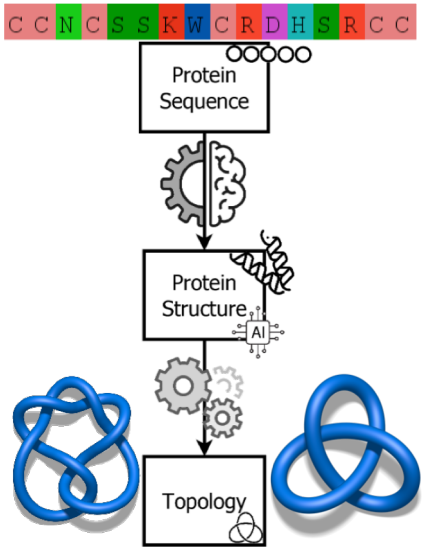
Unveiling Protein Structures: AlphaFold, ESMFold, and other models in Protein Structure Prediction

Joanna Sulkowska

Centre of New Technologies, University of Warsaw

DeepLearning inLifeSciences



AlphaFold Protein Structure Database

ESM Metagenomic Atlas

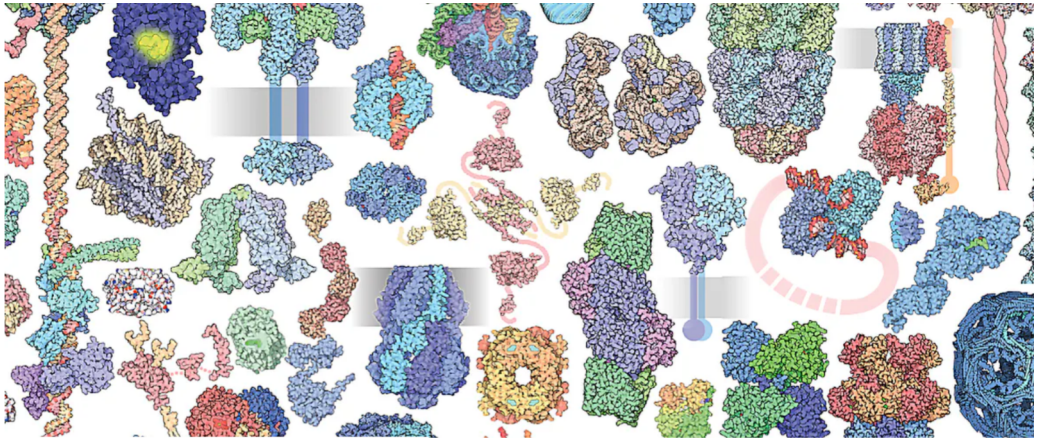
Presentation plan

- Basic information about proteins
- Experimental method to determine 3D protein structure
- Theoretical methods to predict protein 3D conformation
- Intorduction to Neural networks approach in protein 3D prediction
  - AlphaFold: Revolutionizing Protein Structure Prediction
  - ESMFold: Advancements in Protein Structure Prediction
  - RosettaFold and Other Modeling Approaches

