

# KFC/STBI

# Structural Bioinformatics

Structure Prediction

Karel Berka

# Outline

- Structural alignment
- Structure prediction
  - Homology modelling
    - SwissMODEL, Modeller,
  - threading
    - I-TASSER
  - de novo modelling
    - Robbeta, Quark
- molecular mechanics
  - protein folding
  - Folding@Home, FoldIt
- Evolutionary coupling
  - EVcoupling
- Machine learning
  - AlphaFold

# Structural Alignment

# Are Those Structures Similar?

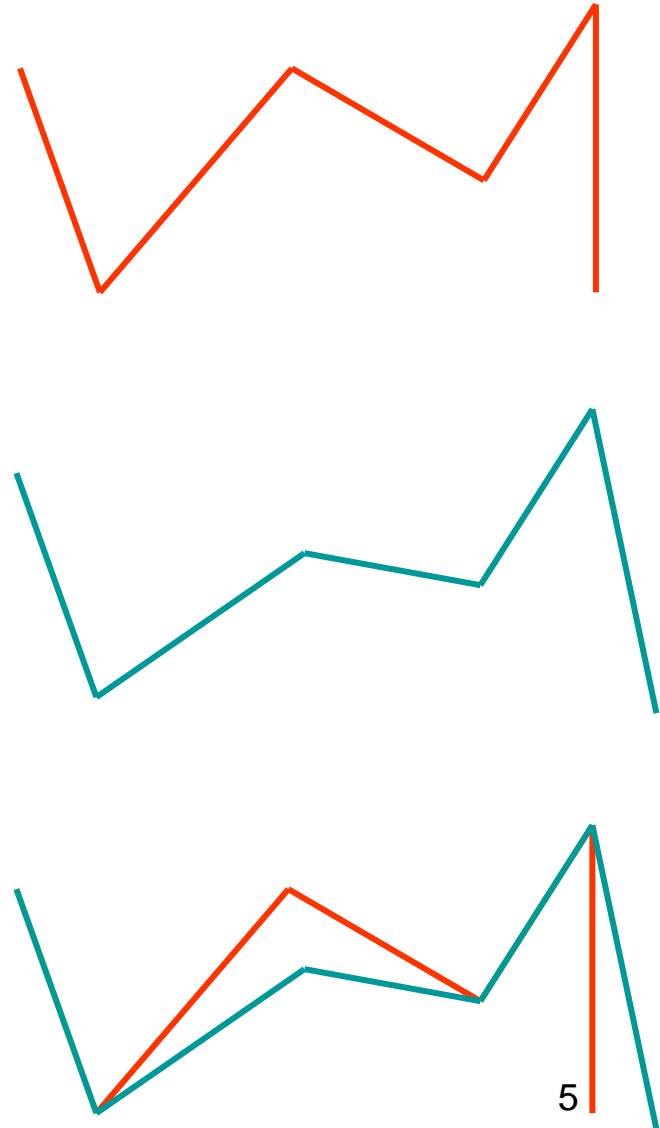
- By eye
- By algorithm:
  - Structural alignment



Structural alignment of [thioredoxins](#) from humans (red) and [Drosophila](#) (yellow)  
PDBID: [3TRX](#) and [1XWC](#).

# Structural Alignment

- To find **the best** pairing between two structures
- The best  
-> “smallest RMSD”
- Problem:
  - Quite often it is possible to find only subset of dissimilar atoms – how to discern them?



# Structural Alignment

- Other problems:
  - Nr of aminoacids in both chains
  - fit and RMSD calculation
  - Identity between „aligned residues“
  - Nr of „gaps“
  - Size of proteins
  - Conserved sequence sites
- There are no universal criteria

# Structural alignment

- Warning:
- It is different to RMSD calculation -
- there is not easy correspondance between atoms. => Z  
Analysis of all possible correspondences
- RMSD is just tool to analyse similarity

# Why to use structural alignment?

Structure is usually more conserved than sequence  
(there is smaller number of structural fold families  
in contrast with sequence clusters)

1. Homologous proteins (same ancestor)
  - „gold standard“ for sequence alignment
2. Nonhomologous proteins
  - similar substructures (domains)
3. Classification to clusters
  - structural similarities (CASP)
  - sequence similarities (Pfam)









# CATH vz Pfam

<http://www.cathdb.info/>

## Latest Release Statistics

 Info

	CATH v4.1		CATH-B	
PDB Release	01-01-2015		about 7 hours ago	
Domains	308999		417837	
Superfamilies	2737		6344	
Annotated PDBs	108378		123625	

	Gene3D v14
Cellular Genomes	19,471
Protein Sequences	43,387,462
CATH Domain Predictions	53,479,436

<http://pfam.xfam.org/>

- [Pfam 30.0](#)  
July 1, 2016
- 16,306 families
  - 22 new and 11 killed families
- 17.7M sequences
  - 11.9M sequences in Pfam 29

Sillitoe I, et al. **CATH: comprehensive structural and functional annotations for genome sequences.** *Nucleic Acids Res.* 2015 [doi: 10.1093/nar/gku947](https://doi.org/10.1093/nar/gku947)

R.D. Finn et al. [The Pfam protein families database: towards a more sustainable future](#); *Nucleic Acids Research* (2016) 44:D279-D285

# Types of Structural Alignment

point methods

- **CE** (Combinatorial Extension)  
rigid structures, start – largest group of sequentially equivalent atoms
- **DALI** (Holm, Sander)  
matrix of distances to search for similar patterns to fit correspondences between atoms (without sequence)

secondary structure methods

- **VAST**  
alignment of secondary structures
- **FATCAT**  
protein is not rigid - hinges can bend

# CE (Combinatorial Extension)

- pairs of protein segments by 8 AA
- comparison by local geometry
- pairs are further enlarged
- results:
  - RMSD
  - z-statistics

(standard) z-score is 
$$z = \frac{x - \mu}{\sigma}$$

where:

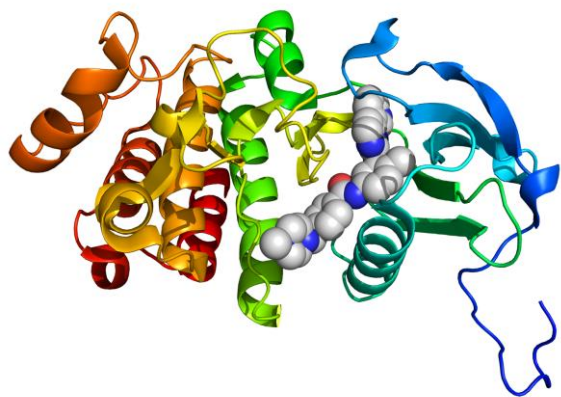
$x$  is a raw score to be standardized;

$\mu$  is the mean of the population;

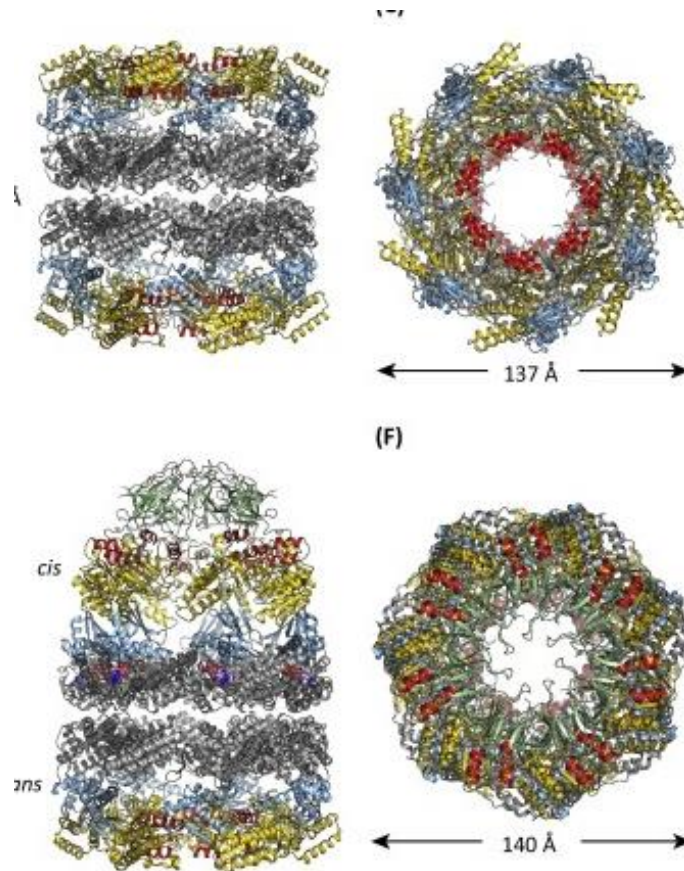
$\sigma$  is the standard deviation of the population

# Structure Prediction

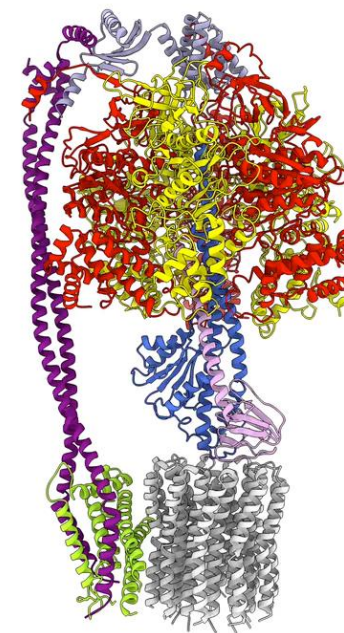
# Knowing structure helps to understand the function



wikipedia

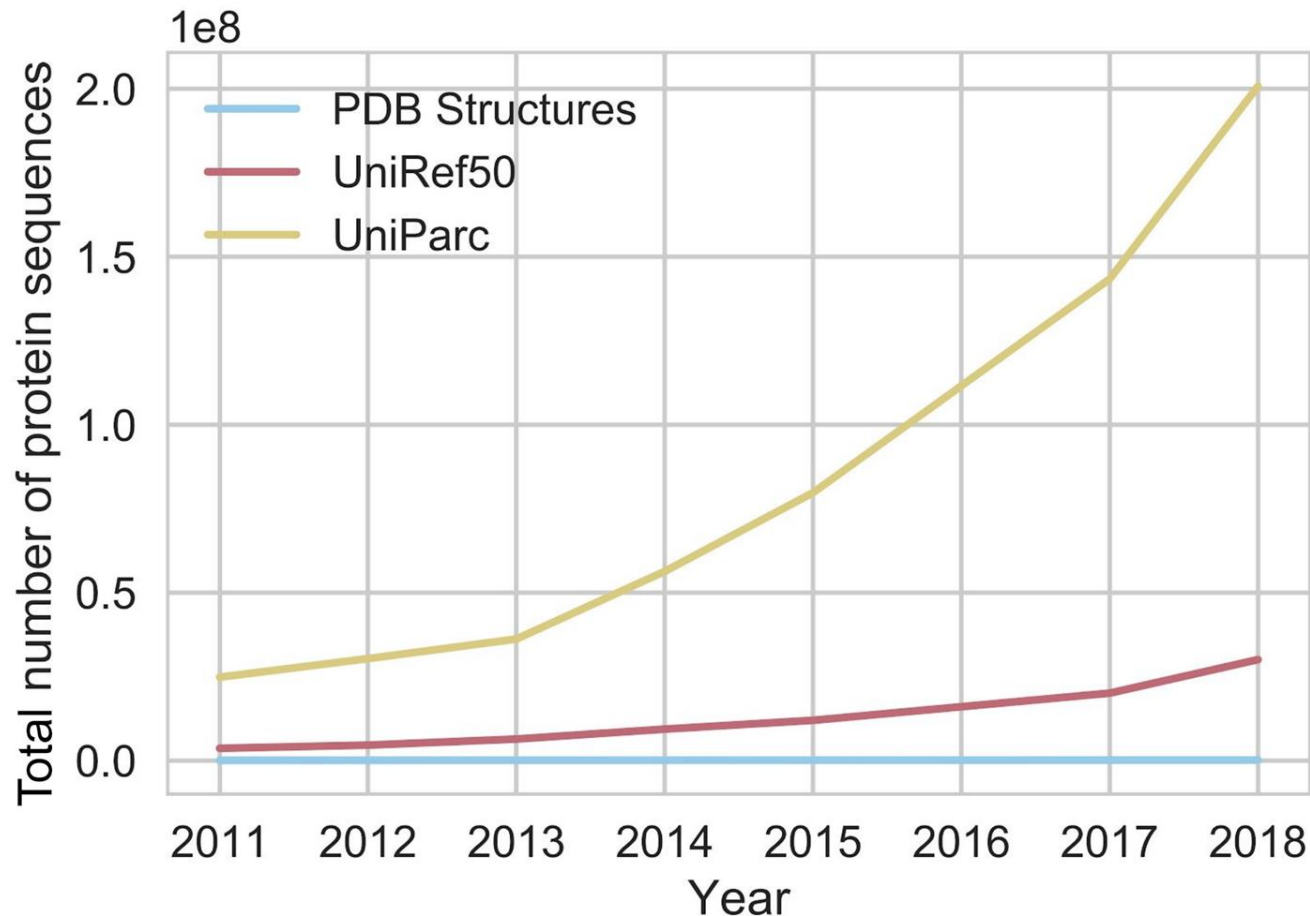


Hayer-Hartl et al., 2015



Guo et al., 2019

# Solving 3D structures is expensive...



<https://bair.berkeley.edu/blog/2019/11/04/proteins/>

The gap between numbers of experimental structures and sequences is increasing over time

# Can we use sequence to predict 3D structure?

- C.B. Anfinsen received Nobel prize in Chemistry (1972) for describing the relationship between sequence and structure



C.B. Anfinsen

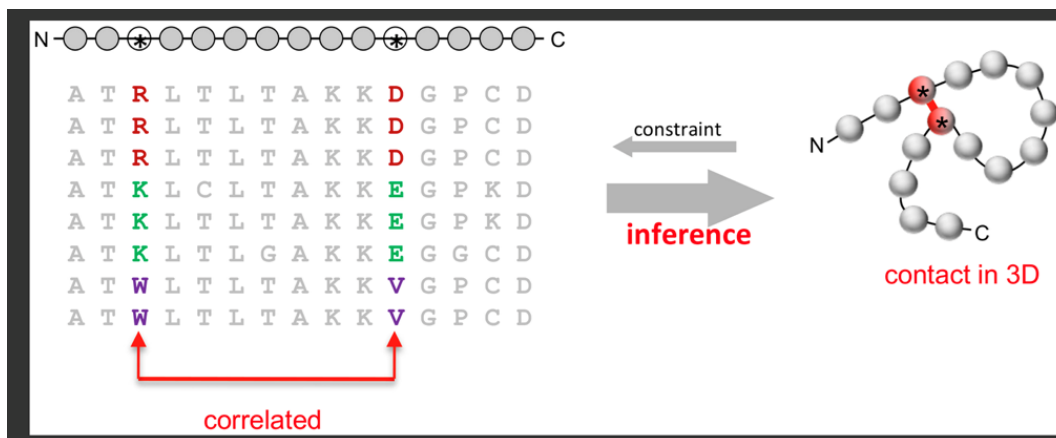
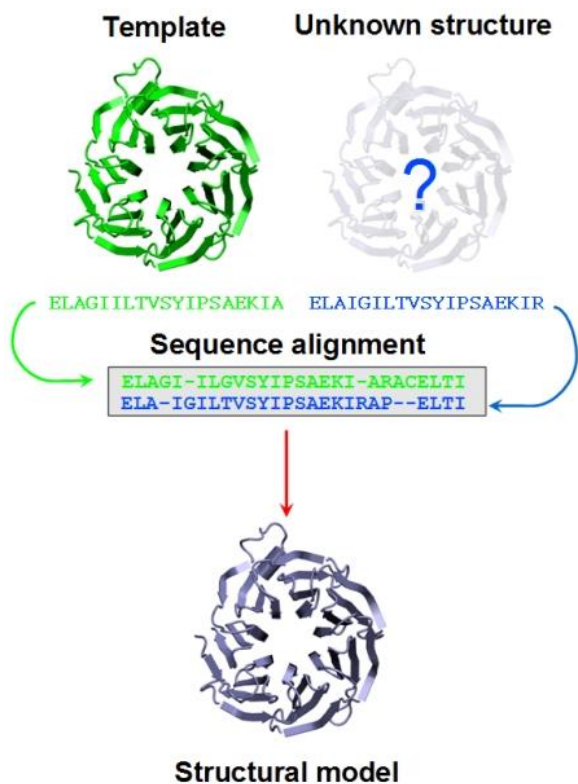
"The native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment."

