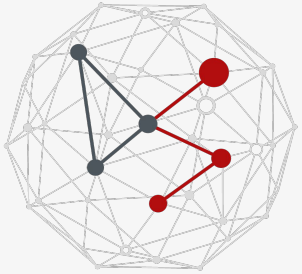
UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
MATEMATICA



DATA SCIENCE
UNIVERSITY OF PADOVA

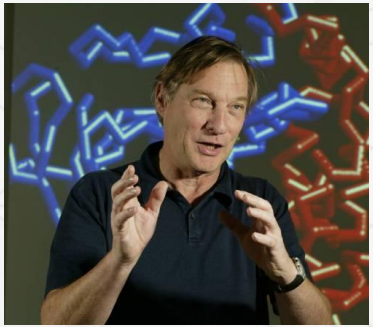
CASP

Master of Science in Data Science

Damiano Piovesan

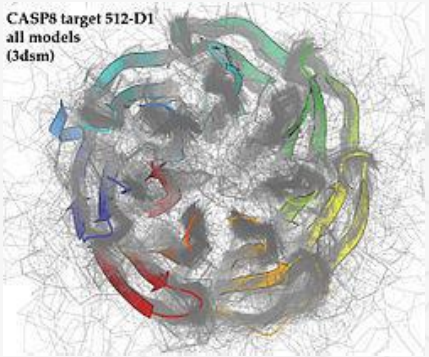
CASP - Critical Assessment of Techniques for Protein Structure Prediction

- 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 2018 - Riviera Maya, Mexico

CASP 14 - December 2020 - Virtual

CASP 15 - December 2022 - Antalya, Turkey

CASP 16 - December 2024 - Caribbean

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

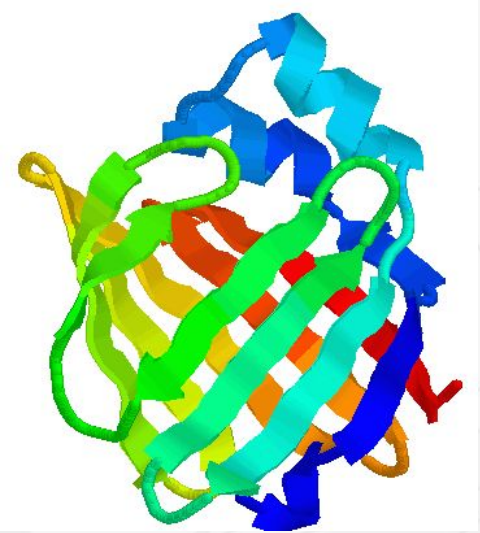
- Models with templates
- Models without templates (“ab initio“)
- Contacts
- Structural domains
- Function
- Model quality
- Disorder

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

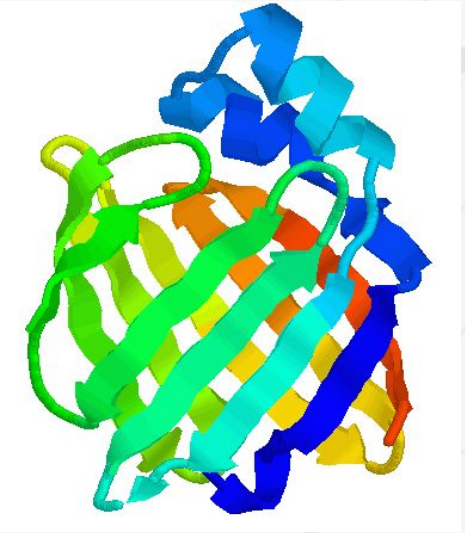
CASP 9

T0137

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



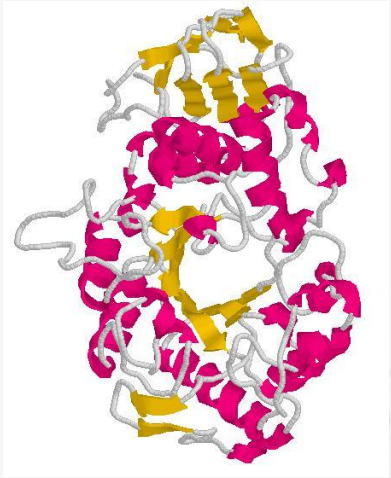
Model



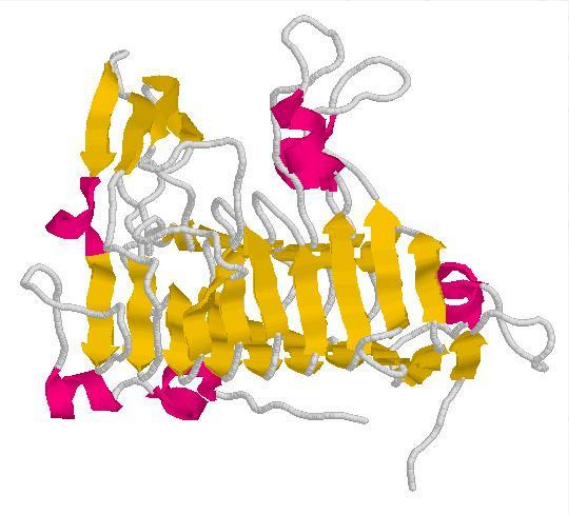
Real Structure

T0100

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



Wrong Prediction from SAM-T99



Real Structure

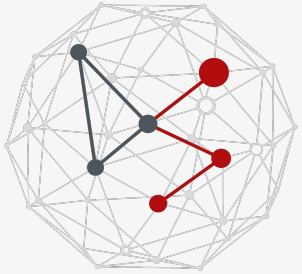
Global distance test total score (GDT_TS)

1. Superimpose the model with the template (after identifying an initial set of equivalent atoms)
2. Identify all atom pairs for which distance is larger than the threshold
3. Re-obtain the transform, excluding those atoms
4. Repeat until the set of atoms used in calculations is the same for two cycles running

In CASP is the average result of cutoffs at 1, 2, 4, and 8 Å and all atoms are considered

Reported as a percentage, ranging from 0 to 100

https://predictioncenter.org/casp13/doc/LCS_GDT.README



DATA SCIENCE
UNIVERSITY OF PADOVA

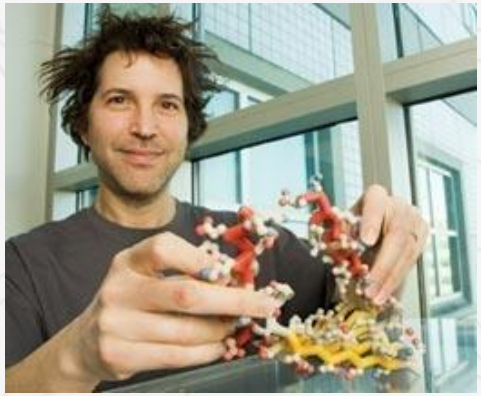
ROSETTA

Master of Science in Data Science

Damiano Piovesan

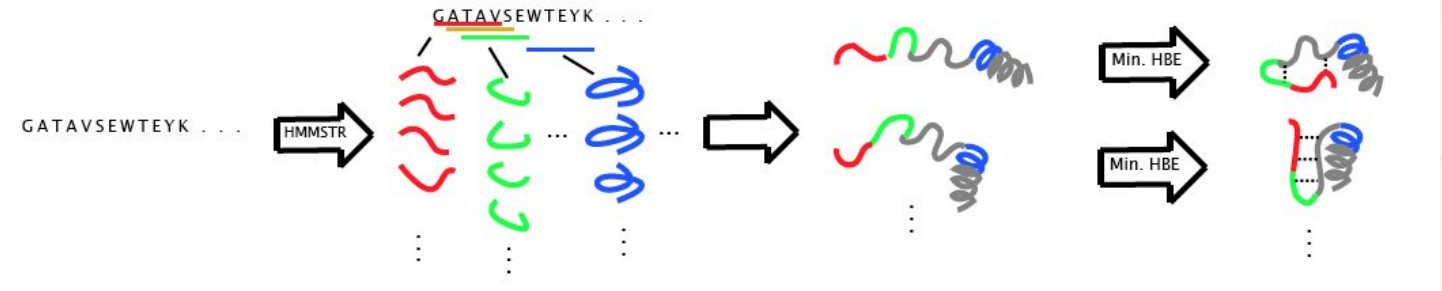
Rosetta

- 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 alternative conformations**. Energy minimization with **Simulated Annealing** using a set of **statistical potentials**
5. Select the most **frequent conformation** among those with similar **(low) energy**



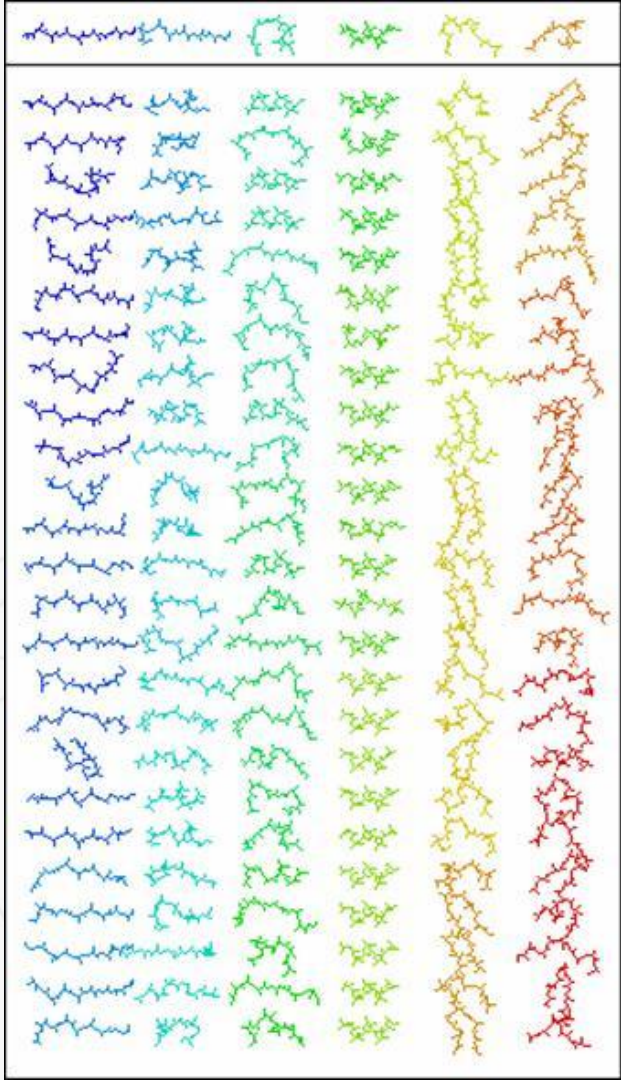
Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions
Simons KT, Kooperberg C, Huang E, Baker D. **(1997) J Mol Biol**

Fragments selection

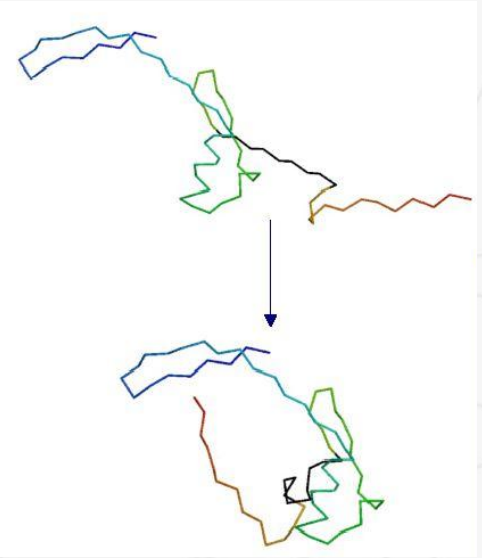
Find top 25 nearest fragment neighbors in the **PDB**

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

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

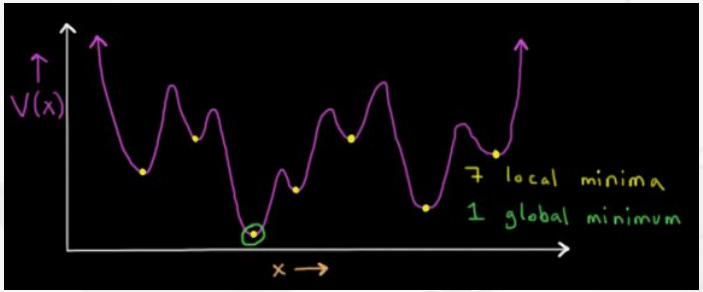


Sample conformations - Simulated annealing



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.)

