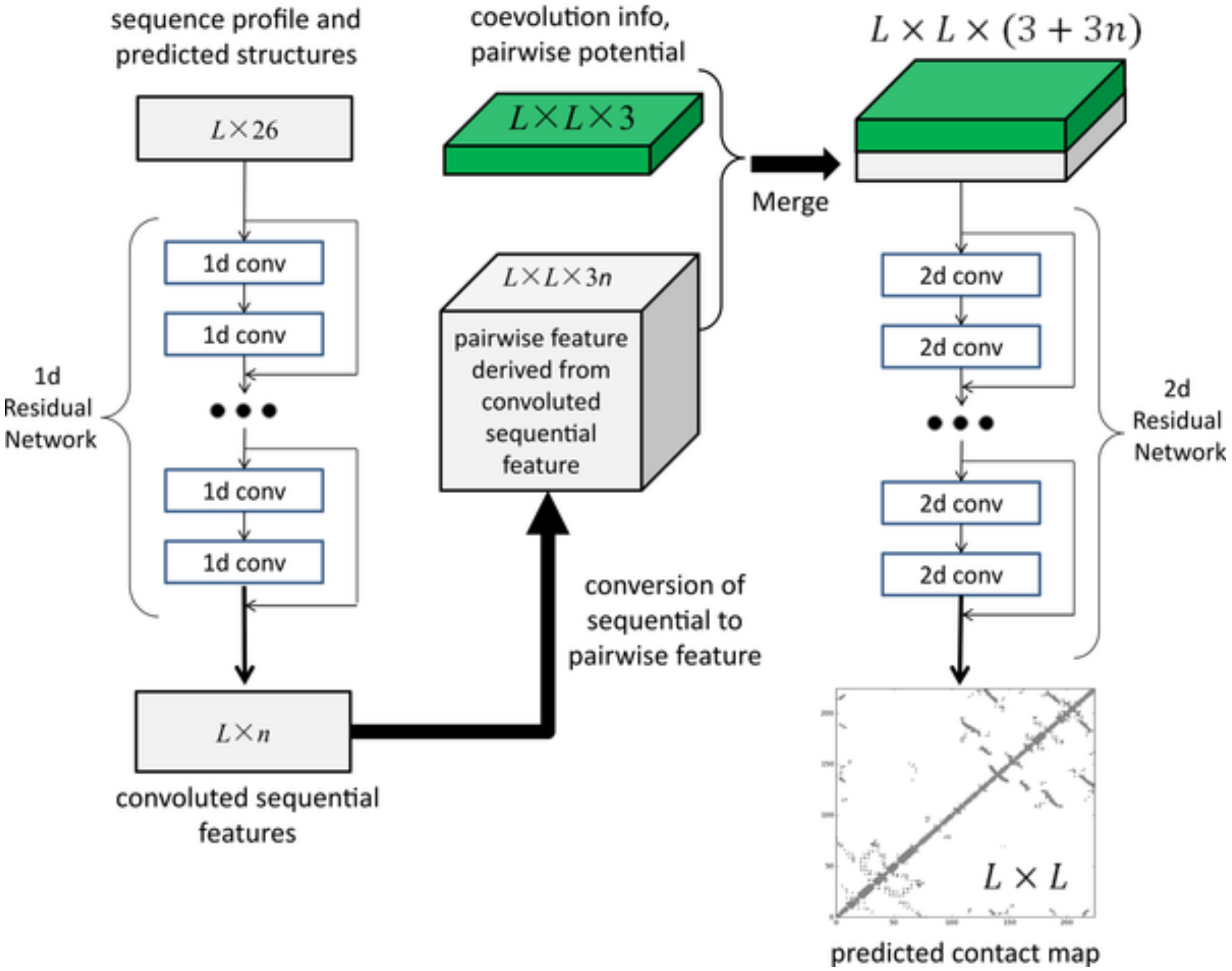
Accuracy of top L/5 predictions VS num. of effective sequences (Neff) in CASP13



Jianlin Cheng - University of Missouri - Columbia

Fig 1. Illustration of our deep learning model for contact prediction where L is the sequence length of one protein under prediction.

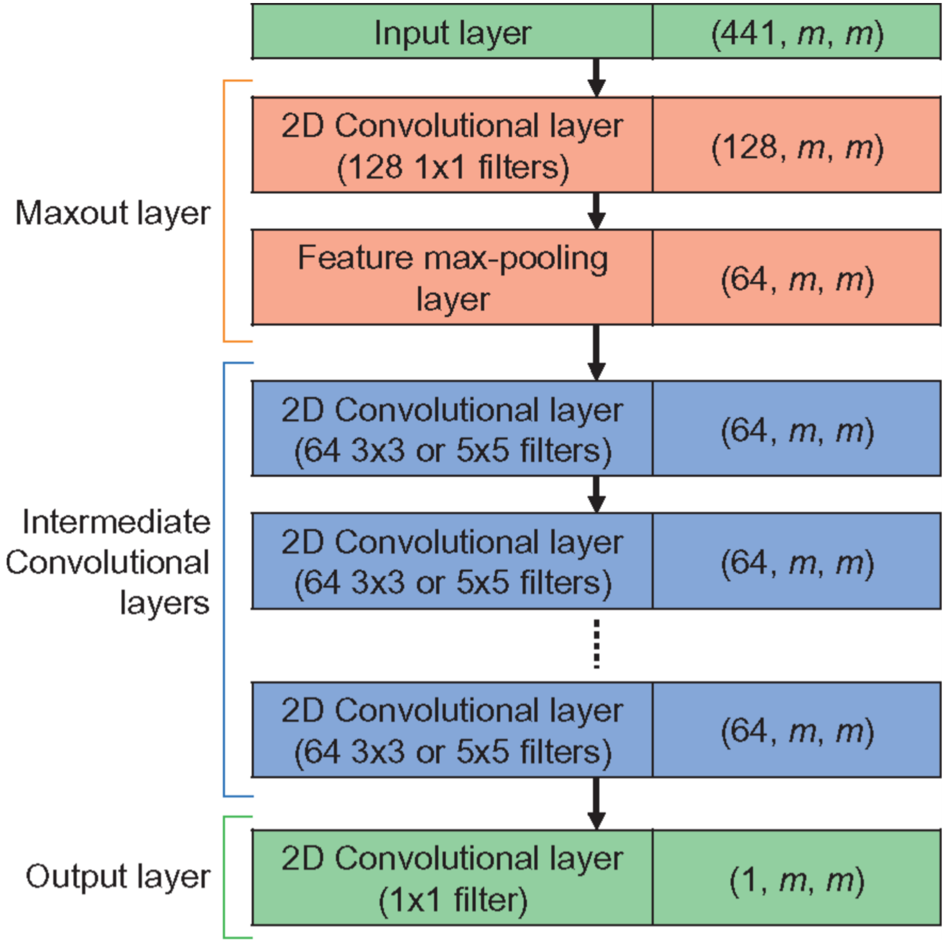
RaptorX



Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLOS Computational Biology 13(1): e1005324. https://doi.org/10.1371/journal.pcbi.1005324
https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005324

PLOS | COMPUTATIONAL BIOLOGY

The architecture of the neural network models used for DeepCov.


Input layer
(441, *m*, *m*)
Maxout layer
2D Convolutional layer (128 1x1 filters)
(128, *m*, *m*)
Feature max-pooling layer
(64, *m*, *m*)
Intermediate Convolutional layers
2D Convolutional layer (64 3x3 or 5x5 filters)
(64, *m*, *m*)
2D Convolutional layer (64 3x3 or 5x5 filters)
(64, *m*, *m*)
2D Convolutional layer (64 3x3 or 5x5 filters)
(64, *m*, *m*)
Output layer
2D Convolutional layer (1x1 filter)
(1, *m*, *m*)

Bioinformatics, Volume 34, Issue 19, 26 April 2018, Pages 3308–3315, https://doi.org/10.1093/bioinformatics/bty341

The content of this slide may be subject to copyright: please see the slide notes for details.

OXFORD
UNIVERSITY PRESS

DeepCov at GitHub: https://github.com/psipred/DeepCov

DMPFold

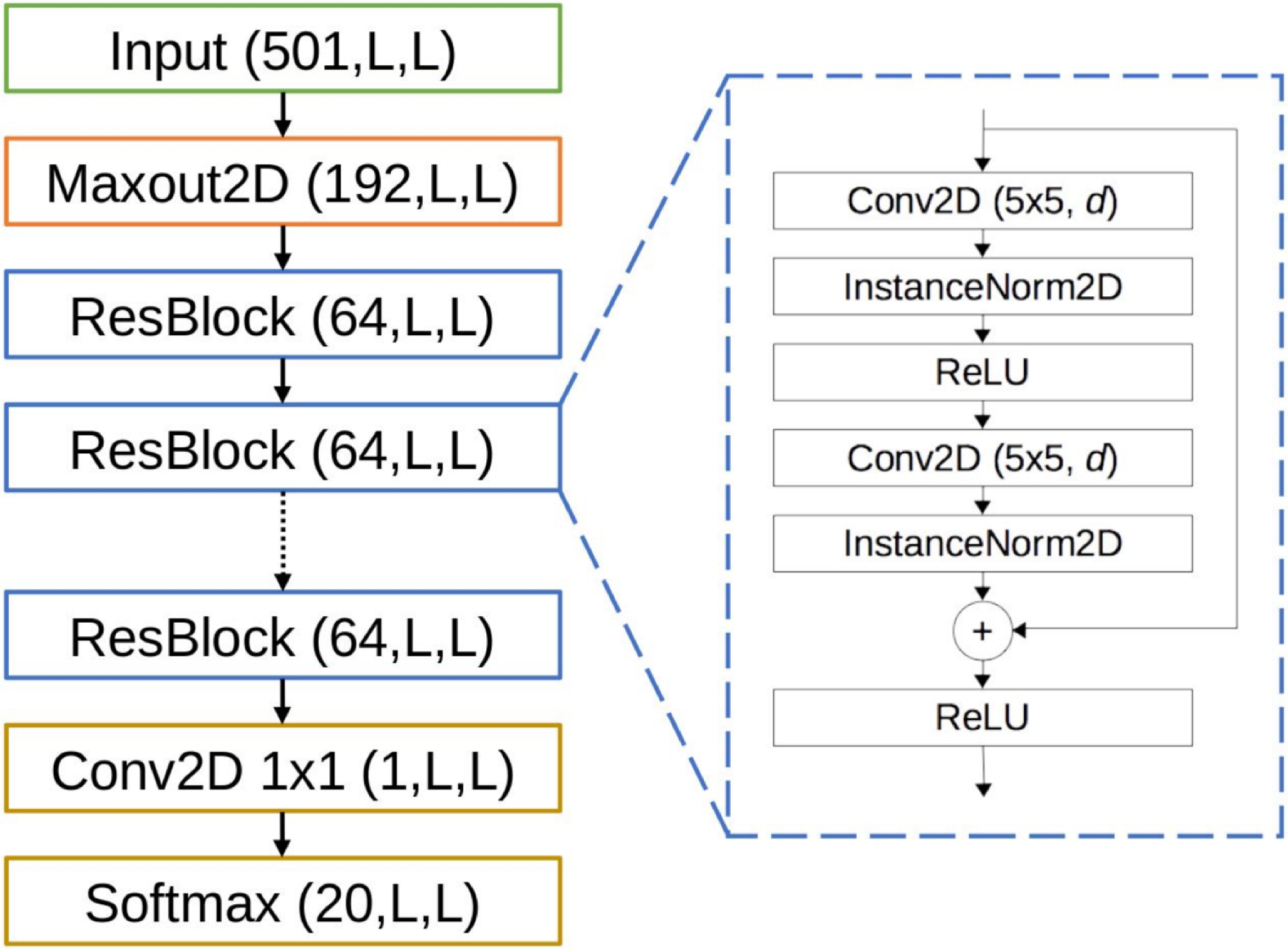
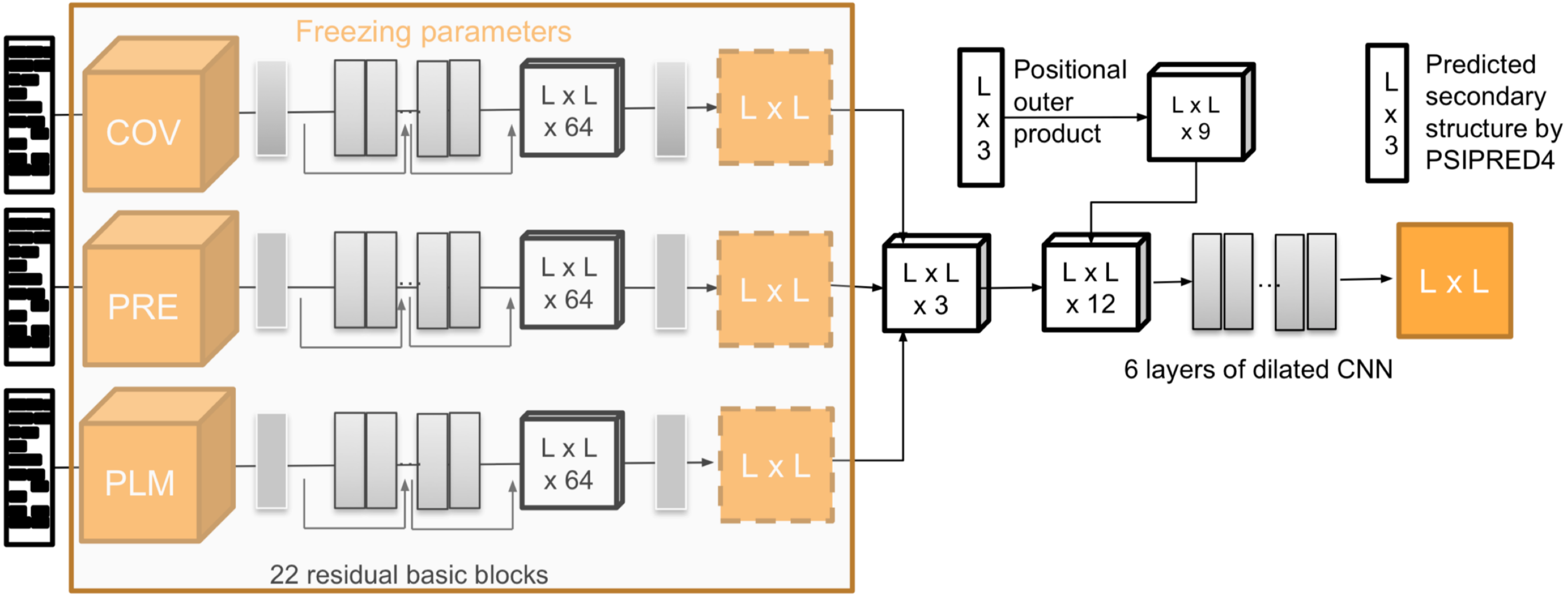


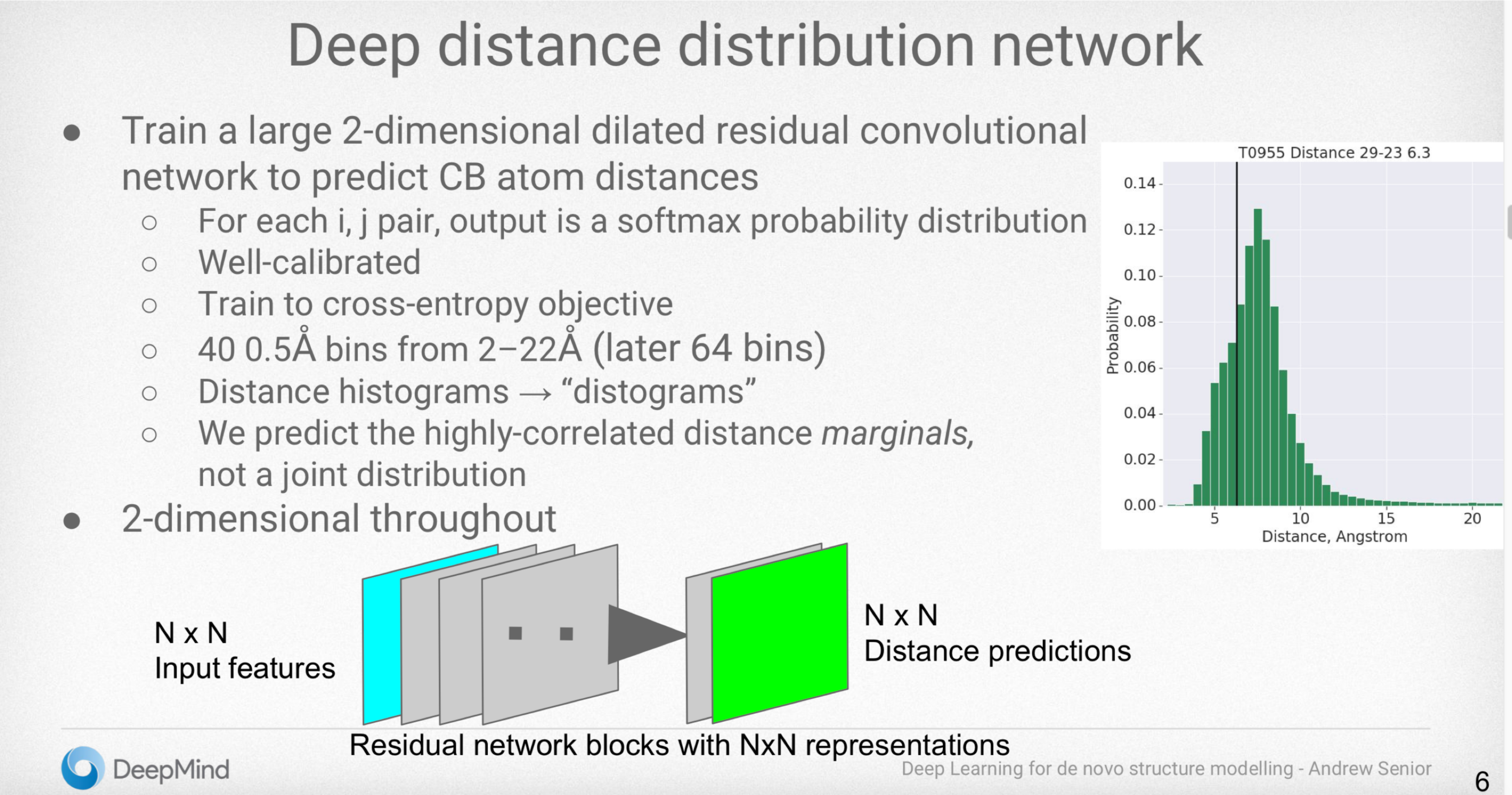
Figure 2 DMPfold model architecture. DMPfold is a deep, fully convolutional residual network. There are a total of 18 residual blocks.