- it shall be possible to give to **predict structure from sequence**

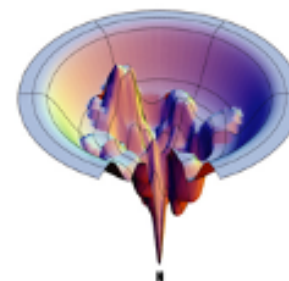


ribonuclease

# Principles of prediction from sequence



<https://www.unil.ch/pmf/en/home/menuinst/technologies/homology-modeling.html>





# Structure prediction = simulation of protein folding?

## Levinthal's paradox

- protein of 100 aa has  $10^{70}$  available conformations
- > it would take  $10^{52}$  years at the speed of  $10^{-11}$ s to sample one conformation to assume its native shape

# How to move the prediction field forward?

- transparent competition
  - provide an “environment” for communication and exchange of experience
  - develop metrics for careful examination of predicted structures
- 
- **CASP** – critical assessment of protein structure prediction
  - once in two years since 1994
  - compare with experimentally solved structures



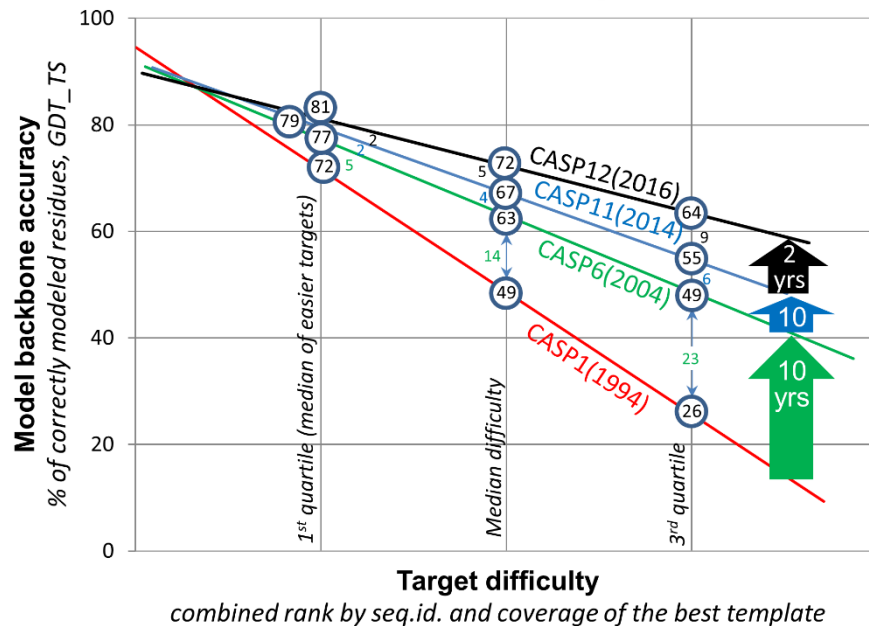
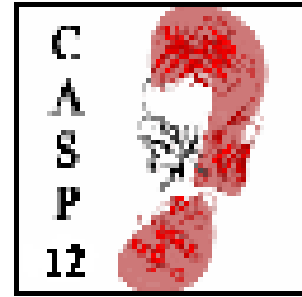
[John Moult](#) - father of CASP

# **CASP**

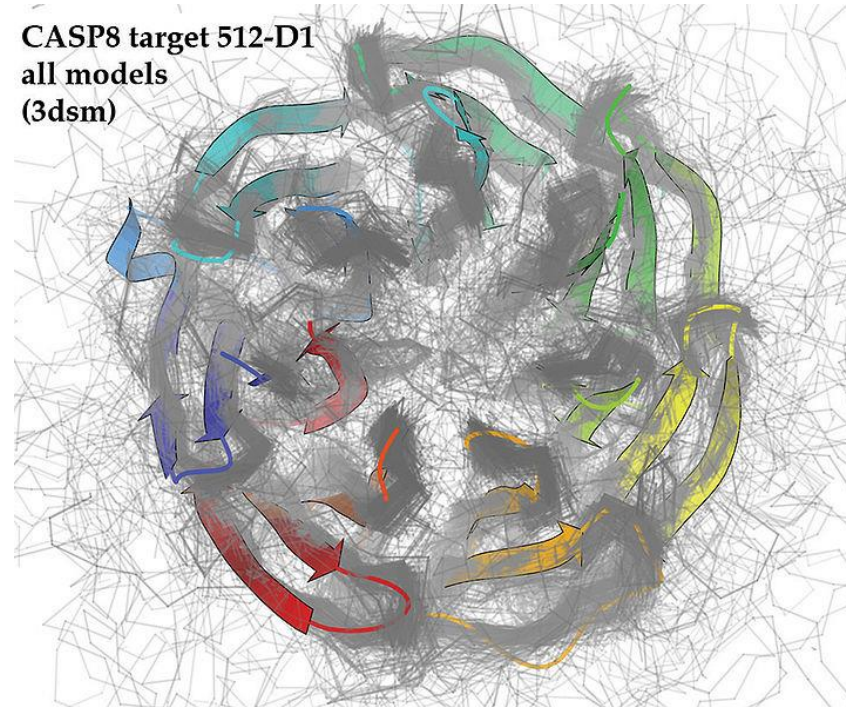
Critical Assessment of protein Structure Prediction

- **Critical Assessment of Techniques for Protein Structure Prediction**
- Comparison with prepublished x-ray data
- no prior information for predictors (double-blind)

# CASP



CASP8 target 512-D1  
all models  
(3dsm)



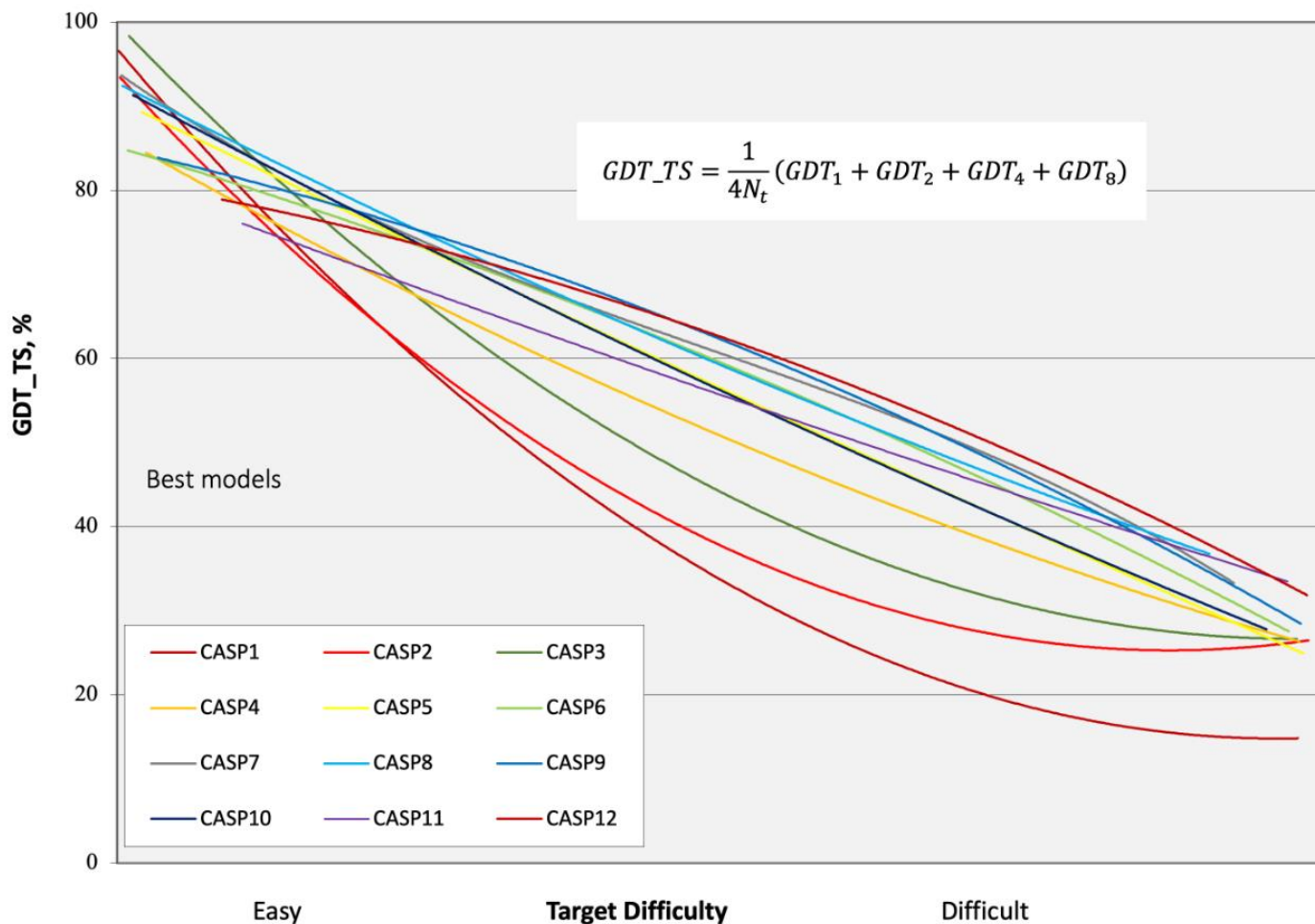
<http://predictioncenter.org/casp9/index.cgi>

Proteins: Structure, Function, and Bioinformatics  
Volume 77, Issue S9, Pages 1-228 (2009)

# CASP

- tertiary structure prediction (all CASPs)
- secondary structure prediction (dropped after CASP5)
- prediction of structure complexes (CASP2 only; a separate experiment - CAPRI - carries on this subject)
- residue-residue contact prediction (starting CASP4)
- disordered regions prediction (starting CASP5)
- domain boundary prediction (CASP6-CASP8)
- function prediction (starting CASP6)
- model quality assessment (starting CASP7)
- model refinement (starting CASP7)
- high-accuracy template-based prediction (starting CASP7)

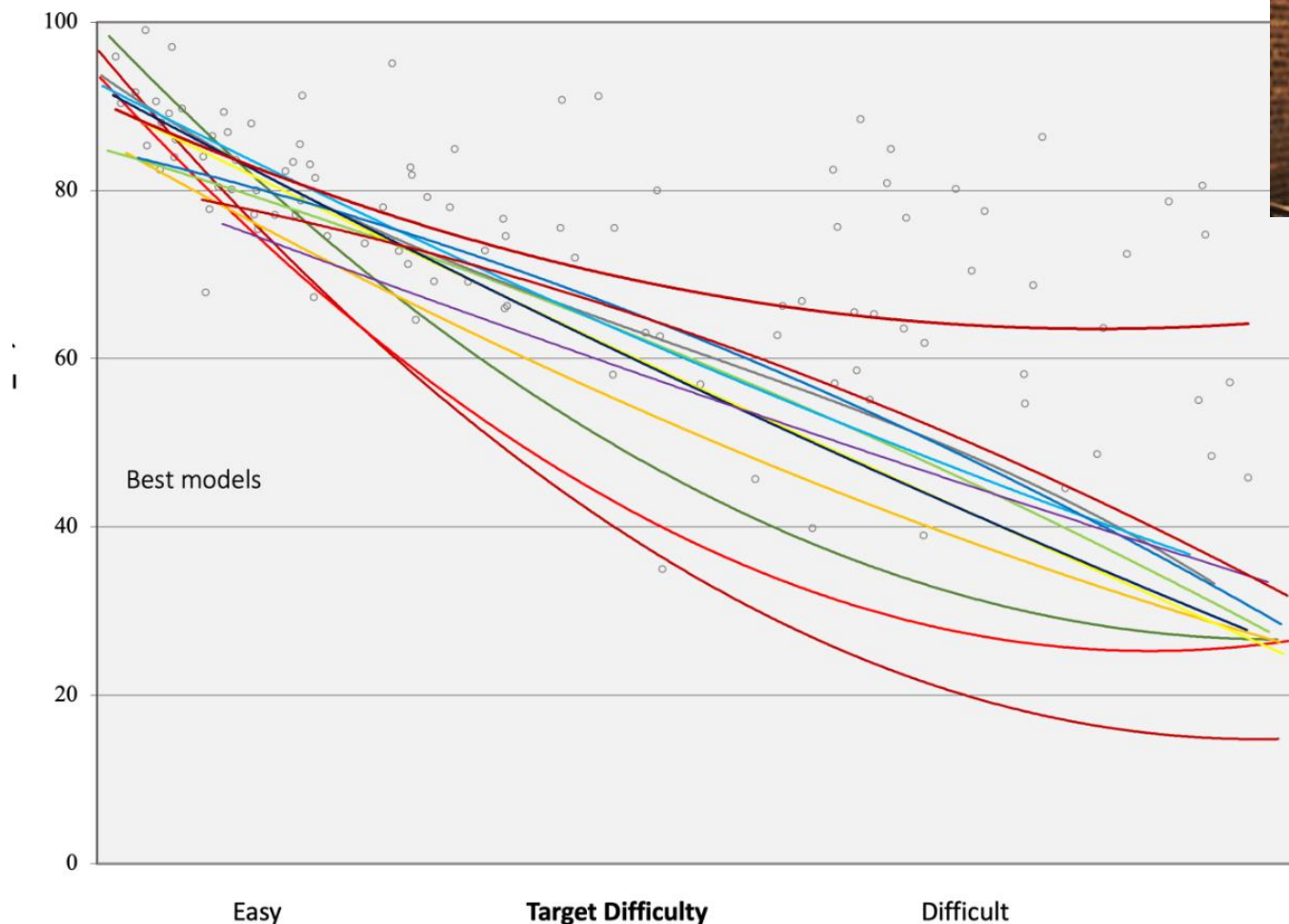
# How to compare structures?



[https://predictioncenter.org/casp14/doc/presentations/2020\\_11\\_30\\_CASP14\\_Introduction\\_Moult.pdf](https://predictioncenter.org/casp14/doc/presentations/2020_11_30_CASP14_Introduction_Moult.pdf)

GDT\_TS = Global distance test - total score (max 100%)  
The conventional GDT\_TS total score in **CASP** is the average result of cutoffs at 1, 2, 4, and 8 Å falling within experimental position

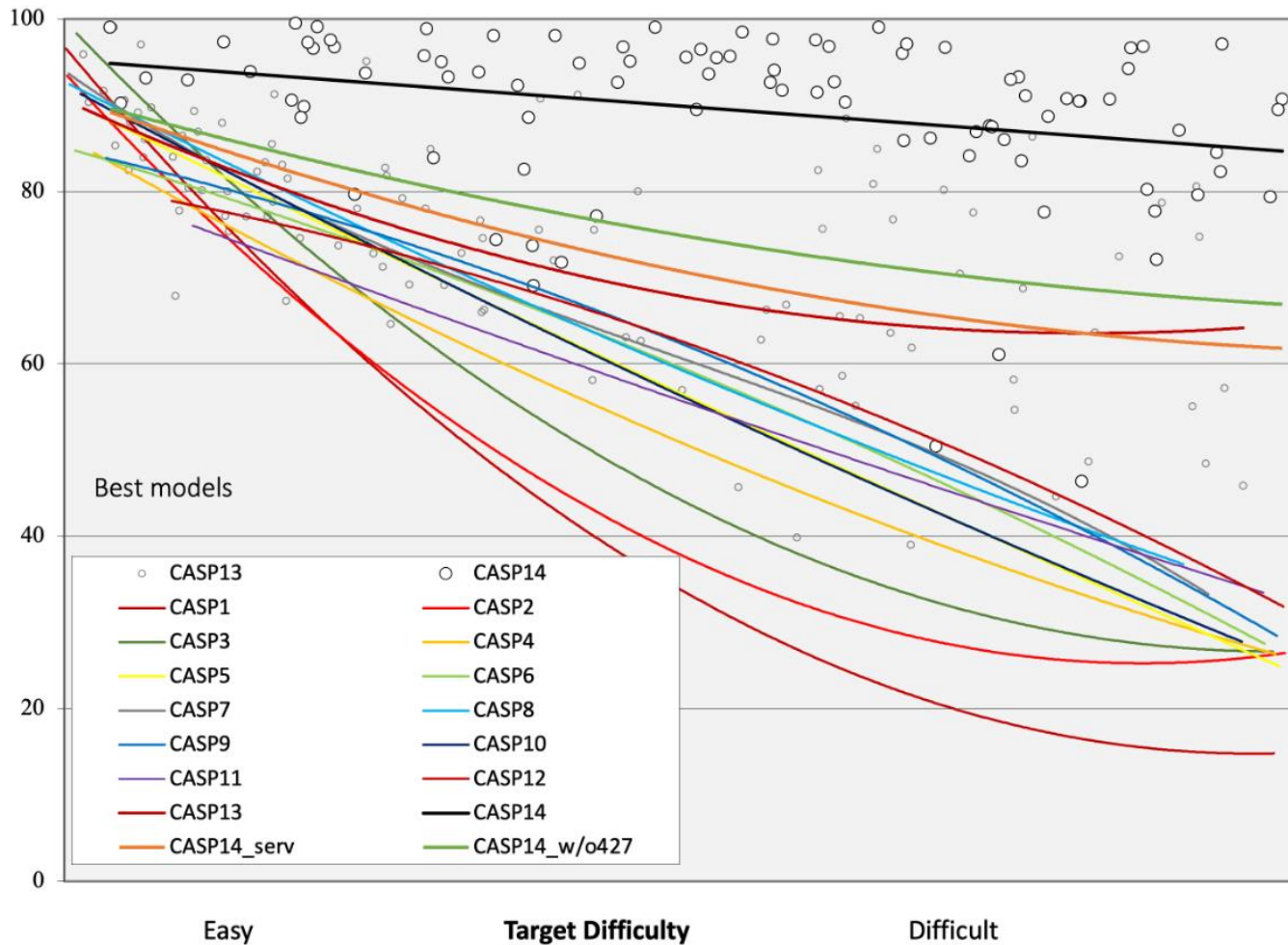
# 2018: AlphaFold enters...