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

Solvation is included implicitly in the pair distributions (see 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)

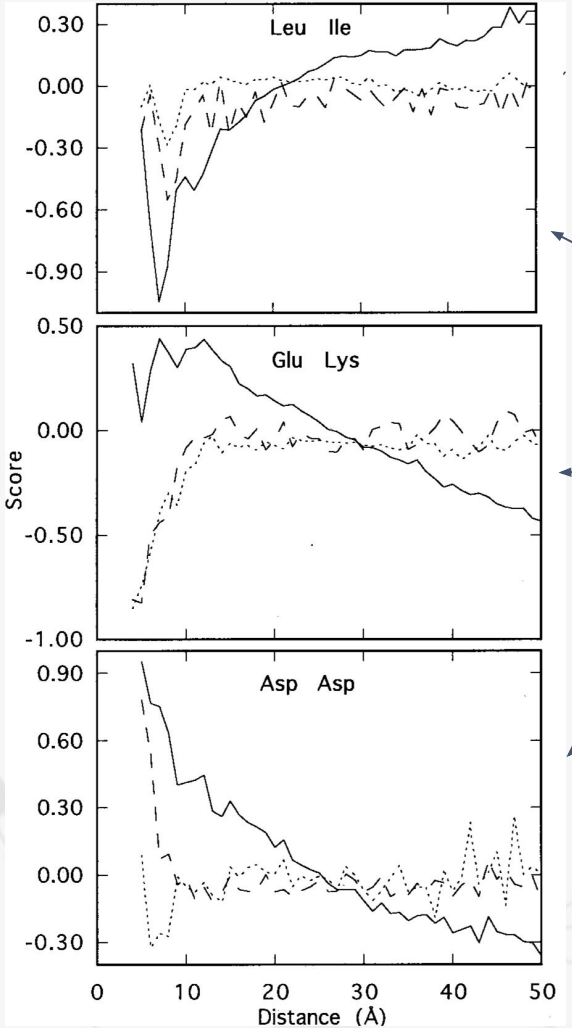
- Fast

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

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

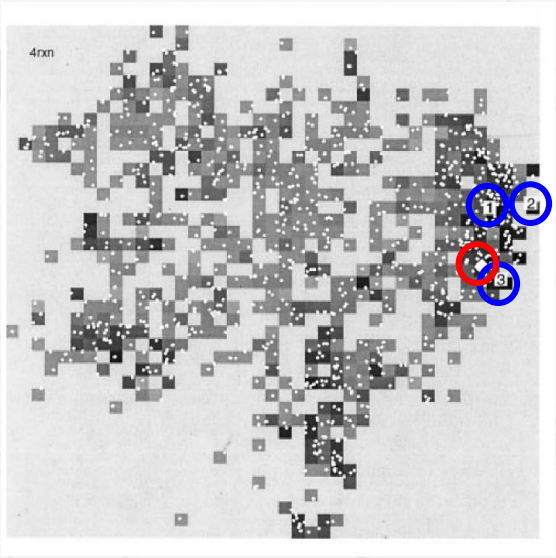
Pairs of charged residues (same charge -)

- Env. ind., repulsive at short distance
- Env. dep
  - **Repulsive** at short distances as expected for surface pairs (broken line)
  - **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

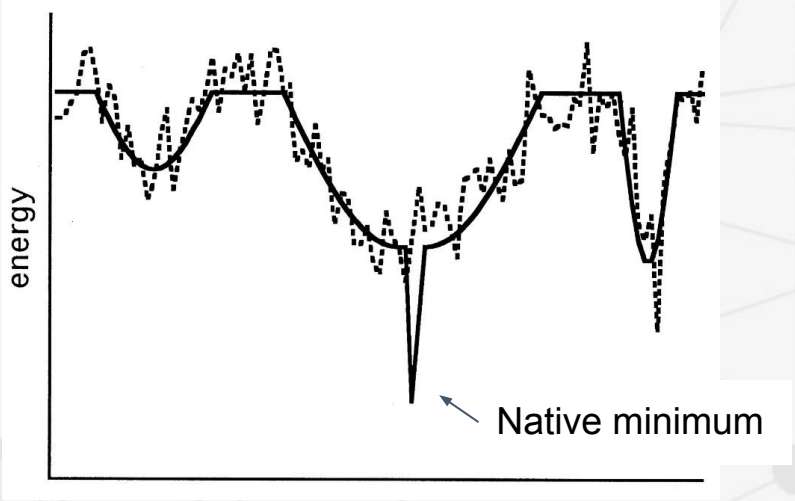
Best predictions

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

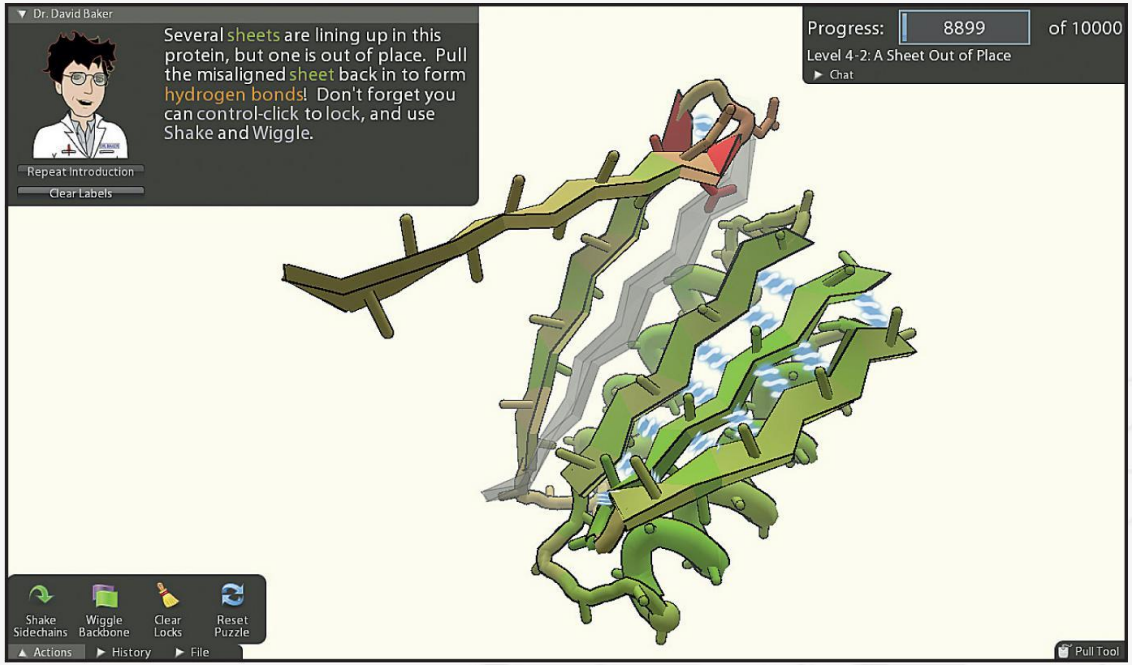
Native minimum

DIPARTIMENTO MATEMATICA

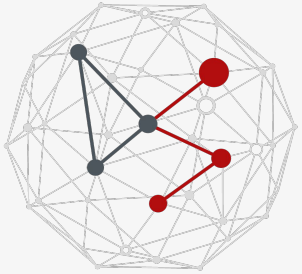
- The **native minimum** is **broader** than any other minimum
- The **breadth** of the native minimum results from the long range character of **hydrophobic interactions**
- The **scoring function** follows the true potential because it is sensitive to hydrophobic burial.
- But produces noise and **fails** 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/



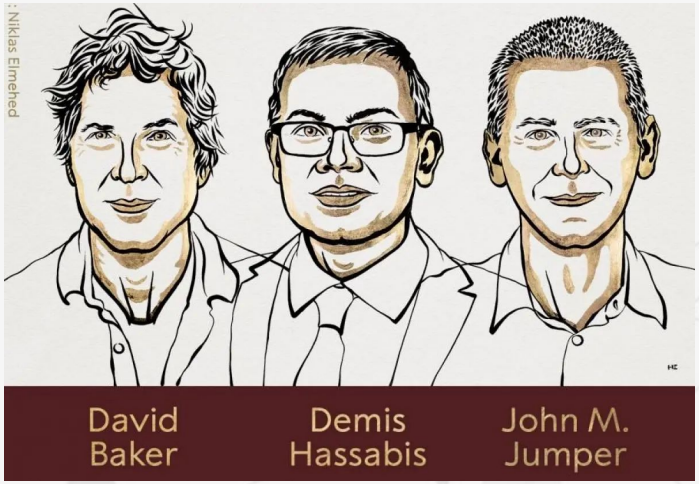
DATA SCIENCE
UNIVERSITY OF PADOVA

ALPHAFOLD

Master of Science in Data Science

Damiano Piovesan

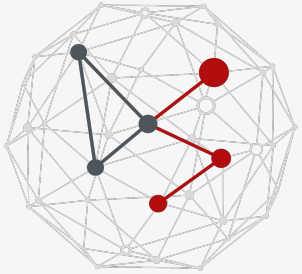
2024 Nobel Prize in Chemistry



“for computational protein design”

“for protein structure prediction”