Basic information about protein structure

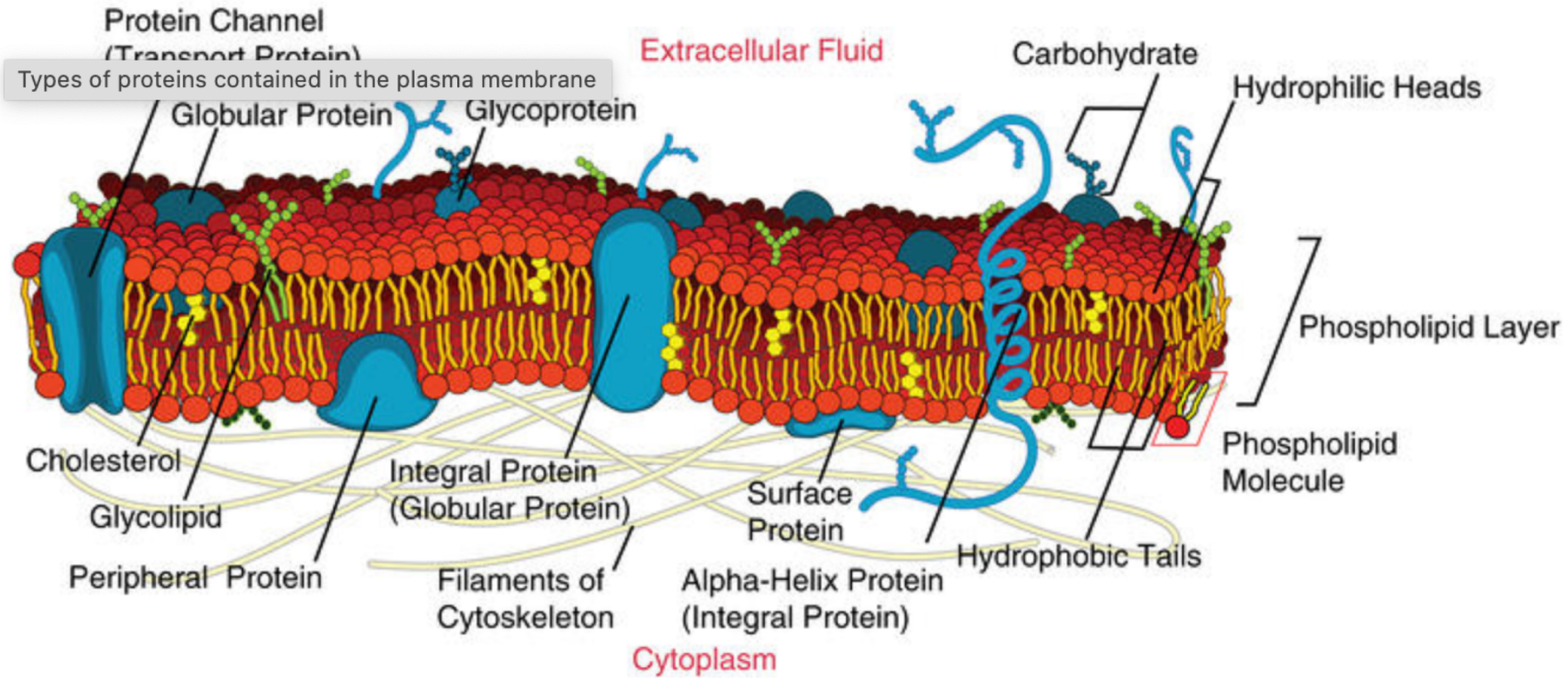
- Introduction to protein structure: primary, secondary, tertiary, and quaternary structures.
- Type of proteins: globular, transmembrane …. disorde proteins
- Proteins foding concept
- Importance of understanding protein structure in drug discovery, disease understanding, and bioengineering.
- Overview of experimental methods of protein structure prediction: X-ray crystallography, NMR spectroscopy, cyro-EM.
- Limitations of traditional methods: time-consuming, expensive, and not suitable for all proteins.
- Theoretical approach: homology modelling, co-evolution, DCA

- Proteins are large, complex molecules essential to all of life. Nearly every function that our body performs (e.g. contracting muscles, sensing light, or turning food into energy), relies on proteins, and how they move and change.
- **What any given protein can do (largely) depends on its unique 3D structure.**
- The recipes for those proteins, called genes, are encoded in our DNA and generated by Ribosome. Many diseases and deaths for an organism, are fundamentally linked to malformed proteins.
- Proteins are composed of **chains of amino acids** (also referred to as amino acid **residues**). But DNA only contains information about the sequence of amino acids, not how they fold into shape.



https://theconversation.com/what-is-a-protein-a-biologist-explains-152870

Location, shape and size of proteins

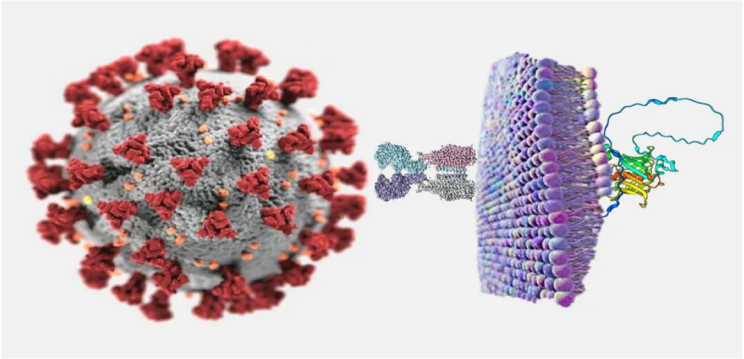


https://bio.libretexts.org/Bookshelves/Introductory_and_General_Biology/Introductory_Biology_%28CK-12%29/02%3A_Cell_Biology/2.06%3A_Membrane_Proteins

Protein folding or prediction

"I think that we shall be able to get a more thorough understanding of the nature of disease in general by investigating the molecules that make up the human body, including the abnormal molecules, and that this understanding will permit...the problem of disease to be attacked in a more straightforward manner such that new methods of therapy will be developed."