[https://predictioncenter.org/casp14/doc/presentations/2020\\_11\\_30\\_CASP14\\_Introduction\\_Moult.pdf](https://predictioncenter.org/casp14/doc/presentations/2020_11_30_CASP14_Introduction_Moult.pdf)



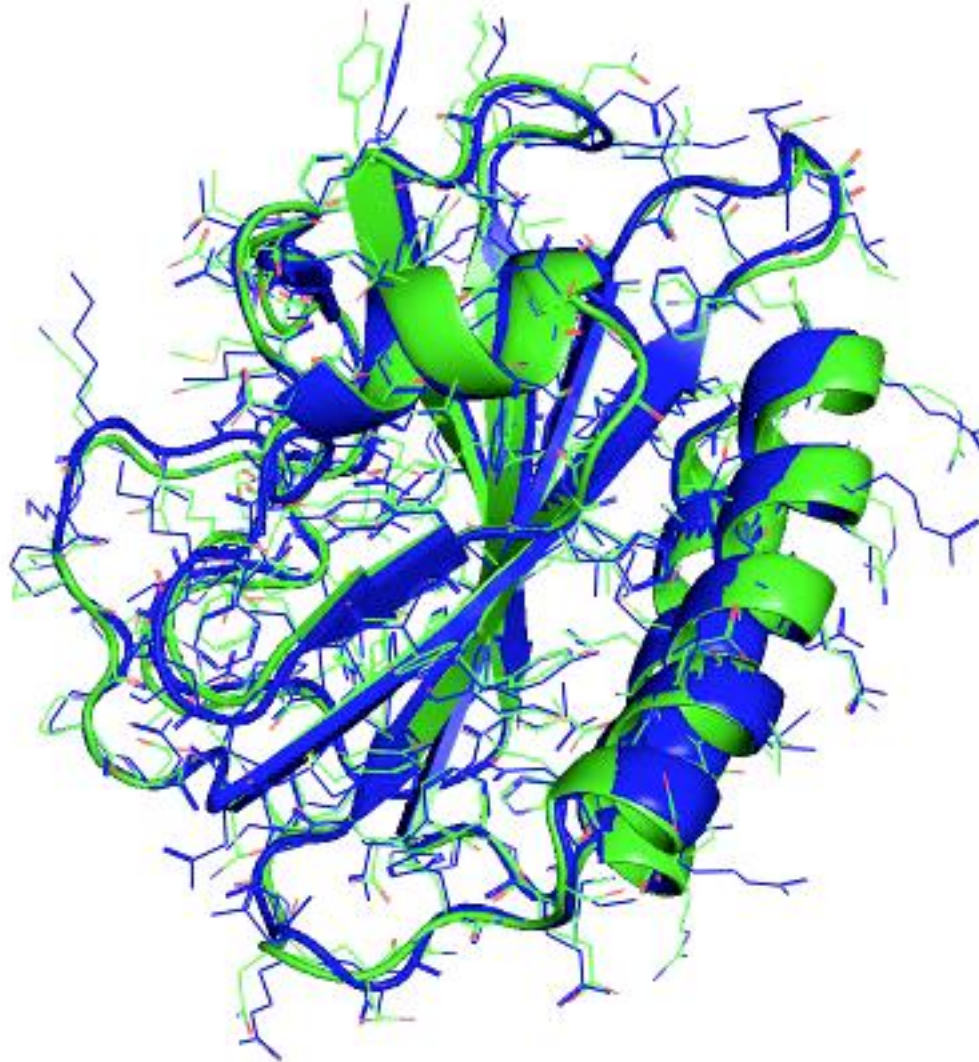
# 2020: AlphaFold2 wins



[https://predictioncenter.org/casp14/doc/presentations/2020\\_11\\_30\\_CASP14\\_Introduction\\_Moult.pdf](https://predictioncenter.org/casp14/doc/presentations/2020_11_30_CASP14_Introduction_Moult.pdf)

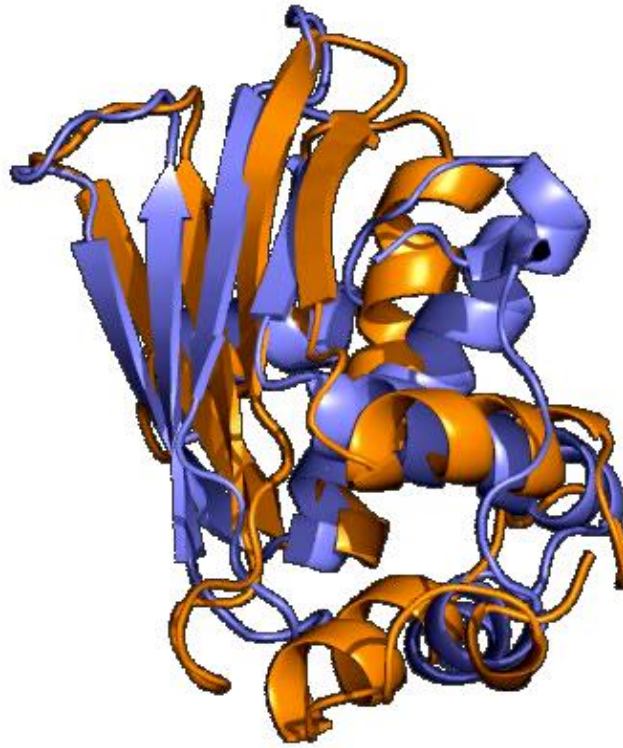


# How does good prediction look like?



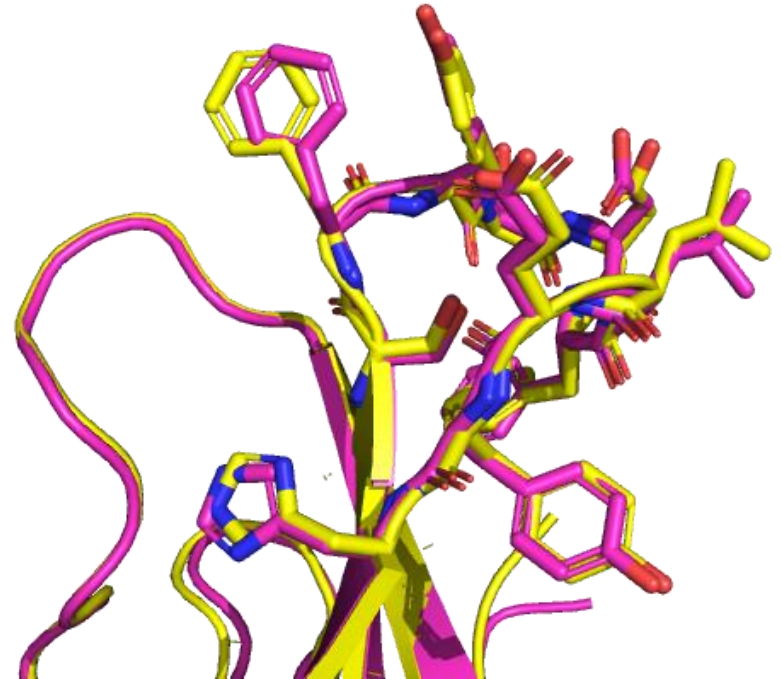
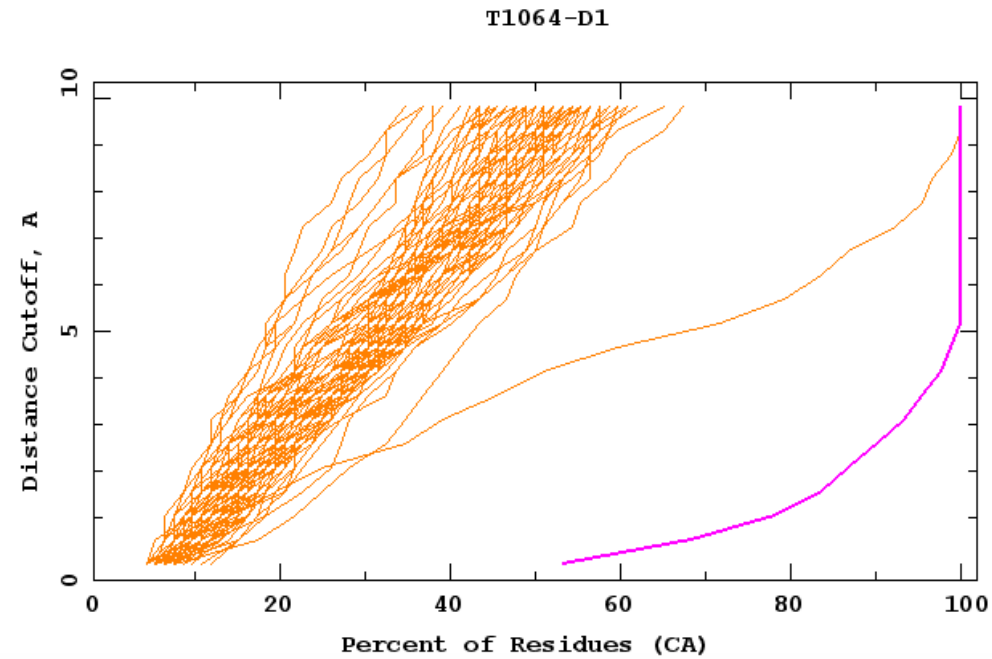
GDT\_TS = 96.5

# The worst prediction of AlphaFold 2 in CASP 14



GDT\_TS = 44.6

# Side chain predictions— orf8 covid19



GDT\_TS= 87

so how it works?

# Prediction of Protein Tertiary Structure

- Structure Prediction
  - from known structures
    - Homology modelling
      - SwissMODEL, I-TASSER
    - threading
      - Modeller
  - from physical models
    - de novo modelling (ab initio)
      - Quark, Robbeta,
      - protein folding
      - Folding@Home, FoldIt
- Protein evolution
  - Evolutionary coupling
- Machine learning
  - AlphaFold
  - Xfolds...

# Homology modelling

- also known as comparative or knowledge-based modelling
- based on template structure

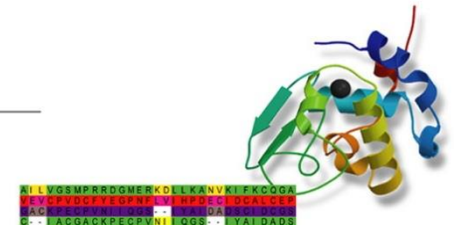
- Swiss-MODEL
  - <http://www.expasy.org/spdbv/>



- Modeller
  - <http://salilab.org/modeller/>

# Modeller

### Program for Comparative Protein Structure Modelling by Satisfaction of Spatial Restraints



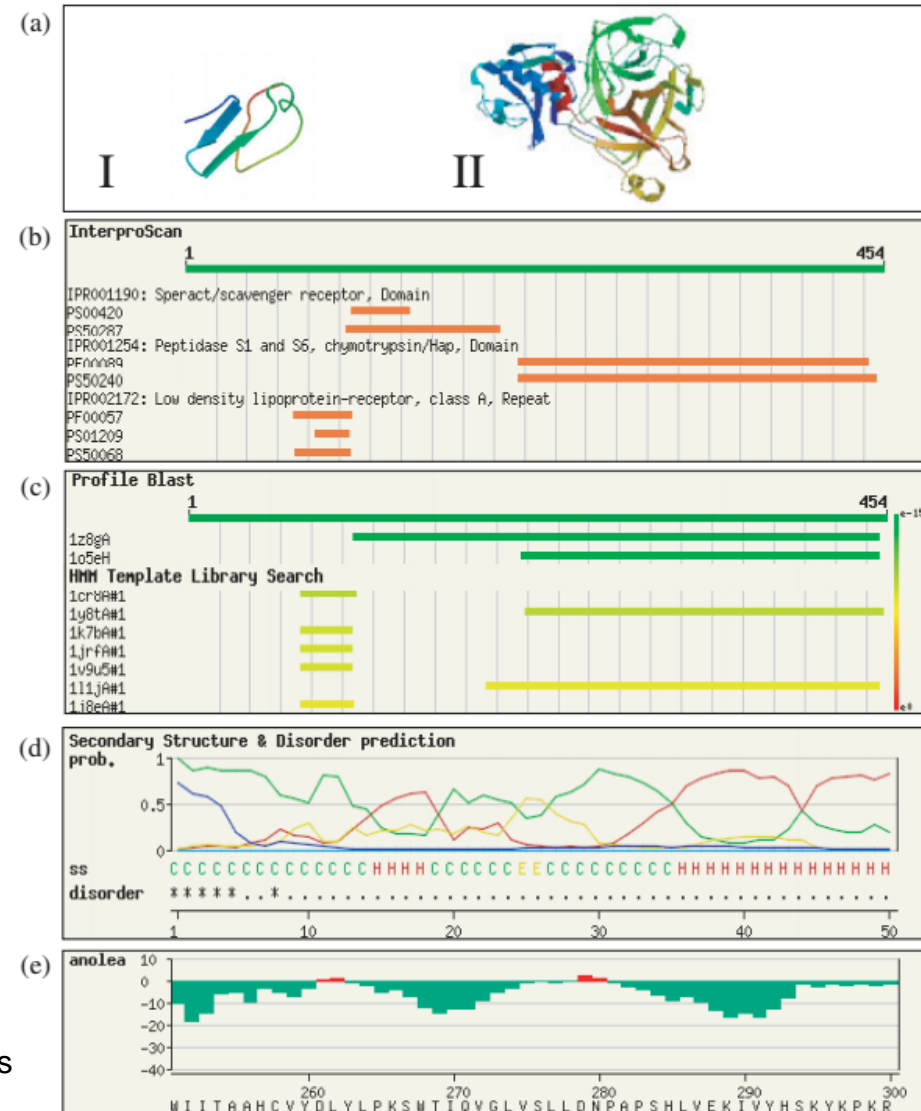
# Typical protocol

- template selection
  - sequence alignment
- target-template alignment
  - pair comparisons (usually iterative refinement of alignment)
- model construction
  - main chain construction
  - loops
  - side chain construction -> rotamers
  - energy minimization
- model assessment
  - stereochemical control (PROCHECK, Ramachandran plot)
  - statistics – scoring function, z-score, probability of failure...

# SwissModel



- Model representation
  - visual
  - InterproScan to detect domains
  - Sequence-based searches of the template library
  - Secondary structure and disorder prediction of the target protein.
  - Anolea mean force potential plot allows for quality assessment





# Modeller

Program for Comparative Protein  
Structure Modelling by Satisfaction  
of Spatial Restraints



A	I	L	V	G	S	M	P	R	R	D	G	M	E	R	K	D	L	L	K	A	N	V	K	I	F	K	C	Q	G	A
Y	E	V	C	P	V	D	C	F	Y	E	G	P	N	F	L	V	I	H	P	D	E	C	I	D	C	A	L	C	E	P
G	A	C	K	P	E	C	P	V	N	I	I	Q	G	S	-	-	Y	A	I	D	A	D	S	C	I	D	C	G	S	
C	-	-	I	A	C	G	A	C	K	P	E	C	P	V	N	I	I	Q	G	S	-	-	I	Y	A	I	D	A	D	S

- homology modelling with constraints (NMR, EM, apod)
  - i-Sites (short pieces with known structure)

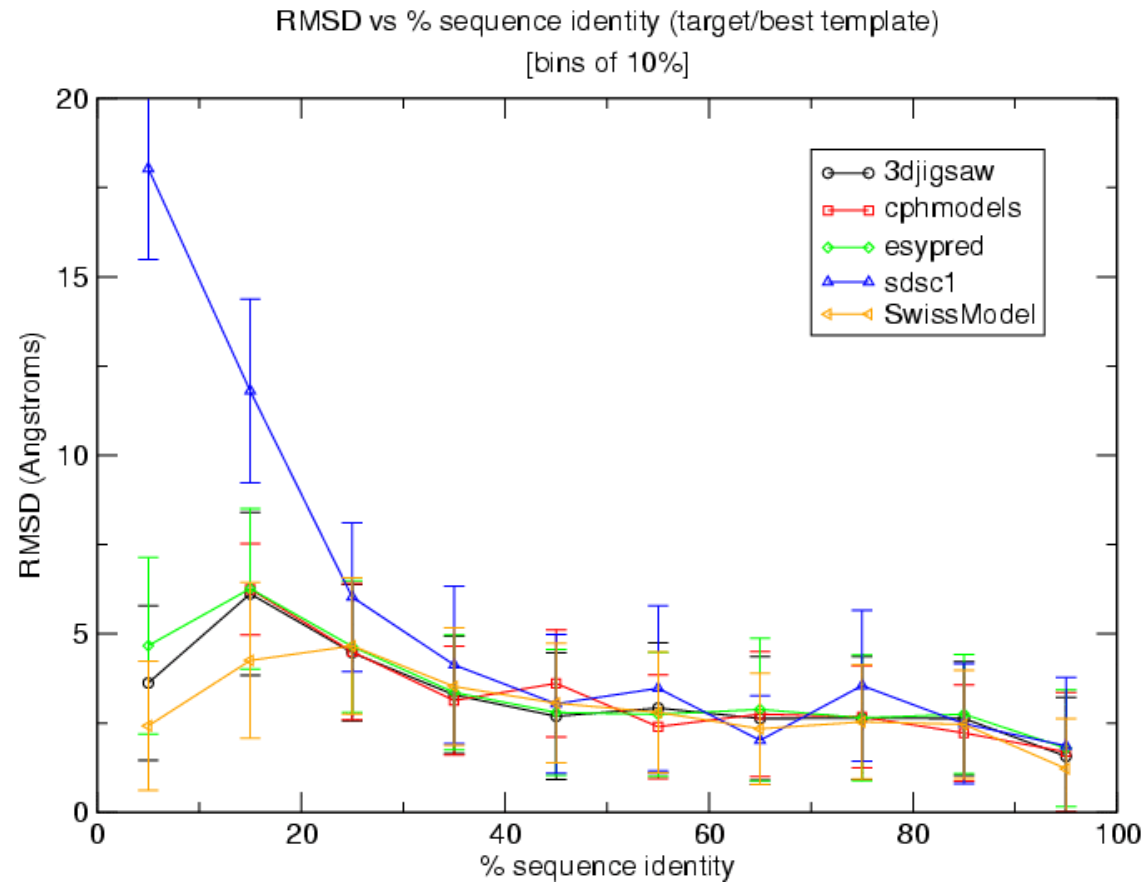
<http://salilab.org/modeller/>



# When to use Homology modelling?

- if there is enough similarity between target and template sequences

- CASP
  - at least
- 30% IDENTITY



# Threading

- or fold recognition
- tries to model onto common folds (not just one target) and tries to find out which one is the best



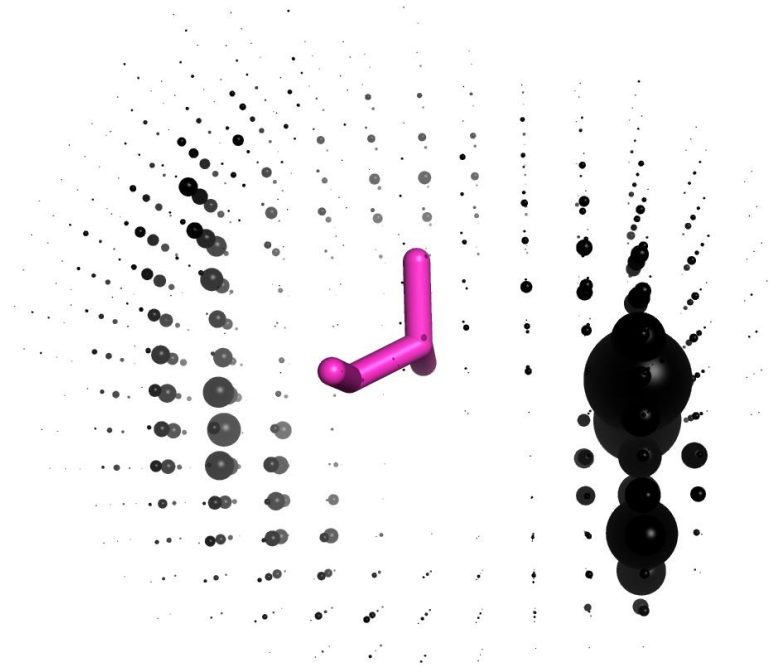
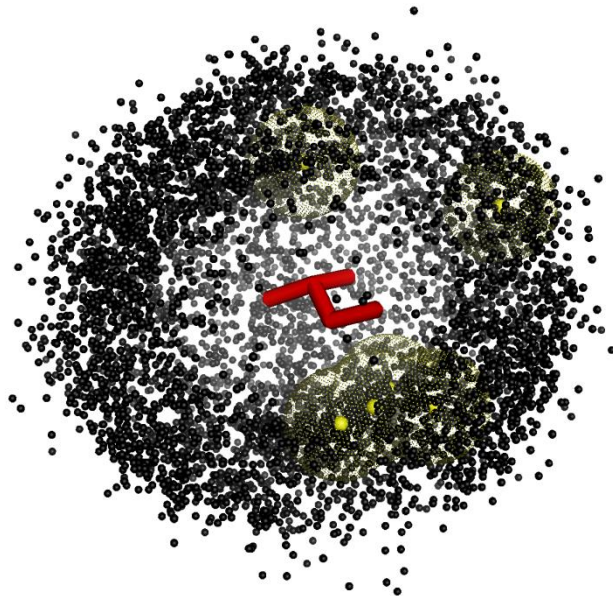
**I-TASSER ONLINE**  
Protein Structure & Function Predictions

