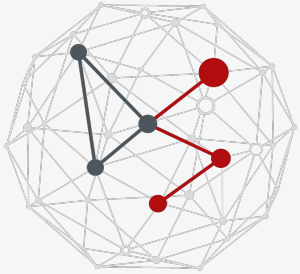
DIPARTIMENTO
MATEMATICA



DATA SCIENCE
UNIVERSITY OF PADOVA

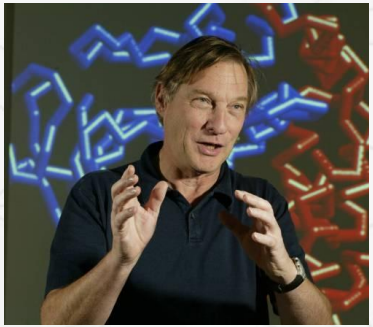
CASP

Master of Science in Data Science

Damiano Piovesan

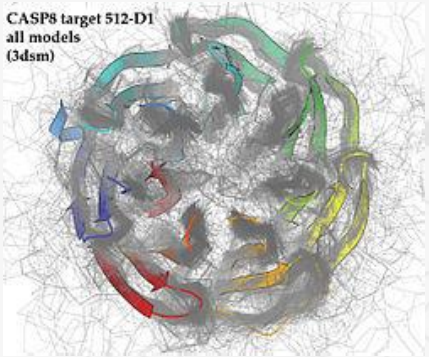
CASP - Critical Assessment of Techniques for Protein Structure Prediction

- Invented by **John Moult**
- **Blind test** that takes place **every two years** (since 1994) and involve the whole community
- Try to measure the state of the art and the improvement in the principal fields of **protein structure prediction**
  - Establishes a ranking of the best groups
  - "…CASP is not science. CASP is sports!" (Barry Honig, CASP-5 conference, 2002)
- Since CASP-3 (1998), CA**FA**SP ("... Fully Automated…")
  - Evaluate the automatic predictors (web servers)
- Since CASP-13 (2018), CAID (Critical Assessment of Intrinsic Disorder)
  - Organized by BioComputingUP lab, University of Padova



How does it work?

- Data collection from experimentalists
- Prediction season (May – August)
- Independent assessment (September – November)
- Conference (November/December)
- Publication (One year later)



CASP 13 - December 1 - 4, 2018

Iberostar Paraiso Maya, Riviera Maya, Mexico

CASP6 Target T0280

1. Protein Name
1wd5
2. Organism Name
Thermus thermophilus
3. Number of amino acids (approx)
208
4. Accession number

5. Sequence Database

6. Amino acid sequence
MRFRDRRHAGALLAEALAPLGLEAPVVLGLPRGGVVVADEVARRLGGELDVVLVRKVGAP
GNPEFALGAVGEGGELVLMPYALRYADQSYLEREAARQRDVLRKRAERYRRVRPKAARKG
RDVVLVDDGVATGASMEAALSVVFQEGPRRVVVAVPVASPEAVERLKARAEVVALSVPQD
FAAVGAYYLDFGEVTDEDVEAILLEWAG
7. Additional information

8. X-ray structure
yes
9. Current state of the experimental work
finished
10. Interpretable map?
no
11. Estimated date of chain tracing completion
completed
12. Estimated date of public release of structure
October 2004

Related Files

Template Sequence file

Template PDB file

CASP categories

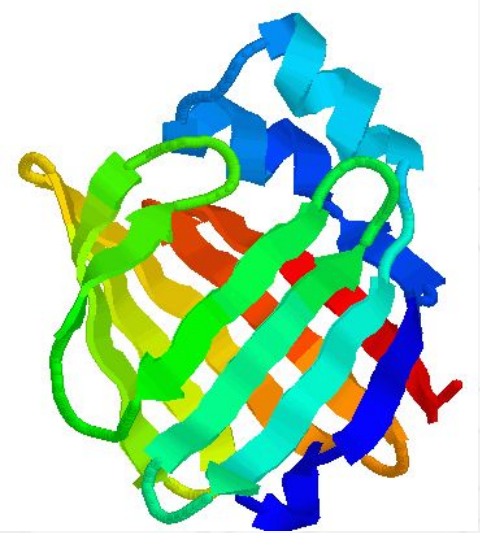
- Models with templates
- Models without templates (“ab initio“)
- Contacts
- Structural domains
- Function
- Model quality
- Disorder

Prediction format	Number of groups/servers contributing (unique)	Number of models designated as 1	Total number of models
TS: 3D coordinates	176 / 79	14659	61665
AL: Alignments to PDB structures	2 / 2	246	1220
RR: Residue-residue contacts	28 / 18	3079	4162
DR: Disordered regions	32 / 22	3955	5210
FN: Binding sites prediction	33 / 15	3044	5666
QA: Quality assessment	46 / 34	5490	7116
TR: Model refinement	37 / 12	416	1709
All:	251 / 140	31032	86891

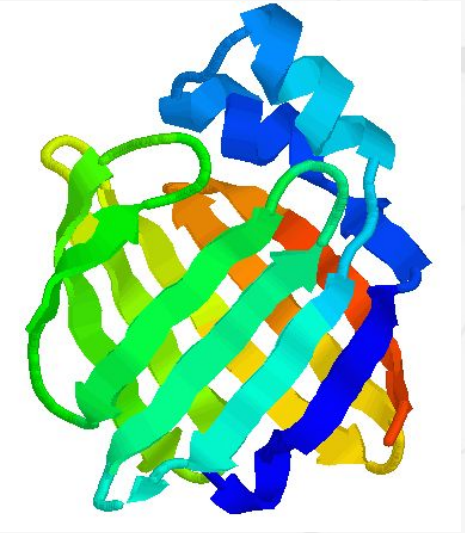
CASP 9

T0137

- Fatty acid binding protein FABP1, E. granulosus (135 residues)
- **40% identity** target/template
- **0.98 Å RMSD** target/template