DIPARTIMENTO
MATEMATICA



DATA SCIENCE
UNIVERSITY OF PADOVA

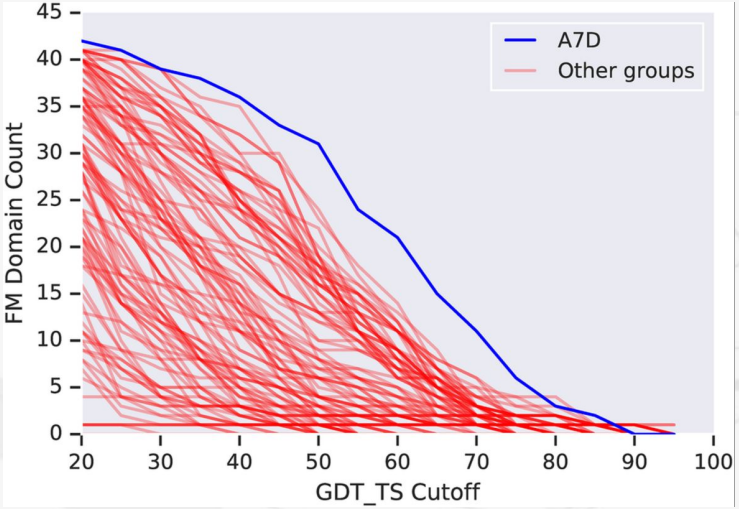
ALPHAFOLD 1

Master of Science in Data Science

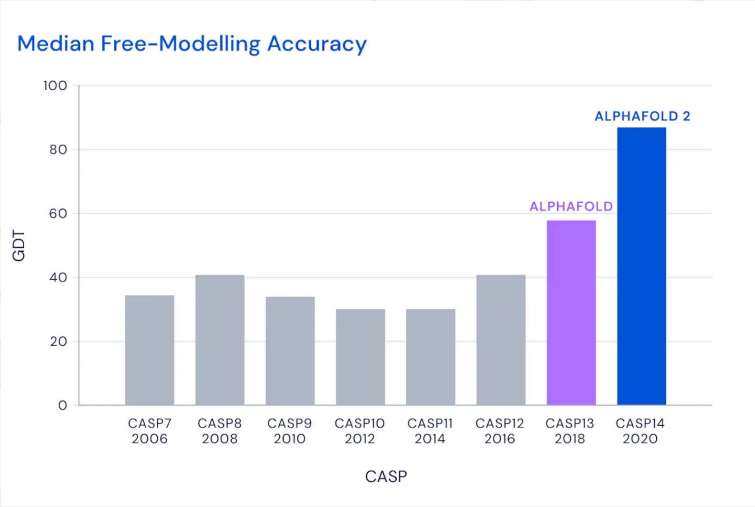
Damiano Piovesan

AlphaFold's success in CASP

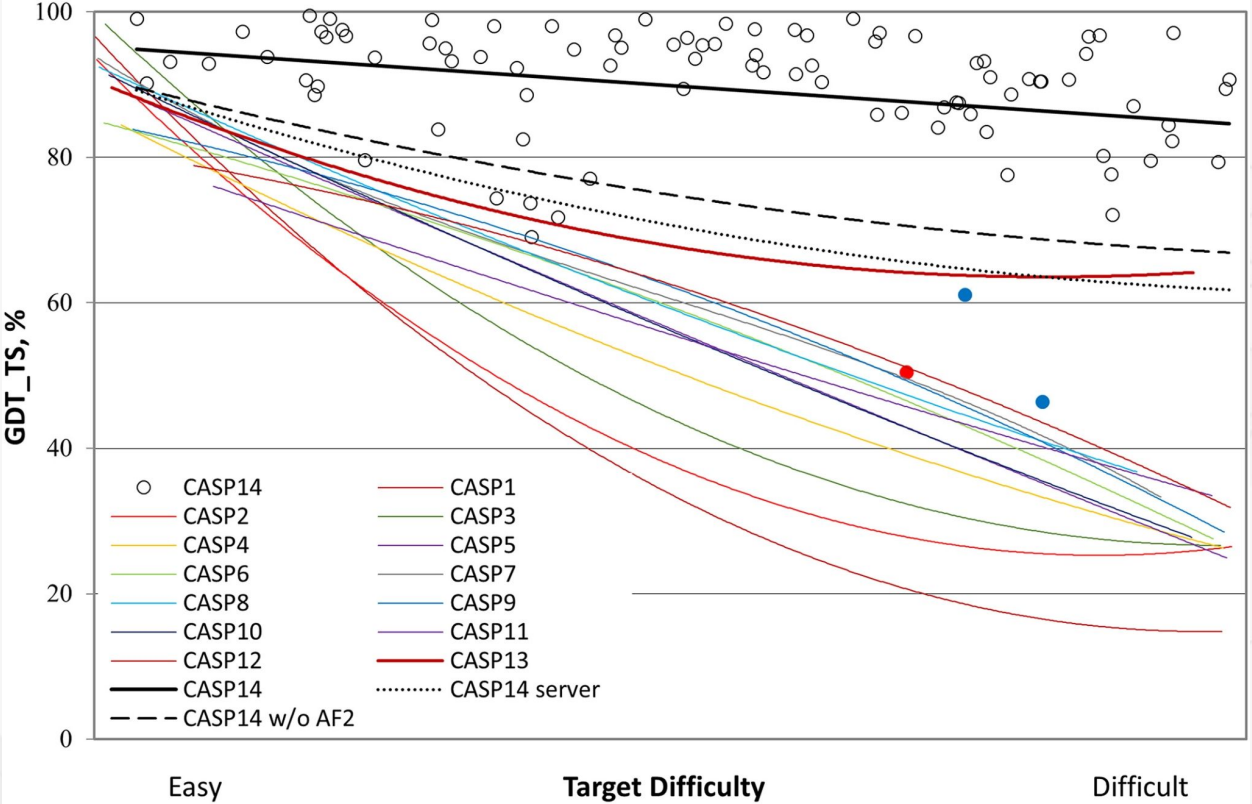
- Google DeepMind **AlphaFold (A7D)** in 2018's CASP13
- **First place** but by a small margin
- Predictions not accurate enough, → structure prediction problem was not considered solved



CASP13 (2018)

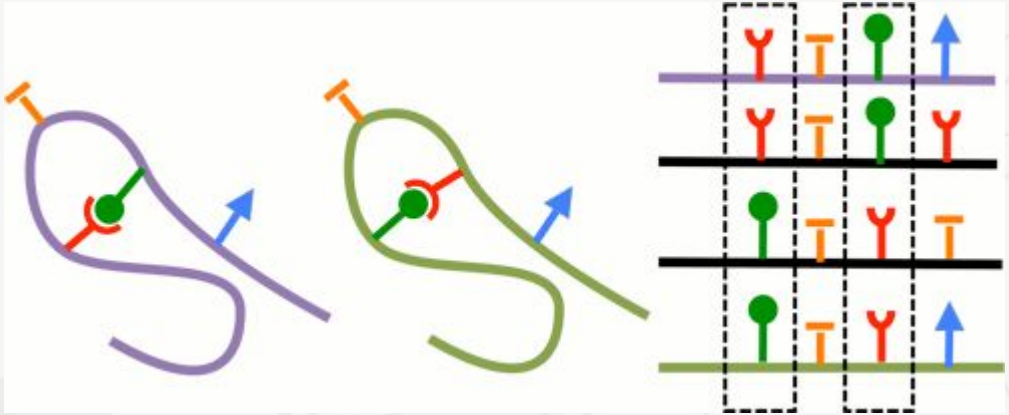


CASP14 (2020)


100
80
60
40
20
0
GDT_TS, %
CASP14
CASP1
CASP2
CASP3
CASP4
CASP5
CASP6
CASP7
CASP8
CASP9
CASP10
CASP11
CASP12
CASP13
CASP14
CASP14 server
CASP14 w/o AF2
Easy
Target Difficulty
Difficult

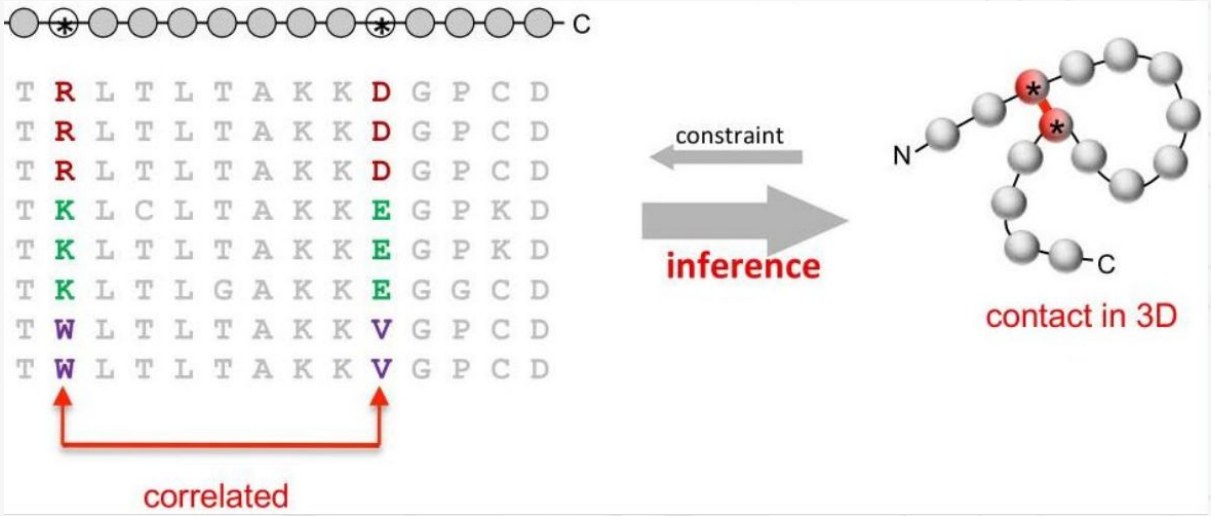
Covariance patterns

- **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.