-- Linus Pauling, 1960

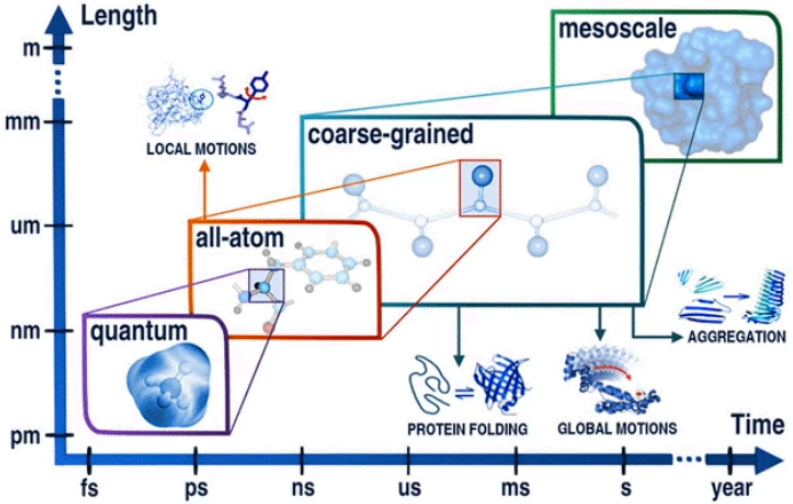


- Knowledge of the 3D structure help in designing new therapeutic agents, optimizing enzyme reactions, and enhancing biotechnological processes.
- Understanding the structure-function relationship enables targeted drug discovery.
- Enzyme engineering: helps in modifying enzymes for enhanced catalytic activity, e.g., engineering enzymes for biofuel production.

Protein folding or prediction

"I think that we shall be able to get a more thorough understanding of the nature of disease in general by investigating the molecules that make up the human body, including the abnormal molecules, and that this understanding will permit...the problem of disease to be attacked in a more straightforward manner such that new methods of therapy will be developed."

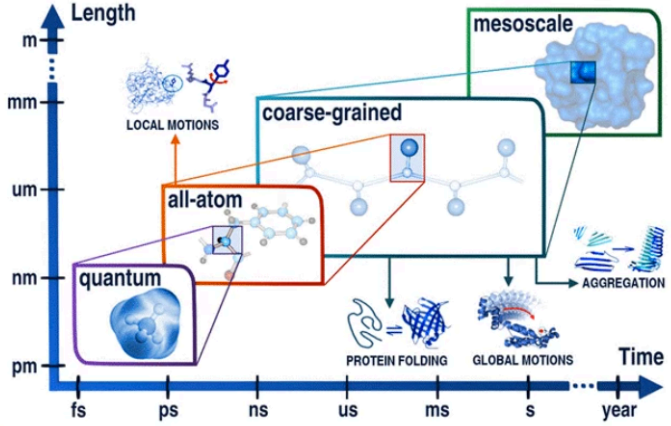
-- Linus Pauling, 1960



Protein folding or prediction

"I think that we shall be able to get a more thorough understanding of the nature of disease in general by investigating the molecules that make up the human body, including the abnormal molecules, and that this understanding will permit...the problem of disease to be attacked in a more straightforward manner such that new methods of therapy will be developed."

-- Linus Pauling, 1960

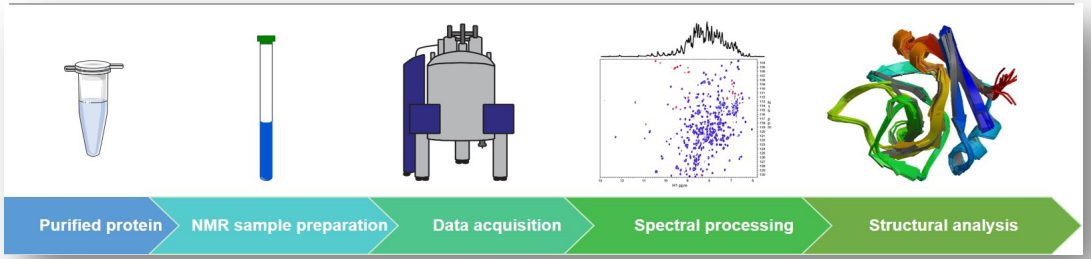


Can we bypass the protein folding mechanism?

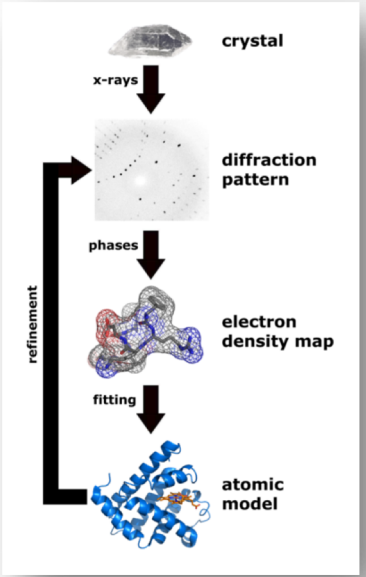
- Scientists have long been interested in determining the structures of proteins because a protein's form is thought to dictate its function.