ResTriplet



http://predictioncenter.org/casp13/doc/presentations/Pred_CASP13_contacts_ResTriplet_TripletRes_Redacted.pdf

AlphaFold of Google DeepMind



Deep distance distribution network

- Train a large 2-dimensional dilated residual convolutional network to predict CB atom distances
  - For each i, j pair, output is a softmax probability distribution
  - Well-calibrated
  - Train to cross-entropy objective
  - 40 0.5Å bins from 2–22Å (later 64 bins)
  - Distance histograms → “distograms”
  - We predict the highly-correlated distance *marginals,* not a joint distribution
- 2-dimensional throughout

Residual network blocks with NxN representations

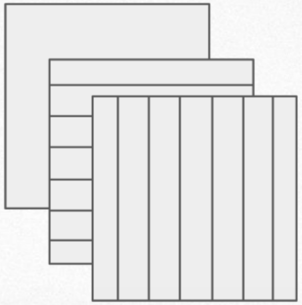
DeepMind

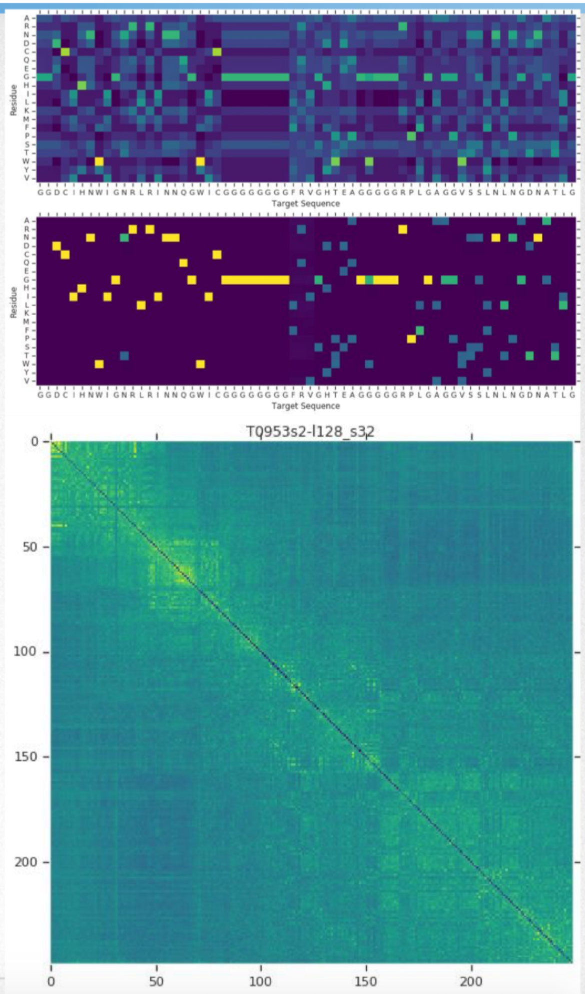
http://predictioncenter.org/casp13/doc/presentations/Pred_CASP13-DeepLearning-AlphaFold-Senior.pdf

Data

- PDB 2018-03-15 / Uniclust30 2017-10
- Train on 29,400 CATH (2018-03-16) s_35 cluster representatives
- MSA features e.g.
  - HHBlits and PSIBLAST profiles
  - 2D features from Potts model fit in TensorFlow
    - Frobenius norm L x L x 1
    - **Raw parameters** L x L x 22 x 22
  - No Mutual Information

Repeat 1D features, tiling in x and y then concatenate with 2D features

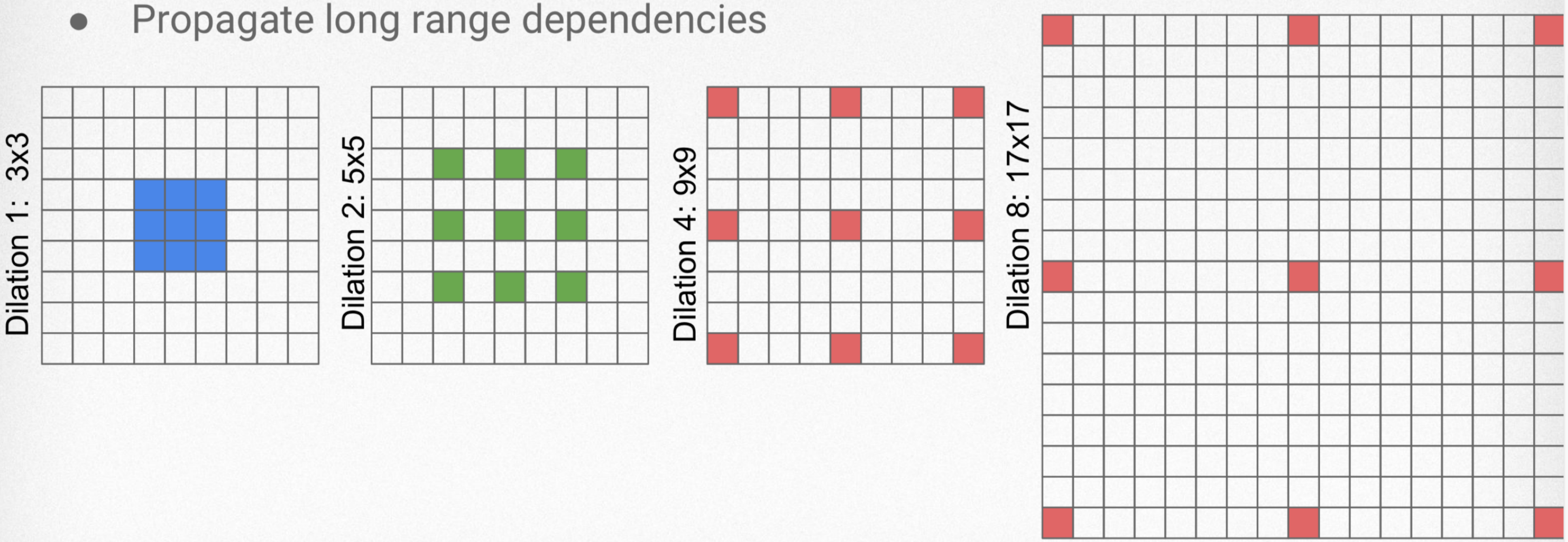




A R N D C Q E G H I L K M F P S T W Y V
Residue
G G D C I H N W I G N R L R I N N Q G W I C G G G G G G G G F R V G H T E A G G G G G R P L G A G G V S S L N L N G D N A T L G
Target Sequence
A R N D C Q E G H I L K M F P S T W Y V
Residue
G G D C I H N W I G N R L R I N N Q G W I C G G G G G G G G F R V G H T E A G G G G G R P L G A G G V S S L N L N G D N A T L G
Target Sequence
T0953s2-l128_s32
0
50
100
150
200
0
50
100
150
200

Dilated convolutions

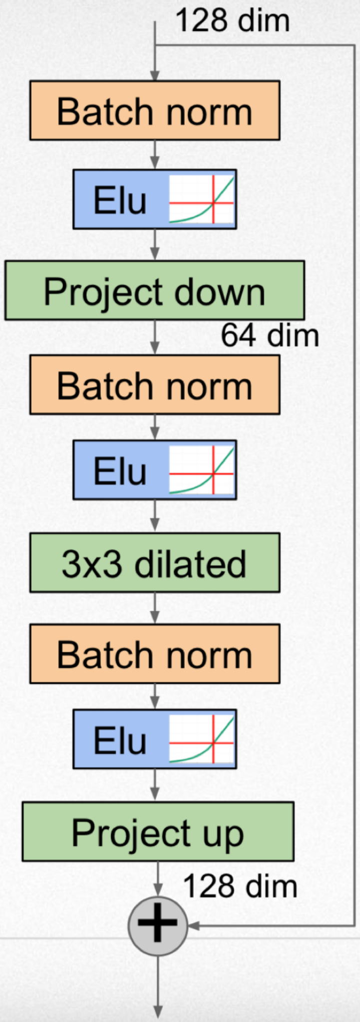
- Dilated convolutions skip pixels
  - Allow wide receptive fields with few parameters and low computation
- Propagate long range dependencies



Residual network

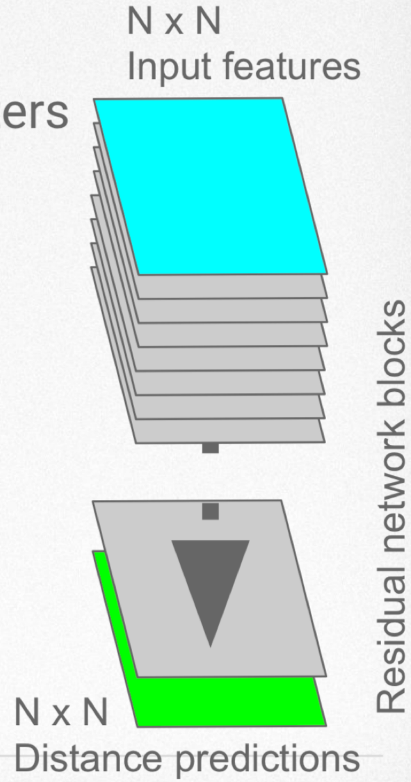
1 residual block

Modifies a 64x64x128 representation from the previous block



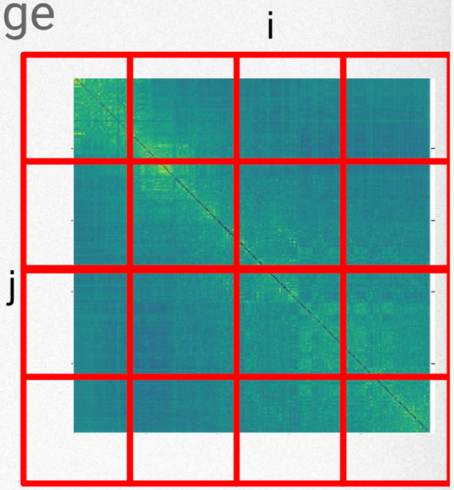
Repeat **220** times, cycling through dilations 1, 2, 4, 8

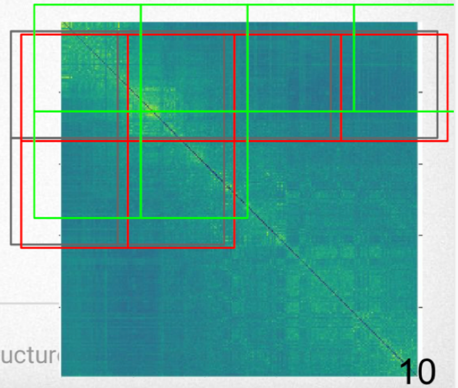
21 million parameters



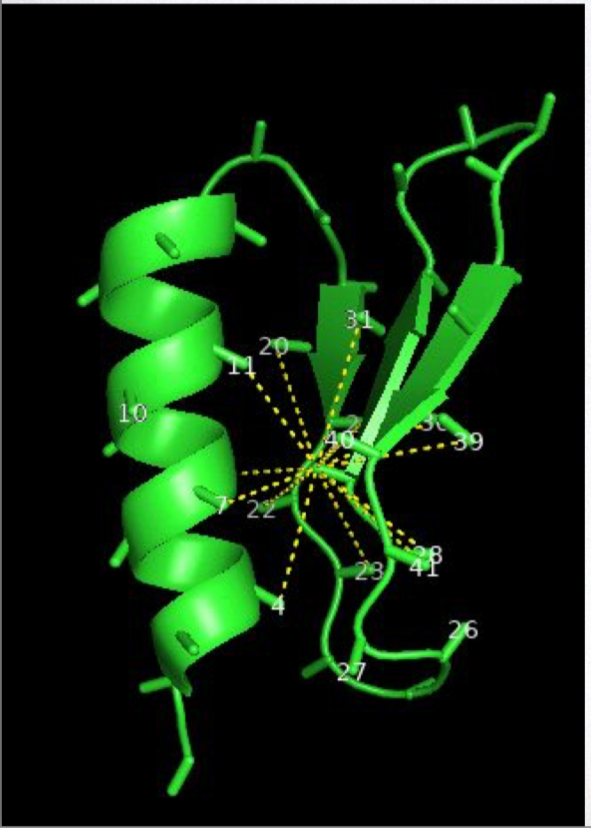
Cropping

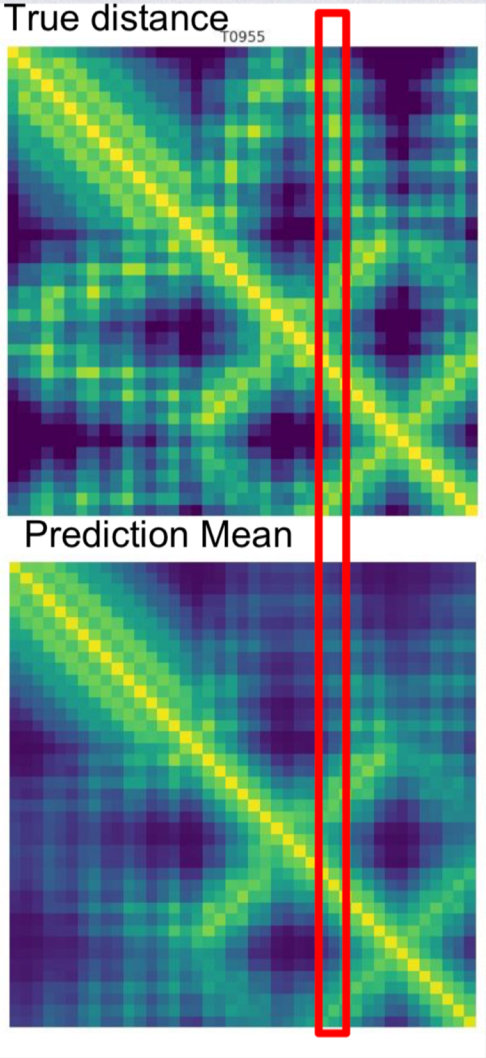
- Handling arbitrary protein length L leads to O(L^2) memory usage
  - Consistent size helps distributed training
- Train on all 64x64 crops from proteins
  - Random offset
  - Including up to 32 residues off-edge
- For a crop (i, i+63)x(j, j+63)
  - Crop corresponding 2D input features
  - Tile corresponding (i, i+63) and (j, j+63) 1D parameters
  - Still allows modelling long range correlations from i to j
- Helps avoid overfitting
  - Data augmentation
  - Each protein leads to many different training examples
- Ensembling:
  - At test time weighted average across alternative offsets
  - Also average across 4 slightly different models