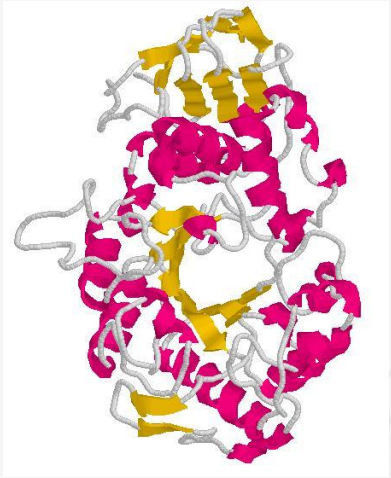
Model



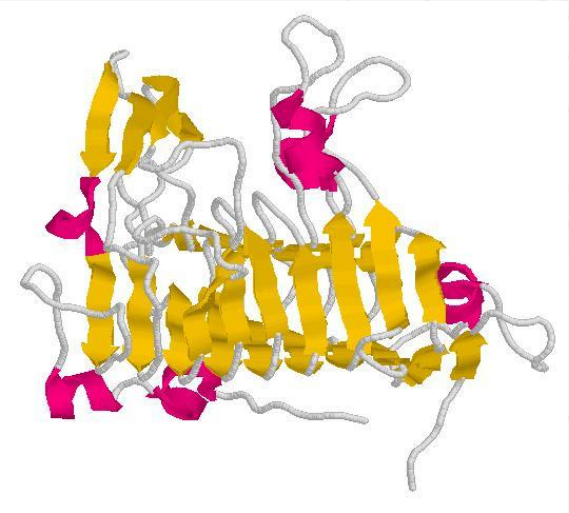
Real Structure

T0100

- Pectin Methylesterase, E. chrysanthemi (342 residues)
- **12% identity** target / template

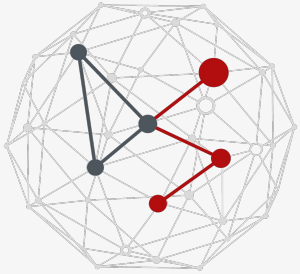


Wrong Prediction from SAM-T99



Real Structure

DIPARTIMENTO
MATEMATICA



DATA SCIENCE
UNIVERSITY OF PADOVA

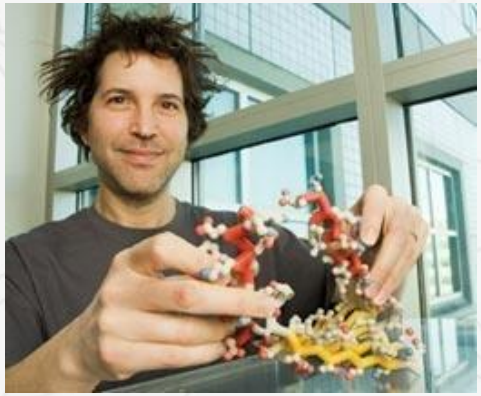
ROSETTA

Master of Science in Data Science

Damiano Piovesan

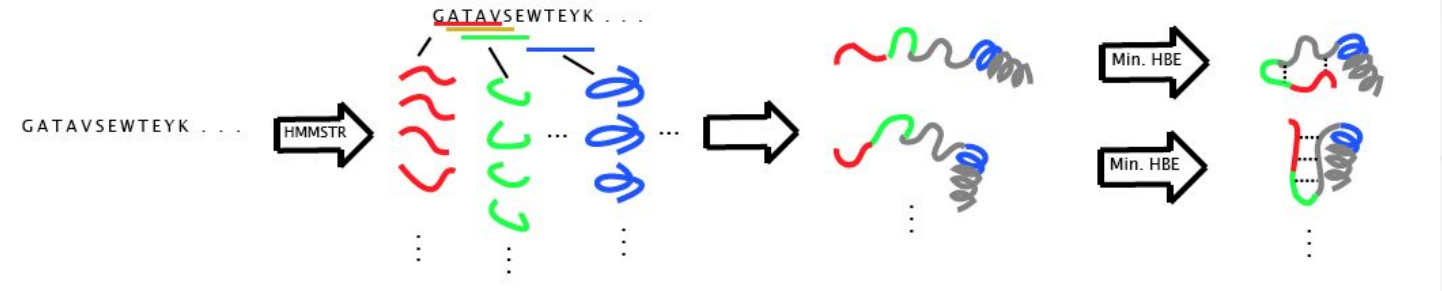
Rosetta

- Developed by **David Baker** (Uni Washington, Seattle)
- Rosetta was first used in **1998** (CASP-3)
- It has dominated every CASP edition since 2000 (CASP-4) **until 2016 (CASP-12)**
- In 2018 **overcomed by AlphaFold** (A7D) by Google DeepMind
- ROSETTA is not pure ab initio as it uses **statistics for local structures**



Algorithm

1. Split the sequence into fragments of **9 residues** (1 per position)
2. Select **similar fragments** from the **PDB** (based on sequence similarity)
3. Combine protein fragments from **unrelated proteins** with **similar local sequences**
4. **Sample conformations**. Energy minimization with **Simulated Annealing** using a set of **statistical potentials**
5. Select the most frequent conformation among those with similar (low) energy



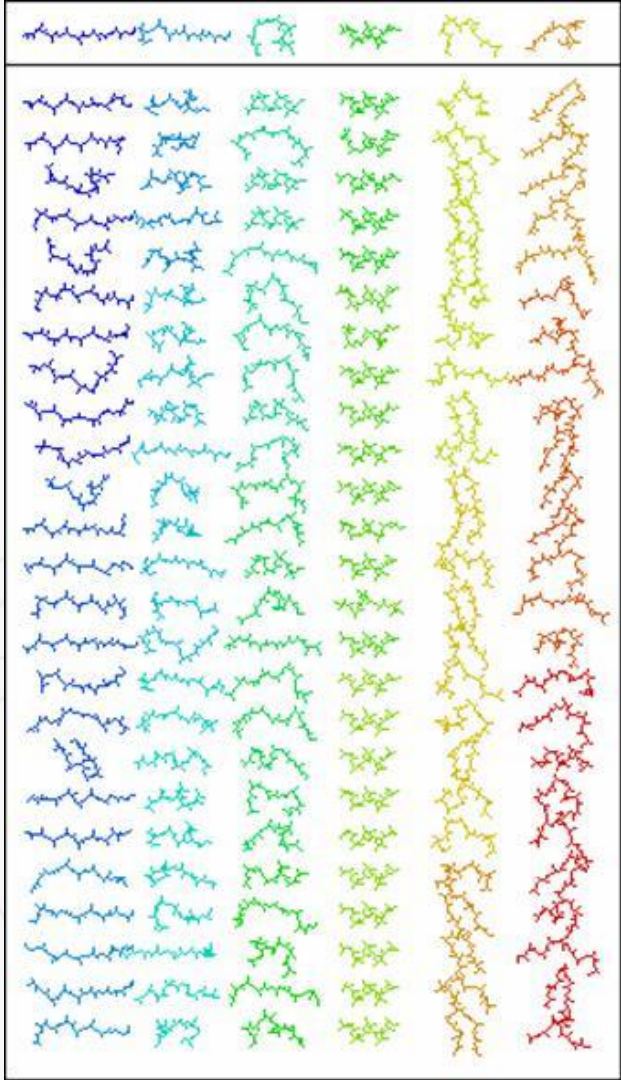
Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions
Simons KT, Kooperberg C, Huang E, Baker D. **(1997) J Mol Biol**