- I-Tasser

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

# threading function

- energy-like function to find out amino acid preference for specific positions

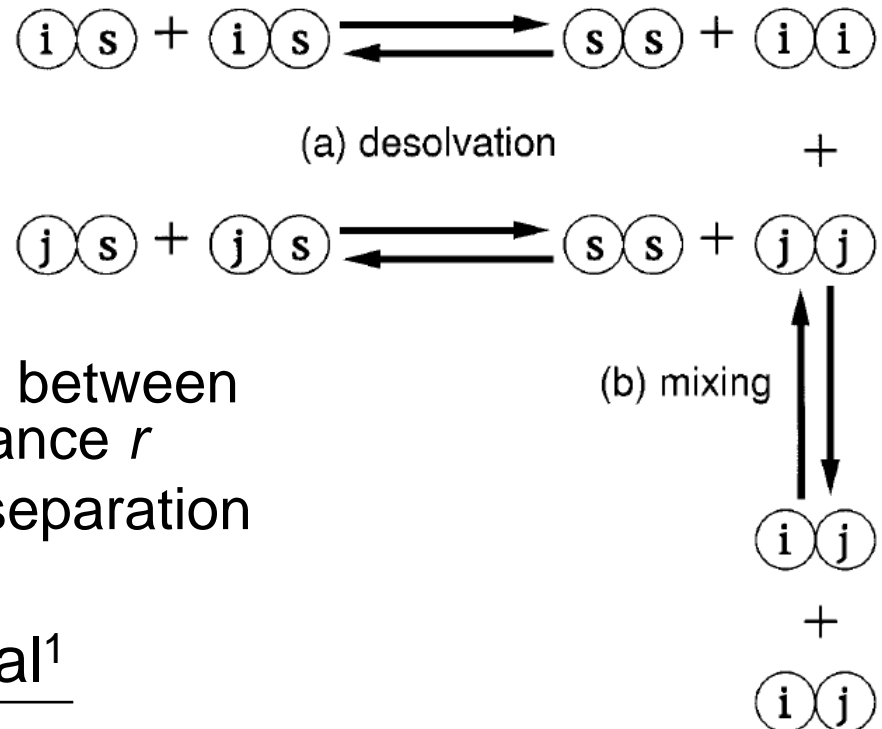


# threading function

- Boltzmann formula

$$w_{ij}(r) = -kT \ln \left( \frac{\rho_{ij}(r)}{\rho^*} \right)$$

- $w_{ij}$  – free energy
- $\rho_{ij}(r)$  – density of contacts between aminoacids  $i$  and  $j$  at distance  $r$
- $\rho^*$  - reference density at separation



- Phenomenological potential<sup>1</sup>

$$e_{ij} = A \cdot \ln \frac{n_{ij} \cdot n_{oo}}{n_{io} \cdot n_{jo}}$$

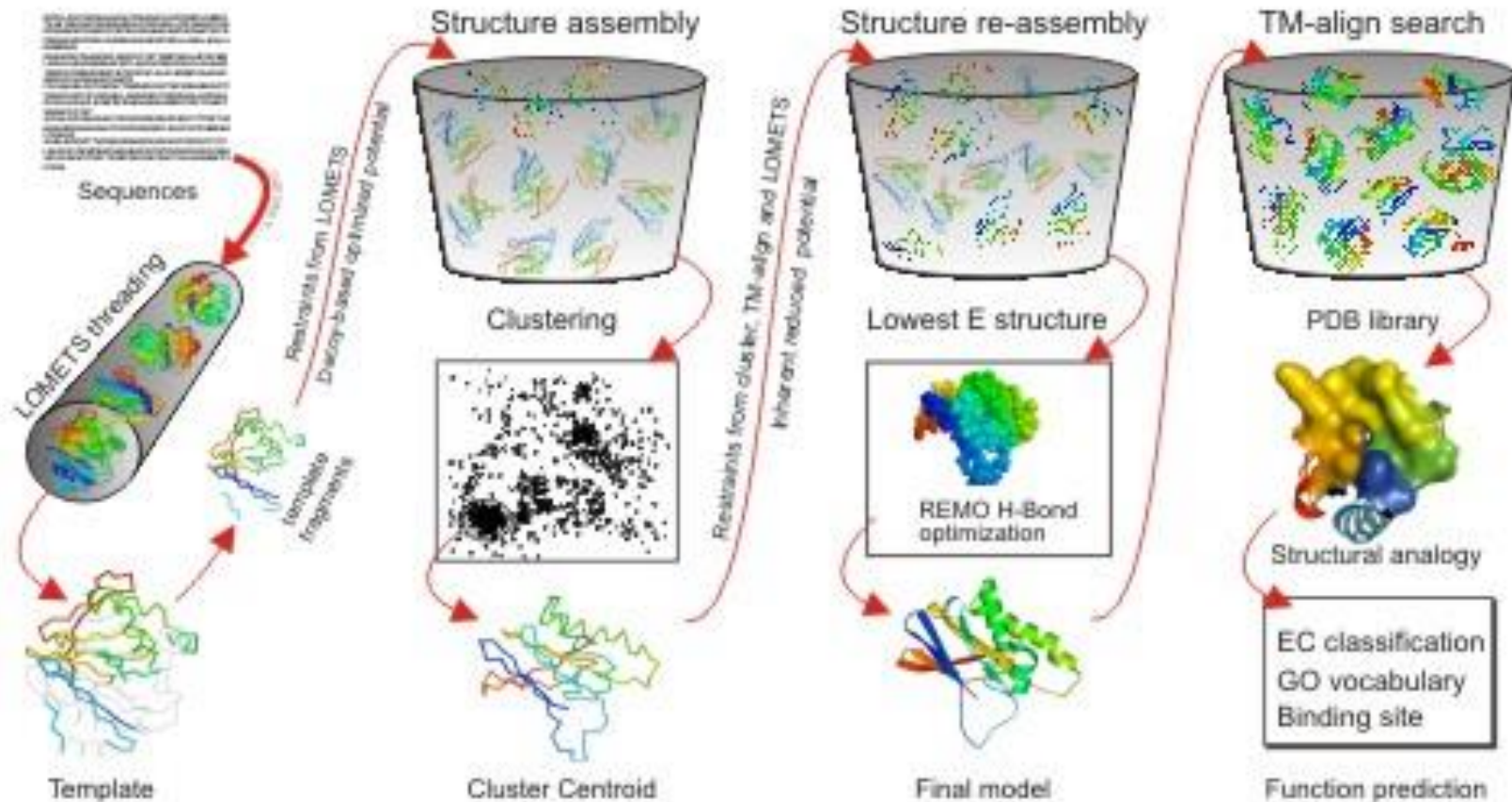
# Protein threading

- **DB of structural templates**
  - from [PDB](#), [FSSP](#), [SCOP](#), or [CATH](#), after **removing** protein structures with high sequence similarities.
- **Scoring function preparation**
  - measure the fitness between target sequences and templates based on the knowledge of the known relationships between the structures and the sequences.
  - A good scoring function should contain **mutation potential, environment fitness potential, pairwise potential, secondary structure compatibilities, and gap penalties**.
  - The quality of the energy function is closely related to the prediction accuracy, especially the alignment accuracy.
- **Threading alignment**
  - Align the target sequence with each of the structure templates by optimizing the designed scoring function.
  - solving the optimal alignment problem derived from a scoring function considering pairwise contacts.
- **Threading prediction**
  - statistically most probable alignment => threading prediction
  - construct a structure model for the target by placing the **backbone atoms** of the target sequence at their aligned backbone positions of the selected structural template.



# I-TASSER

- Best automated server for prediction of 3D structure



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



# I-TASSER

1. LOMETS
  - **metaserver** pro 8 methods of prediction for tertiary structure of fragments
2. use of fragments from identified templates
  - replica-exchange Monte Carlo simulations
  - threading of not homologous regions (loops)  
with ab initio modeling
3. SPICKER – clustering of best results
4. again LOMETS on individual clusters
5. TM-align - sequence-order independent protein structure alignment



# energy terms

[contact\\_cut.comm](#): Residue contact cutoff parameters

[contact\\_profile.comm](#): Side-chain contacts environment profile

[contact3.comm](#): Orientation-dependent side-chain contact potential

[CA13.comm](#): Short-range C-alpha correlation of (i,i+2)

[CA14.comm](#): Short-range C-alpha correlation of (i,i+3)

[CA15.comm](#): Short-range C-alpha correlation of (i,i+4)

[CA14s.comm](#): Short-range C-alpha correlation of (i,i+3) for strands

[CA14h.comm](#): Short-range C-alpha correlation of (i,i+3) for helices

[CA15s.comm](#): Short-range C-alpha correlation of (i,i+4) for strands

[CA15h.comm](#): Short-range C-alpha correlation of (i,i+4) for helices

[CB.comm](#): C-beta positions

[sidechain.comm](#): Sidechain center positions



# When to use threading?

- If there is not enough sequence identity to one template
- by individual domains
- fold recognition by consensus from several programs – meta methods
- use as much as experimental evidence as possible – to discern which fold is true

# Ab initio modeling

- ab initio = without template
- masive search for right conformation
- (pseudo-)physical energy function for free energy



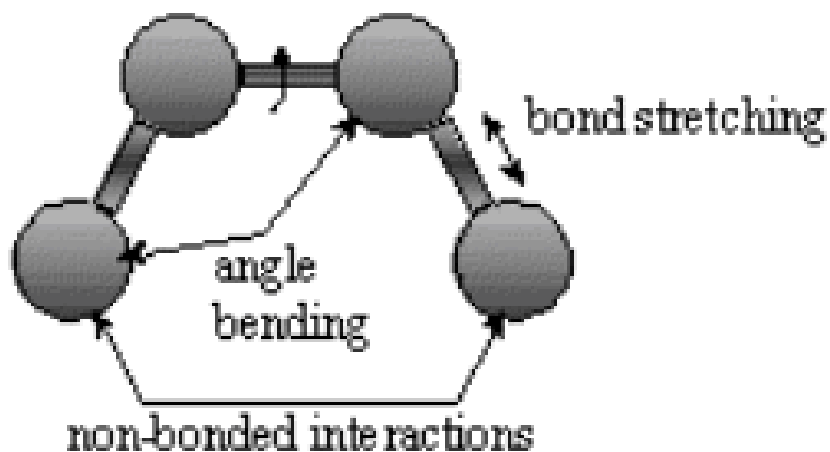
[www.bakerlab.org](http://www.bakerlab.org)

- <http://robetta.bakerlab.org/>
- *ab initio* and comparative models of protein domains
- The least precise, but the only one which can be use when no template is known

# Molecular mechanics

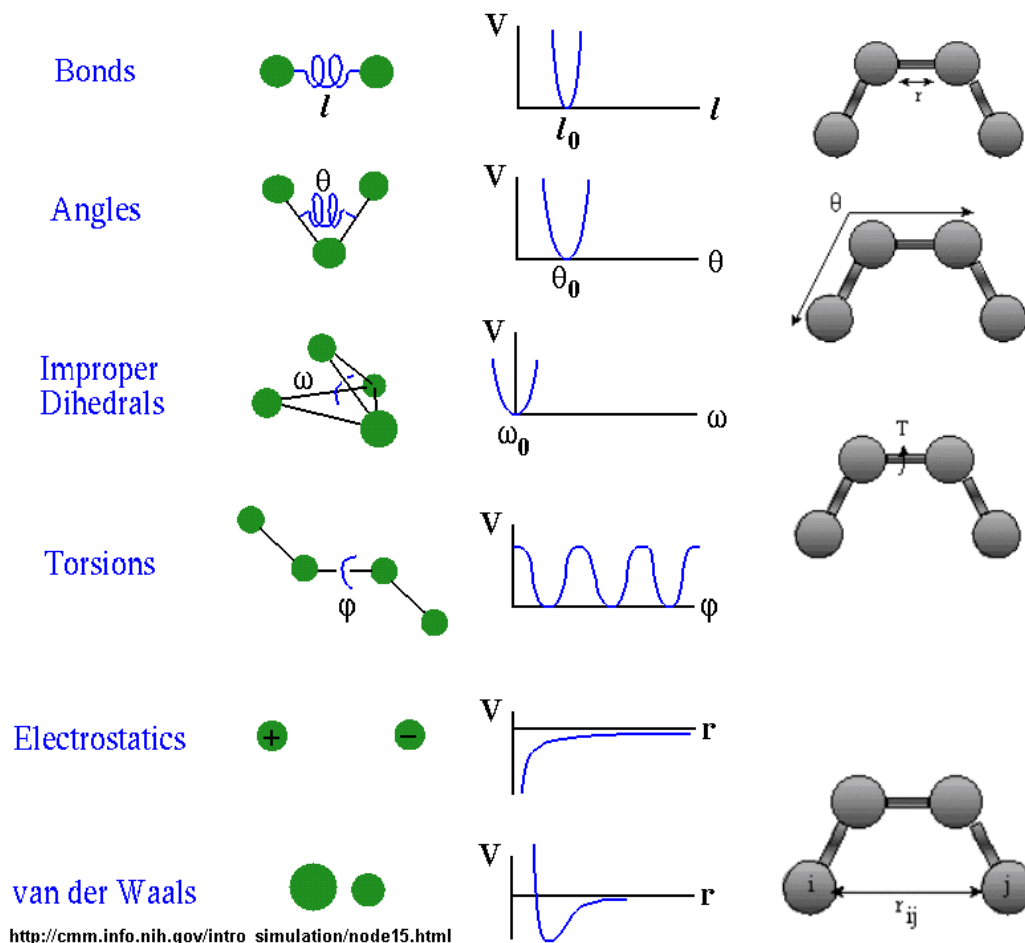
total energy is function of atom positions

$$E = f(\mathbf{R}) = E_b + E_a + E_t + E_c + E_{vdw}$$



# Force-field

## Empirical Potential Energy Function



[http://cmm.info.nih.gov/intro\\_simulation/node15.html](http://cmm.info.nih.gov/intro_simulation/node15.html)

$$E_b = \frac{k_r}{2} (r - r_0)^2$$

$$E_a = \frac{k_\theta}{2} (\theta - \theta_0)^2$$

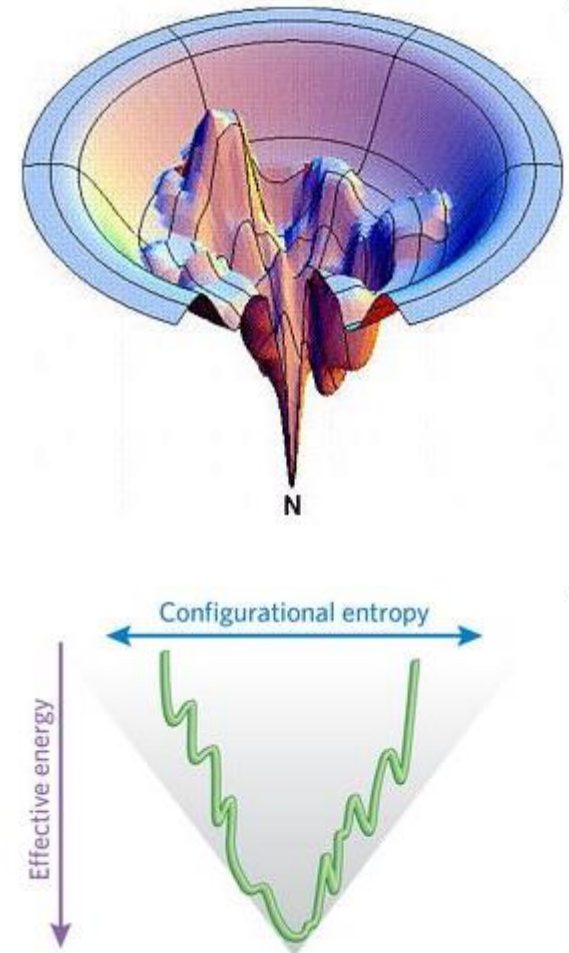
$$E_t = \frac{k_\phi}{2} (1 + \cos(n\phi - \phi_0))$$

$$E_c = \frac{1}{4\pi\epsilon_0\epsilon_r} \frac{q_i q_j}{r_{ij}}$$

$$E_{vdw} = -2\epsilon_{ij} \left( \frac{r_{ij}^*}{r_{ij}} \right)^6 + \epsilon_{ij} \left( \frac{r_{ij}^*}{r_{ij}} \right)^{12}$$

# Conformations

- Potential energy surface (PES)
  - barriers, minima
  - global vz. local
- folding -> movement over PES
  - folding funnel
  - methods:
    - energy optimization
    - molekular dynamics
    - simulated heating
    - metadynamics
    - Monte Carlo



# Use of MM/MD

- **protein folding**
  - search for global minima
  - too resource intensive
  - usable only for small proteins
  - + can be used to study process of folding and its kinetics
  - + model raffination