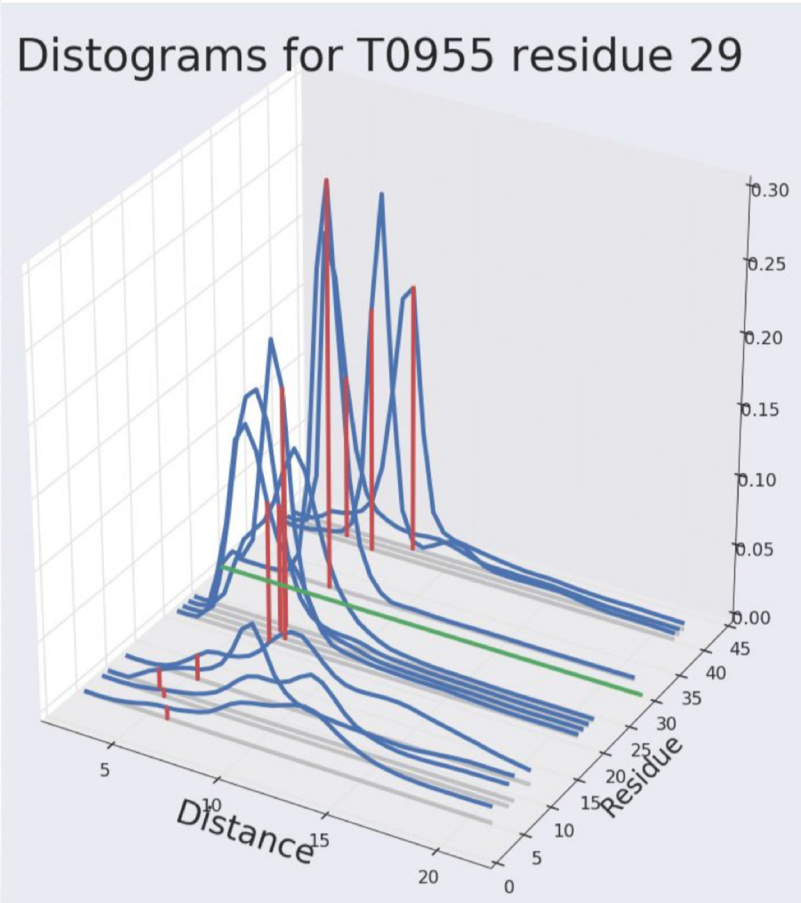
Deep distance distribution Network (D^{3}N)

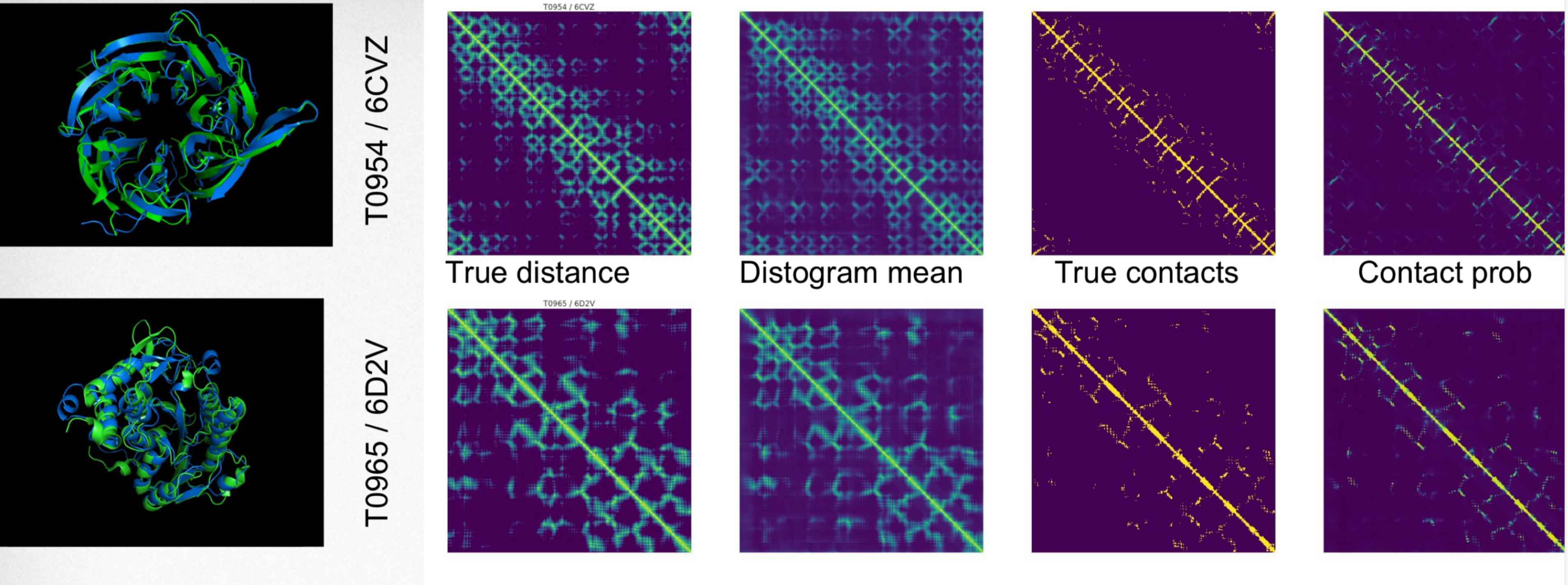



True distance

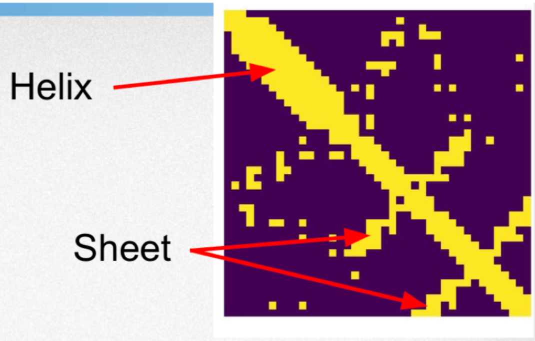
Prediction Mean

Distograms for T0955 residue 29

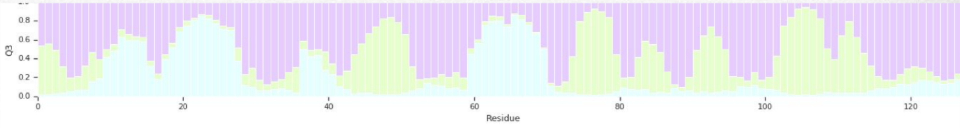


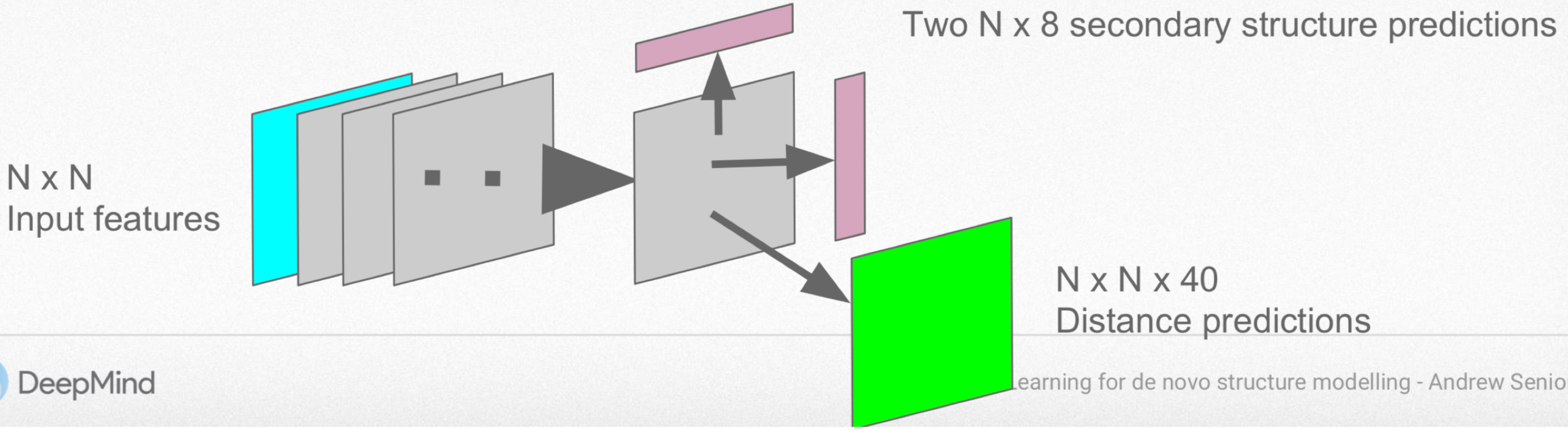

T0954 / 6CVZ
T0954 / 6CVZ
True distance
Distogram mean
True contacts
Contact prob
T0965 / 6D2V
T0965 / 6D2V

Auxiliary losses

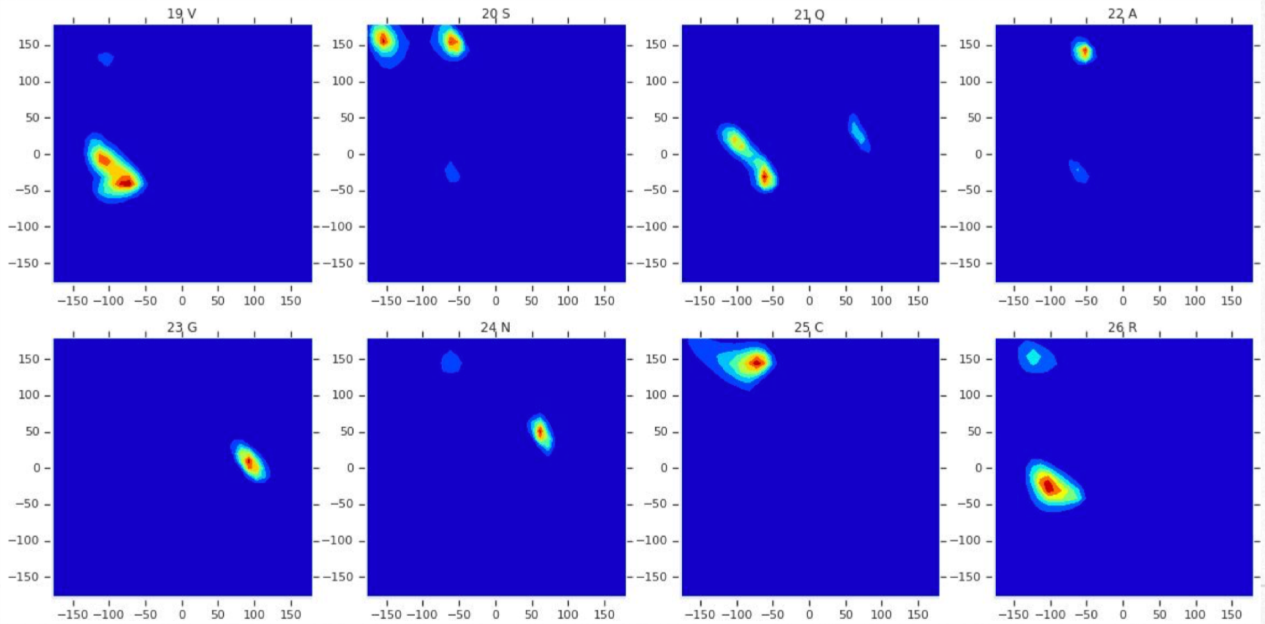


- We know the contact map encodes secondary structure
    - A distance network should be good at predicting it
- *Auxiliary loss* of secondary structure from 1D reductions for **both** (i, i+63) and (j, j+63)
    - Ensembled across all 2D crops
- Q3 Accuracy on CASP11 ~84%
- Predicting secondary structure **improves** contact prediction





- For repeated gradient descent, we need torsion predictions
  - From 1D reduction also predict a joint (phi, psi) Ramachandran probability distribution for each residue (10 degree bins)
  - Again marginal distributions

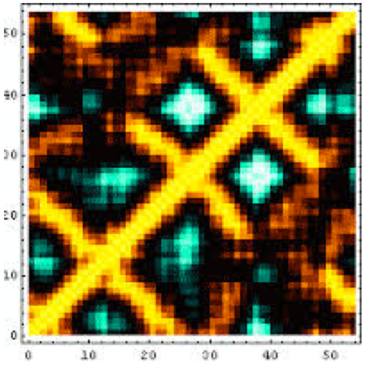


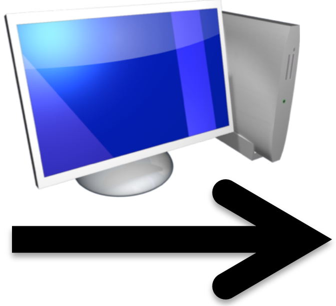
T0954

Reconstruct 3D protein structures from contacts / distances

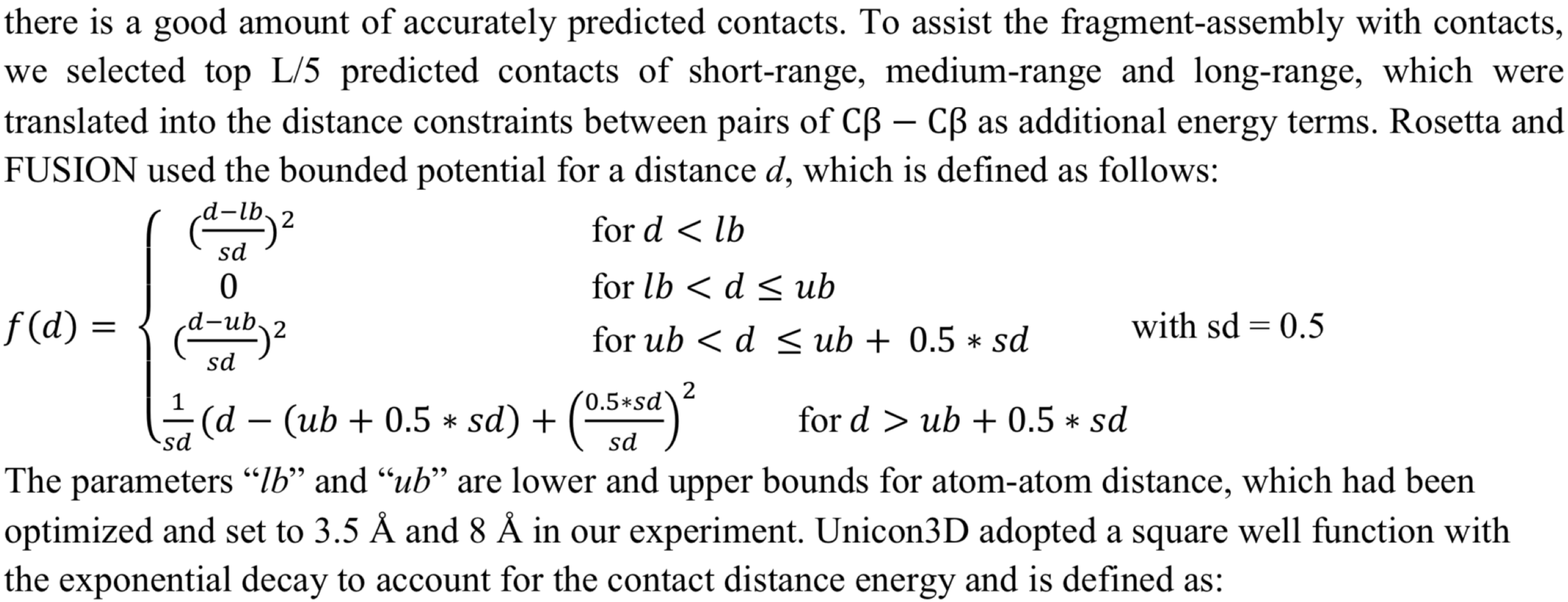
- Fragment Assembly + Contact Distances (Rosetta, FUSION, UniCon3D)
- CONFOLD
- DMPfold
- AlphaFold