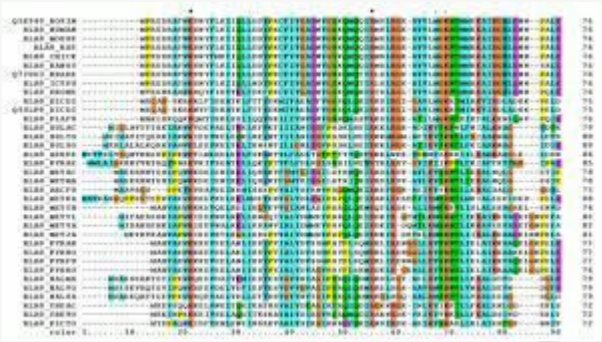


Foldit blog - https://fold.it/portal/node/2008706

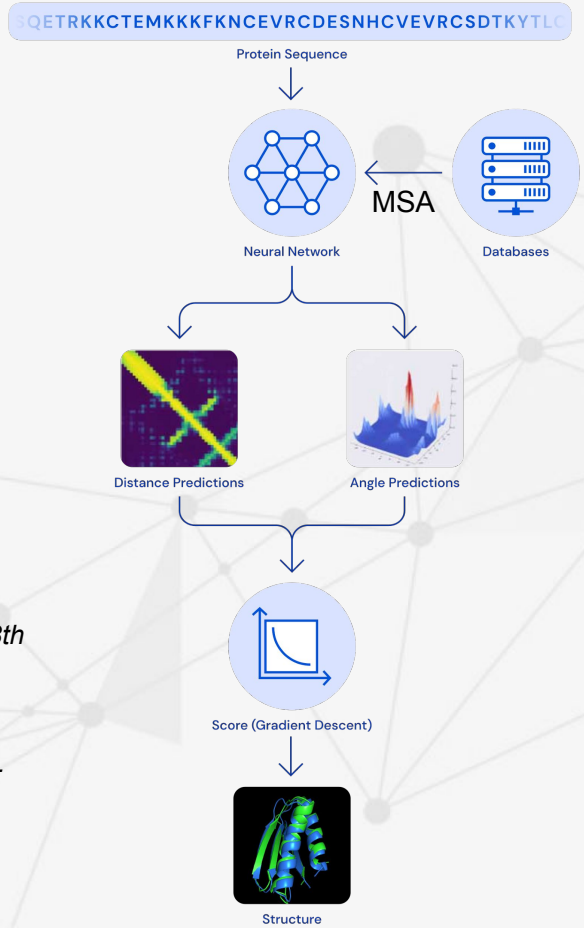
Covariance patterns



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

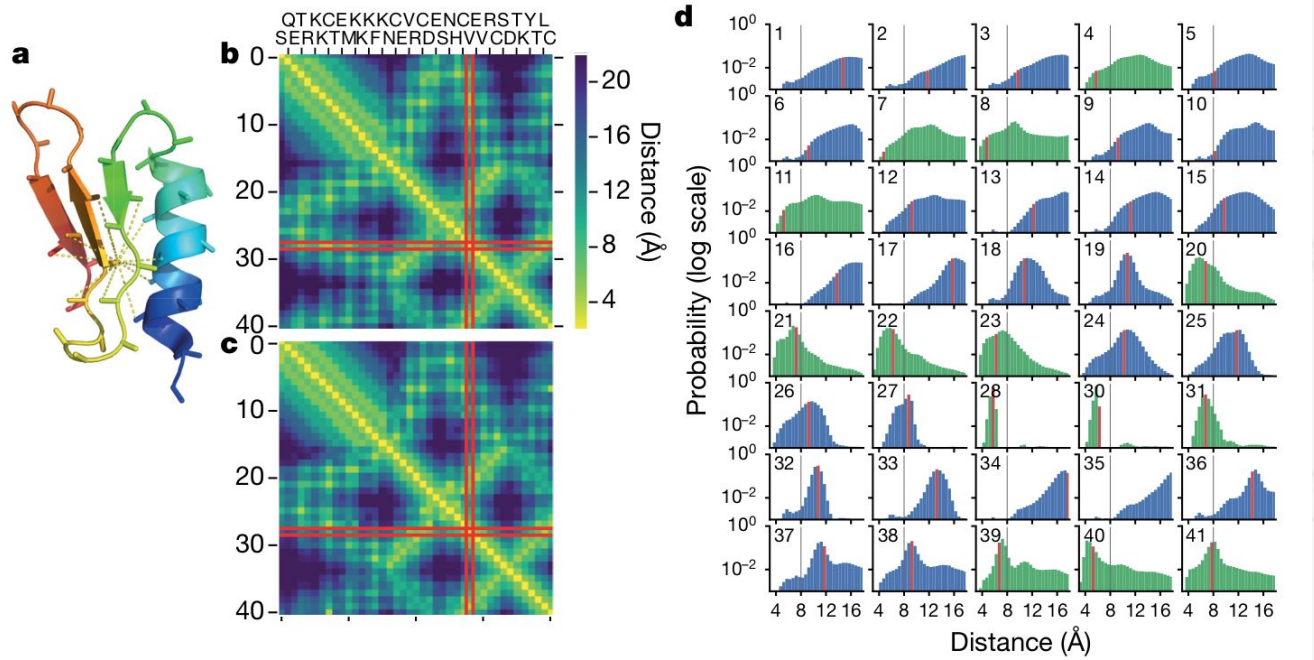
DeepMind blog https://deepmind.com/blog/alphafold/