# Distributed computing projects

## Folding@Home

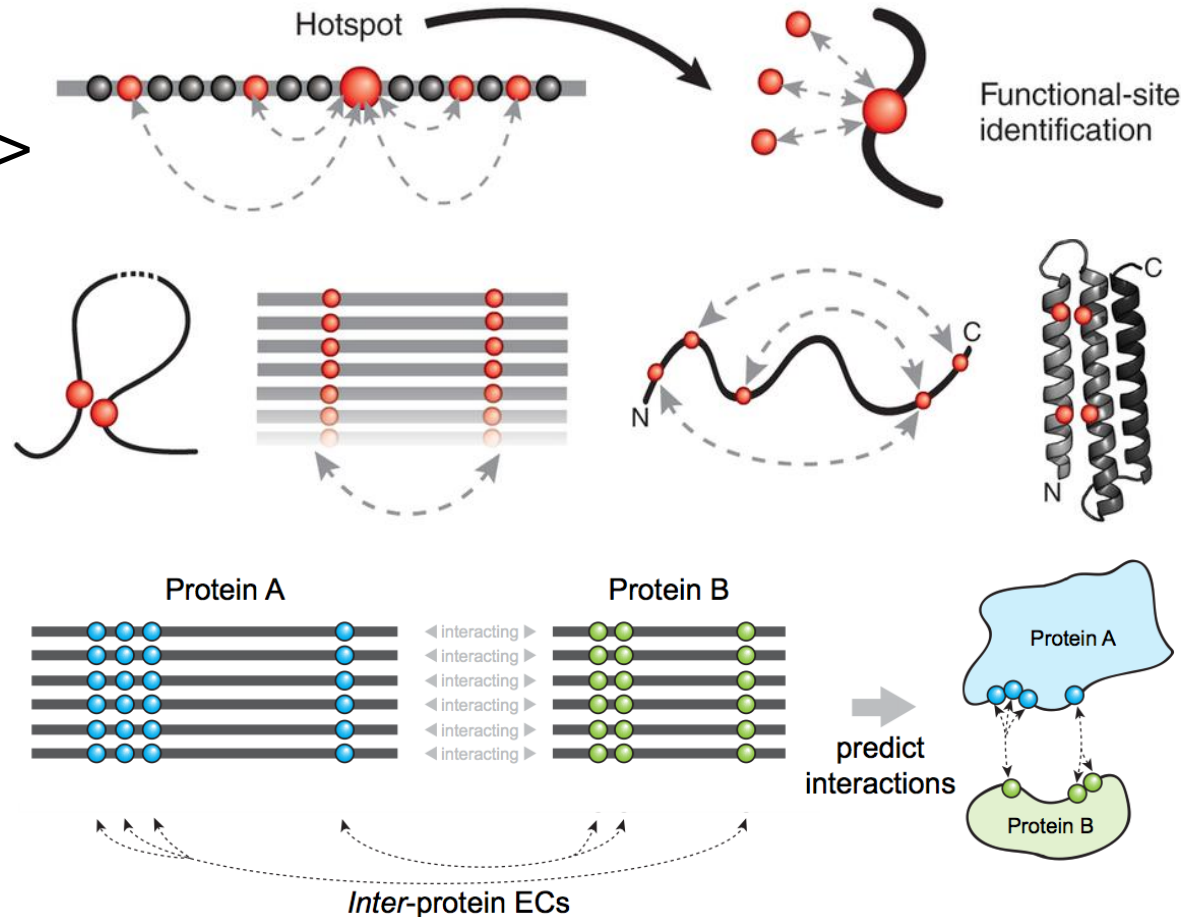
- <http://foldingathome.stanford.edu/>
- simulates protein folding, computational drug design, and other types of molecular dynamics
- determine the mechanisms of protein folding
- Pande Lab at Stanford U
- 100 petaFLOPS on May 11, 2016

## Rosetta@Home

- <http://boinc.bakerlab.org/>
- protein structure prediction
- on the Berkeley Open Infrastructure for Network Computing (BOINC)
- Baker lab at UWashingon
- predict protein–protein docking and design new proteins with the help of
- ~60 000 active computers over 210 teraFLOPS on average as of July 29, 2016

# Evolutionary Couplings

- EVcouplings
  - multiple alignment -> functional site
- EVfold
  - Folds the protein when unknown structure
- EVcomplex
  - Finds and joins partner sequences from the two MSAs



<http://evfold.org/>



# **ALPHAFOLD2**

under the hood

# Průlom v biologii. Umělá inteligence „vyřešila“ šmodrchání proteinů na 92 %

NEWS • 30 NOVEMBER 2020

**‘It will change everything’: DeepMind’s AI makes gigantic leap in solving protein structures**

**Umělá inteligence AlphaFold dosáhla vědeckého průlomu. Dovede stanovit tvar molekul proteinů**

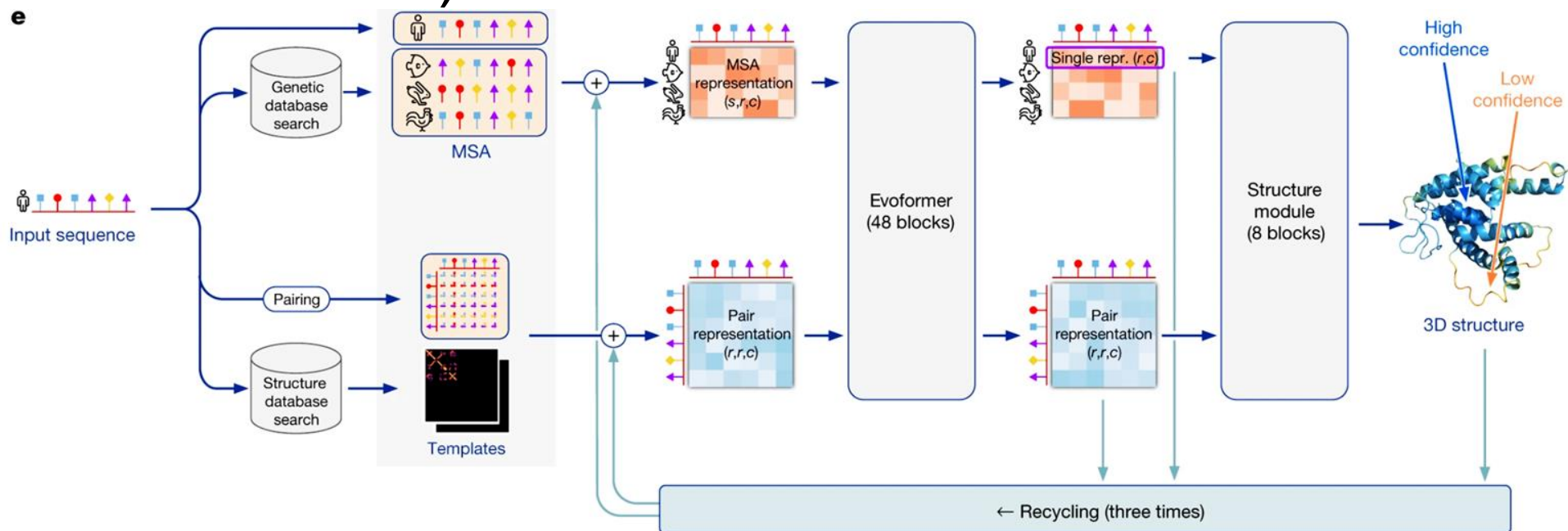
‘The game has changed.’ AI triumphs at solving protein structures

# AlphaFold2

**Input:** sequence

extended by MSA + structural templates

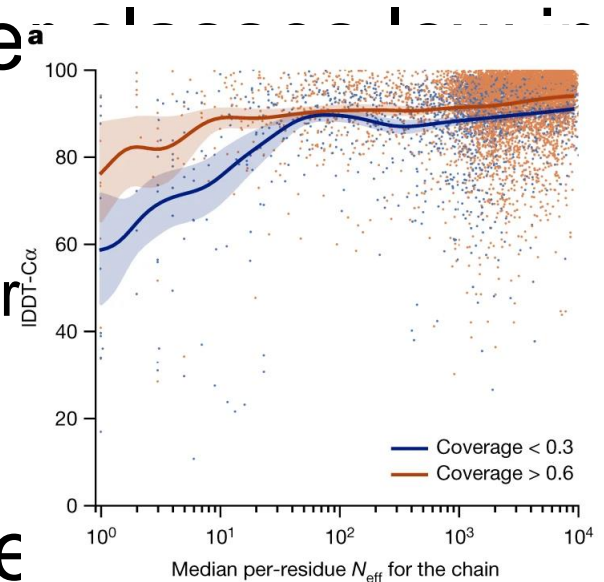
Evoformer and Structure model (w MD simulation)



# MSA - multiple sequence alignment

using standard tools - jackhammer, HHBlits

- sequence DBs:
  - *UniRef90*
    - reference sequences from UniProt
  - *UniClust30*
    - for sequence self-distillation
- metagenomicsDBs - fully covered  
UniRef90
  - *Big Fantastic database (BFD)*
    - 66M protein families from 2.2G proteomes
  - clustered *MGnify*
    - more than 260k genomes



<https://www.nature.com/articles/s41586-021-03819-2> needed at least 33 sequences per

# Training

PDB database + PDB70 clusters

training db:

40% identity clusters, crop to 258 residues,  
batches by 128 per Tensor processing unit (TPU)

enhance accuracy by noisy student self-distillation

predict 350000 structures from UniRef30 using  
trained network

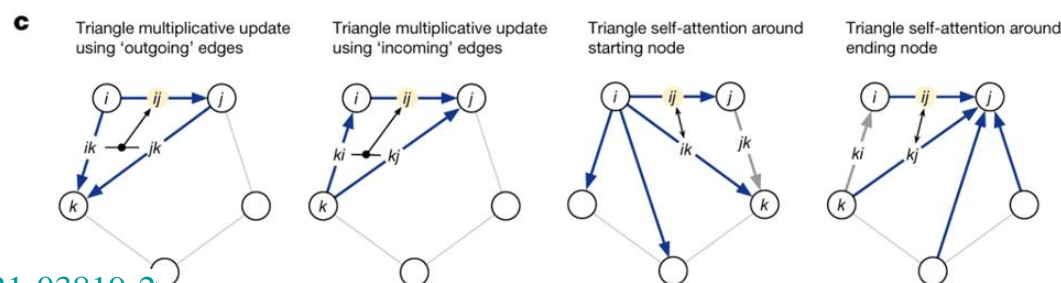
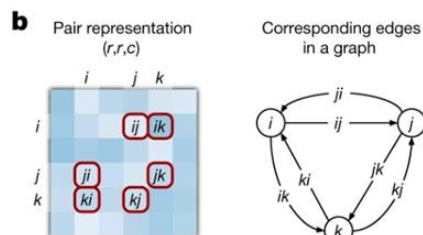
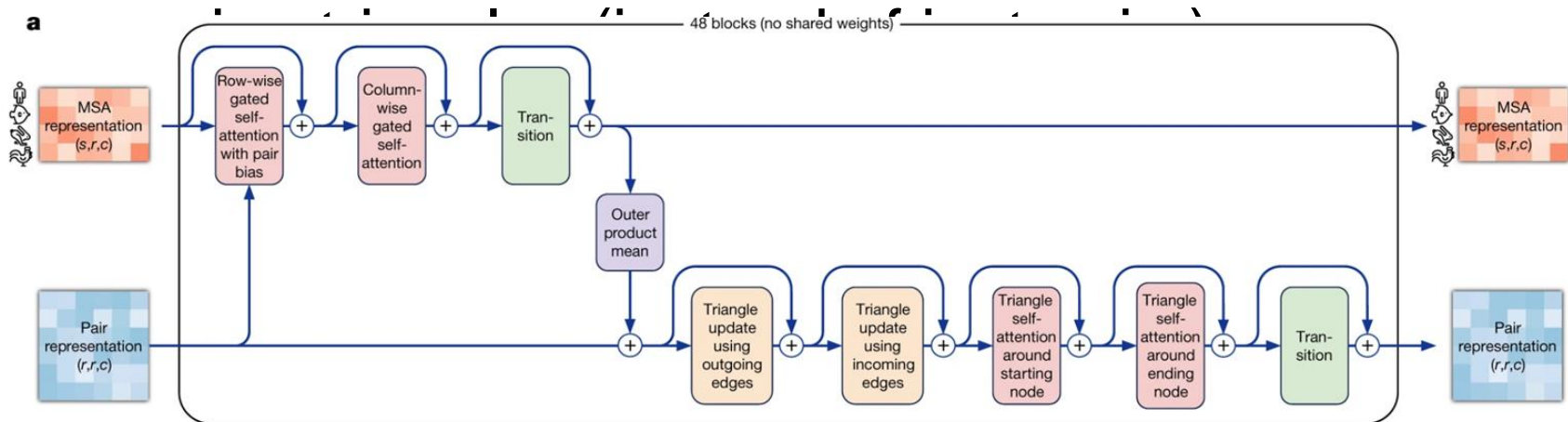
filter to high confidence subset

then train again from scratch with mixture of  
PDB and UniRef30

=> effective use of unlabelled sequence data

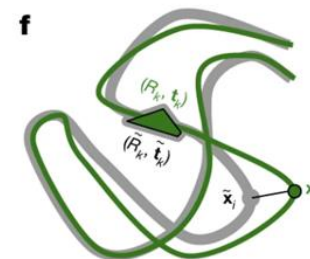
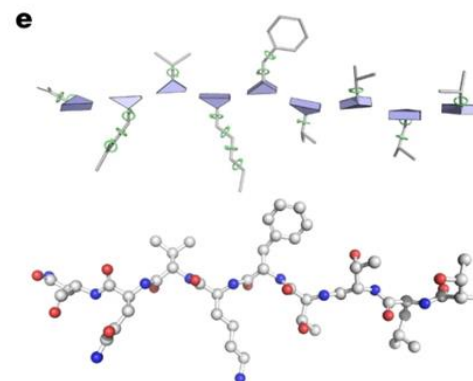
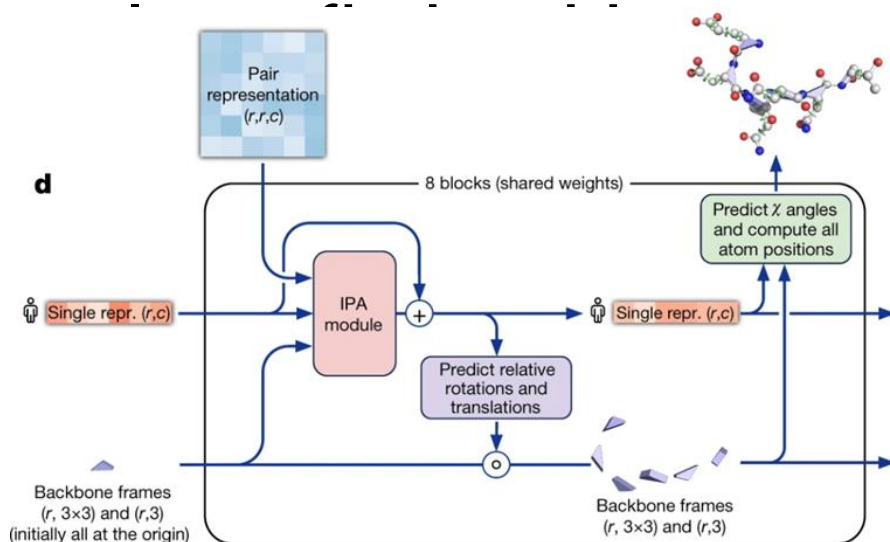
# EvoFormer

- mixing MSA and pairs via updates
- graph inference problem in 3D space
  - edges = residues in proximity
  - updates per each block (48 blocks) separately (AF1 updated all network at once)



# Structure model

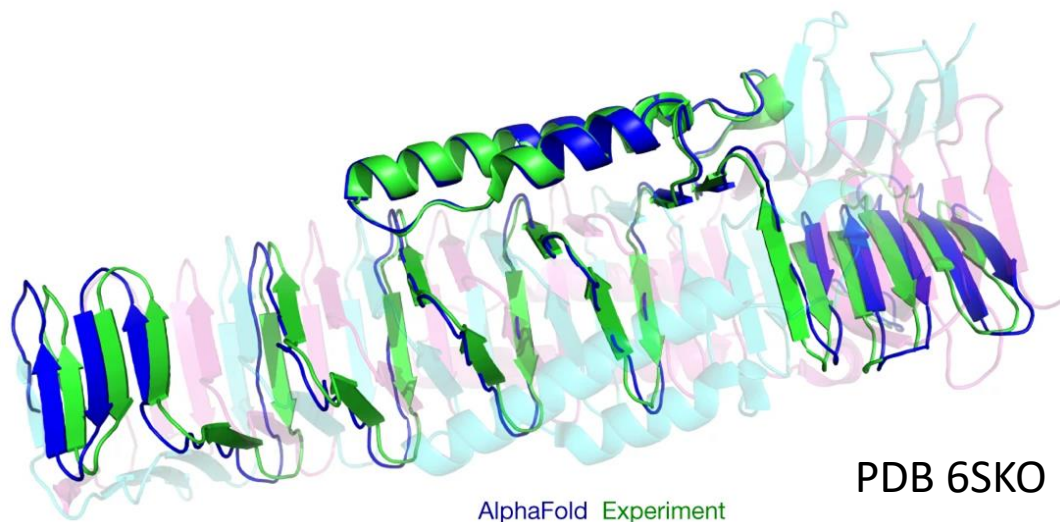
- prioritize backbone positions+orientations
  - residue gas - free floating rigid body rotations and translation
  - updates
    - IPA (invariant point attention) - neural activations only in rigid 3D
    - equivariant update using updated activations



# Effect of cross-chain contacts.

prediction is worse for heterotropic contacts (large complexes where 3D structure is dictated by other chains in complex)

homotropic . . . . .  
when ch





# Timings

one GPU minute per model with 384 residues

=> allows proteome-scale studies

1500 residues trimer (SARS-CoV2 S protein) - about a day on ELIXIR CZ Metacentrum pipeline

**ALPHAFOLDDDB**

# AlphaFold Protein Structure Database

Developed by DeepMind and EMBL-EBI

Search for protein, gene, UniProt accession or organism

BETA





















Search

Examples: [Free fatty acid receptor 2](#) [At1g58602](#) [Q5VSL9](#) [E. coli](#) [Help: AlphaFold DB search help](#)

AlphaFold DB provides open access to protein structure predictions for the human proteome and 20 other key organisms to accelerate scientific research.