Contact-Based Structure Prediction





Fragment Assembly + Contact Distances



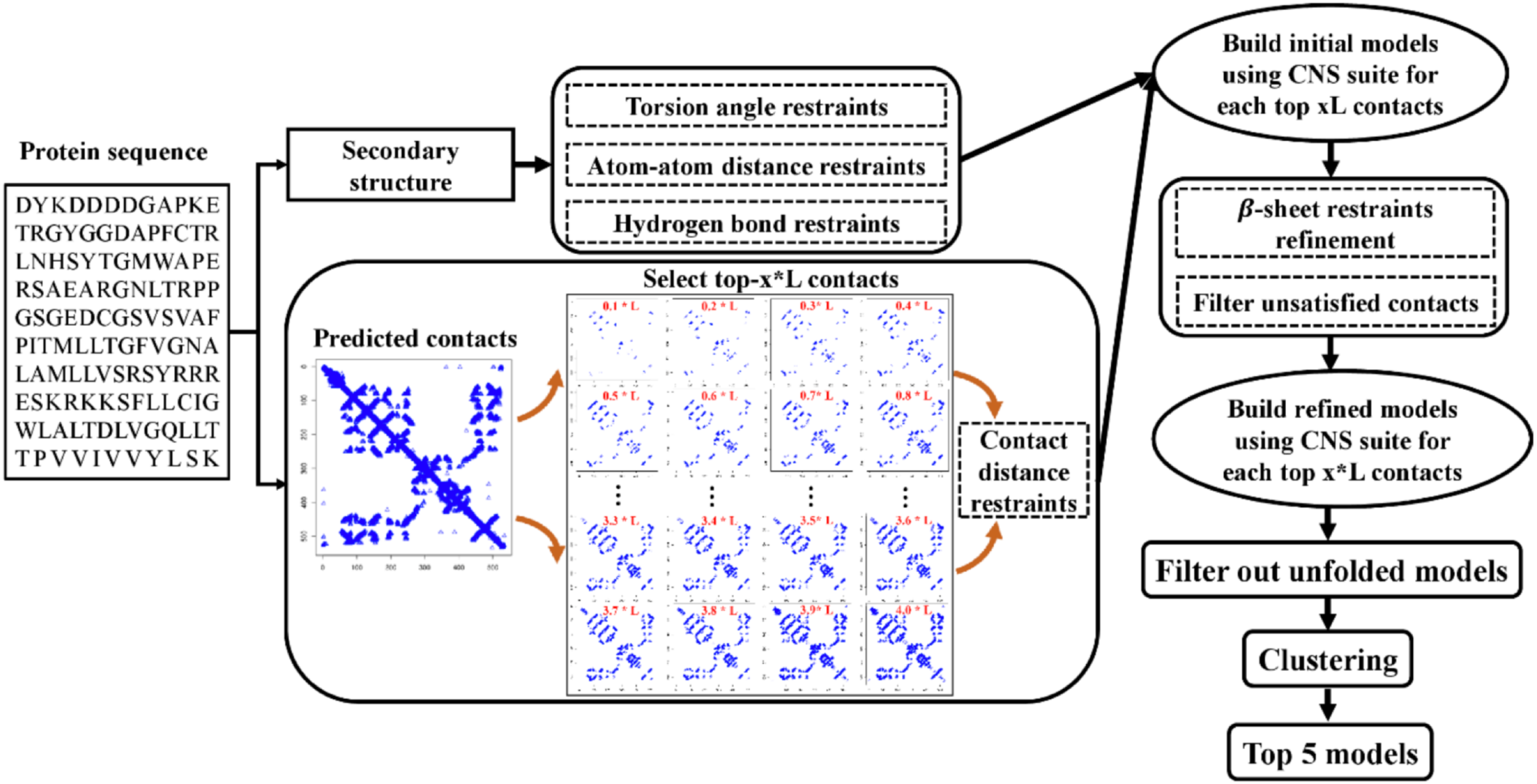
there is a good amount of accurately predicted contacts. To assist the fragment-assembly with contacts, we selected top L/5 predicted contacts of short-range, medium-range and long-range, which were translated into the distance constraints between pairs of $C\beta - C\beta$ as additional energy terms. Rosetta and FUSION used the bounded potential for a distance *d*, which is defined as follows:

$$f(d) = \begin{cases} (\frac{d-lb}{sd})^2 & \text{for } d < lb \\ 0 & \text{for } lb < d \leq ub \\ (\frac{d-ub}{sd})^2 & \text{for } ub < d \ \leq ub + \ 0.5 * sd \\ \frac{1}{sd}(d - (ub + 0.5 * sd) + \left(\frac{0.5*sd}{sd}\right)^2 & \text{for } d > ub + 0.5 * sd \end{cases} \quad \text{with sd = 0.5}$$

The parameters "*lb*" and "*ub*" are lower and upper bounds for atom-atom distance, which had been optimized and set to 3.5 Å and 8 Å in our experiment. Unicon3D adopted a square well function with the exponential decay to account for the contact distance energy and is defined as:

Advantage: using fragment information
Disadvantage: contact distance plays an indirect role; sampling fails for large/complicated protein structures

CONFOLD



Protein sequence

DYKDDDDGAPKE
TRGYGGDAPFCTR
LNHSYTGMWAPE
RSAEARGNLTRPP
GSGEDCGSVSVAF
PITMLLTGFVGNA
LAMLLVSRSYRRR
ESKRKKSFLLCIG
WLALTDLVGQLLT
TPVVIVVYLSK

Secondary structure

Torsion angle restraints

Atom-atom distance restraints

Hydrogen bond restraints

Predicted contacts

Select top-x*L contacts

Contact distance restraints

Build initial models using CNS suite for each top xL contacts

β-sheet restraints refinement

Filter unsatisfied contacts

Build refined models using CNS suite for each top x*L contacts

Filter out unfolded models

Clustering

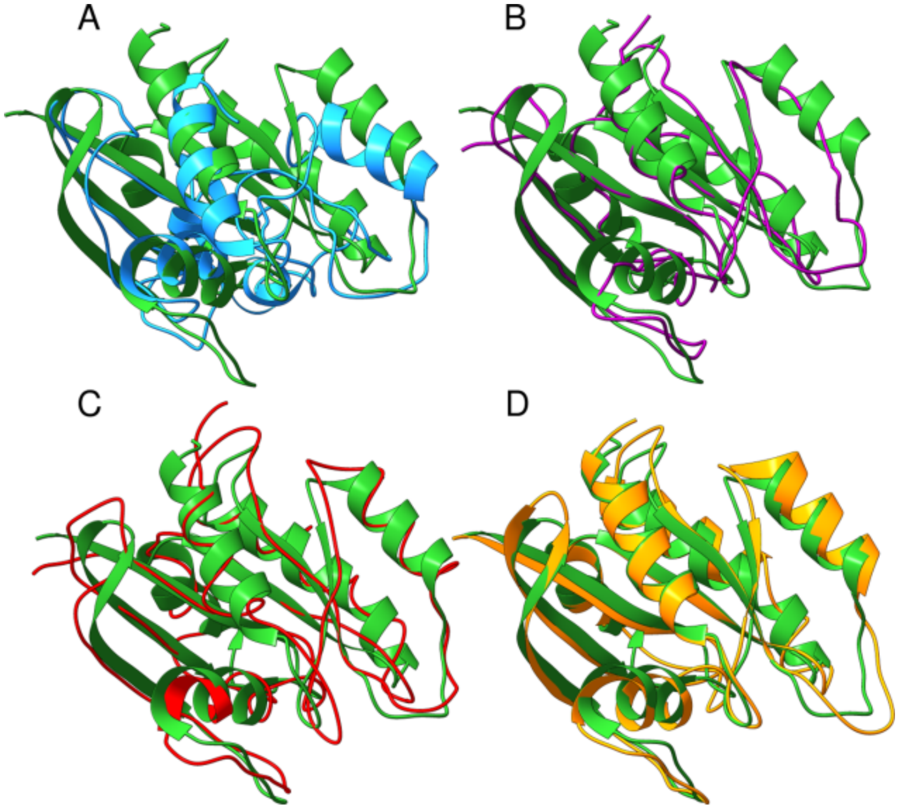
Top 5 models

Figure 3. Automated contact distance-based *ab initio* protein structure prediction by CONFOLD2.

Advantage: directly translating distances into structures; contact distances play dominant role

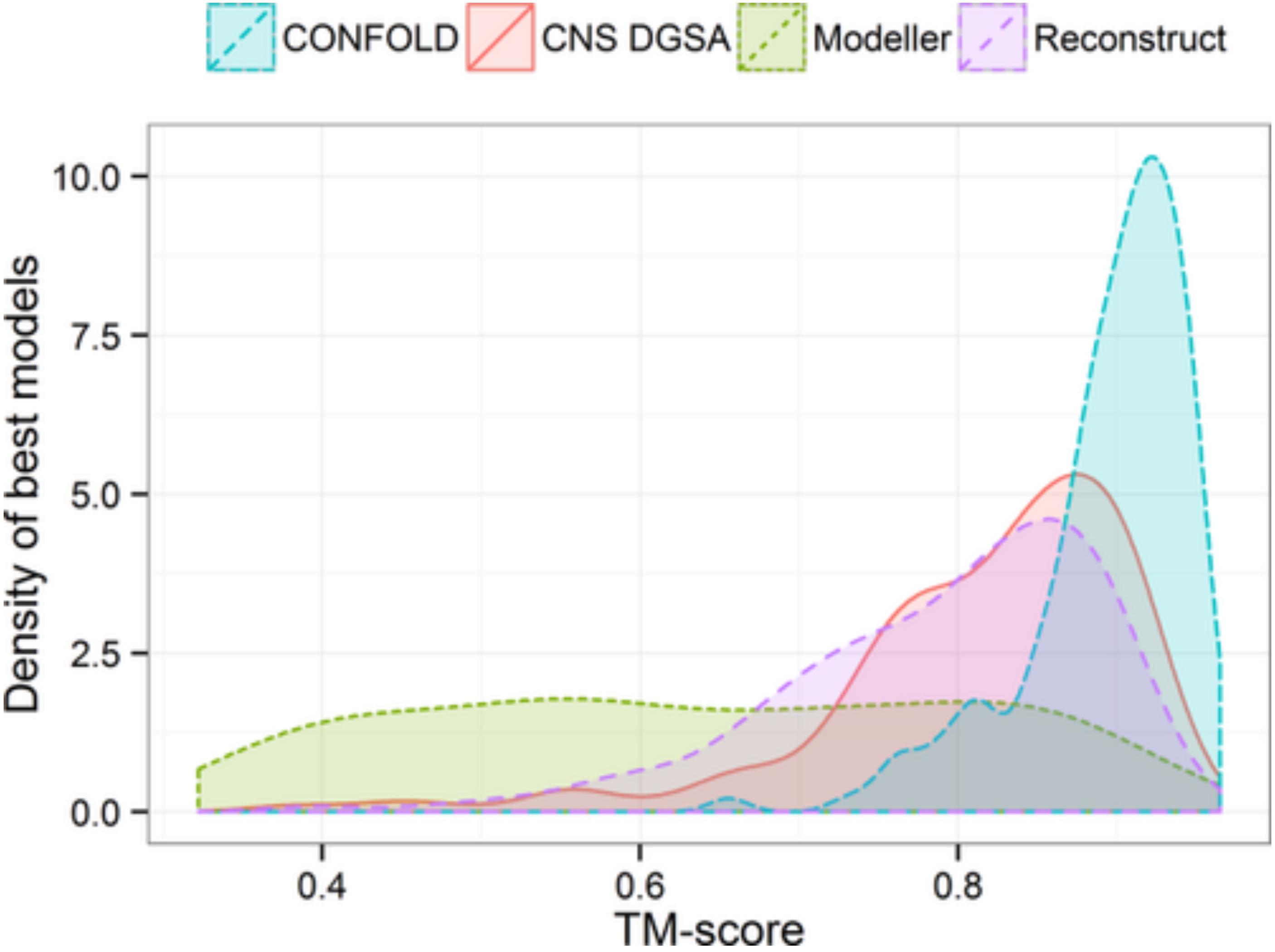
Disadvantage: fail if there is no sufficient amount of accurate distances

CONFOLD: Residue-residue contact-guided ab initio protein folding



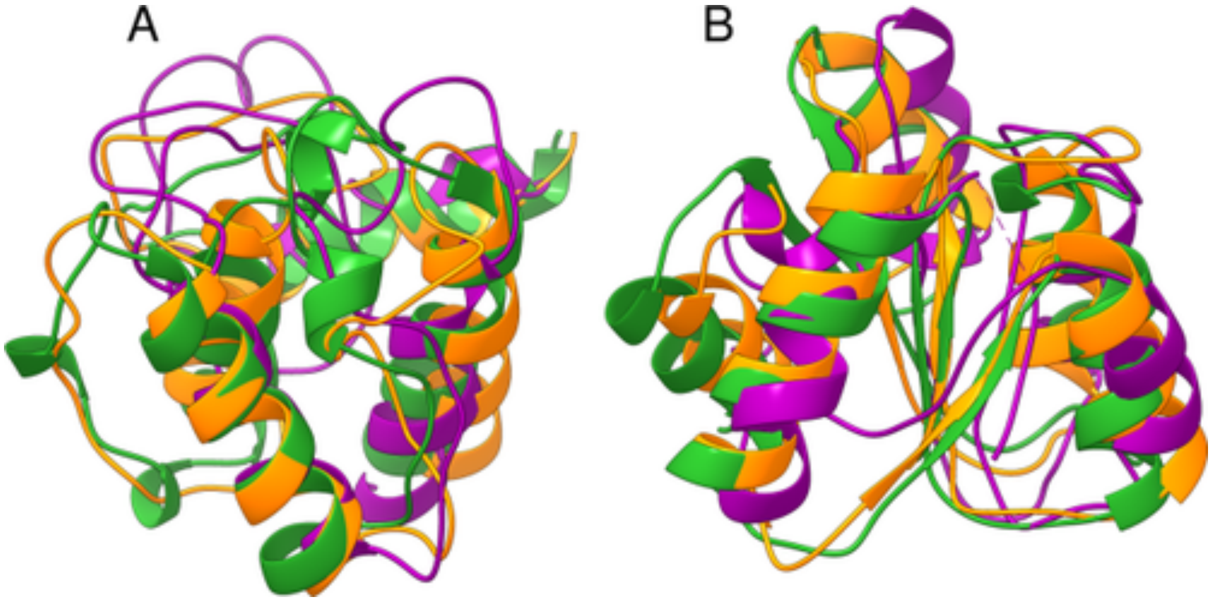
Best models reconstructed for the protein 5p21 using Modeler (**A**), reconstruct (**B**), customized CNS DGSA protocol (**C**), and CONFOLD (**D**). All models are superimposed with native structure (green). The TM-scores of Models A, B, C, and D are 0.53, 0.86, 0.88, and 0.94, respectively. Model D reconstructed by CONFOLD has higher TM-score and also much better secondary structure quality than the other models.

Proteins: Structure, Function, and Bioinformatics, Volume: 83, Issue: 8, Pages: 1436-1449, First published: 13 May 2015, DOI: (10.1002/prot.24829)

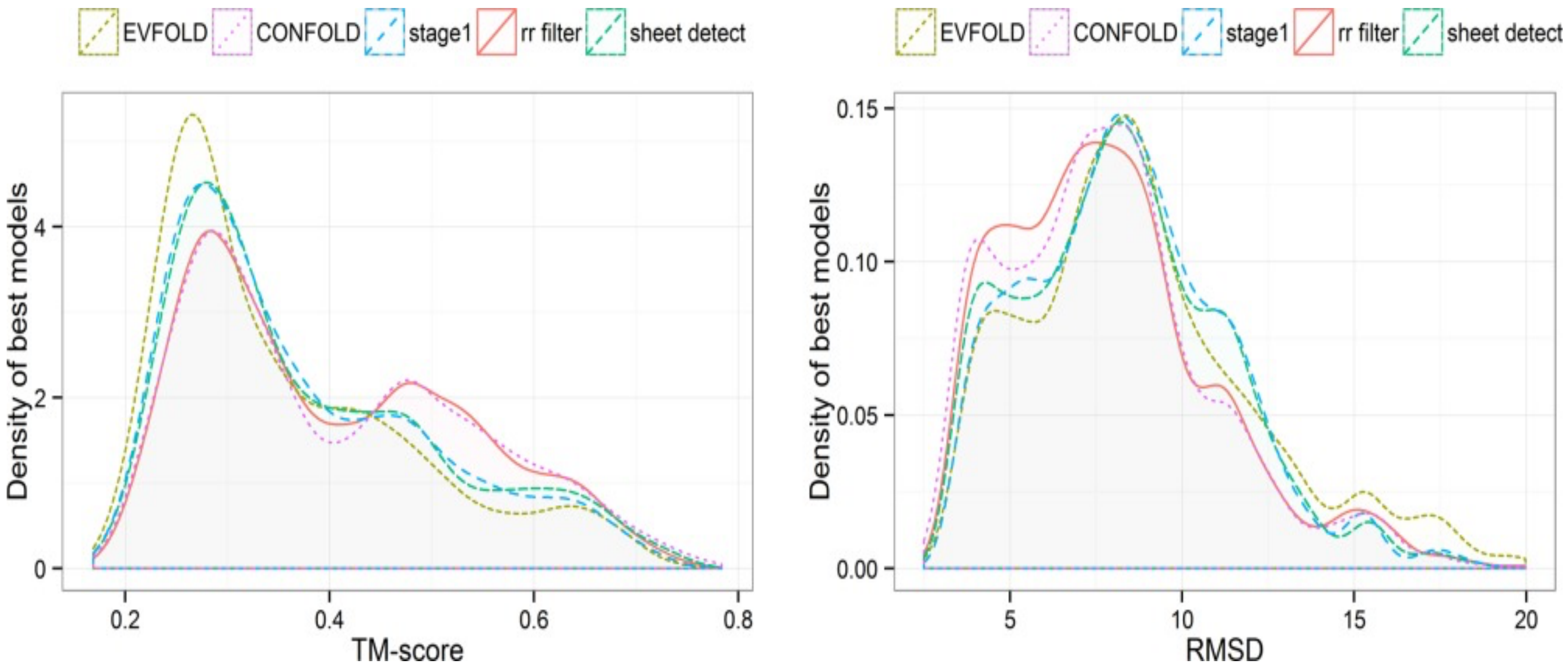


Distribution of TM-scores of the best models reconstructed by the four methods for 150 FRAGFOLD proteins.