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

Probability of the distance prediction between one residue and all others



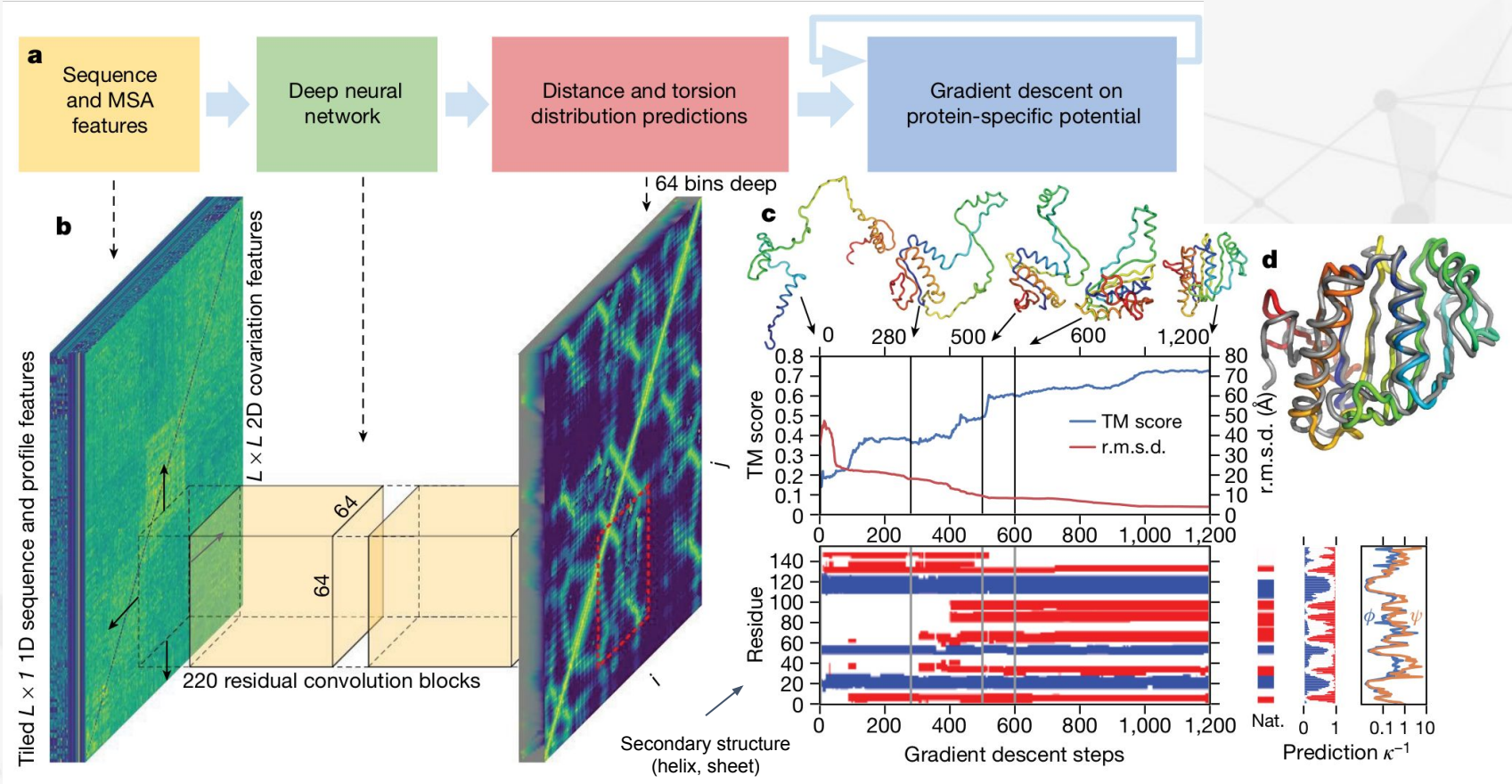
Range of values between 2 and 20 Å

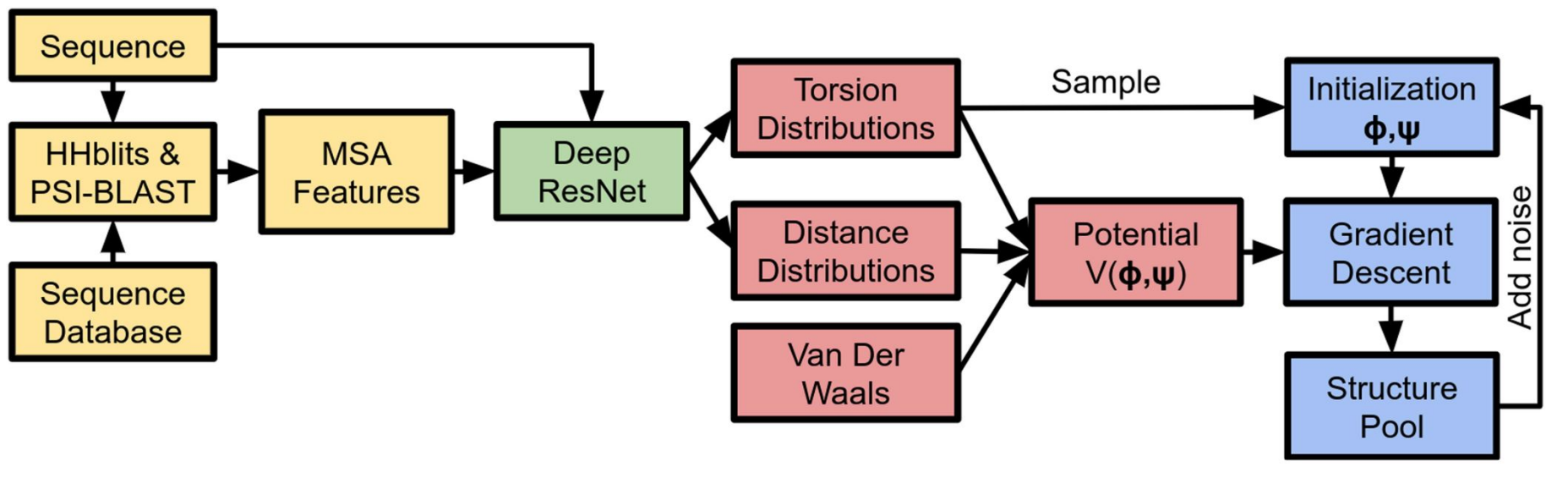
Red lines indicate the true distance

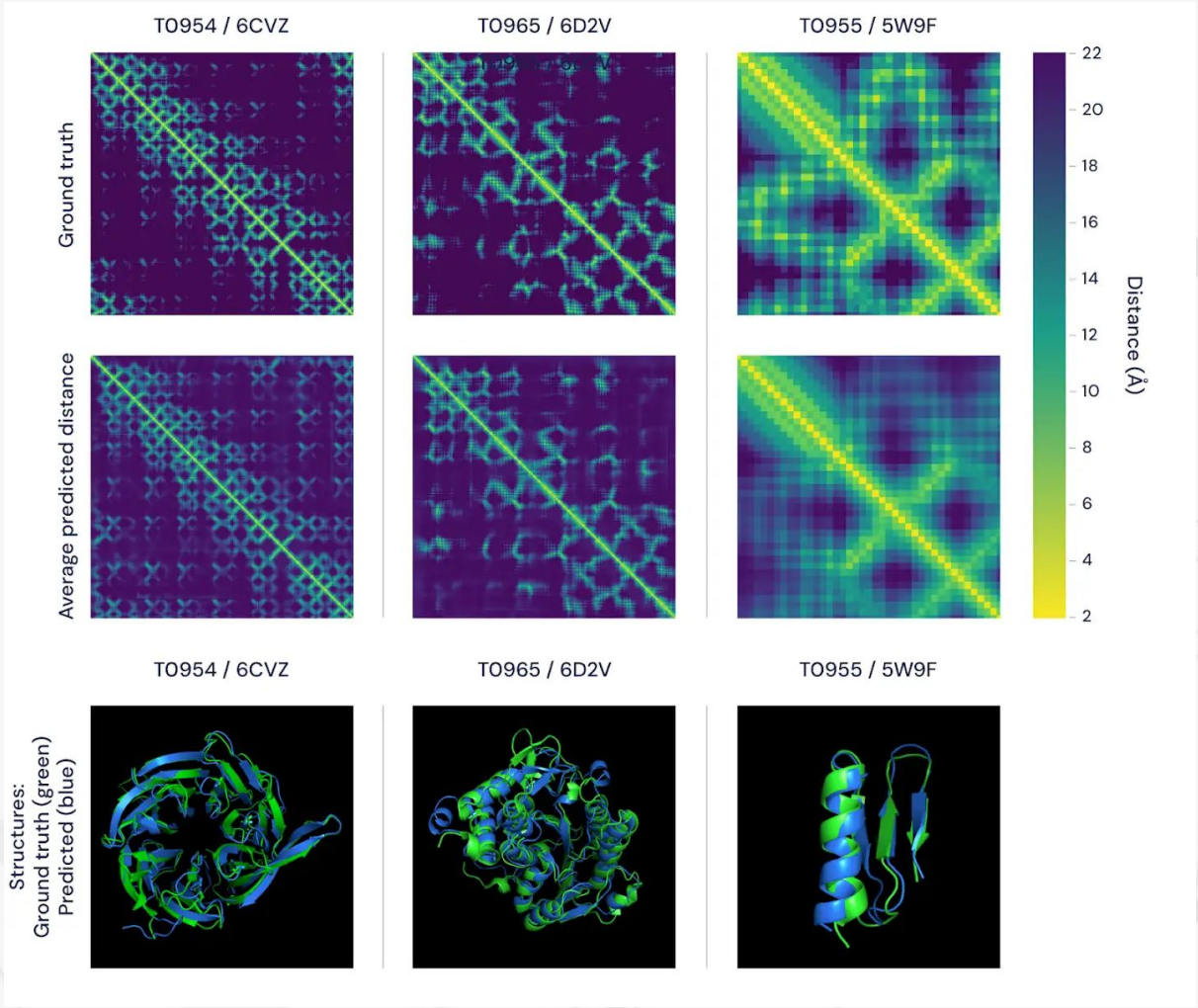
Distribution in green are true contacts

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

DIPARTIMENTO MATEMATICA

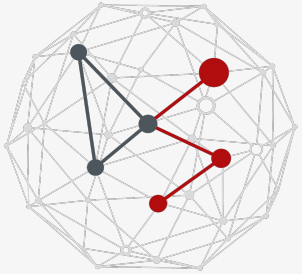



Sequence
HHblits & PSI-BLAST
Sequence Database
MSA Features
Deep ResNet
Torsion Distributions
Distance Distributions
Van Der Waals
Sample
Potential V(ϕ,ψ)
Initialization ϕ,ψ
Gradient Descent
Structure Pool
Add noise


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)

UNIVERSITÀ
DEGLI STUDI
DI PADOVA

DIPARTIMENTO
MATEMATICA



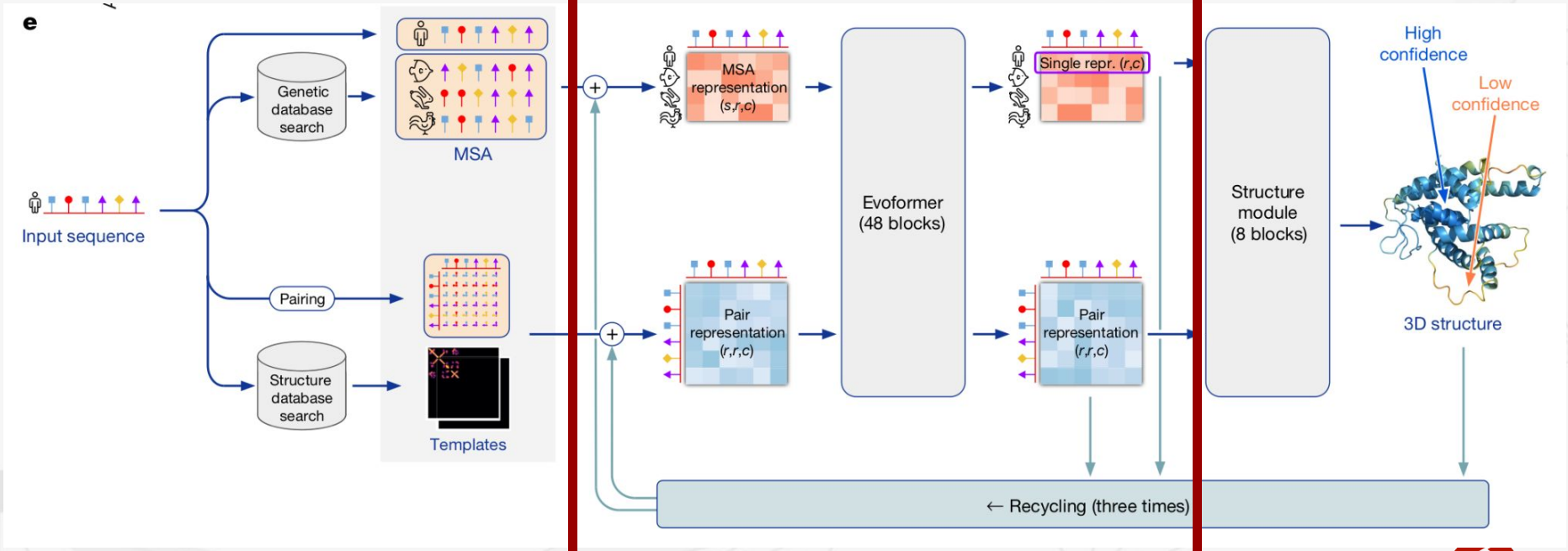
DATA SCIENCE
UNIVERSITY OF PADOVA

ALPHAFOLD 2

Master of Science in Data Science

Damiano Piovesan

End-to-end network



Harness additional sequence-based priors

Learn structural features

Generate structures

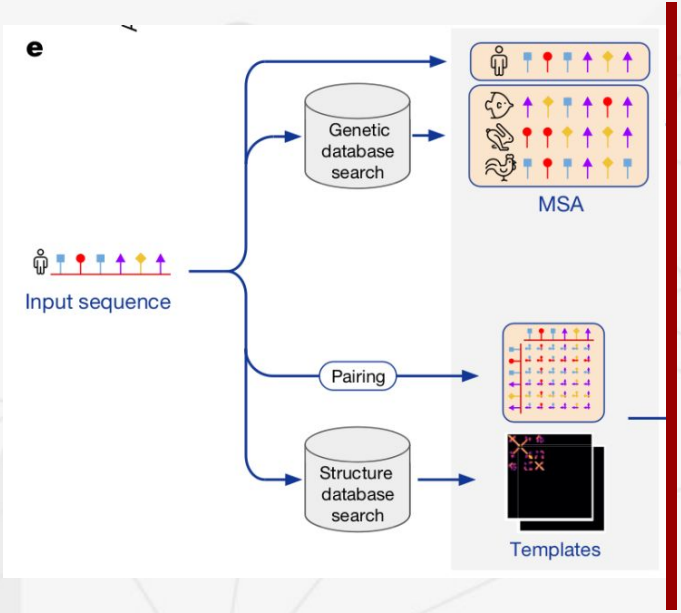
Inputs

MSA Representation

- stores each residue position across homologous sequences, capturing evolutionary signals
- Identify parts that are more likely to mutate
- Identify correlations (rows and columns)

Pair Representation

- models pairwise residue interactions
- central in predicting the spatial proximity and orientation between residues
- uses templates when available (proteins with a similar structure)



DIPARTIMENTO MATEMATICA

AlphaFold2 training data

PDB (X-ray crystallography, NMR, cryo-EM)

- Resolution < 9 Å
- 40% sequence identity clusters with MMSeqs2
- Removed sequences with >80% of a single amino acid

Self-Distillation. High-confidence predictions from earlier model iterations as pseudo-labels

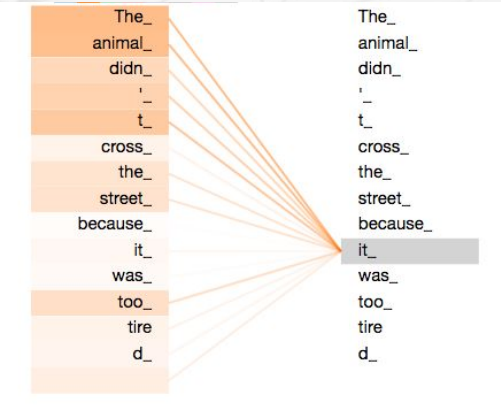
- An "undistilled" model trained exclusively on PDB to predict the structure of ~300K diverse sequences from Uniclust30
- New distillation dataset created (high-confident predictions + PDB)
- Train from scratch, using split (75% self-distillation dataset; 25% PDB)

Review of transformers

- **Deep learning** model architecture that uses **self-attention** mechanisms to process and **generate** sequences of data efficiently
- Self-attention allows the model to weigh the **importance** of each **part of an input sequence** when processing it, making it effective for tasks like translation and text generation

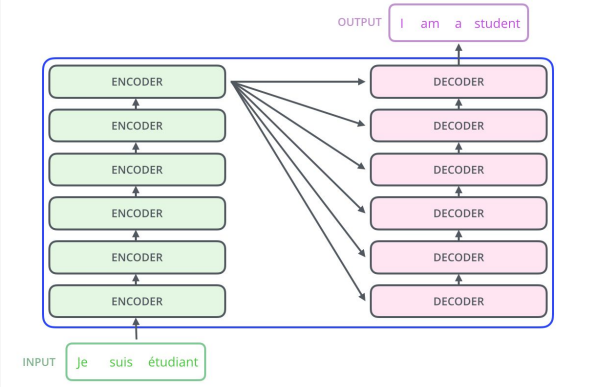
For each word

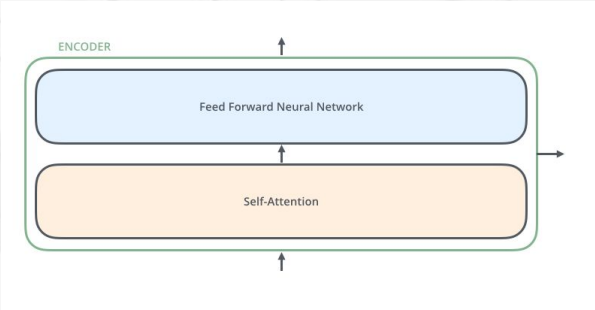
- Input vector (Embedding)
- **Q**ueries: tokens to care about
- **K**eys: characteristics used for matching
- **V**alues: information each token needs to transmit