Fragments selection

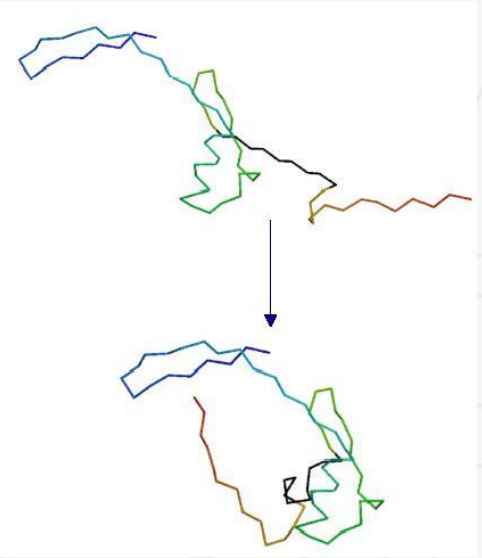
Find top 25 nearest fragment neighbors in the **PDB**

$$DISTANCE = \sum_{i}^{9} \sum_{aa}^{20} | S(aa, i) - X(aa, i) |$$

- ***S(aa, i)*** frequencies of amino acid *aa* at position *i* in a multiple sequence alignment (MSA) of the **fragment to be folded**
- ***X(aa, i)*** frequencies in the MSA of **one sequence of the PDB**
- If they have identical sequence the distance is 0

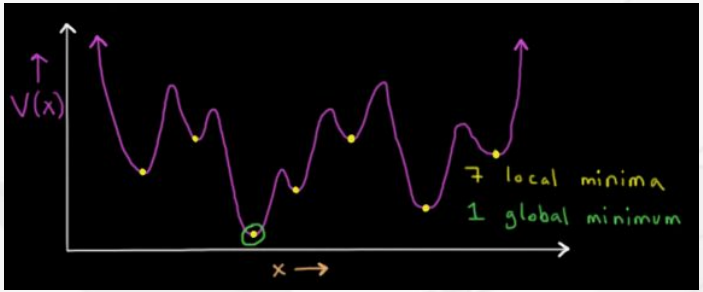


Sample conformations - Simulated annealing



A move consists of substituting the **torsional angles** of a randomly chosen neighbor at a randomly chosen position (10K cycles)

- Moves which bring two atoms within **2.5 A** are immediately rejected
- Other moves are evaluated with the **Metropolis Montecarlo** criterion using an **energy function** (statistical potential)



1. Assign initial X_0
2. Propagate $X_i \rightarrow X_{i+1}$
3. Decrease T
4. Repeat until $T_i = 0$

Molecular Dynamics, Molecular Mechanic

$T_{i+1} = T_0(1 - \alpha)$

Discriminatory functions (1)

DIPARTIMENTO MATEMATICA

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

Constant

Radius of gyration

- Able to **distinguish** random chains from **folded** structures

$$P(structure) \sim \exp(-radius\ of\ gyration^2)$$

Profile method

- Independence of positions
- $E_i \rightarrow$ **structural environment** (SS or solv. acc.)

$$P(sequence \mid structure) \cong \prod_i P(aa_i \mid E_i)$$

Not used
Solvation is included implicitly in the pair distributions (below)

Distance method

- Independence of pairs of positions (neglect chain connectivity)

$$P(sequence \mid structure) \cong \prod_{i<j} P(aa_i, aa_j \mid r_{ij})$$

$$P(aa_i, aa_j \mid r_{ij}) = \cancel{P(aa_i, aa_j)} \times \frac{P(r_{ij} \mid aa_i, aa_j)}{P(r_{ij})}$$

Independent of structure

Discriminatory functions (2)

DIPARTIMENTO MATEMATICA

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

Constant

Rosetta (generation step)

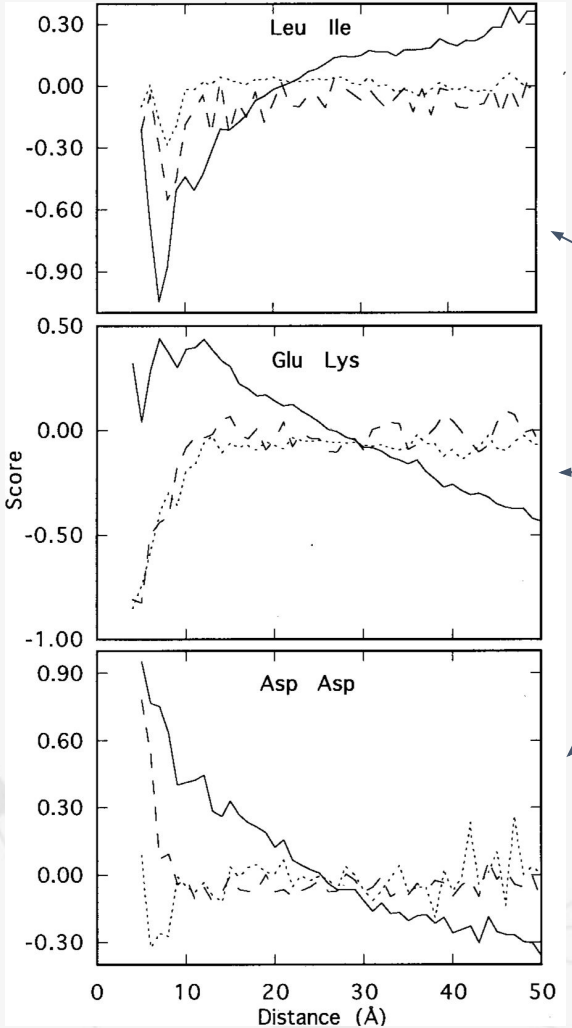
- Fast

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

Rosetta (evaluation step)

- **Decoupling** of the **distance** and **environment** dependencies
- Incorporation of solvation and residue pair interactions in a non-redundant manner
- Avoid blurring specific residues interactions with the overall partitioning of residues into the protein core

$$P(aa_1, aa_2, \ldots, aa_n \mid structure) \cong \prod_i P(aa_i \mid E_i) \times \prod_{i<j} \frac{P(aa_i, aa_j \mid r_{ij}, E_i, E_j)}{P(aa_i \mid r_{ij}, E_i, E_j) P(aa_j \mid r_{ij}, E_i, E_j)}$$



Environment independent	Environment dependent
$P(aa_i, aa_j) \times \frac{P(r_{ij} \mid aa_i, aa_j)}{P(r_{ij})}$	$\prod_i P(aa_i \mid E_i) \times \prod_{i<j} \frac{P(aa_i, aa_j \mid r_{ij}, E_i, E_j)}{P(aa_i \mid r_{ij}, E_i, E_j) P(aa_j \mid r_{ij}, E_i, E_j)}$
———	$E_x \rightarrow$ Surface – – – Buried . . . .