<https://www.alphafold.ebi.ac.uk/>

# Complete structures of 20 model organisms

Species	Common Name	Reference Proteome	Predicted Structures	Download
<i>Arabidopsis thaliana</i>	<i>Arabidopsis</i>	<a href="#">UP000006548</a> 	27,434	<a href="#">Download (3642 MB)</a>
<i>Caenorhabditis elegans</i>	Nematode worm	<a href="#">UP000001940</a> 	19,694	<a href="#">Download (2601 MB)</a>
<i>Candida albicans</i>	<i>C. albicans</i>	<a href="#">UP000000559</a> 	5,974	<a href="#">Download (965 MB)</a>
<i>Danio rerio</i>	Zebrafish	<a href="#">UP000000437</a> 	24,664	<a href="#">Download (4141 MB)</a>
<i>Dictyostelium discoideum</i>	<i>Dictyostelium</i>	<a href="#">UP000002195</a> 	12,622	<a href="#">Download (2150 MB)</a>
<i>Drosophila melanogaster</i>	Fruit fly	<a href="#">UP000000803</a> 	13,458	<a href="#">Download (2174 MB)</a>
<i>Escherichia coli</i>	<i>E. coli</i>	<a href="#">UP000000625</a> 	4,363	<a href="#">Download (448 MB)</a>
<i>Glycine max</i>	Soybean	<a href="#">UP000008827</a> 	55,799	<a href="#">Download (7142 MB)</a>
<i>Homo sapiens</i>	Human	<a href="#">UP000005640</a> 	23,391	<a href="#">Download (4784 MB)</a>
<i>Leishmania infantum</i>	<i>L. infantum</i>	<a href="#">UP000008153</a> 	7,924	<a href="#">Download (1481 MB)</a>
<i>Methanocaldococcus jannaschii</i>	<i>M. jannaschii</i>	<a href="#">UP000000805</a> 	1,773	<a href="#">Download (171 MB)</a>
<i>Mus musculus</i>	Mouse	<a href="#">UP000000589</a> 	21,615	<a href="#">Download (3547 MB)</a>
<i>Mycobacterium tuberculosis</i>	<i>M. tuberculosis</i>	<a href="#">UP000001584</a> 	3,988	<a href="#">Download (421 MB)</a>
<i>Oryza sativa</i>	Asian rice	<a href="#">UP000059680</a> 	43,649	<a href="#">Download (4416 MB)</a>
<i>Plasmodium falciparum</i>	<i>P. falciparum</i>	<a href="#">UP000001450</a> 	5,187	<a href="#">Download (1132 MB)</a>
<i>Rattus norvegicus</i>	Rat	<a href="#">UP000002494</a> 	21,272	<a href="#">Download (3404 MB)</a>
<i>Saccharomyces cerevisiae</i>	Budding yeast	<a href="#">UP000002311</a> 	6,040	<a href="#">Download (960 MB)</a>
<i>Schizosaccharomyces pombe</i>	Fission yeast	<a href="#">UP000002485</a> 	5,128	<a href="#">Download (776 MB)</a>
<i>Staphylococcus aureus</i>	<i>S. aureus</i>	<a href="#">UP000008816</a> 	2,888	<a href="#">Download (268 MB)</a>
<i>Trypanosoma cruzi</i>	<i>T. cruzi</i>	<a href="#">UP000002296</a> 	19,036	<a href="#">Download (2905 MB)</a>

# SNW domain-containing protein 1

AlphaFold structure prediction

Download

PDB file

mmCIF file

Predicted aligned error

## Information

Protein	SNW domain-containing protein 1
Gene	SNW1
Source organism	Homo sapiens <a href="#">go to search</a>
UniProt	Q13573 <a href="#">go to UniProt</a>
Experimental structures	17 structures in PDB for Q13573 <a href="#">go to PDBe-KB</a>
Biological function	(Microbial infection) Proposed to be involved in transcriptional activation by EBV EBNA2 of CBF-1/RBPJ-repressed promoters. <a href="#">go to UniProt</a>

## 3D viewer

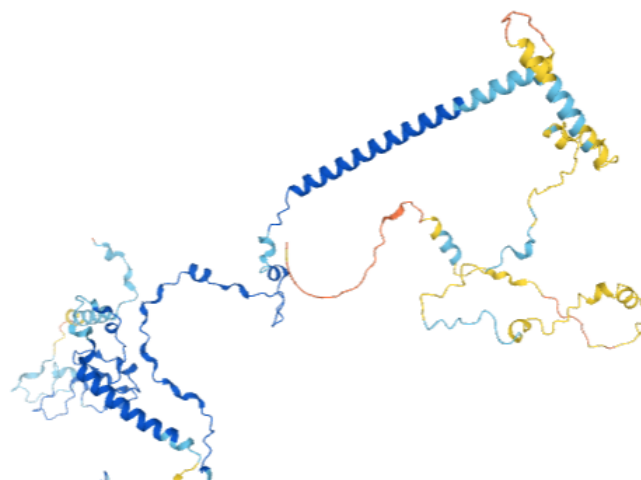
### Model Confidence:

- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

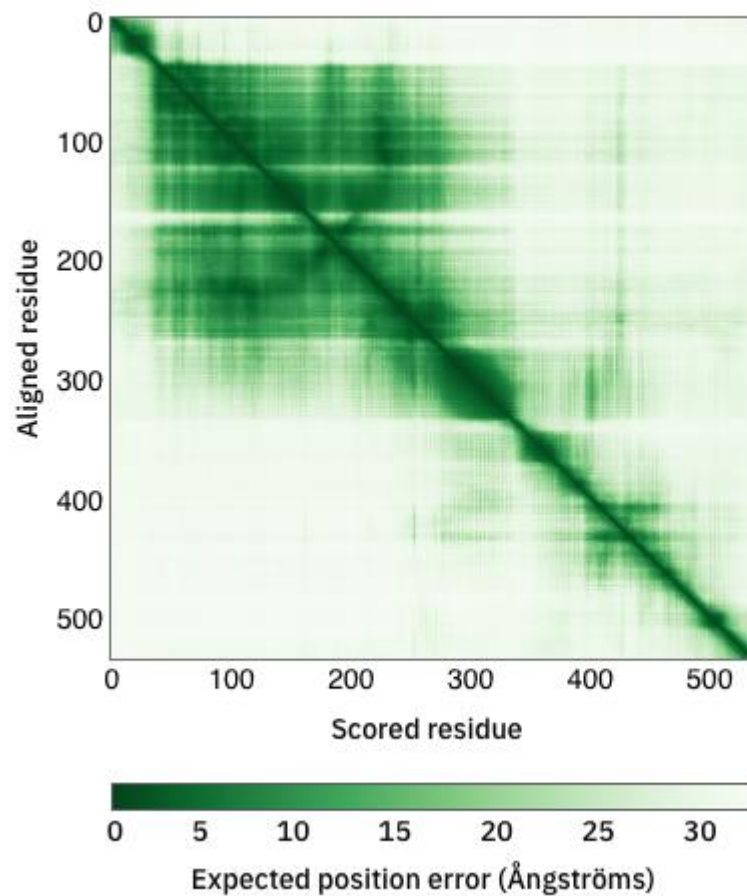
AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions below 50 pLDDT may be unstructured in isolation.

Sequence of AF-Q13573-... 1: SNW do... A

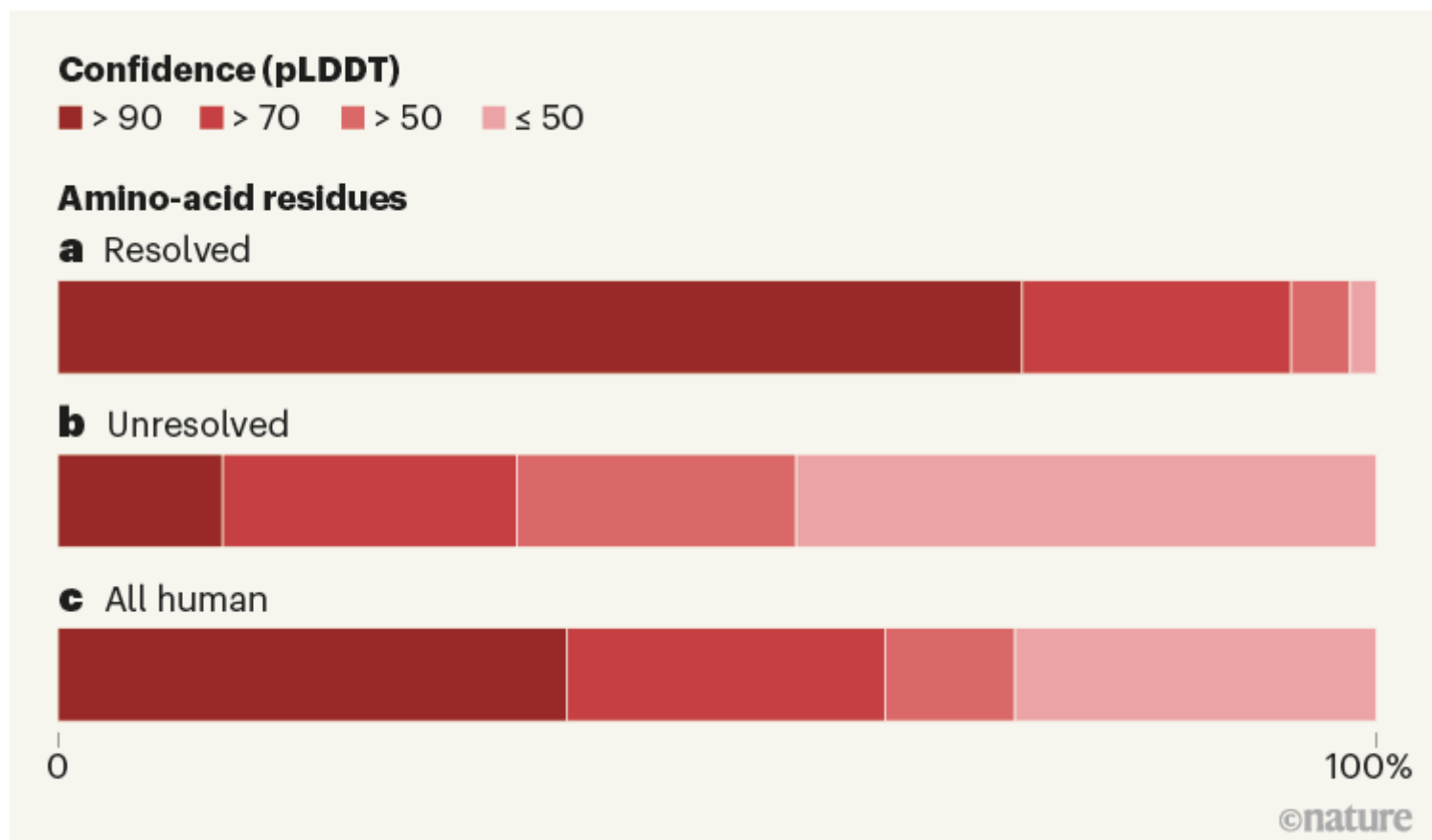
1	11	21	31	41	51	61	71	81	91	101	111	121
M	A	L	T	S	F	L	P	A	P	T	Q	L
S	D	Q	L	E	A	E	E	K	A	R	S	R
S	R	Q	T	S	L	V	S	R	R	E	P	P
P	P	P	Y	G	Y	R	K	G	W	I	P	R
L	L	E	D	F	G	D	G	A	F	E	I	H
V	A	Q	Y	P	L	D	M	G	R	K	K	M
S	N	A	L	I	Q	V	D	S	E	G	K	I
K	Y	D	A	I	A	R	Q	G	S	K	D	K
V	I	S	K	Y	T	D	L	V	P	K	E	V
M	N	A	D	D	P	L	Q	R	P	D	E	E
A	I	K	E	I	T	E	K	T	R	V	A	L
E	K	S	V	S	Q	K	V	A	A	M	P	V
R	A	A	D	K	L	A	P	A	Q	Y	I	R
T	P	S	Q	Q	G	V	A	F	N	S	G	A
K	Q	R	V	I	R	M	V	E	M	Q	K	D
P	M	E	P	P	R	F	K	I	N	K	K	I
P	R	G	P	P	S	P	P	A	P	V	M	H
S	P	S	R	K	M	T	V	K	E	Q	Q	E
W	K	I	P	C	I	S	N	W	K	N	A	K
G	Y	T	I	P	L	D	K	R	L	A	A	D
G	R	G	L	T	V	H	I	N	E	N	F	A
K	L	A	E	A	L	Y	I	A	D	R	K	A
E	A	V	E	M	R	A	Q	V	E	R	K	M
A	Q	K	E	K	E	K	H	E	E	K	L	R
E	M	A	Q	K	A	R	E	R	R	A	G	I
K	T	H	V	E	K	D	G	E	A	R	E	D
E	I	R	H	D	R	R	K	E	R	Q	H	D
N	L	S	R	A								



# AlphaFold tells you where is it right!



# How good are the predictions of human proteins?



**pLDDT** - per-residue estimate of its confidence on a scale from 0 - 100 model's predicted score on the [IDDT-C \$\alpha\$  metric](#) (local superposition-free score for comparing protein structures and models using distance difference tests).

# USAGES



# AlphaFold in Google Colab

Github enabled  
JupyterNotebooks  
running in Google Colab  
environment

limitation in size



Repozitář: [\[link\]](#)

sokrypton/ColabFold

Větev: [\[link\]](#)

main

Cesta



AlphaFold2.ipynb



AlphaFold2\_complexes.ipynb



RoseTTAFold.ipynb



batch/AlphaFold2\_batch.ipynb

[Mirdita M, Ovchinnikov S, Steinegger M. ColabFold - Making protein folding accessible to all. bioRxiv, 2021](#)

<https://colab.research.google.com/github/sokrypton/ColabFold/>

# Alphafold on ELIXIR CZ

- Alphafold needs good GPU/TPU to run  
-> not many people have it on their PC
- Alphafold has been installed on ELIXIR CZ hardware
  - `/storage/brno11-elixir/projects/alphafold`
- Elixir is accessible through Metacentrum
  - <https://wiki.metacentrum.cz/wiki/AlphaFold>
- speed is dependent on size of predicted protein but can be in order of tens of minutes

# Alphafold is just a start...

- use Alphafold ideas for development of their own 3D structure predictions
  - RoseTTAfold
- prediction of designed proteins
- prediction of RNA structures
- prediction of orphan proteins
  - molecular replacement
  - interpretation of cryoEM
  - pLDDT can act as IDP predictor
- ...

Search worldwide, life-sciences

alphafold

[Coronavirus articles and preprints](#) Search

[Recent history](#)

[Saved searches](#)

## Search only

### Type ?

☐ Research articles (111)

☐ Reviews (88)

☐ Preprints (38)

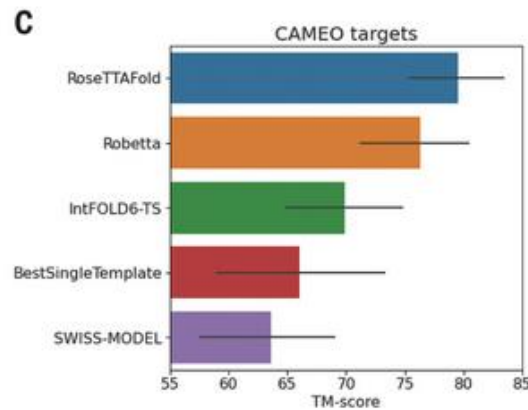
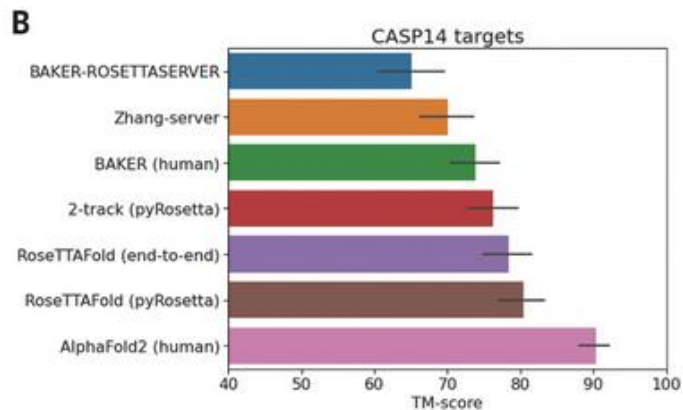
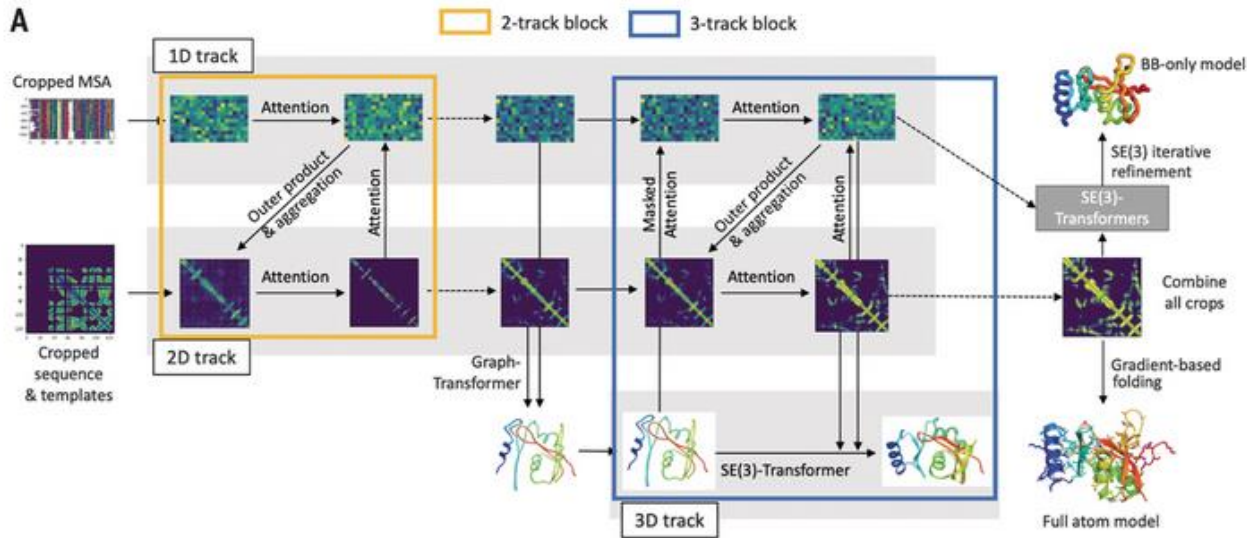
### Free full text ?