Evoformer

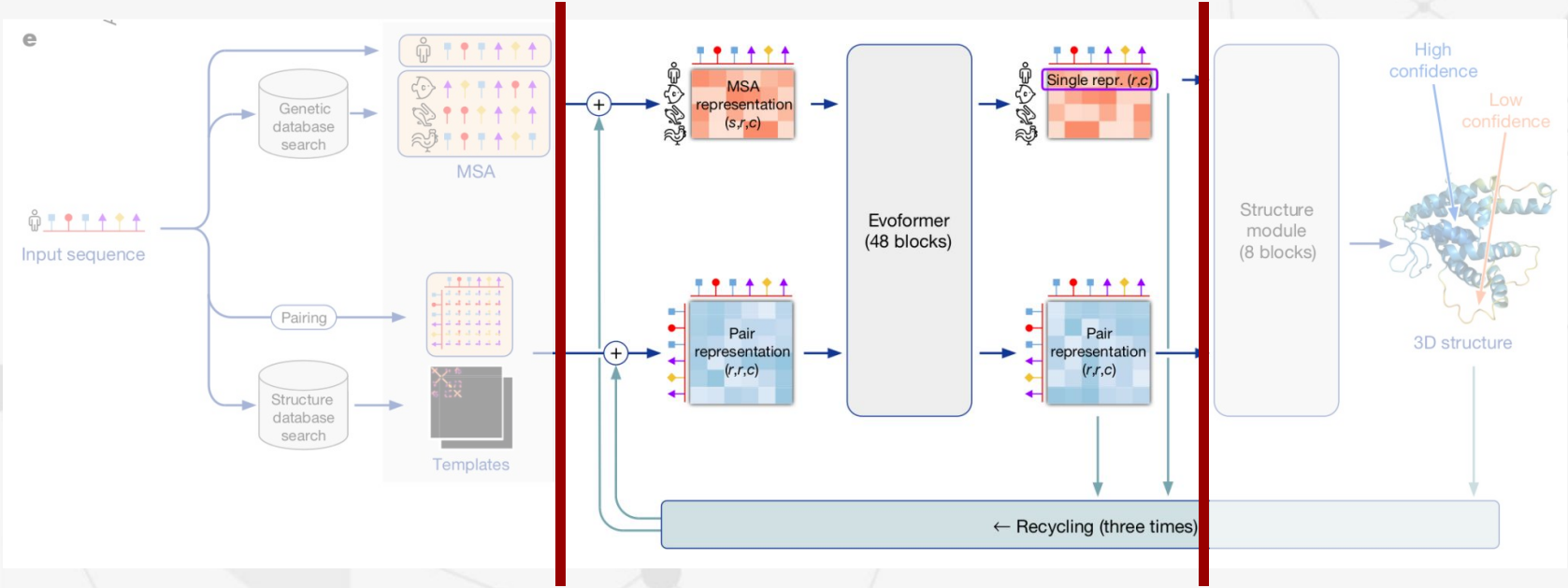
- Variation of the “transformer” deep learning architecture
- Invented by Google Brain
- Based on “Attention” → Identify which parts of the input are more important
- Attention matrix → quadratic memory cost
- https://jalammar.github.io/illustrated-transformer/



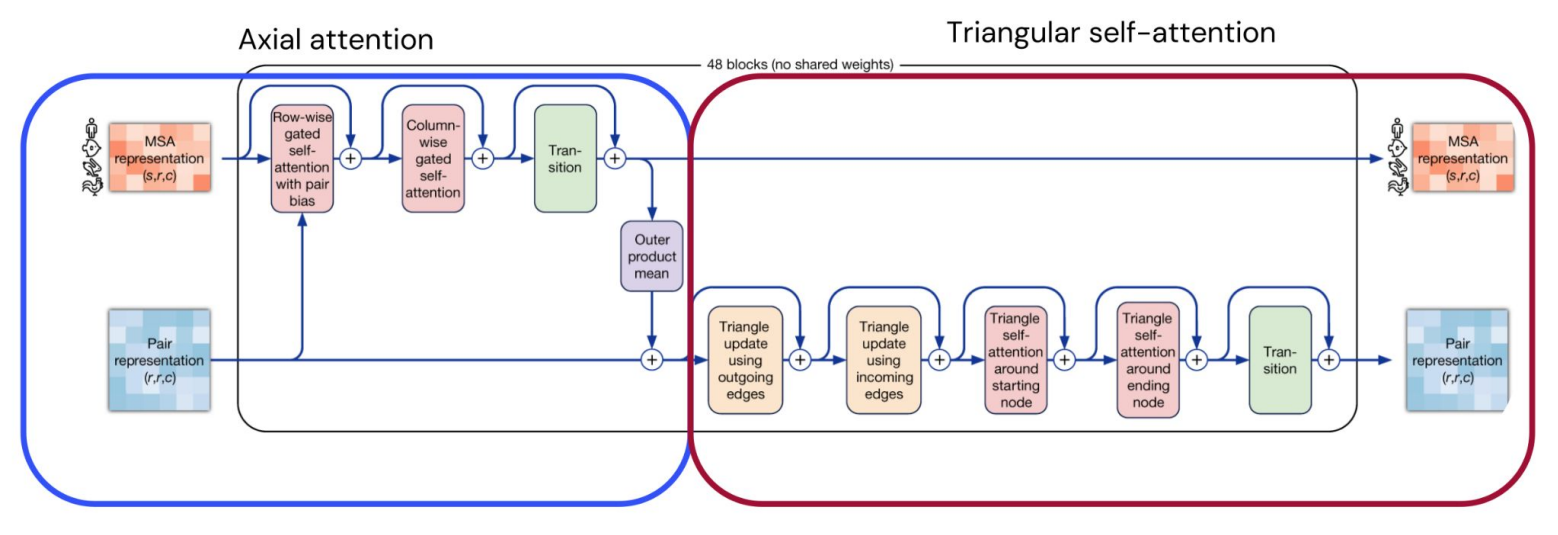


Evoformer

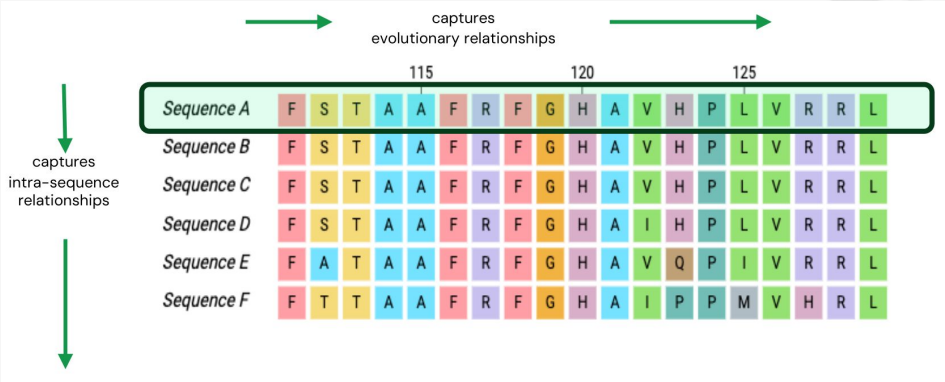
Refine both the MSA and pair representations with clear communication channels between the two sections that update both representations



Inputs


Axial attention
Triangular self-attention
48 blocks (no shared weights)
MSA representation (s,r,c)
Row-wise gated self-attention with pair bias
Column-wise gated self-attention
Tran-sition
Outer product mean
Pair representation (r,r,c)
Triangle update using outgoing edges
Triangle update using incoming edges
Triangle self-attention around starting node
Triangle self-attention around ending node
Tran-sition
MSA representation (s,r,c)
Pair representation (r,r,c)

Encoding evolution with attention

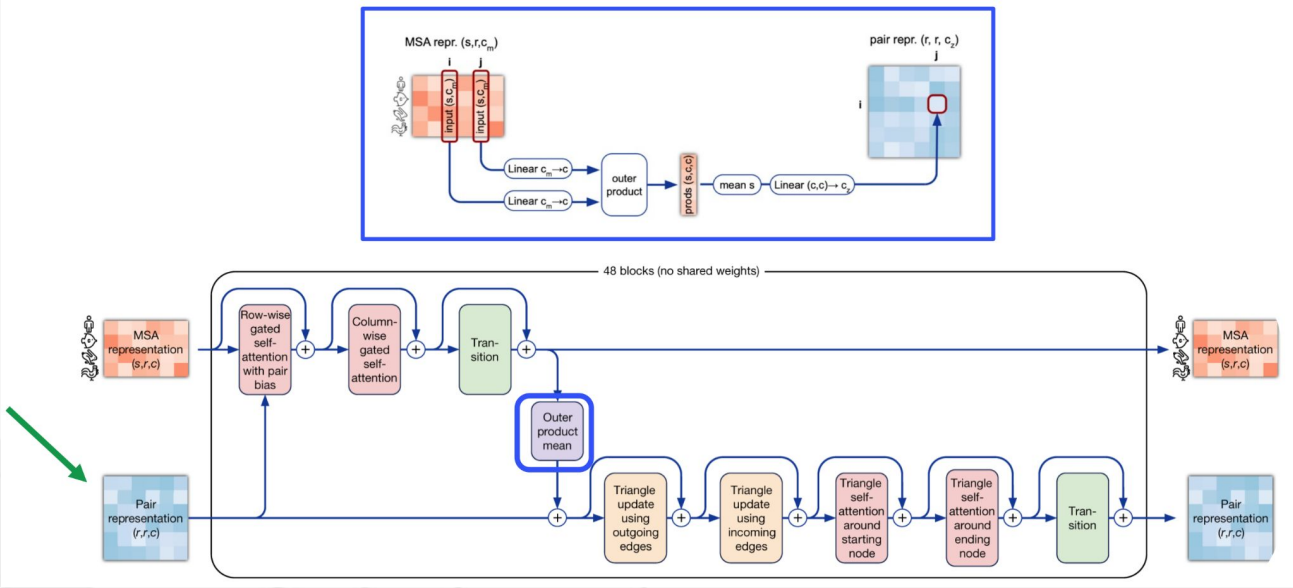


MSA transformer

- Row-wise, identify which pairs of amino acids are more related
- Column-wise, determining which sequences are more informative

MSA-to-pair-update

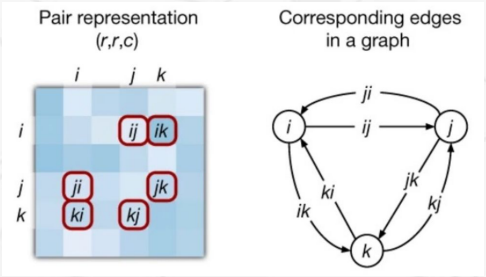
- Translates evolutionary information (from the MSA representation) into pairwise residue relationships
- Columns are averaged over the sequences and projected to a dimension C_z to obtain an update for entry *ij* in the pair representation