Pairs of hydrophobic residues

- Env. ind., attractive at short distance and repulsive at long distances
- Env. dep., weekly attractive at ~8A and decay rapidly to zero

Pairs of charged residues (opposite charge +/-)

- Env. ind., attractive at large distances → partitioning of polar residues to protein surfaces
- Env. dep., closer to physical intuition, attractive at short distance

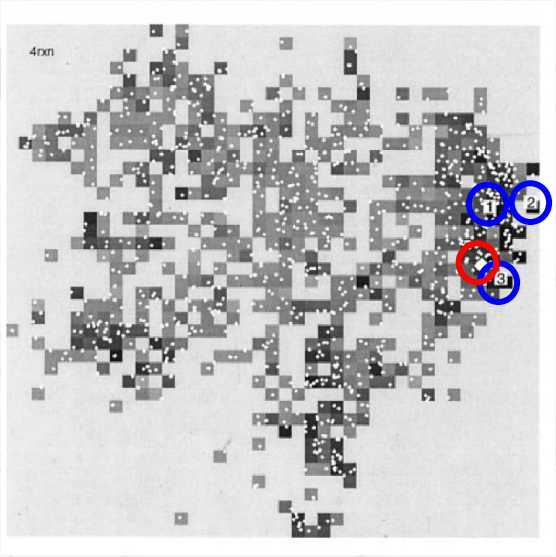
Pairs of charged residues (same charge -)

- Env. ind., repulsive at short distance
- Env. dep. **Repulsive** at short distances as expected for surface pairs (broken line).
- Env. dep. **Weakly attractive** at short distance →buried metal binding sites and enzyme active sites (dotted line)

Clustering of conformations

- The **native state** of a protein is an **ensemble** of many **similar conformations**
- Proteins participate (sample) a second much larger ensemble, the "**denatured state**" (or low resolution structures)
- Many of the **global topological features** of the native state are retained in the "**denatured state**" (burial of hydrophobic surface)
- Atomic forces contribute little to the properties of denatured proteins

2D energy space



Native

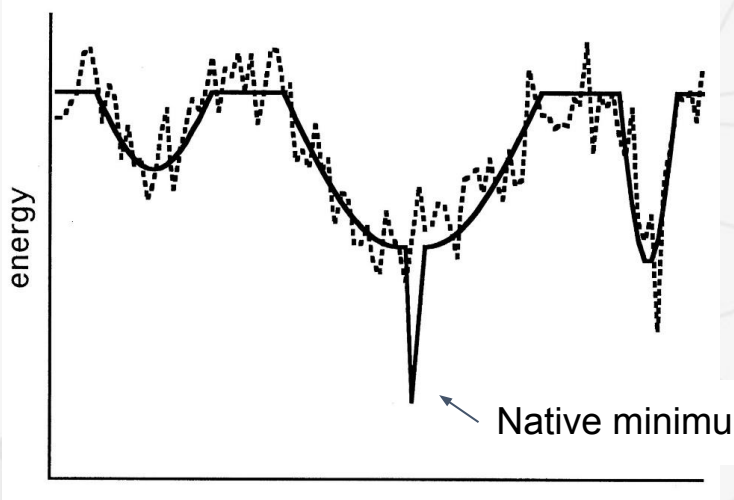
Best predictions

Clustering of low-energy conformations near the native structures of small proteins
Shortle, Simons and Backer. **PNAS. 1998**

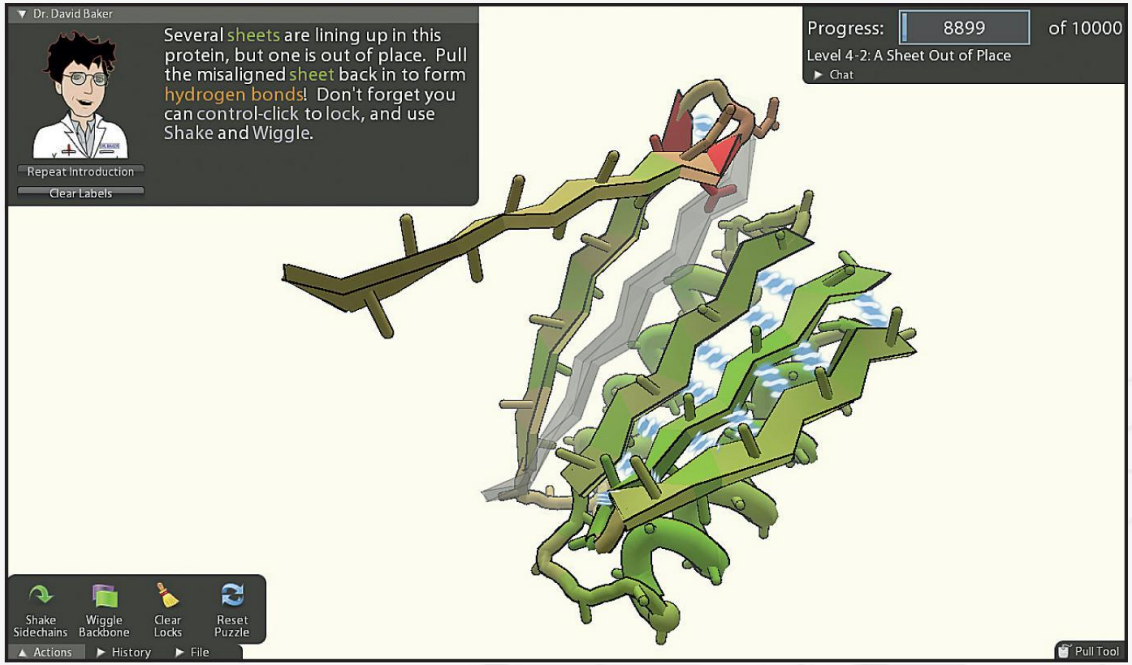
Native minimum

DIPARTIMENTO MATEMATICA

- The **native minimum** is **broader** than any other minimum
- The **breadth** of the native minimum results from the long range character of **hydrophobic interactions**
- The **scoring function** follows the true potential because it is sensitive to hydrophobic burial.
- But produces noise and **fails** to detect the sharp drop of the **native state**
- Inaccuracies in quantifying **hydrogen bonds**, **electrostatic** and **van der Waals** interactions
- However, the scoring function is able to **detect** the **higher density of low-energy states** in the broad region surrounding the native state