CONFOLD VS EVFOLD

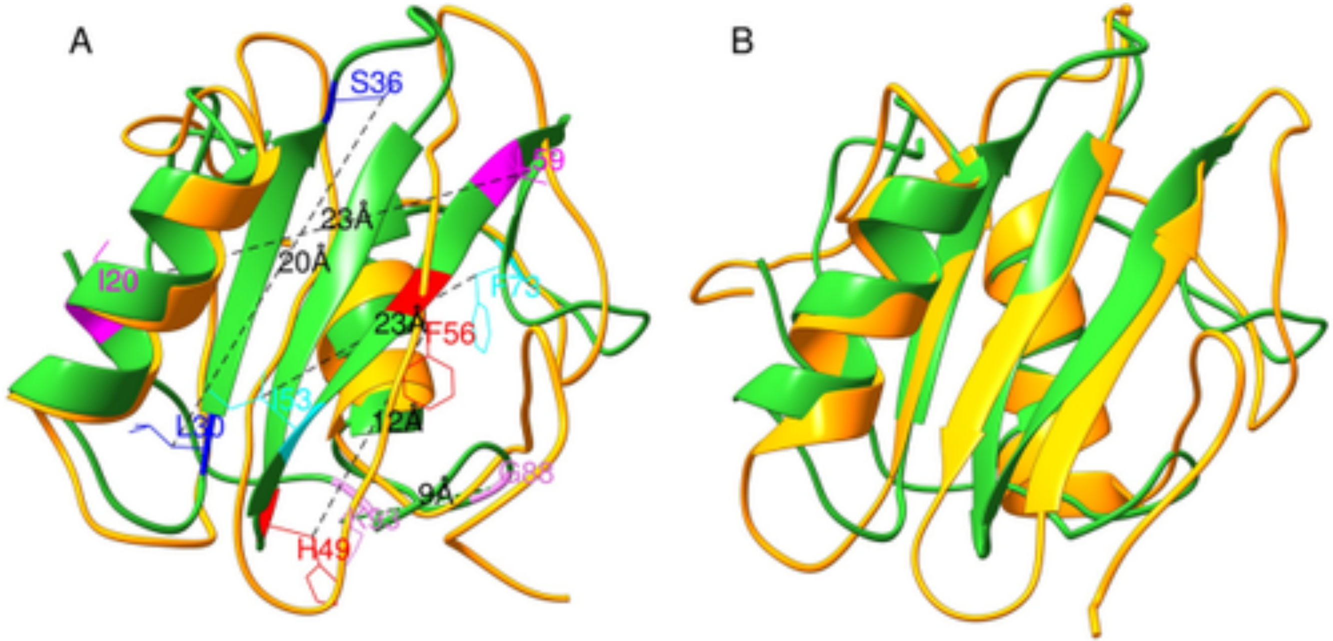


Best predicted models for the proteins RNH_ECOLI (**A**) and SPTB2_HUMAN (**B**) using EVFOLD (purple) and CONFOLD (orange) superimposed with native structures (green). The TM-scores of these models are reported in Table IV. CONFOLD models have higher TM-score and better secondary structure quality than EVAFOLD.



Distribution of model quality of the EVFOLD models and the models built by CONFOLD. Distribution of models built in first stage of CONFOLD (Stage 1), second stage with contact filtering only (rr filter), and second stage with β-sheet detection only (sheet detect) are also presented. Each curve represents the distribution of 400 times 15 models.

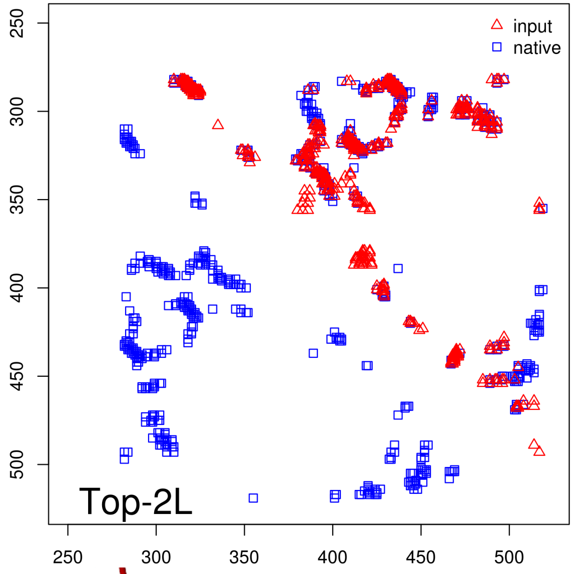
Contact Filtering



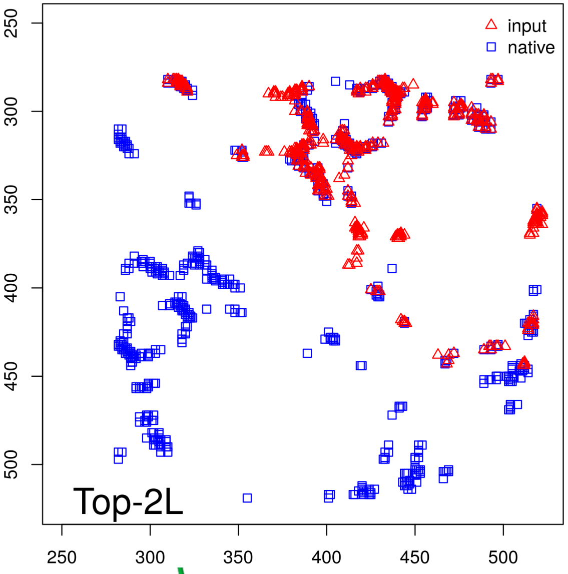
Contact filtering from Stages 1 to 2 for the protein 1NRV. (**A**) Superimposition of the best model in stage 1 reconstructed with top-0.6 L contacts by CONFOLD (orange) with the native structure (green). The model has TM-score of 0.50. Among the top-0.6 L (60) contacts, 5 out of 8 erroneous contacts that were removed in Stage 2 are visualized in the native structure along with the distance between their Cβ-Cβ atoms. The filtered, predicted contacts (20–59, 53–73, 30–36, 49–56, and 88–93) have Cβ-Cβ distances of 23, 23, 20, 12, and 9 Å, respectively, in the native structure. Each pair of residues predicted to be in contact is denoted by the same color. (**B**) Superimposition of the best model in Stage 2 reconstructed with reduced/filtered top-0.6 L contacts by CONFOLD (orange) with the native structure (green). TM-score of the model is 0.61.

Comparison on T1000 – FM Domain (residues: 282-523)

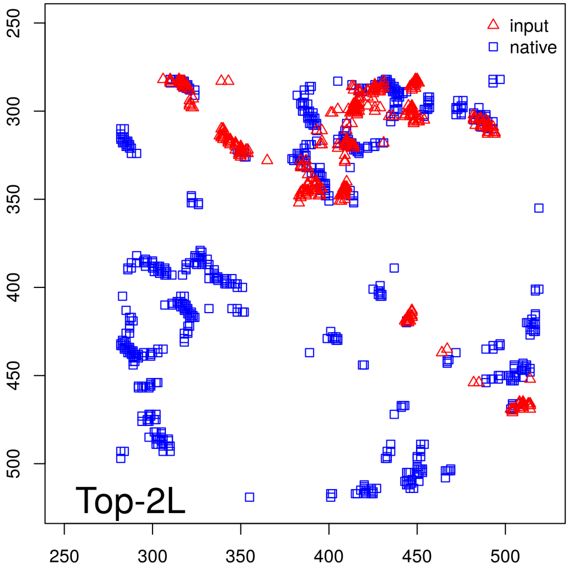
DNCON2 (red) VS Native (blue)
(L/5: 100%, L: 79%, 2L: 50%)



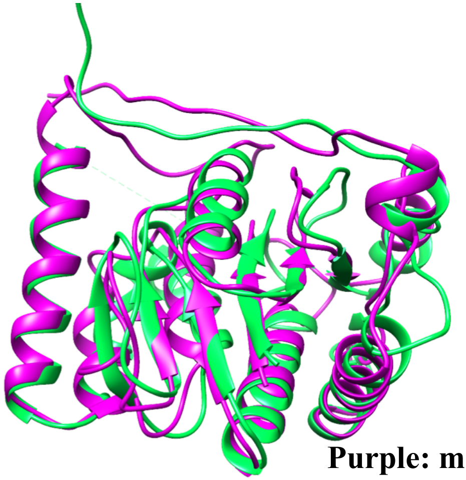
CONFOLD (red) VS Native
(L/5: 67%, L: 65%, 2L: 55%)



Rosetta-Con (red) VS Native
(L/5: 20%, L: 18%, 2L: 17%)



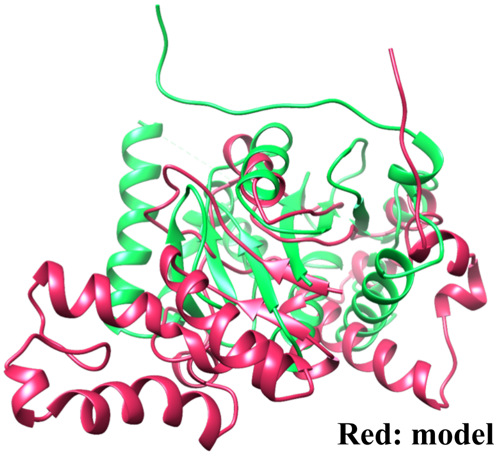
Top L/5 contacts on native structure



Purple: model
Green: native

TM-score: 0.80

GDT-TS-score: 0.64



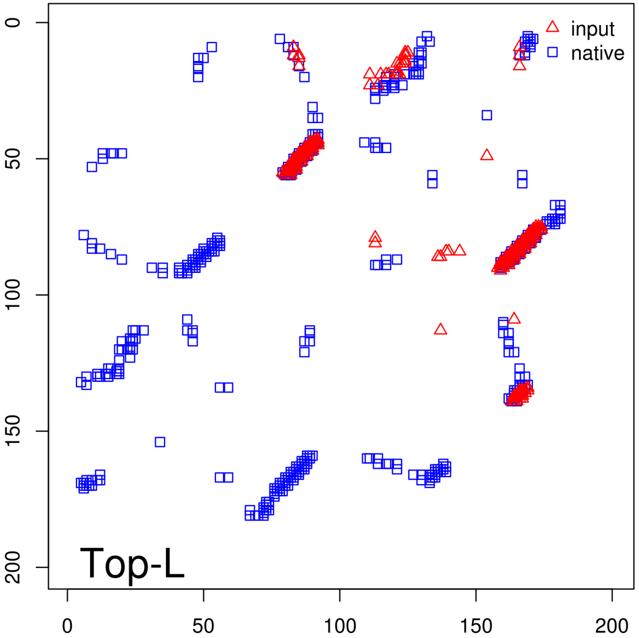
Red: model
Green: native

TM-score: 0.33

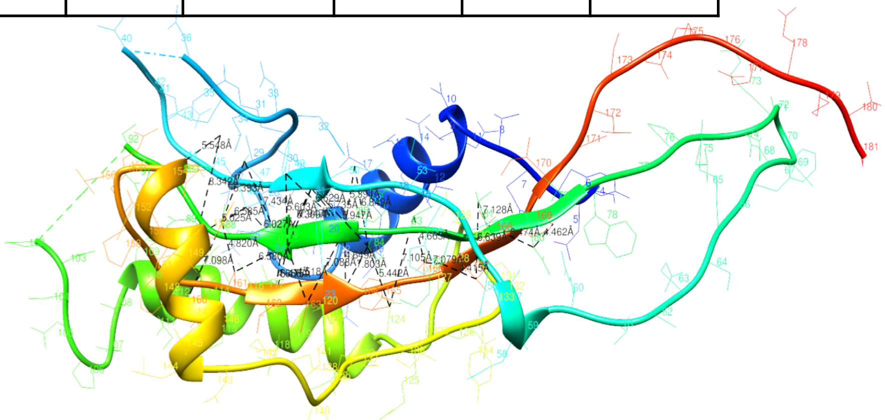
GDT-TS-score: 0.23

(1) Success of Building Models for T1021s3-D1 (FM) by CONFOLD

NCON2 (red) VS Native (blue)

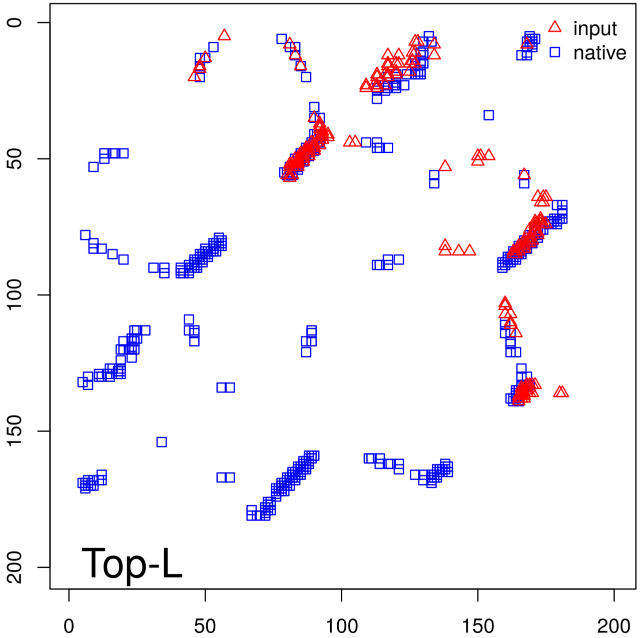


	Top 5	Top L/10	Top L/5	Top L/2	Top L
Acc.	100%	94%	97%	88%	61%

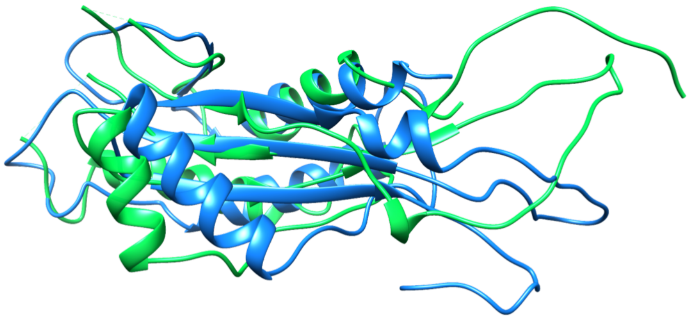


Top L/5 long-range contacts on native structure

CONFOLD (red) VS Native (blue)



	Top 5	Top L/10	Top L/5	Top L/2	Top L
Acc.	80%	47%	52%	51%	46%

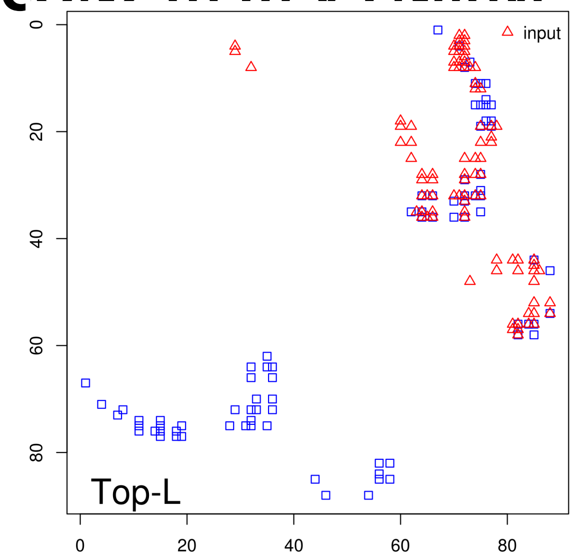


Blue: predicted; Green: native

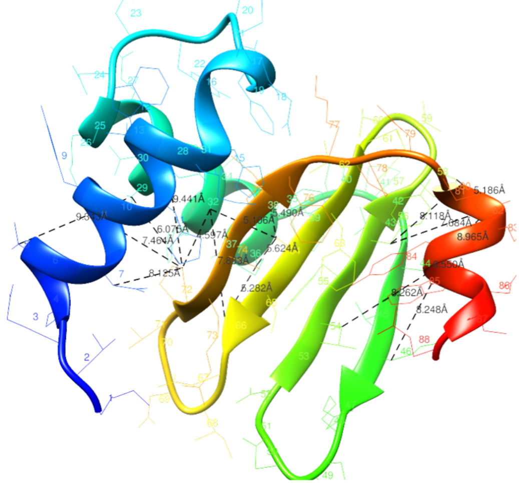
TM-score: 0.50 GDT-TS-score: 0.41

(2) Success of Building Models from Contacts with Rosetta When Failing to Identify Templates for T1019s2 (TBM)