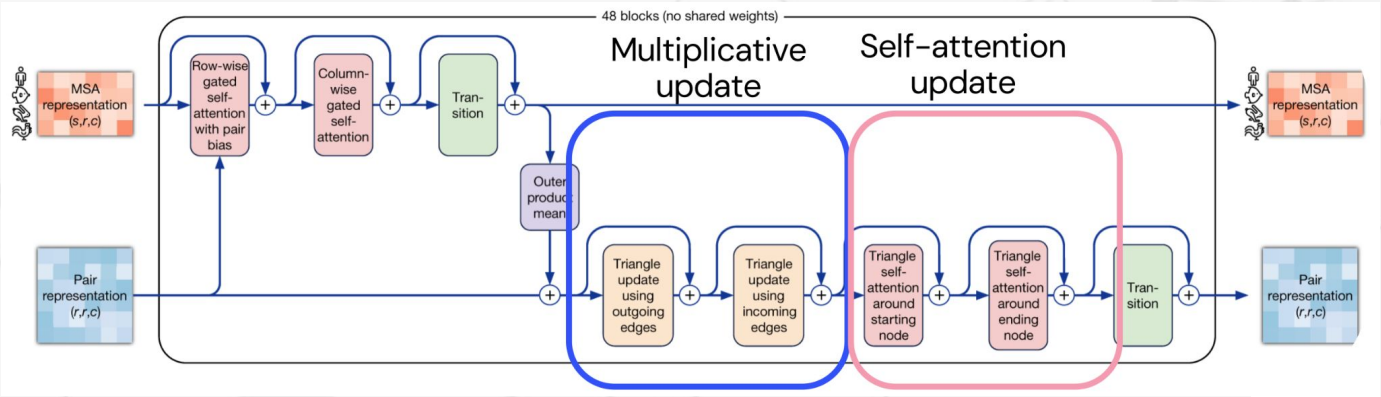


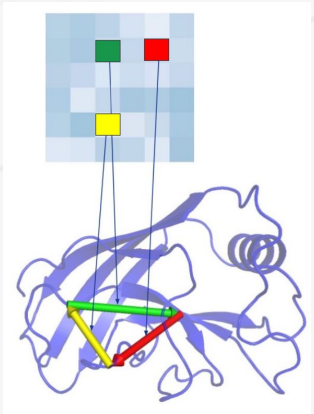
Triangular self-attention: encoding graph structure

For a pairwise description of amino acids to be representable as a single 3D structure, many constraints must be satisfied

- Designed to propagate geometric constraints and ensure consistency in residue pair relationships
- Graph structures represent these edge relationships well → Aim: update transformer learning to implicitly capture graph relationships





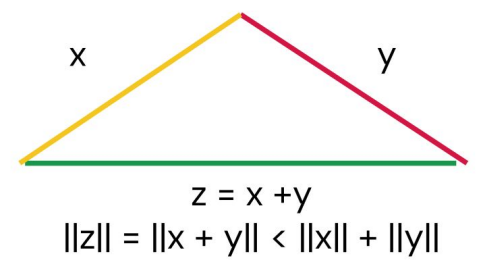


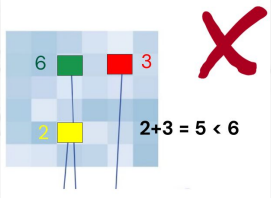
Triangular self-attention: enforcing 3D common sense

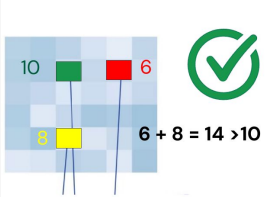
Basic idea: adhere to triangle inequality

- Note: the inequality refers to **distances** between coordinates rather than the 3D coordinate positions

Update an edge based on all the triangles it is involved in

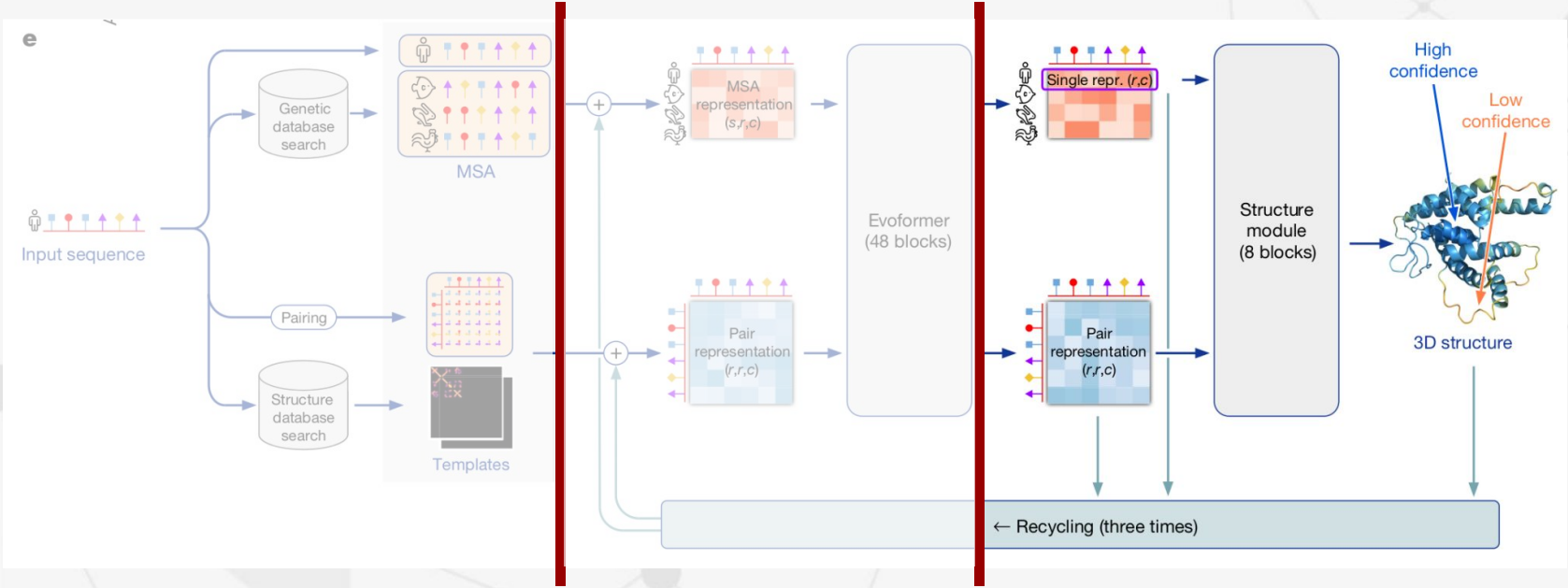






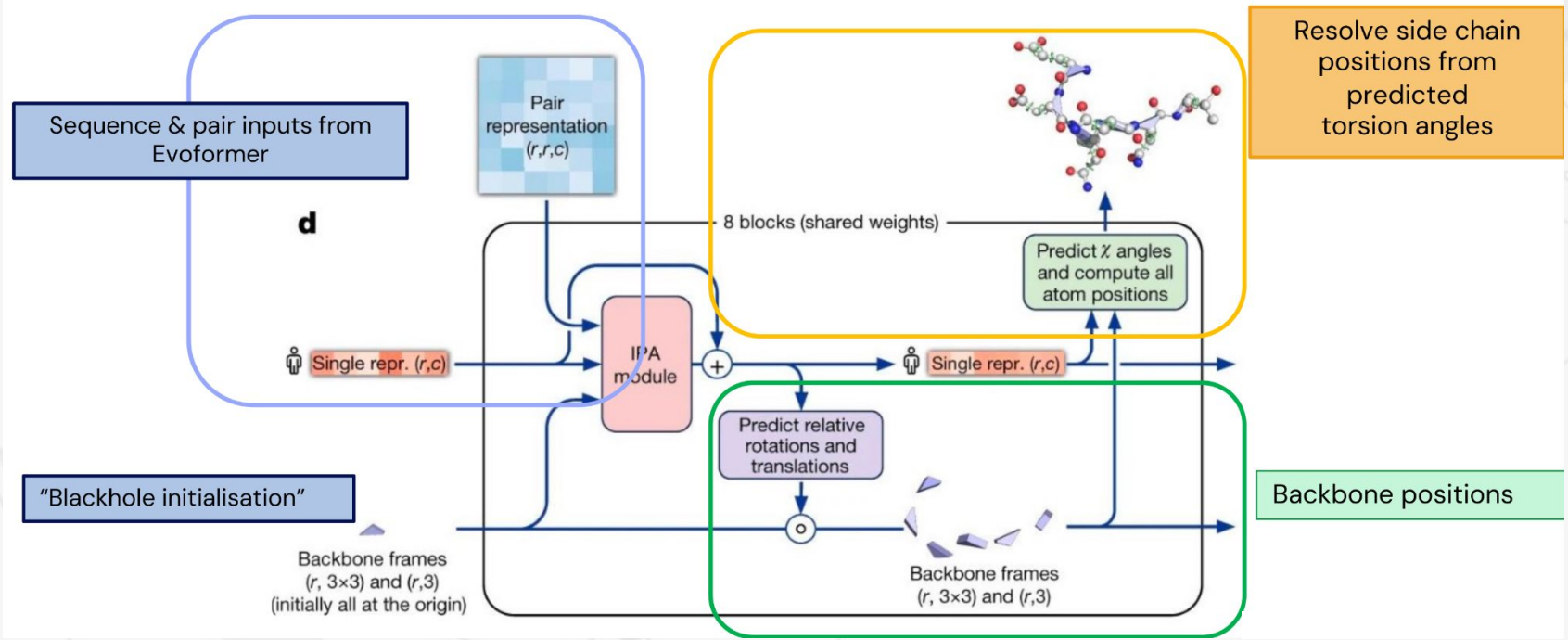
Structure module

Maps the abstract representation of the protein structure (created by the Evoformer stack) to concrete 3D atom coordinates



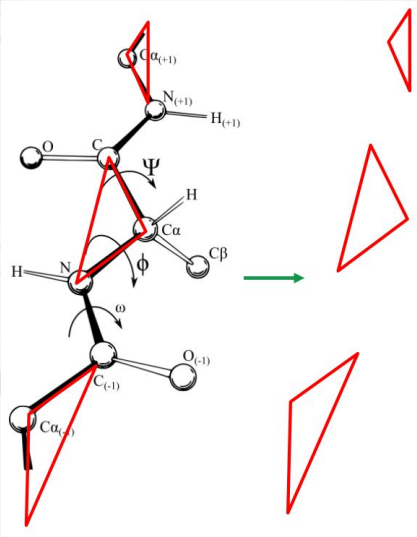
Structure module: Overview

Converting 2D pairwise and 1D residue representations from the Evoformer into a 3D atomic structure