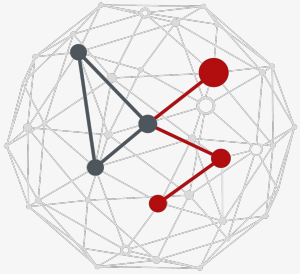


Protein folding as a game, FOLD IT



https://fold.it/portal/

DIPARTIMENTO
MATEMATICA



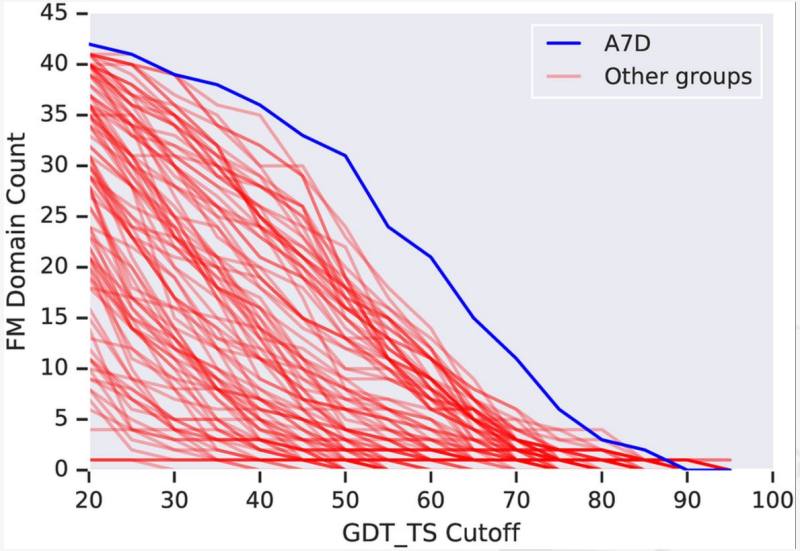
DATA SCIENCE
UNIVERSITY OF PADOVA

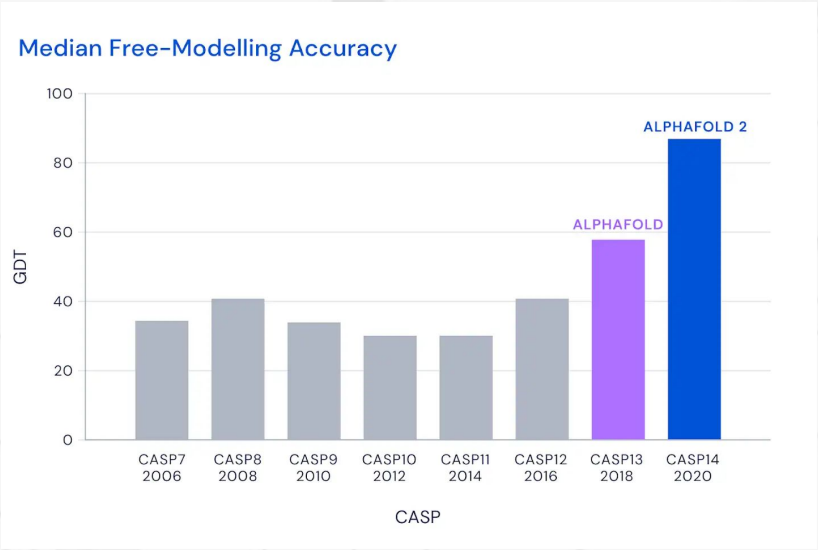
ALPHAFOLD

Master of Science in Data Science

Damiano Piovesan

CASP results



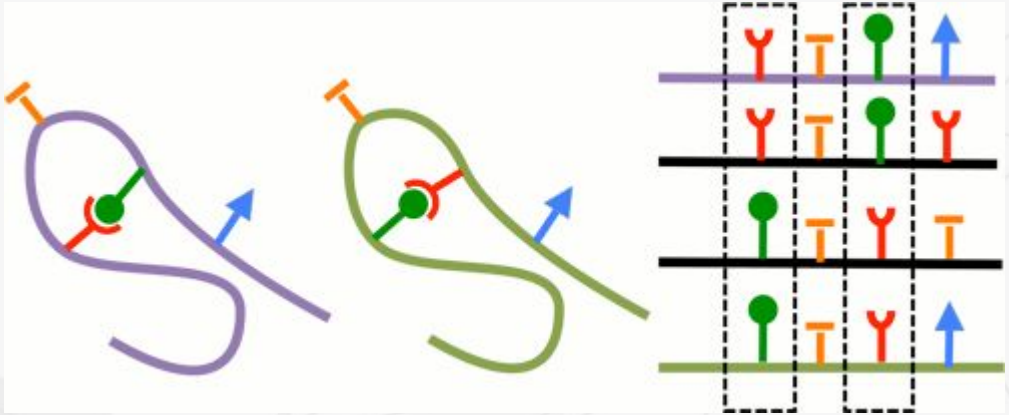


CASP-13 (2018)

CASP-14 (2020)

Covariance patterns

- **Positions that tend to co-vary**, i.e. two residues that seem to change together, as if they depend on one another
- Strong covariance between two residues usually suggests that those **residues interact with one another in the folded structure**, through side-chain packing, H-bonding, electrostatics, etc.