NCON2 (red) VS Native (blue)

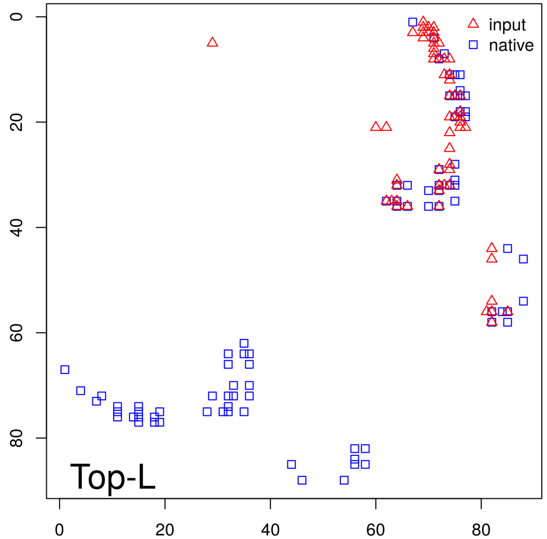


	Top L/10	Top L/5	Top L/2	Top L
Acc.	78%	61%	39%	26%

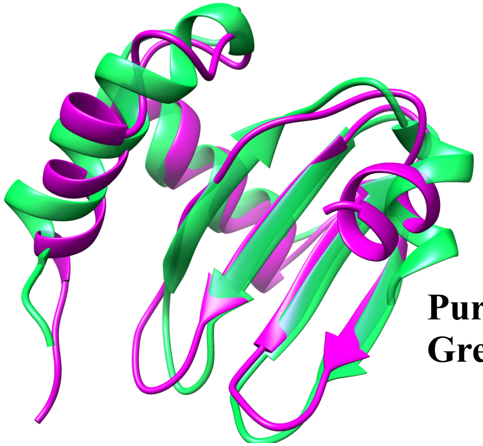


Top L/5 long-range contacts on native structure

Rosetta-Con (red) VS Native (blue)



	Top L/10	Top L/5	Top L/2	Top L
Acc.	56%	56%	39%	36%

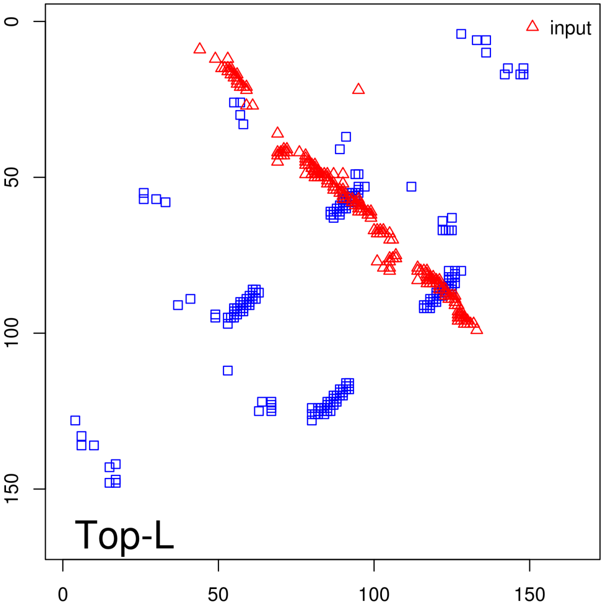


Purple: predicted
Green: native

TM-score: 0.68 GDT-TS-score: 0.67

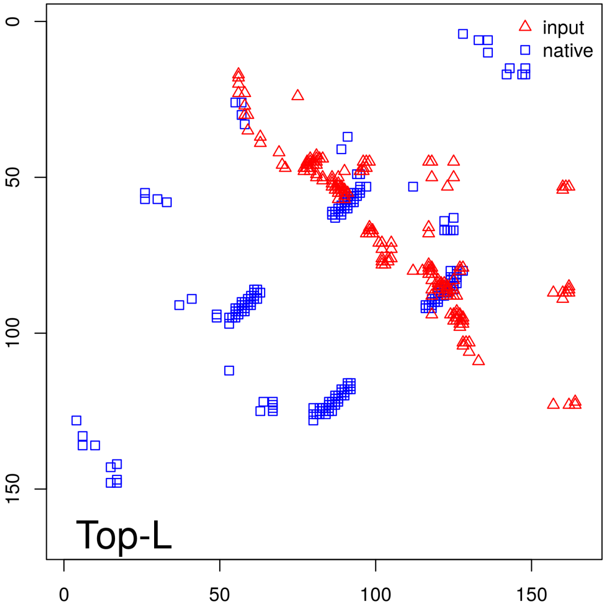
(2) Failure of predicting / using contacts (T0998 FM)

NCON2 (red) VS Native (blue)



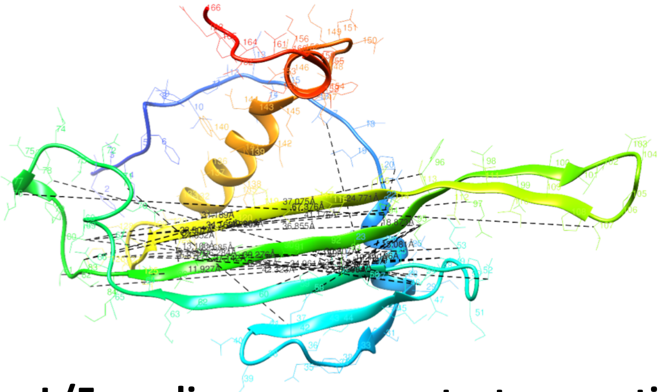
# of effective sequences = 2

Model (red) VS Native (blue)

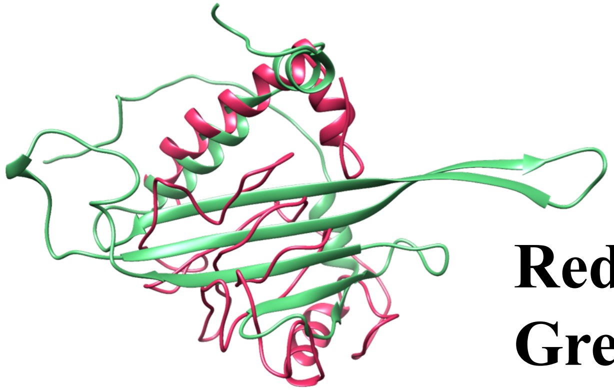


	Top L/10	Top L/5	Top L/2	Top L
Acc.	6%	6%	5%	5%

	Top L/10	Top L/5	Top L/2	Top L
Acc.	6%	6%	6%	4%



Top L/5 medium-range contacts on native structure



Red: model

Green: nativ

TM-score: 0.21 GDT-TS-score: 0.15

DMPfold

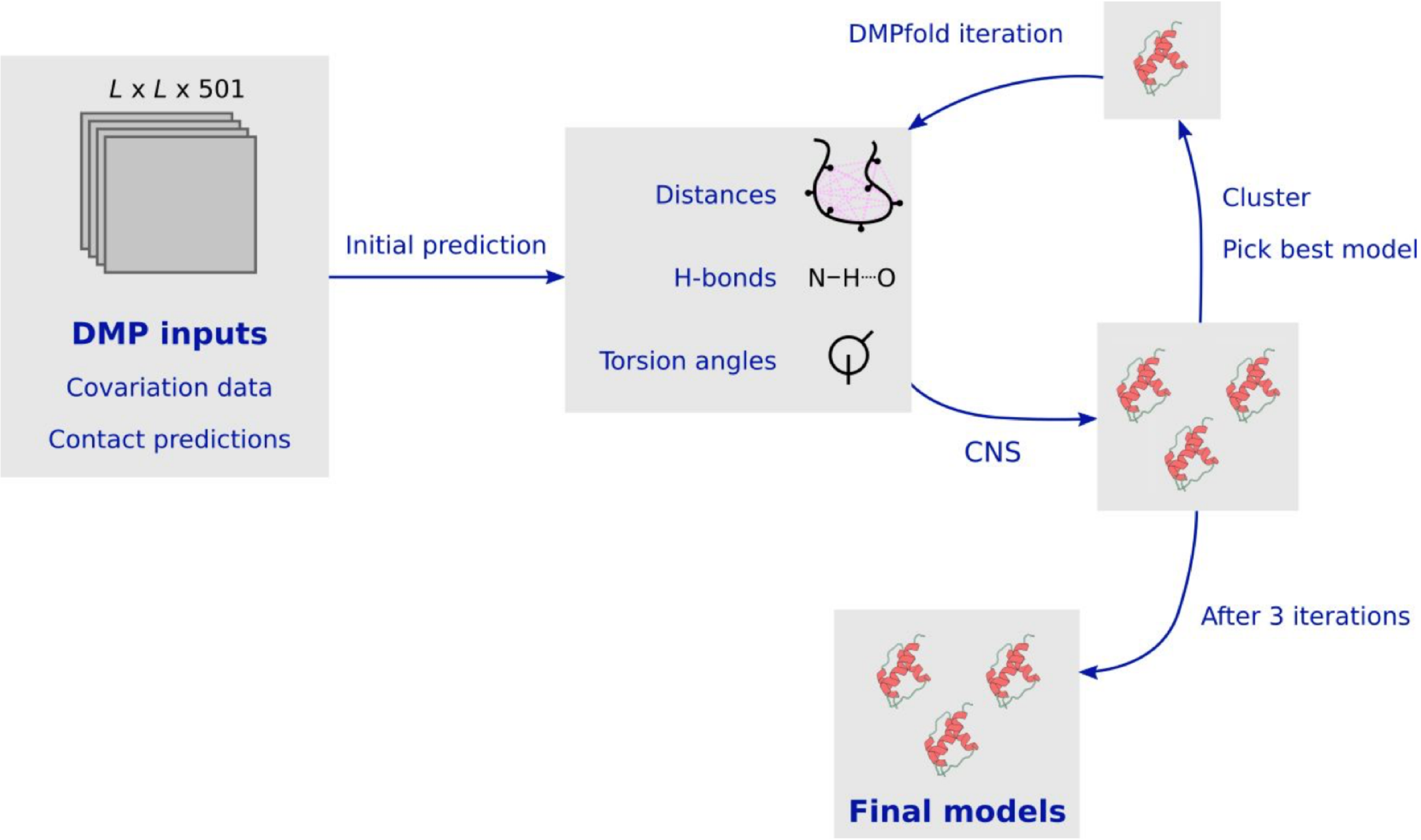
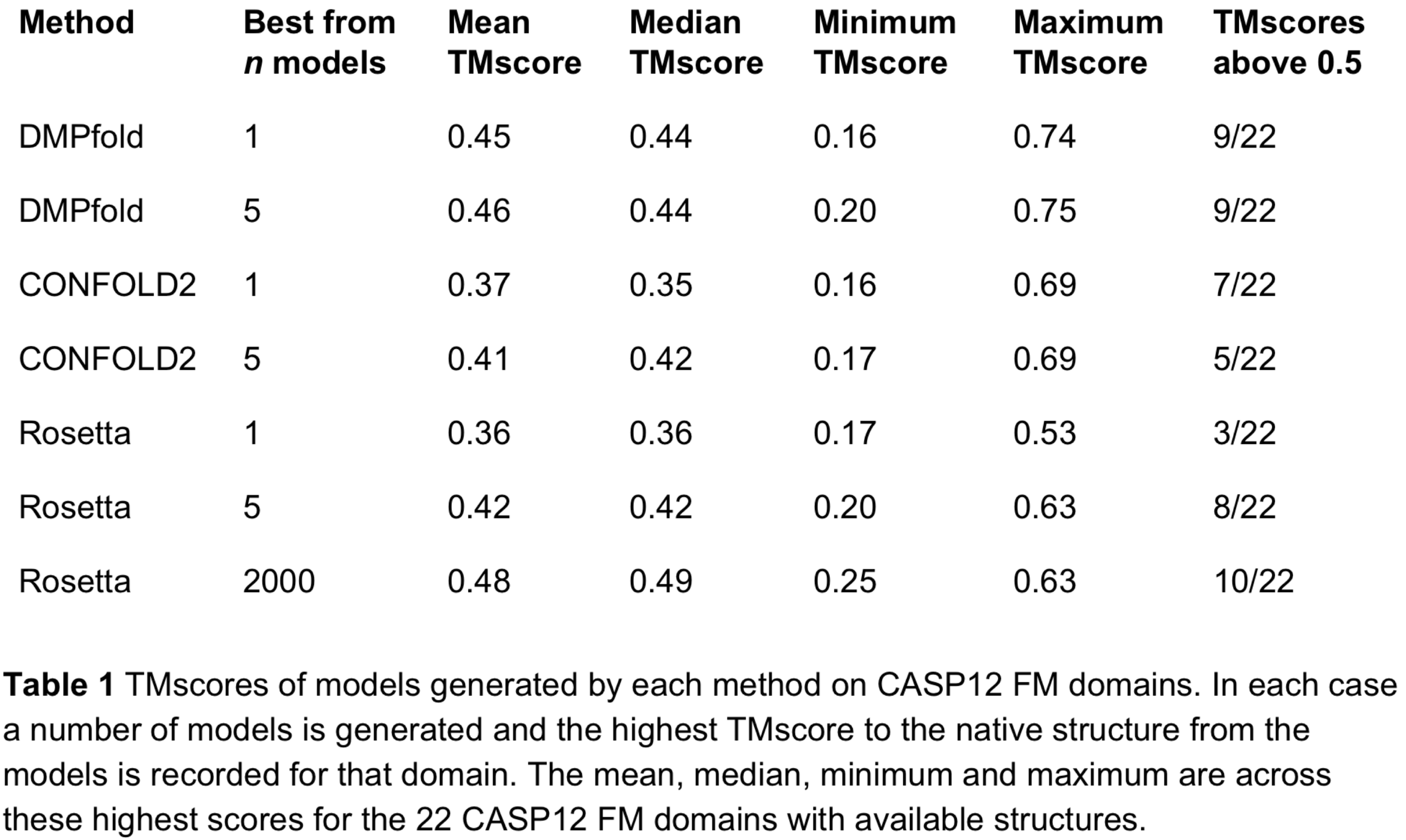


Figure 1 Overview of the DMPfold pipeline. Initially interatomic distances, H-bonds and torsion angles are predicted from DMP inputs. These are used to generate models with CNS, and a single model is used as additional input to refine the distances and H-bonds. After 3 iterations a final set of models is returned.

https://arxiv.org/pdf/1811.12355.pdf



Method	Best from *n* models	Mean TMscore	Median TMscore	Minimum TMscore	Maximum TMscore	TMscores above 0.5
DMPfold	1	0.45	0.44	0.16	0.74	9/22
DMPfold	5	0.46	0.44	0.20	0.75	9/22
CONFOLD2	1	0.37	0.35	0.16	0.69	7/22
CONFOLD2	5	0.41	0.42	0.17	0.69	5/22
Rosetta	1	0.36	0.36	0.17	0.53	3/22
Rosetta	5	0.42	0.42	0.20	0.63	8/22
Rosetta	2000	0.48	0.49	0.25	0.63	10/22

Table 1 TMscores of models generated by each method on CASP12 FM domains. In each case a number of models is generated and the highest TMscore to the native structure from the models is recorded for that domain. The mean, median, minimum and maximum are across these highest scores for the 22 CASP12 FM domains with available structures.