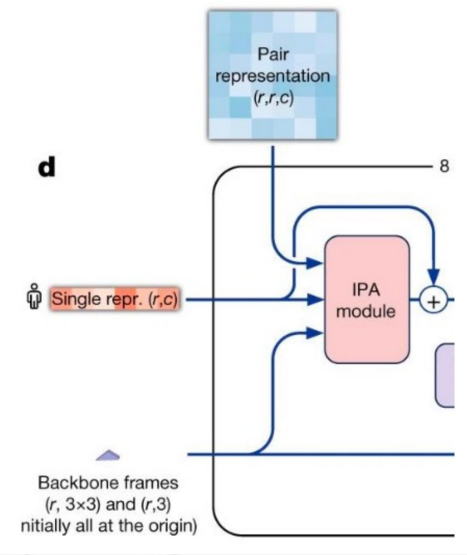
3D residue gas

- Backbone is a series of tiled triangles in 3D
- “Chunk up” the structure and consider each residue as “independently floating”
    - Represent each residue as a triangle with {*N, C_α, C*} as vertices
- All frames are initialized at the origin, “black hole initialization”. No enforcement of chain



Invariant Point Attention (IPA)

- **Geometry-aware** attention operation. A specialised attention mechanism that processes both sequence and 3D spatial information
- Refines the orientation and position of residues by considering their geometric context. Bias by attending to frame transformations for residues *i* and *j*
- Each residue in the protein has a set of reference points defined in its local coordinate frame. These points are **invariant to rotation and translation**, meaning that they do not change if the protein as a whole rotates or moves in space
- Next: Predict backbone frames from IPA module output



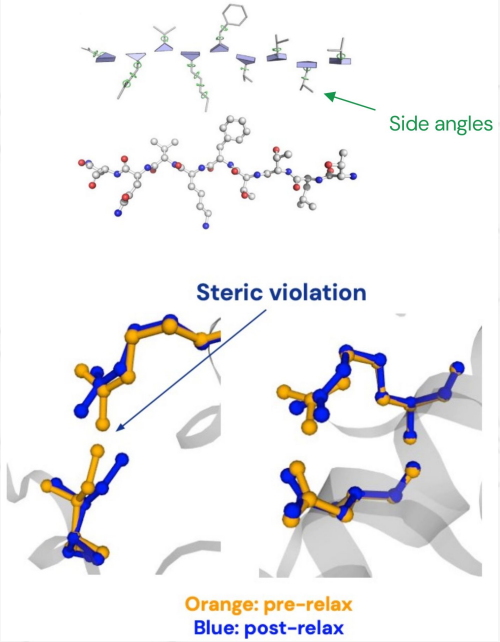
Placing side chain angles

- To place the side chains, AF2 also predict the torsion angles
- The torsion angles are the only degrees of freedom, while all bond angles and bond length are fully rigid
    - Consider proteins as having “moveable joints”, while other components are kept rigid
- A shallow ResNet predicts the torsion angles

Refinement in structure module

- In order to resolve any remaining structural violations and clashes, models are relaxed by an iterative restrained energy minimisation procedure
- Violations of these constraints are resolved with coordinate-restrained gradient descent, Amber ff99SB force field with OpenMM
    - Keep the system near its input, restraints are applied independently to heavy atoms
    - Determine which residues still contain violations → remove restraints from all atoms within these residues → Perform restrained minimisation again
    - This process is repeated until all violations are resolved

The end result of iterative refinement is not guaranteed to obey all stereochemical constraints

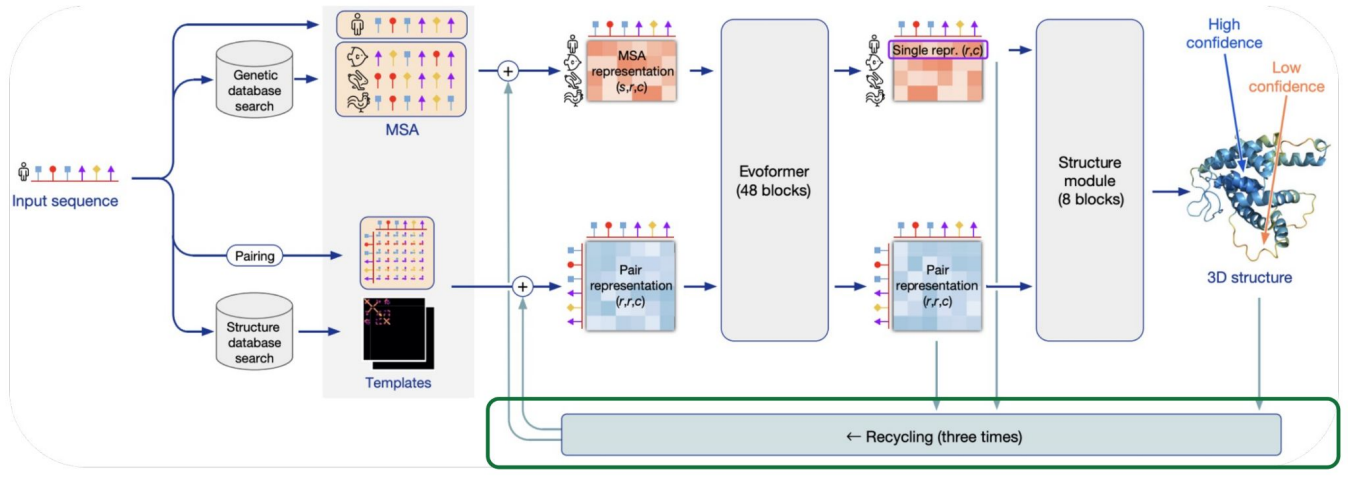


Some outputs

- **Predicted 3D Coordinates**. The 3D atomic coordinates of the predicted protein structure. This includes the positions of C_α atoms (for the backbone) and the side-chain atoms for each amino acid
- **Predicted Aligned Error (PAE)** in ‘pickle’ format (.pkl, contains a Python dictionary) - not readable from the command line. (e.g. results_model_1_pred_0.pkl)
- Per-Residue Local Distance Difference Test (**pLDDT**). Provides a per-residue measure of the model's confidence in the local 3D structure

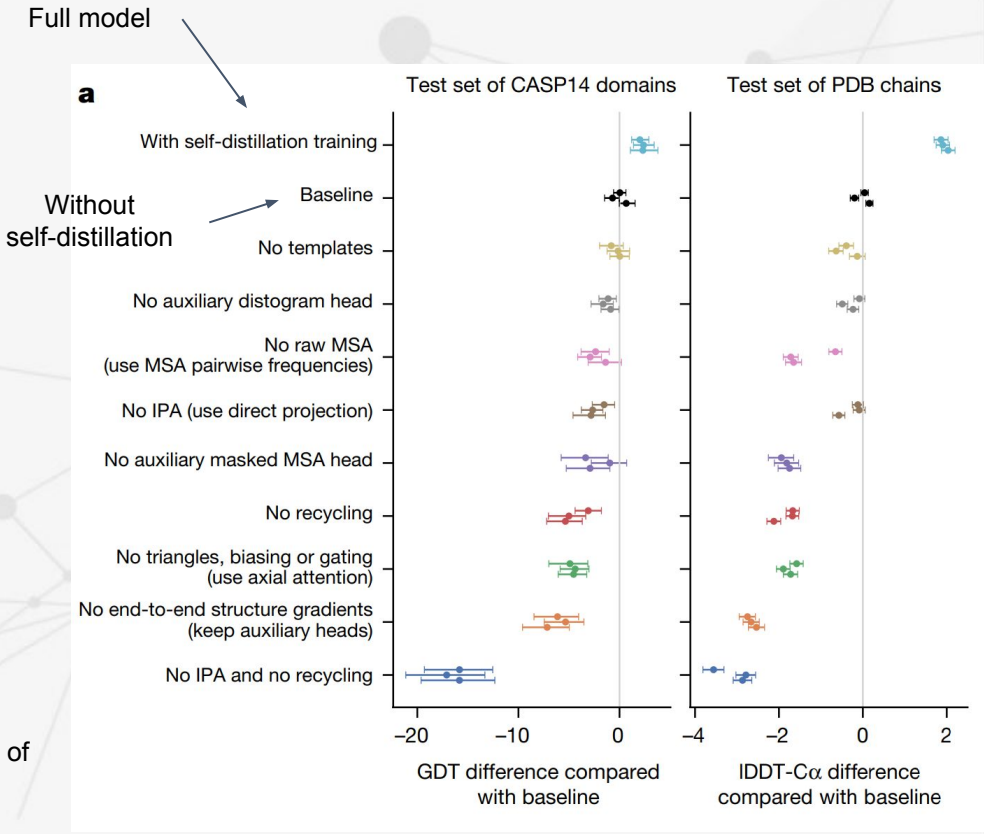
Recycling

- The whole network is executed sequentially, where the outputs of the former execution are recycled as inputs for the next execution
- The first row of the MSA representation and the full pair representation are updated by the recycled outputs from the previous iteration - for the first iteration the recycled outputs are initialized to zero



Ablations, which part matters?

- “Knockdown study” to see which network components are most important
- No single improvement is dominant



Ablation results on two target sets: the CASP14 set of domains (n = 87 protein domains) and the PDB test set of chains with template coverage of ≤30% at 30% identity (n = 2,261 protein chains)

Strengths of AlphaFold

- AlphaFold can predict protein structures **based only on amino acid sequences**
  - It is a very good starting point for creating testable hypotheses
  - AlphaFold models are routinely used as starting models for X-ray crystallography and cryo-EM

- If available, **AlphaFold can use templates** from the PDB
  - But even if templates are available, AlphaFold may discard them and use only sequence data

- AlphaFold produces three independent outputs
  - Predicted atomic coordinates (PDB and mmCIF)
  - Confidence measure per amino acids (pLDDT)
  - Predicted Aligned Error (PAE)

Limitations

- AlphaFold **only accepts the 20 standard amino acids in its input**
- AlphaFold (by default) predicts five models per run: → However, these models are generally very similar. In other words, AlphaFold usually cannot predict **conformational variability** in a protein
- Cannot predict assemblies (AlphaFold-Multimer was trained to do this)
- **Fails to predict the effects of point mutations**
- Not designed to fold nucleic acid structures or model protein-DNA and protein-RNA complexes
- Not designed to predict the **binding of ligand molecules** or post-translational modifications (AlphaFill adds small molecules)
- Not aware of the membrane plane for the transmembrane proteins
- Rarely predicts correct antigen-antibody interactions
- Reduced performance for **orphan proteins**

AlphaFold 2

- Trained on **16 TPUv3s**
  - 128 TPUv3 cores or roughly equivalent to **~100-200 GPUs**
  - Run over **a few weeks**

AlphaFold 2

- **Physical insights** are built into the **network structure**, not just a process around it
- **End-to-end** system directly producing a structure instead of inter-residue distances
- Inductive biases reflect our knowledge of protein physics and geometry
  - The **positions of residues** in the sequence are de-emphasized
  - Instead residues that are **close in the folded protein** need to communicate
  - The network iteratively learns a graph of which residues are close, while reasoning over this implicit graph as it is being built

DIPARTIMENTO MATEMATICA

References

- BLOG - Oxford Protein Informatics Group

  https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/

  - The Illustrated Transformer

    https://jalammar.github.io/illustrated-transformer/

- ColabFold

  https://colab.research.google.com/github/sokrypton/ColabFold/blob/main/AlphaFold2.ipynb

- Nature paper

  https://www.nature.com/articles/s41586-021-03819-2