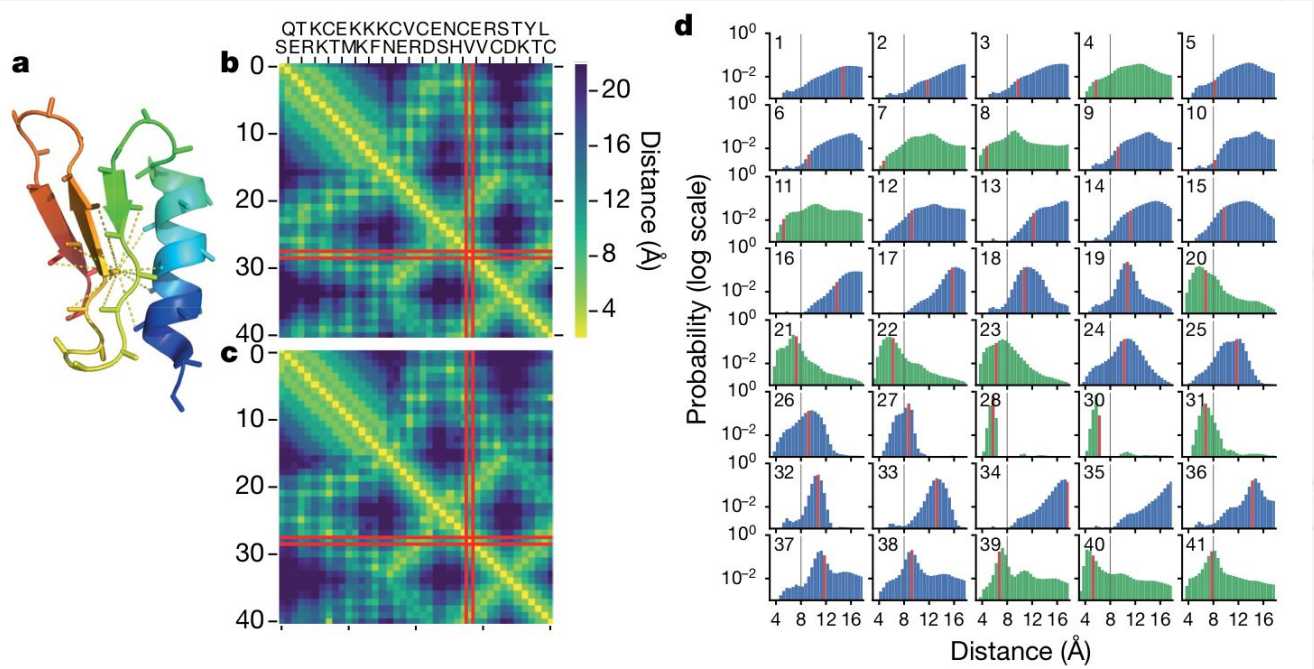
Foldit blog - https://fold.it/portal/node/2008706

Contacts Vs. Distances prediction

DIPARTIMENTO MATEMATICA

- **AlphaFold** does **not** use **covariance** to predict **contacts** (a simple yes/no)
- **AlphaFold** predicts the **distance** between the two residues

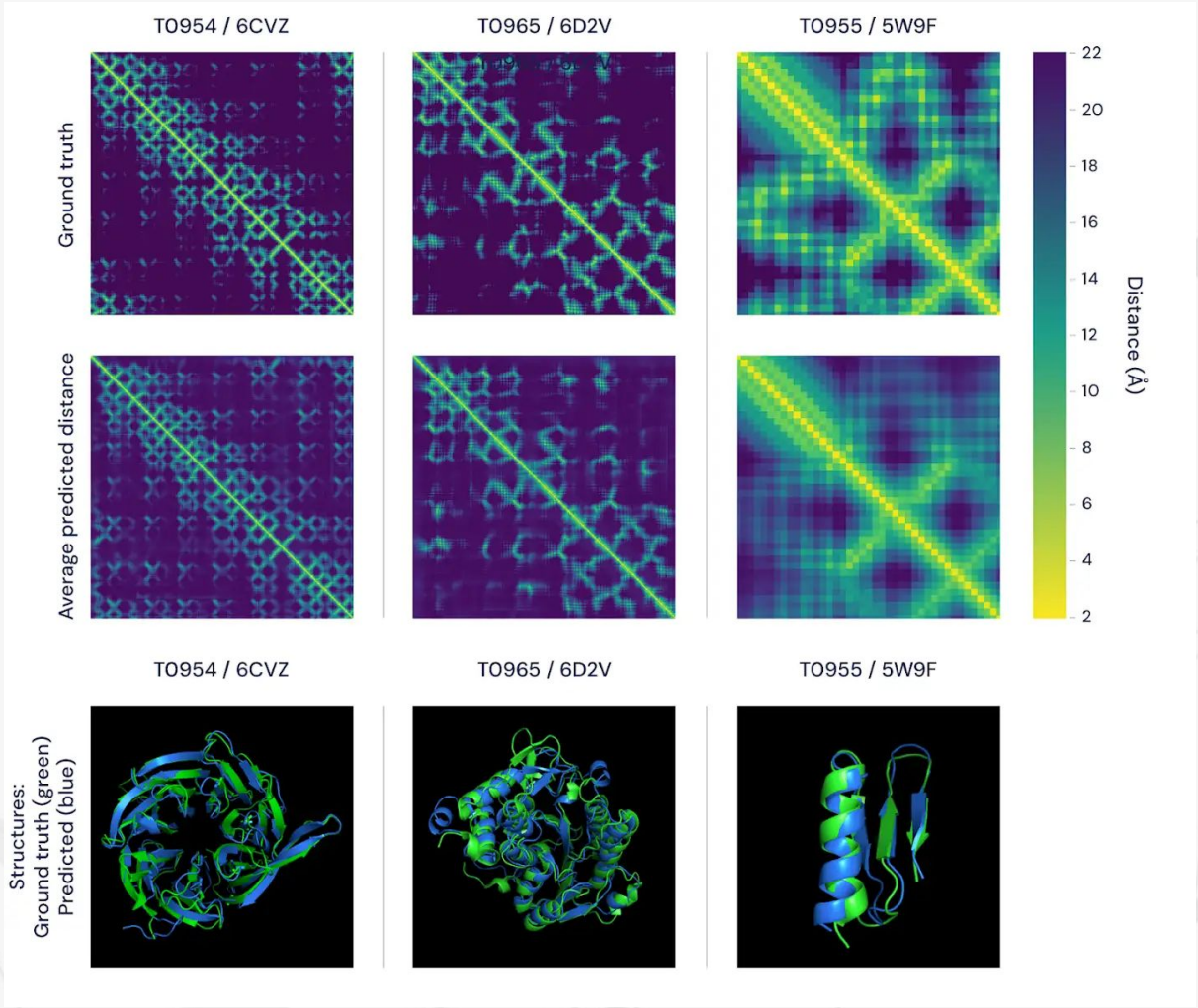
Predicted distance probability of one residue against all the others