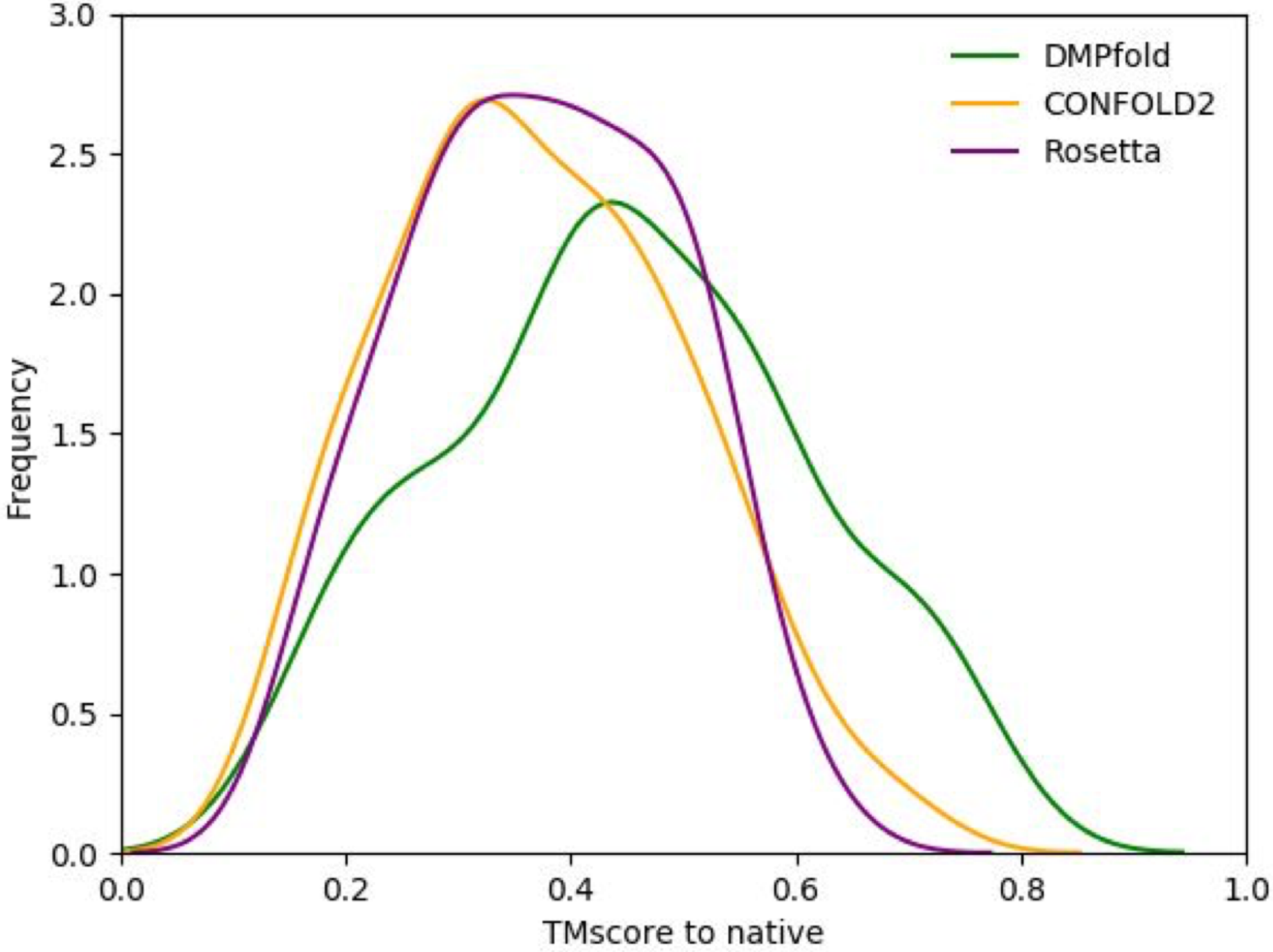
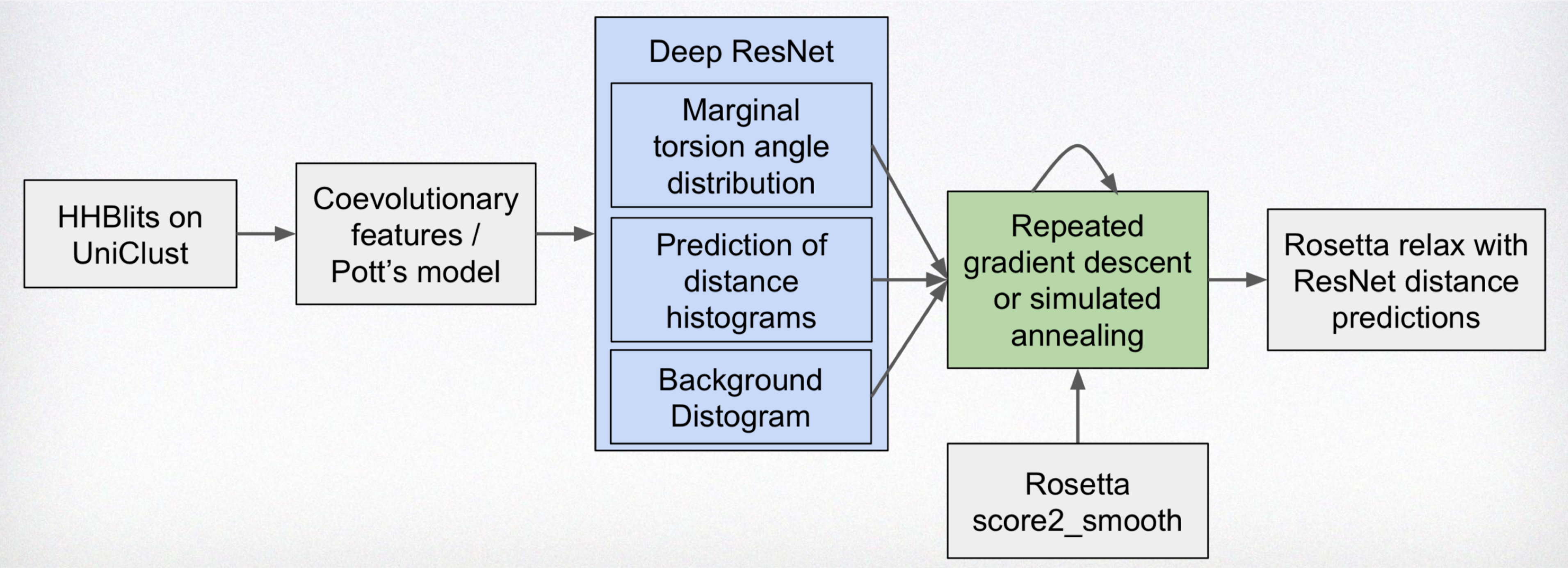


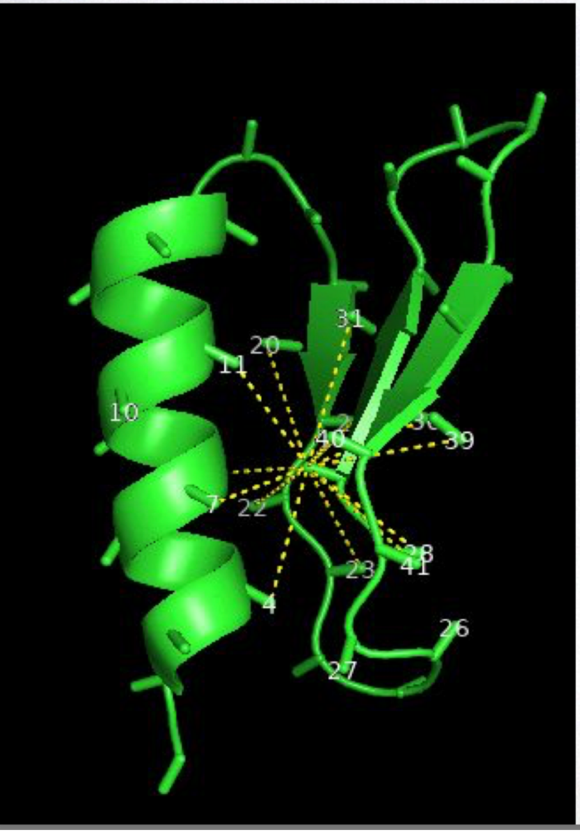
Figure 4 Distribution of TMscores across 5 models for each CASP12 FM domain.

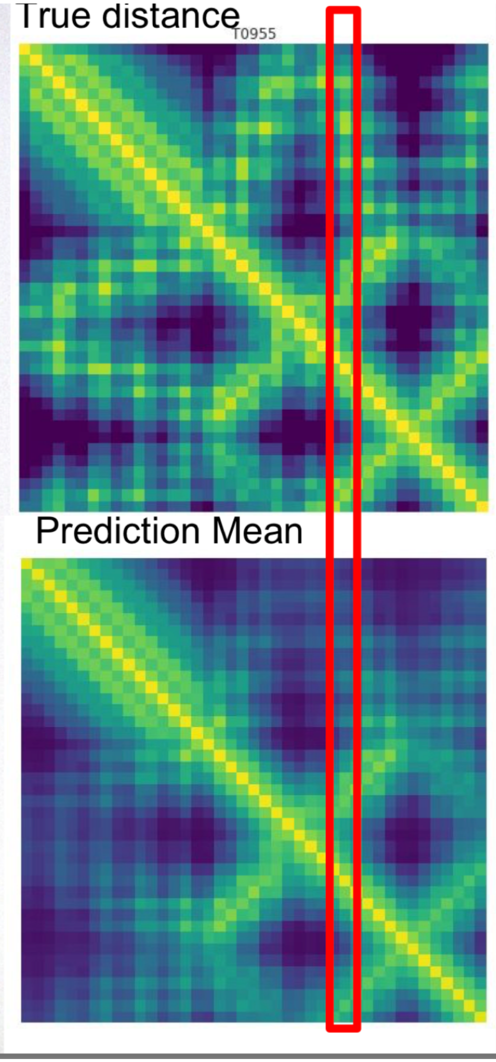
AlphaFold


HHBlits on UniClust
Coevolutionary features / Pott's model
Deep ResNet
Marginal torsion angle distribution
Prediction of distance histograms
Background Distogram
Repeated gradient descent or simulated annealing
Rosetta score2_smooth
Rosetta relax with ResNet distance predictions

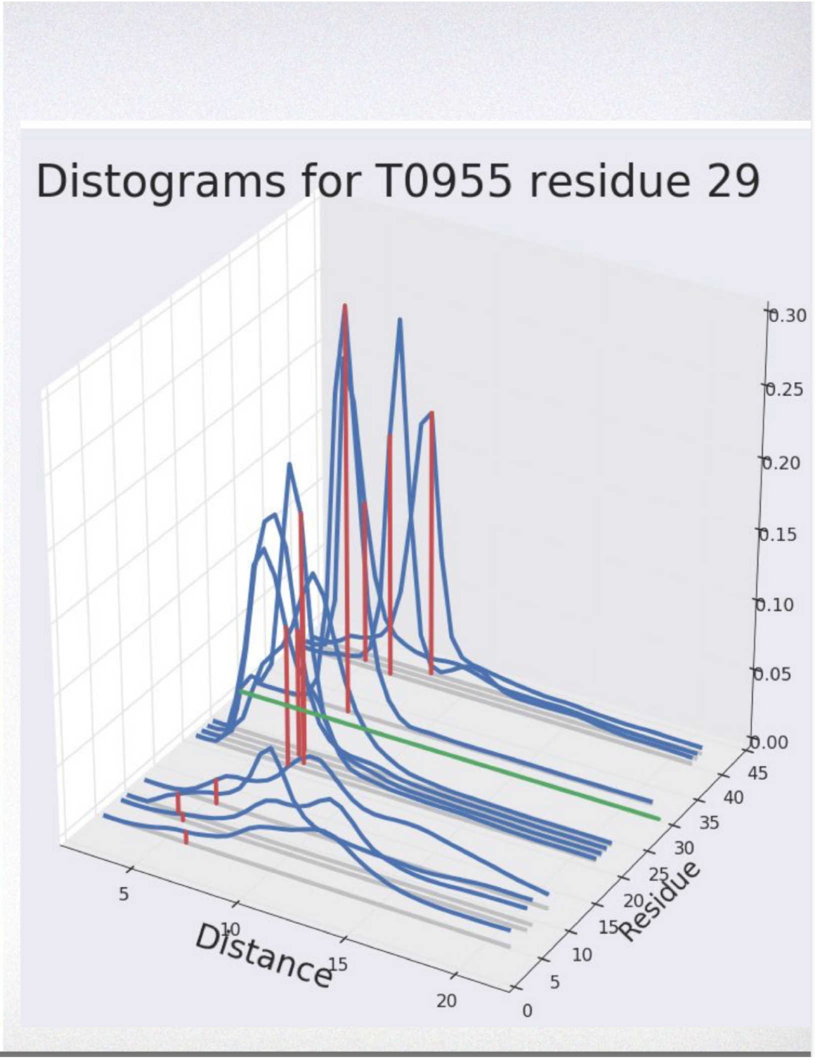
http://predictioncenter.org/casp13/doc/presentations/Pred_CASP13-Structure-AlphaFold-Jumper.pdf

Deep distance distribution Network (D^{3}N)




True distance

Prediction Mean


Distograms for T0955 residue 29

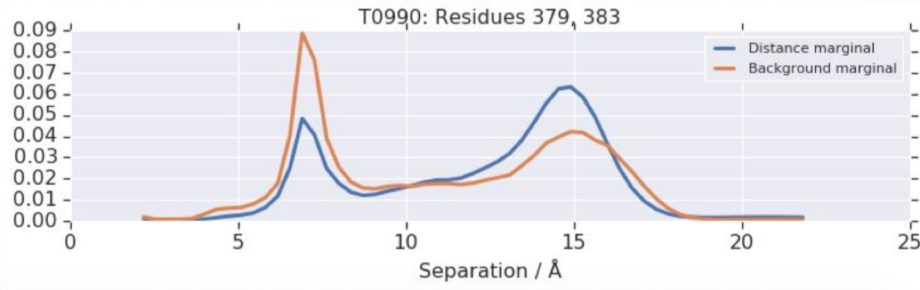
Using deep learning to construct a reference state

The outputs of the distance prediction network are analogous to raw counts in a tabular knowledge-based potential

To obtain a potential, we must apply a reference state correction

We train a neural network to produce reference state distance distributions

- Only input features are i, j, N, and is_glycine
- No other sequence or MSA information



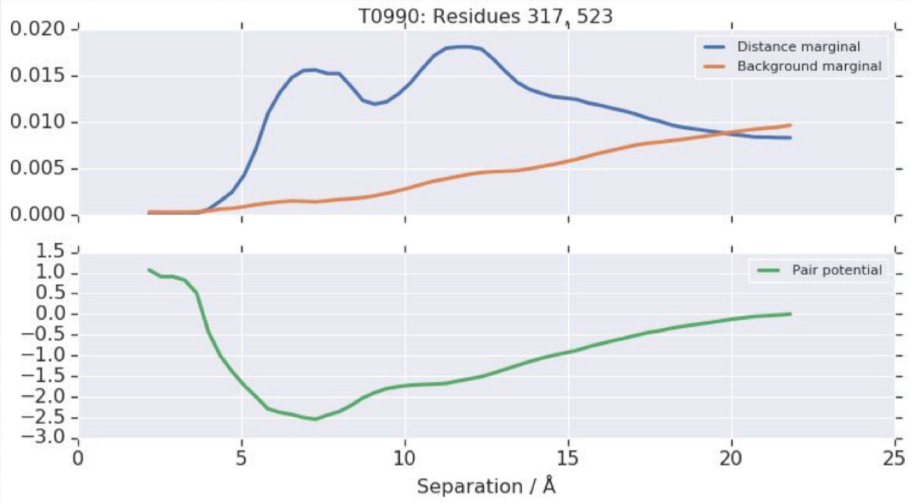
Potential construction

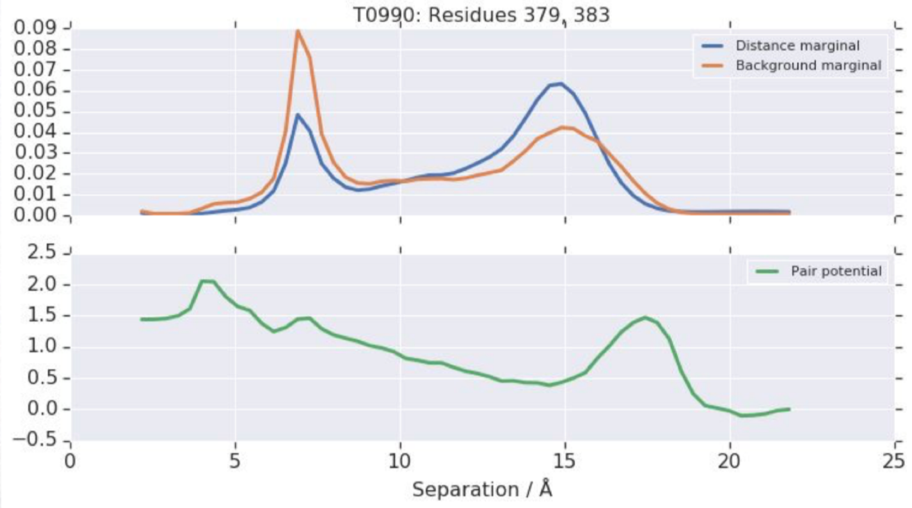
The log ratio tends to be more convex than the distance predictions

$$V_{ij}(d_{ij}) = -\log\left(\frac{\Pr(d_{ij}|i,j,N,\text{sequence},\text{co-evolution})}{\Pr(d_{ij}|i,j,N,\text{is\_glycine})}\right)$$