☐ Free to read (201)

☐ Free to read & use (183)

as of

# Accurate prediction of protein structures and interactions using a three-track neural network



---

## USING ALPHAFOLD FOR RAPID AND ACCURATE FIXED BACKBONE PROTEIN DESIGN

---

🌱 **Lewis Moffat**

Department of Computer Science  
University College London  
Gower St, London WC1E 6BT  
[lewis.moffat@cs.ucl.ac.uk](mailto:lewis.moffat@cs.ucl.ac.uk)

🌱 **Joe G. Greener**

Department of Computer Science  
University College London  
Gower St, London WC1E 6BT  
[j.greener@ucl.ac.uk](mailto:j.greener@ucl.ac.uk)

🌱 **David T. Jones\***

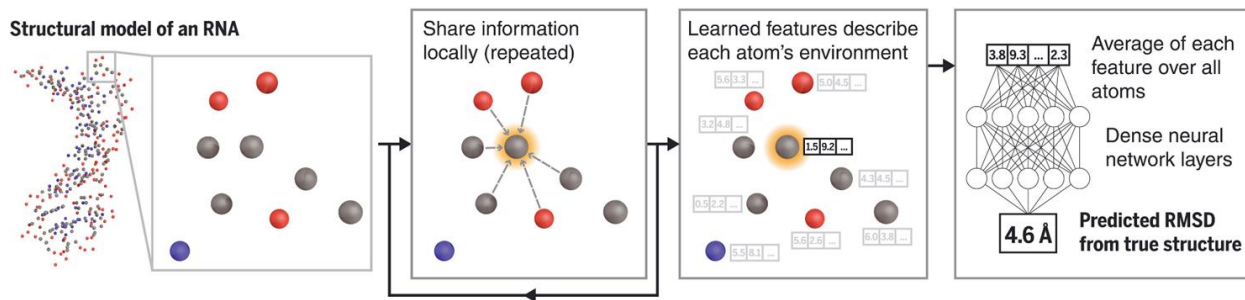
Department of Computer Science  
University College London  
Gower St, London WC1E 6BT  
[d.t.jones@ucl.ac.uk](mailto:d.t.jones@ucl.ac.uk)

### ABSTRACT

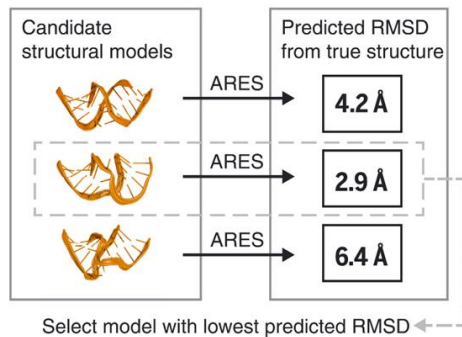
The prediction of protein structure and the design of novel protein sequences and structures have long been intertwined. The recently released AlphaFold has heralded a new generation of accurate protein structure prediction, but the extent to which this affects protein design stands yet unexplored. Here we develop a rapid and effective approach for fixed backbone computational protein design, leveraging the predictive power of AlphaFold. For several designs we demonstrate that not only are the AlphaFold predicted structures in agreement with the desired backbones, but they are also supported by the structure predictions of other supervised methods as well as *ab initio* folding. These results suggest that AlphaFold, and methods like it, are able to facilitate the development of a new range of novel and accurate protein design methodologies.

# Geometric deep learning of RNA structure

**A** ARES predicts the accuracy of a structural model, given only atomic coordinates and element types



**B** RNA structure prediction with ARES



**C** Training set: 18 older, smaller RNA structures

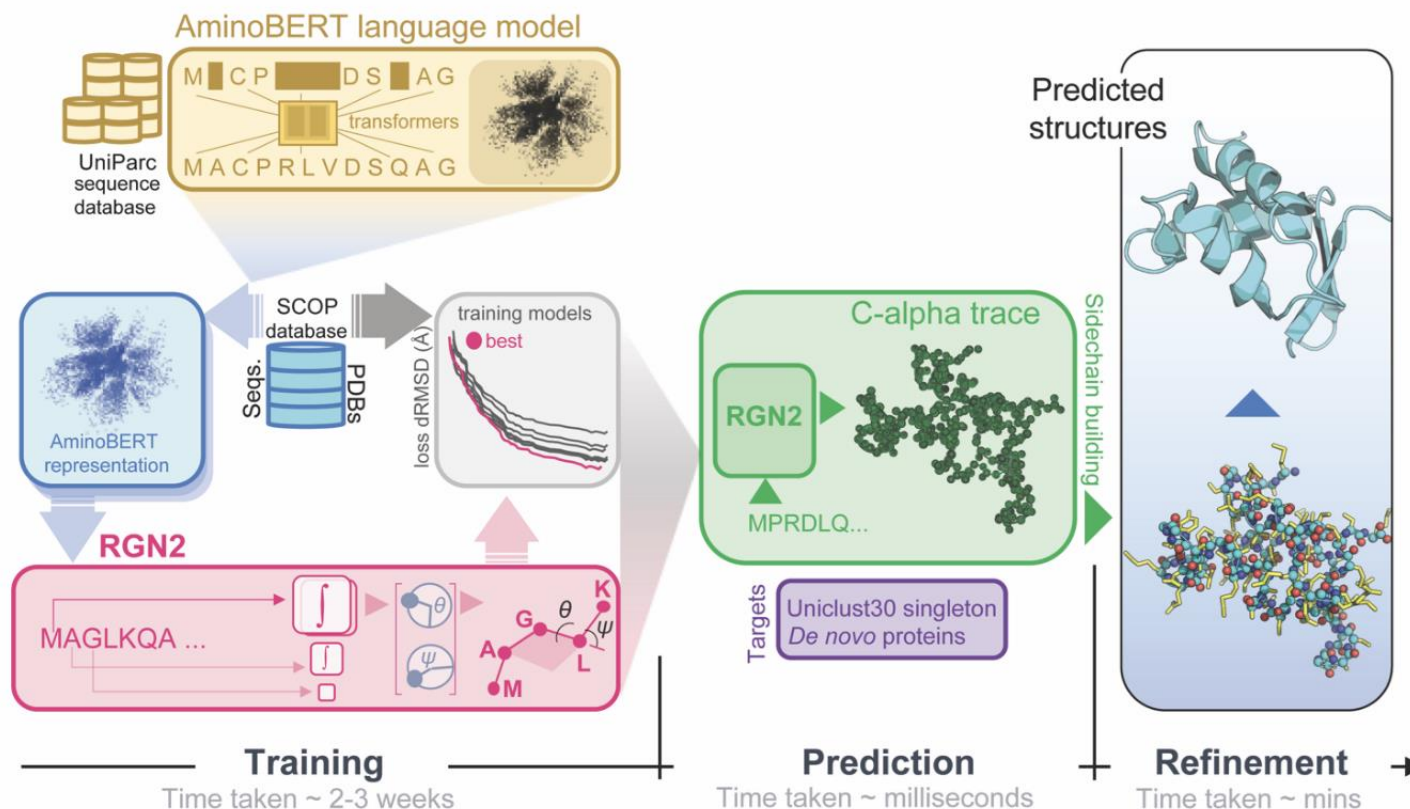


**D** Benchmark sets: newer, larger RNA structures



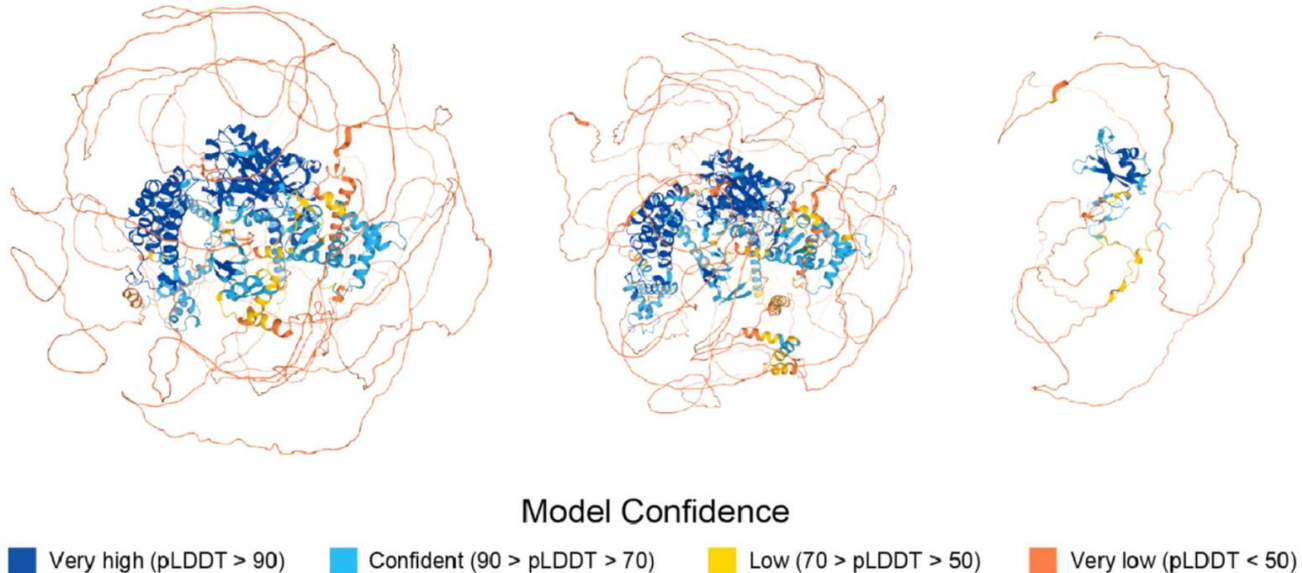
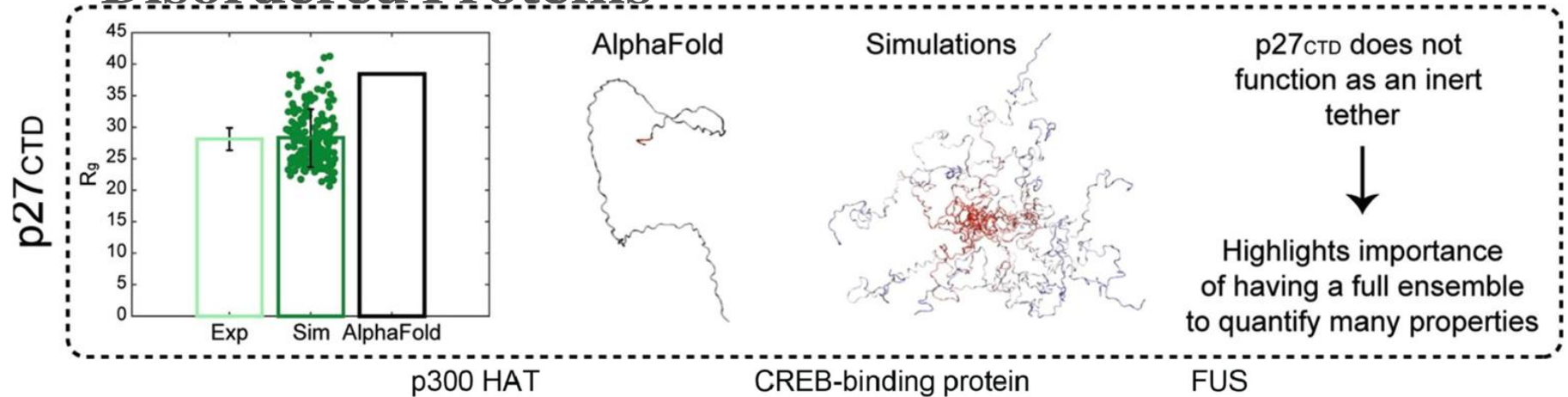


# Single-sequence protein structure prediction using language models from deep learning



**Figure 1. Organization and application of RGN2.** RGN2 combines a Transformer-based protein language model (AminoBERT) with a recurrent geometric network that utilizes Frenet-Serret frames to generate the backbone structure of a protein. Placement of side chain atoms and refinement of hydrogen-bonded networks are subsequently performed using the Rosetta energy function.

# AlphaFold and Implications for Intrinsically Disordered Proteins





# AlphaFold in MobiDB



# MrParse: Finding homologues in the PDB and the EBI AlphaFold database for Molecular Replacement and more



## MrParse Analysis

Version: 0.2.1

MrParse: a program to find and analyse search models for crystallographic Molecular Replacement. The program is being developed by [Dan Rigden's group](#) at the University of Liverpool.

MrParse is currently under development and we are keen to make it as useful to the community as possible. If you have any suggestions for it's development, or ideas on how we could improve it, please [get in touch](#).

## IKL Info

Name	Resolution	Space Group	Has NCS?	Has Twinning?	Has Anisotropy?
<a href="#">7dry-sf</a>	1.44	P41212	false	false	true

## Experimental structures from the PDB

Name	PDB	Resolution	Region	Range	Length	eLLG	Mol. Wt.	eRMSD	Seq. Ident.
<a href="#">2cvi_B_1</a>	<a href="#">2cvi</a>	1.50	1	158-230	71	43.5	8676	1.085	0.31

## Structure predictions from the EBI AlphaFold database

Name	model	Date Made	Region	Range	Length	Avg. pLDDT	H-score	Seq. Ident.
<a href="#">Q12362.1</a>	<a href="#">Q12362</a>	01-JUL-21	1	2-180	177	90.15	85	0.41
<a href="#">P87241.1</a>	<a href="#">P87241</a>	01-JUL-21	1	4-176	171	91.55	85	0.38

## Visualisation of Regions

## Sequence Based Predictions



## Visualisation of Regions



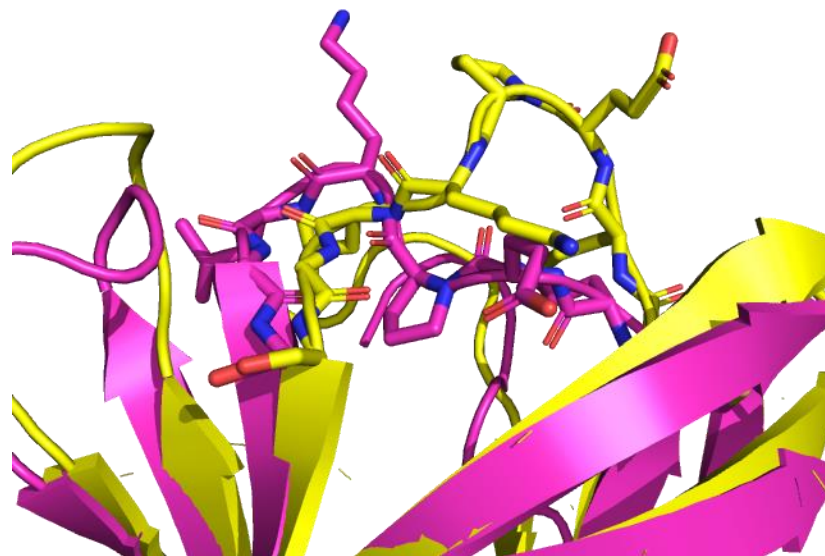
# **LIMITATIONS**

# Are structural biologists and bioinformaticians on the job market?

- Alphafold can not do **multi-protein complexes** – interactions
- Alphafold can not do **point mutations** - design of functions
- Alphafold can not do **conformational changes or dynamics**
- Alphafold can not do effects of **post-translational protein modifications**
- Alphafold can not do **ligand effects**
- Alphafold is not good with **orphan sequences**
- Alphafold does not tell much about **folding process**

# Are the models good enough for drug design?

- we do not know yet
- average RMSD for AlphaFold2 models is 1.3 Å
- average RMSD of X-Ray structures is 0.3 - 0.5 Å
- best AlphaFold prediction has RMSD 0.6 Å
- locally AlphaFold2 might be there



T1064

# Summary

- Alphafold2 made a huge leap in **prediction accuracy**
- Role of **open science** and publicly available data can not be overstated
- CASP competition was a driver of the change
- Alphafold2 is **publicly available** and can be run from many places including ELIXIR CZ
- Alphafold has inspired many tools already
- Alphafold limits are yet to be fully described