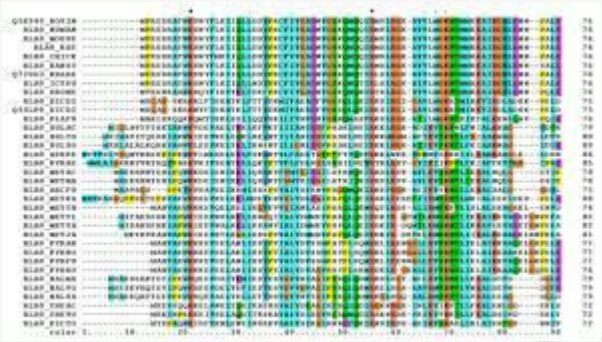
Range of values between 2 and 20 Å

Red lines indicate the true distance

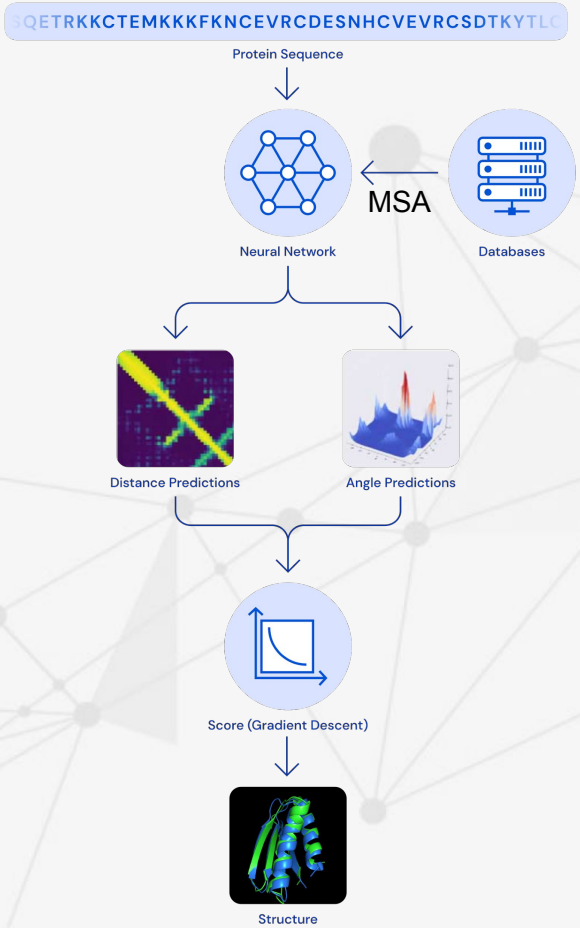
Distribution in green are true contacts


T0954 / 6CVZ
T0965 / 6D2V
T0955 / 5W9F
Ground truth
Average predicted distance
Distance (Å)
22
20
18
16
14
12
10
8
6
4
2
T0954 / 6CVZ
T0965 / 6D2V
T0955 / 5W9F
Structures:
Ground truth (green)
Predicted (blue)

AlphaFold



Multiple Sequence Alignment (**MSA**)



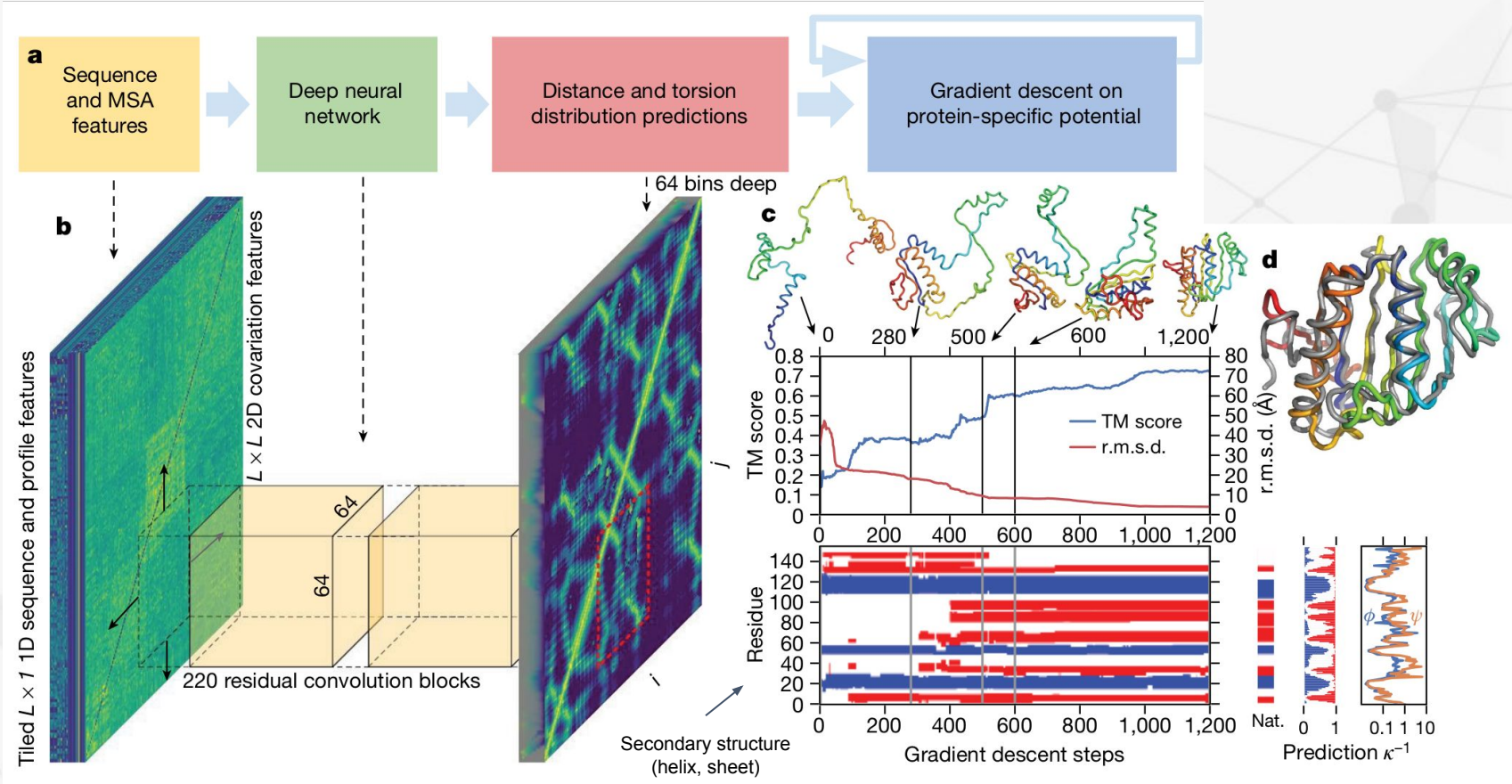
Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13).
Senior et al. (Oct 2019) Proteins