Potential is score2 + distance potential

Alternatively, can train a scoring network to predict GDT





Optimizing the statistical potential

Two methods

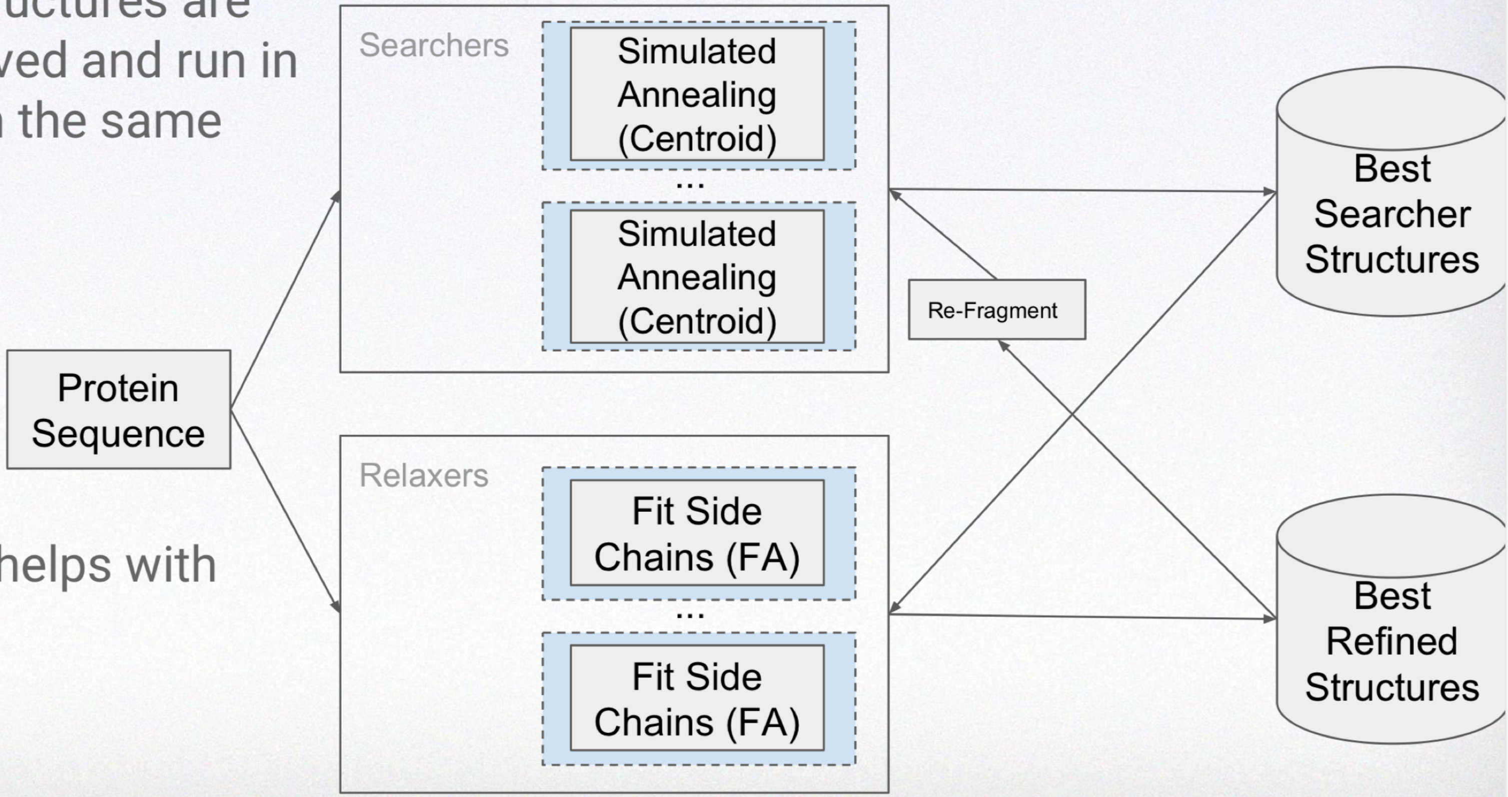
- Simulated annealing with fragment insertion
    - Domain segmented
    - Generative model of protein fragments
    - Higher diversity

- Repeated gradient descent
    - Full chains
    - Lower diversity

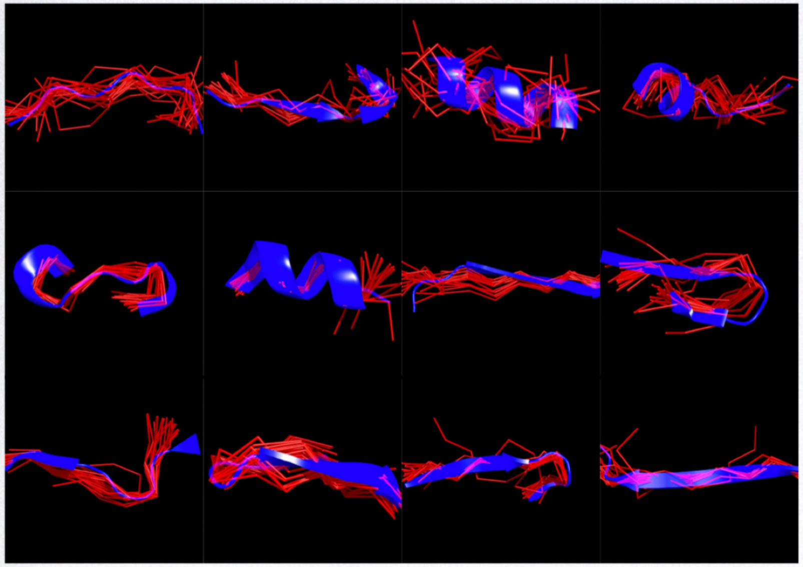
Simulated annealing with fragment insertion

Lowest energy structures are periodically removed and run in Rosetta relax with the same pairwise energy.

Refragmentation helps with accuracy



Generative model of fragments



End-to-end trained model of 32-residue fragments

Based on VAE (variational auto-encoder) with recurrent “canvas”

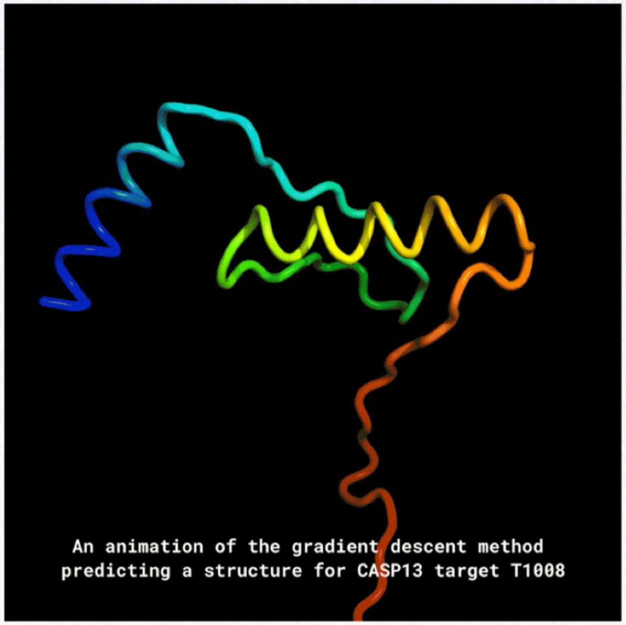
Cut into 9-residue fragments for fragment insertion

Repeated gradient descent

With a smooth Rama, the potential minimizes using repeated gradient descent (initialize from corruptions of best results)

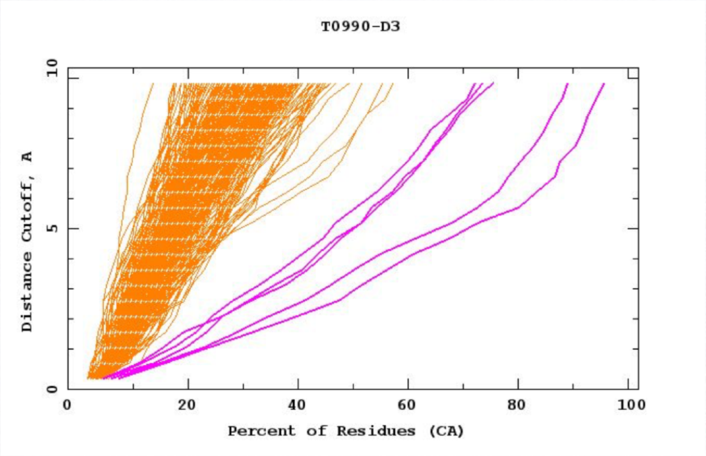
Instead of using fragments, we will use a Rama energy term smoothed to a single von Mises

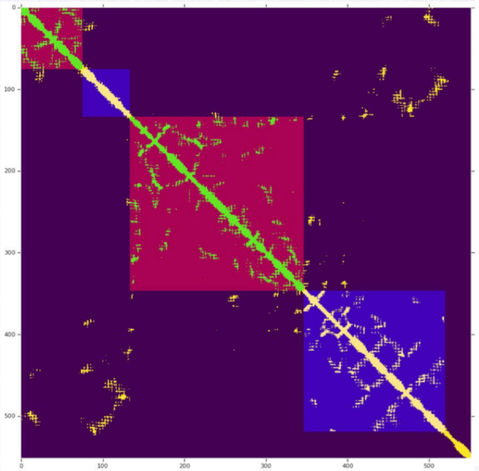
No domain segmentation (except T0999)

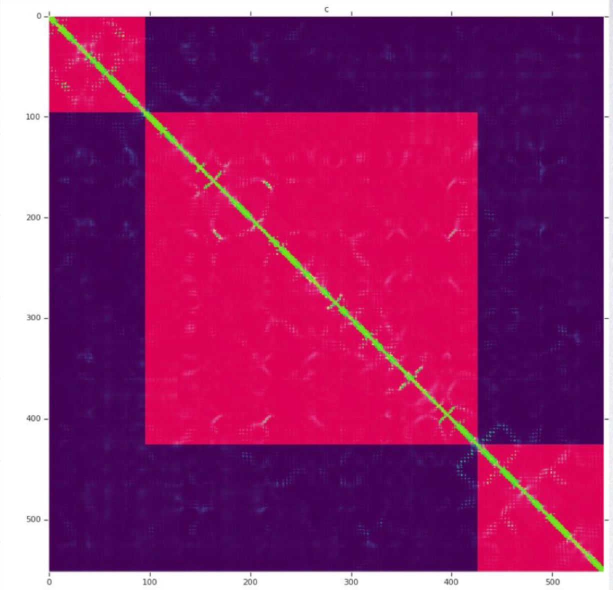


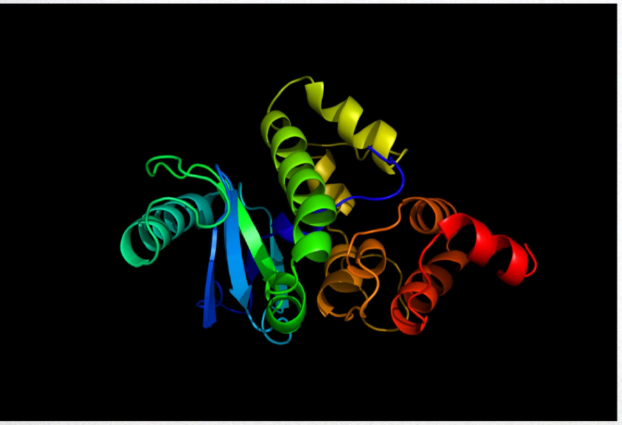
An animation of the gradient descent method predicting a structure for CASP13 target T1008

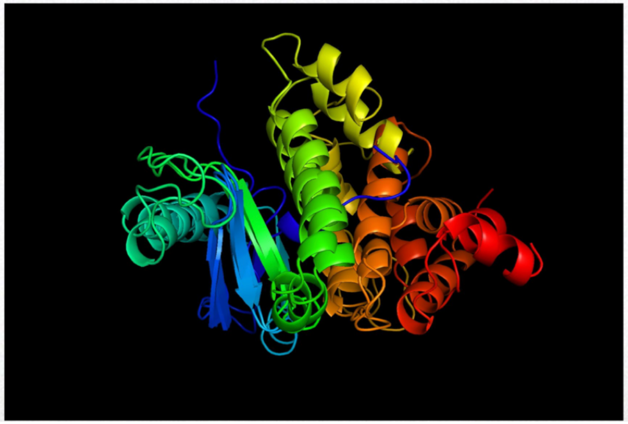
T0990-D3


T0990-D3
Distance Cutoff, A
0
5
10
0
20
40
60
80
100
Percent of Residues (CA)

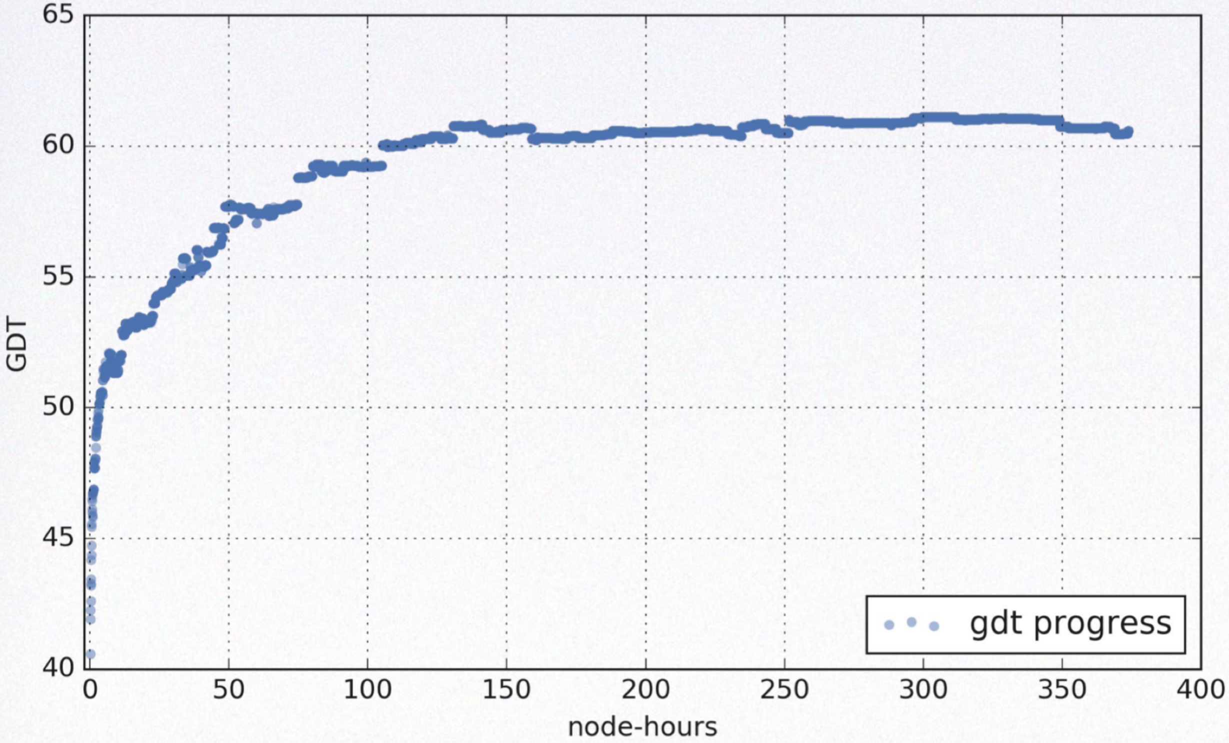

0
100
200
300
400
500


c
0
100
200
300
400
500





Accuracy vs computational cost



(for a subset of targets, on CPU nodes)

Repeated gradient descent

Using simple vdW instead of score2

Highly parallelizable

Project 2

- Develop a simple prototype of contact distance-based *ab initio* protein structure prediction system
- You may use existing contact prediction tools and distance-based model reconstruction tools or develop you own tools (e.g. gradient descent based model construction tools).
- Test it on three CASP12 or CASP13 targets

Timeline

- March 18: discussion of the plan
- March 20: presentation of the plan
- April 3rd, presentation of the results
- April 8th, report due

Discussion of Project Plan

- Select targets (two easy, one hard?)
- Contact prediction (co-evolution-based methods, deep learning methods (DNCON2, DeepCov))
- Contact-based modeling (CONFOLD2, Rosetta, UniCon3D, Modeller, your own gradient descent)
- Model Refinement
- Evaluation and Analysis
- Visualization (contact map, 3D structures, modeling movies)
- Project management / task assignment

Technical Resources

Contact prediction

- DNCON2: https://github.com/multicom-toolbox/DNCON2
- DeepCov: https://github.com/psipred/DeepCov
- CCMpred: https://github.com/soedinglab/CCMpred

Contact Visualization

- ConEVA: http://iris.rnet.missouri.edu/coneva/index.php

Technical Resources

Model reconstruction

CONFOLD2: https://github.com/multicom-toolbox/CONFOLD2

Rosetta: https://www.rosettacommons.org/manuals/archive/rosetta3.4_user_guide/index.html

UniCon3D: **https://github.com/multicom-toolbox/UniCon3D**

Model Refinement (both software and web servers)

3DRefine: http://sysbio.rnet.missouri.edu/3Drefine/index.html
i3DRefine: http://protein.rnet.missouri.edu/i3drefine/