# Quality Control

CAMEO, CASP

CAPRI

CAFA

# Quality control of protein models

## - CASP, CAMEO

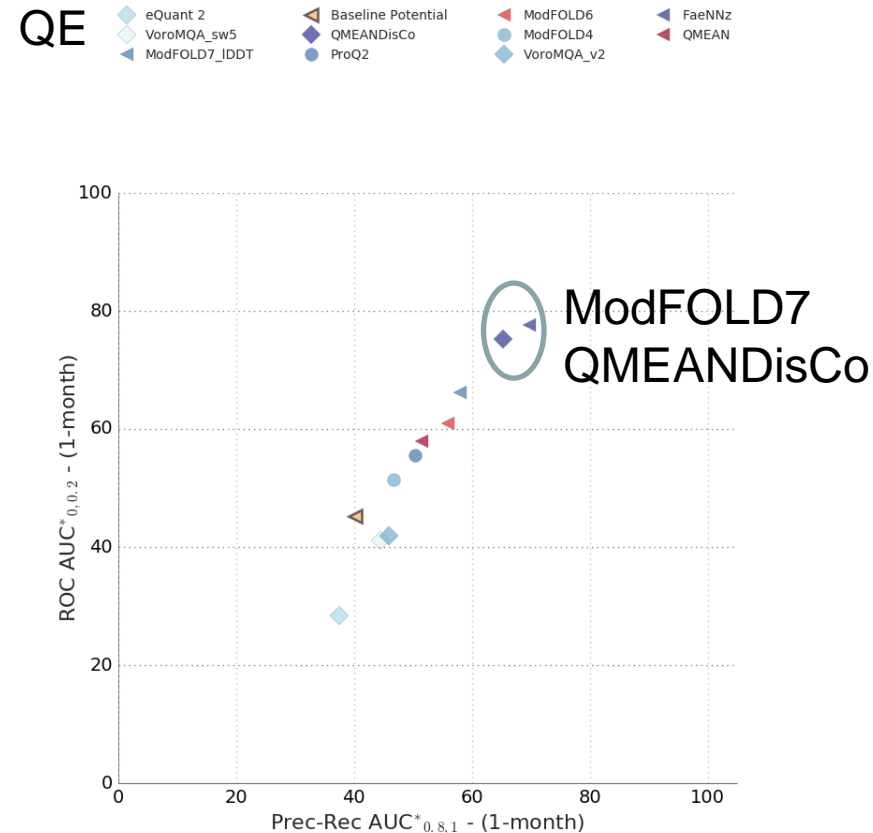
- **CASP** - Critical Assessment of techniques for protein Structure Prediction
  - runs every two years - <http://www.predictioncenter.org/casp13>
  - large QA for whole protein modelling field - establishes the criteria and ranks prediction teams, programs and servers
  - last round not yet published - special Proteins 2019 issue
- **CAMEO** - Continuous Automated Model EvaluatiOn
  - runs every week - <https://www.cameo3d.org/>
  - tests **3D structure** prediction, **Quality of Model** Estimation, **Contact Prediction**
  - 3D - based on IDDT (Local Distance Difference Test - model v<sub>z</sub> exp. 0-100(best)
  - QE - predicted IDDT>60



# CAMEO - best 3D and QE

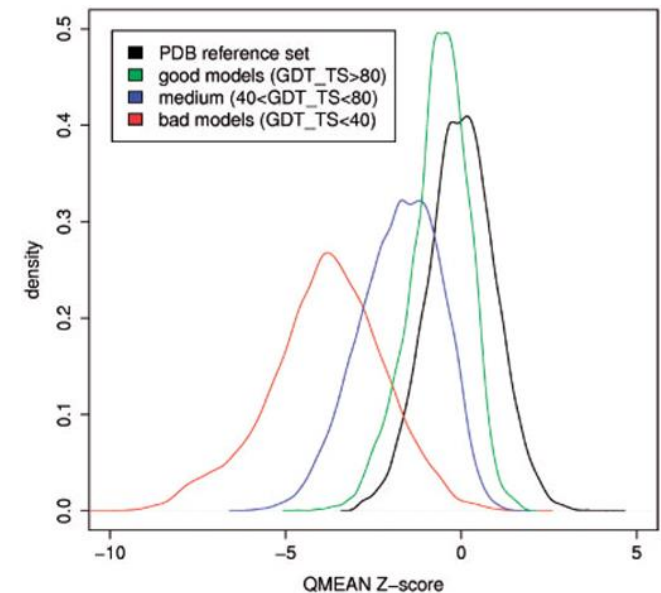
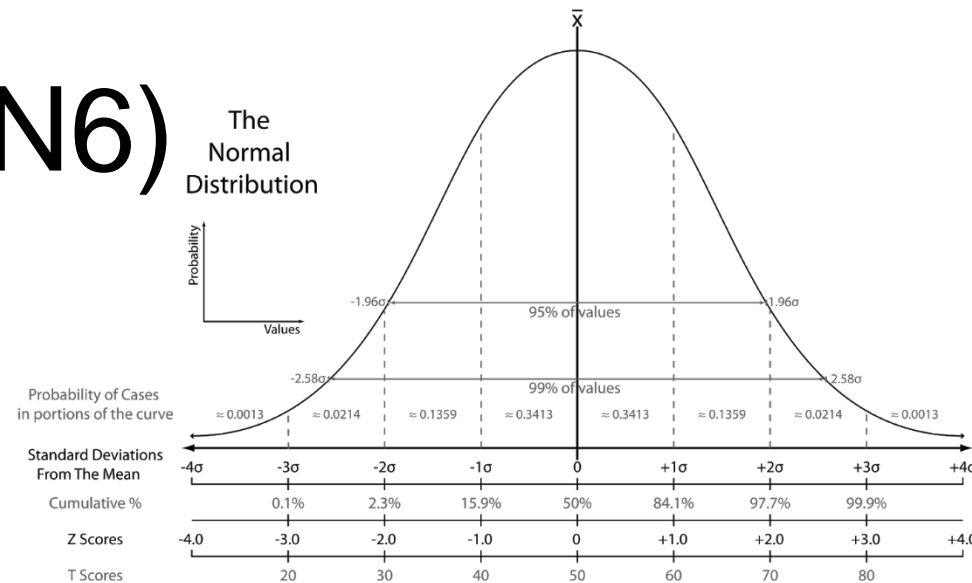
- 3D

Server	time	IDDT	IDDT-BS
Robetta	23 h	<b>69.1 ± 13.6</b>	65.77
IntFOLD5-TS	35 h	67.7 ± 15.5	<b>71.61</b>
RaptorX	10 h	66.8 ± 15.3	67.87
HHpredB	38 min	63.9 ± 17.2	67.14
SwissModel	<b>7 min</b>	63.1 ± 21.1	69.24
...			
Phyre2	2 h	52.9 ± 21.6	65.12



# QMEAN4 (QMEAN6)

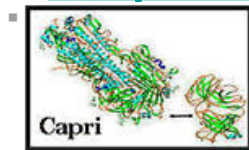
- composite scoring of aspects of QA of model
  - all-atoms -
  - $C\beta$  -
  - solvation -
  - torsions -
  - (ss\_agreement) -
  - (acc\_agreement) -
- Benchmarked to PDB reference set (cross validated)
  - z-score - good models  $\sim -1$ , bad models  $< -4$



Bioinformatics, 27(3) 2011, 343–350,  
<https://doi.org/10.1093/bioinformatics/btq662>

# CAPRI

- Critical Assessment of Prediction of Interactions
- <http://www.ebi.ac.uk/msd-srv/capri/>



Databases > PDBe > Services > Capri-Home > Round 34

Community wide experiment on the comparative evaluation of protein-protein docking for structure prediction

Hosted By EMBL/EBI-PDB Group

## Round 35

[Round 35 ID mapping from group number to Accessor Number for Target 107](#)

[Round 35 UPLOADER ID mapping for Target 107](#)

[Round 35 SCORER ID mapping for Target 107](#)

Full results:

## Target 107

[Results T107 - click here to see the results](#)

Clash threshold	53.61
average	16.81
std dev	18.40

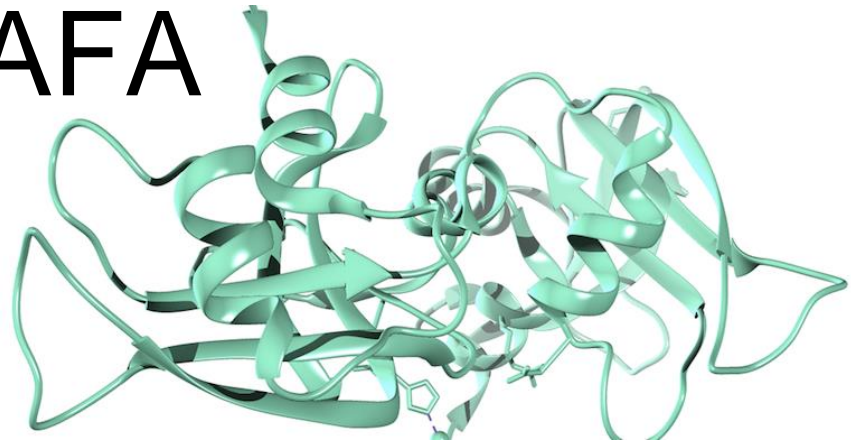
	Predictor	Uploader	Scorer
Nr groups	25	14	14
High Accuracy (***)	0 ( 0)	0 ( 0)	0 ( 0 <- 0)
Medium (**)	0 ( 0)	0 ( 0)	0 ( 0 <- 0)
Acceptable (*)	0 ( 0)	0 ( 0)	0 ( 0 <- 0)
Incorrect	239 (25)	1075 (13)	137 (14 <- 13)
Clashes	11 ( 4)	135 ( 5)	2 ( 2 <- 2)
Low ID	0 ( 0)	0 ( 0)	0 ( 0 <- 0)
Total	250 (25)	1210 (13)	139 (14 <- 13)

Numbers in parentheses refer to the number of different source groups producing these models (ie z scorer groups recognize y out of x uploader groups)

- Call For Targets
- Exp. Description
- Management
- Formats
- ROUND 35
- ROUND 34
- ROUND 33
- ROUND 32
- ROUND 31
- ROUND 30
- ROUND 29
- ROUND 28
- ROUND 27
- ROUND 26
- ROUND 25
- ROUND 24
- ROUND 23
- ROUND 22
- ROUND 21
- ROUND 20
- ROUND 19
- ROUND 18
- ROUND 17
- ROUND 16
- ROUND 15
- ROUND 14

**Protein  
Function  
Prediction**

CAFA



- Critical Assessment of Function Annotation
  - large-scale assessment of computational methods dedicated to [predicting protein function](http://biofunctionprediction.org/).
  - <http://biofunctionprediction.org/>

P. Radivojac et al A large-scale evaluation of computational protein function prediction. Nat Methods, 10(3):221–227, 2013.

# Summary

- Alphafold2 made a huge leap in **prediction accuracy**
- Role of **open science** and publicly available data can not be overstated
- CASP competition was a driver of the change
- Alphafold2 is **publicly available** and can be run from many places including ELIXIR CZ
- Alphafold has inspired many tools already
- Alphafold limits are yet to be fully described

# Acknowledgement

Marian Novotný



PŘÍRODOVĚDECKÁ  
FAKULTA  
Univerzita Karlova

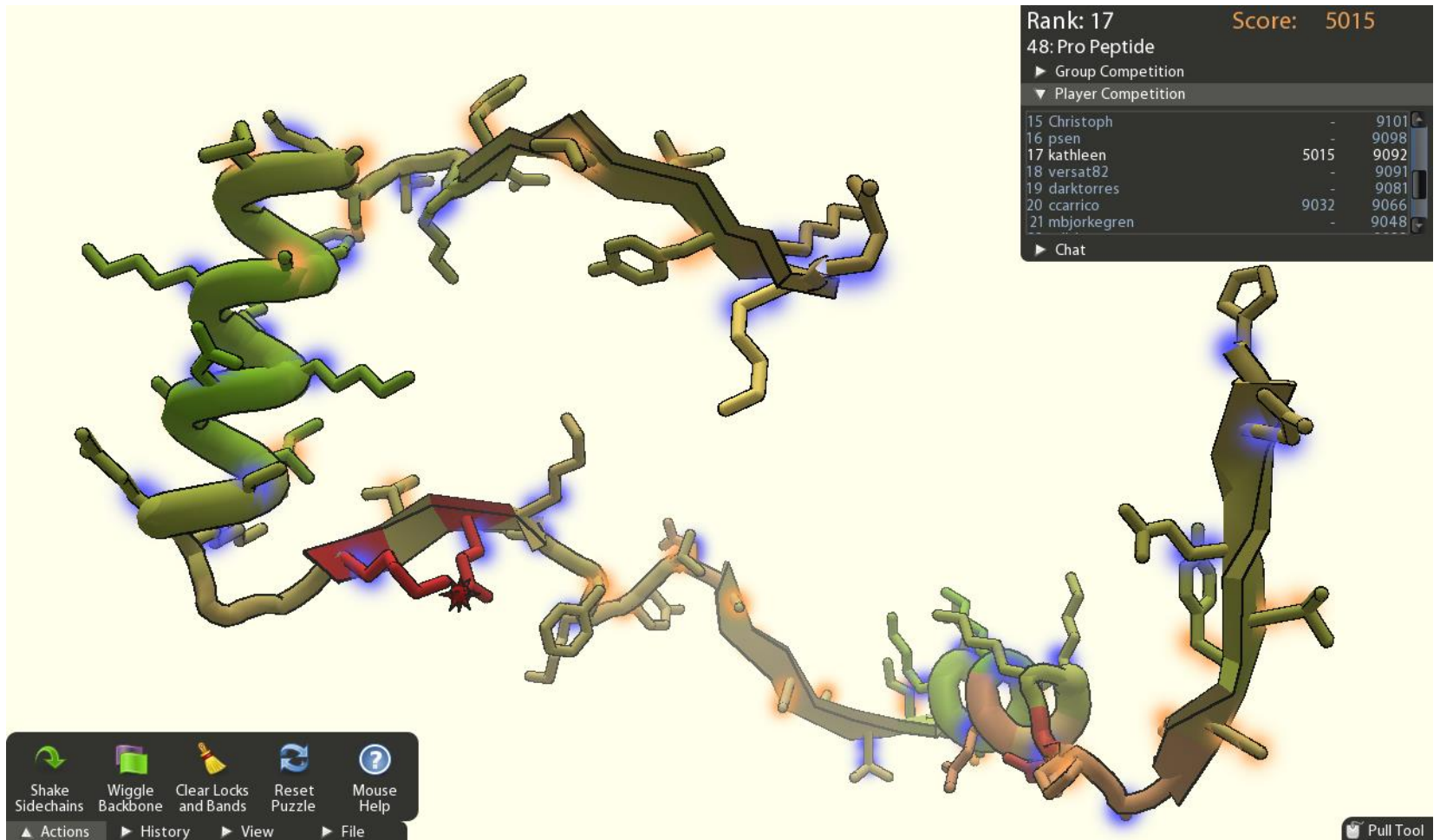
And now something  
completely different...





# FoldIt

- protein folding as a game



The screenshot displays the FoldIt game interface. The main area shows a protein structure composed of green and yellow segments, with some segments highlighted in red and blue. The interface includes a top right panel with a leaderboard, a bottom left panel with action buttons, and a bottom right panel with a pull tool.

**Rank: 17** **Score: 5015**  
48: Pro Peptide

► Group Competition  
▼ Player Competition

15	Christoph	-	9101
16	psen	-	9098
17	kathleen	5015	9092
18	versat82	-	9091
19	darktorres	-	9081
20	ccarrico	9032	9066
21	mbjorkegren	-	9048

► Chat

Shake Sidechains Wiggle Backbone Clear Locks and Bands Reset Puzzle Mouse Help

▲ Actions ► History ► View ► File

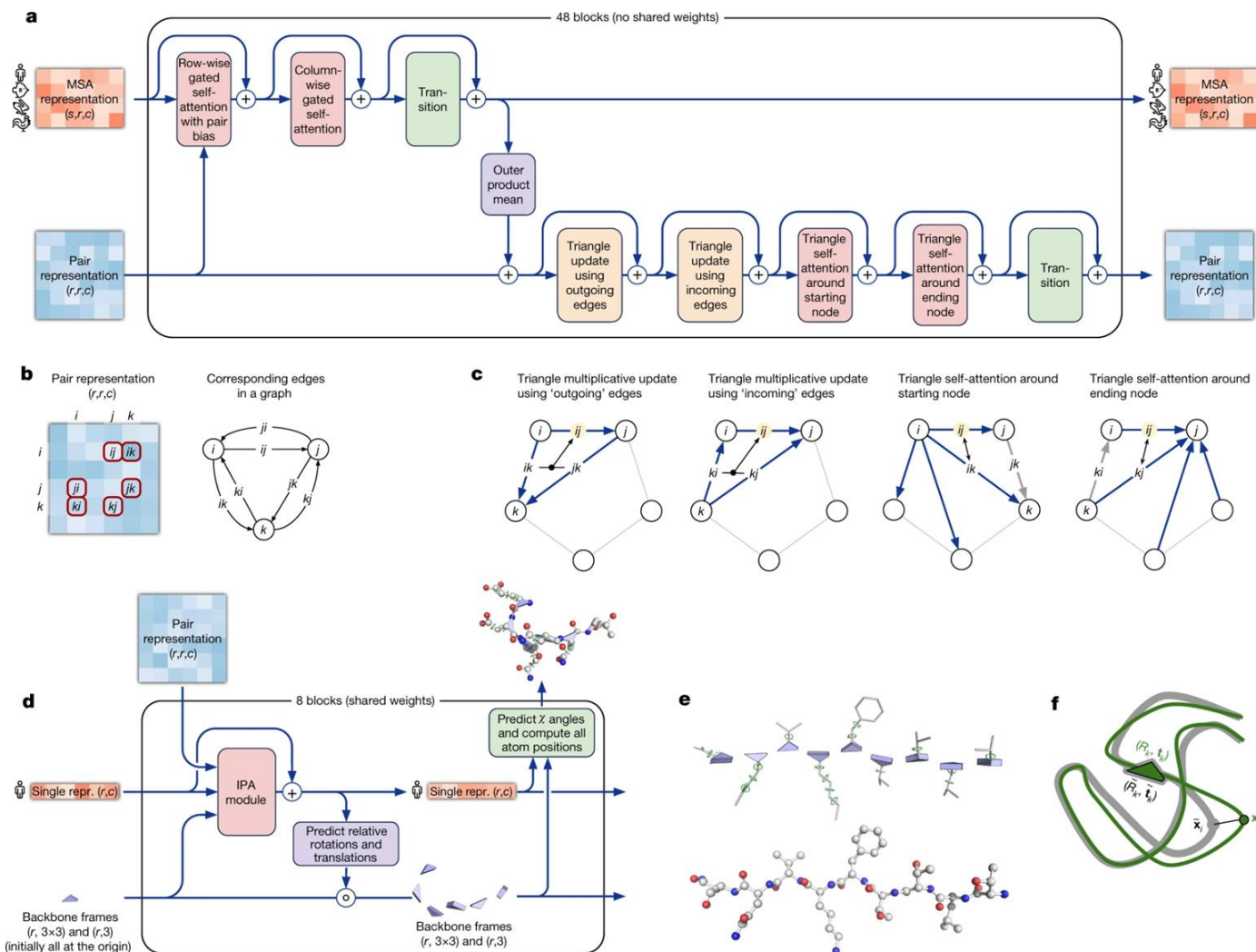
Pull Tool

- <http://fold.it/portal/>

# EXTRA SLIDES

AlphaFold

# Architectural details.



# Interpreting the neural network