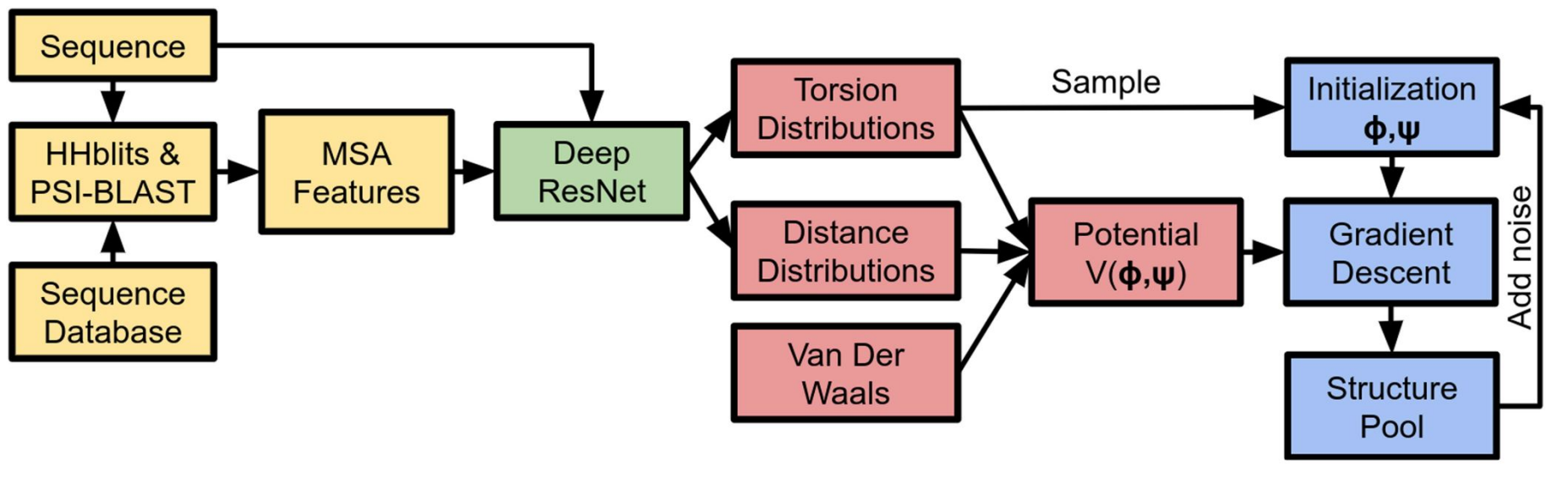
Improved protein structure prediction using potentials from deep learning.
Senior et al. (2020) Nature

DeepMind blog https://deepmind.com/blog/alphafold/

$$P(d_{ij}|S, \mathrm{MSA}(S))$$

DIPARTIMENTO MATEMATICA

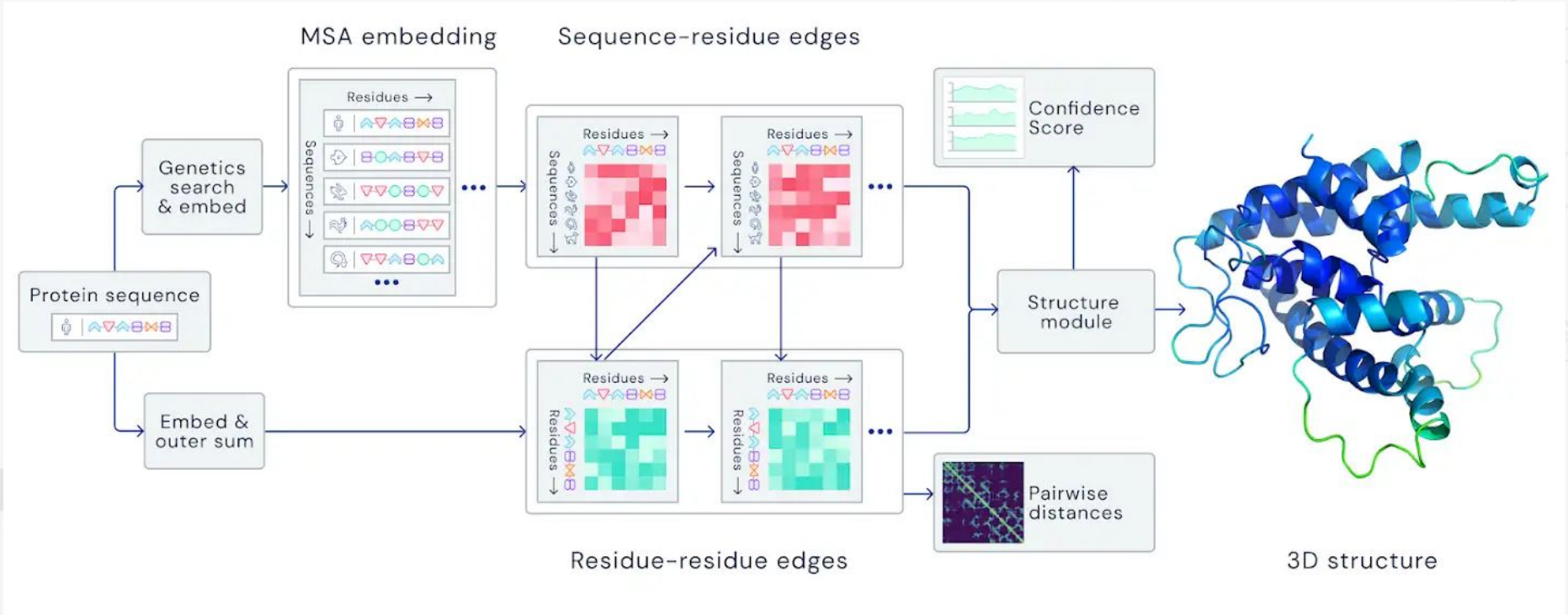



Sequence
HHblits & PSI-BLAST
Sequence Database
MSA Features
Deep ResNet
Torsion Distributions
Distance Distributions
Van Der Waals
Sample
Potential V(ϕ,ψ)
Initialization ϕ,ψ
Gradient Descent
Structure Pool
Add noise

AlphaFold 2

- Folded protein as “spatial graph”, residues are the nodes and edges connect the residues in close proximity
- Attention-based neural network system, trained end-to-end
- It uses evolutionarily related sequences, multiple sequence alignment (MSA), and a representation of amino acid residue pairs to refine the graph
- 16 TPUv3s (which is 128 TPUv3 cores or roughly equivalent to ~100-200 GPUs) run over a few weeks

AlphaFold 2


MSA embedding
Sequence-residue edges
Genetics search & embed
Protein sequence
Embed & outer sum
Residues →
Sequences →
Confidence Score
Structure module
Pairwise distances
Residue-residue edges
3D structure