- Once a protein's shape is understood, its role within the cell can be guessed at, and scientists can develop drugs that work with the protein's unique shape, two opitions:
  - traditional methods – are *time consuming, high cost*
  - AI methods as an alternative to this long and laborious process for difficult proteins.

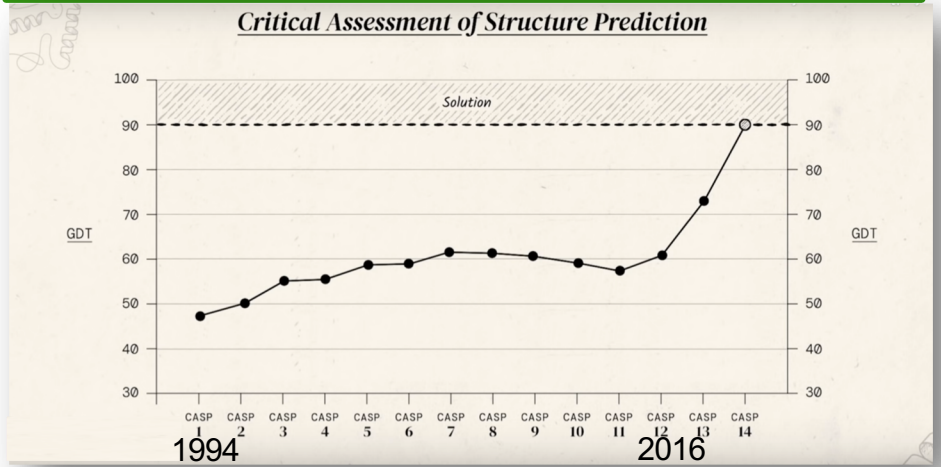
Why to determine 3D protein structure



- Protein NMR
- X-ray crystallography
- Cryoelectron microscopy
- PREDICTION

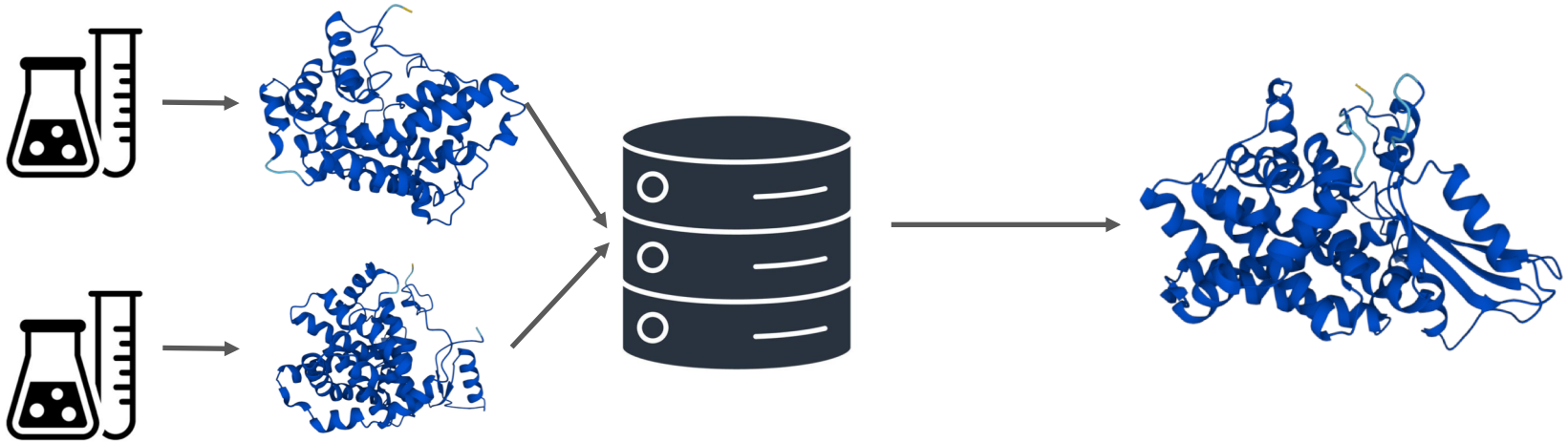


50-year-old grand challenge of protein structure prediction, often referred to as the **'protein folding problem'**



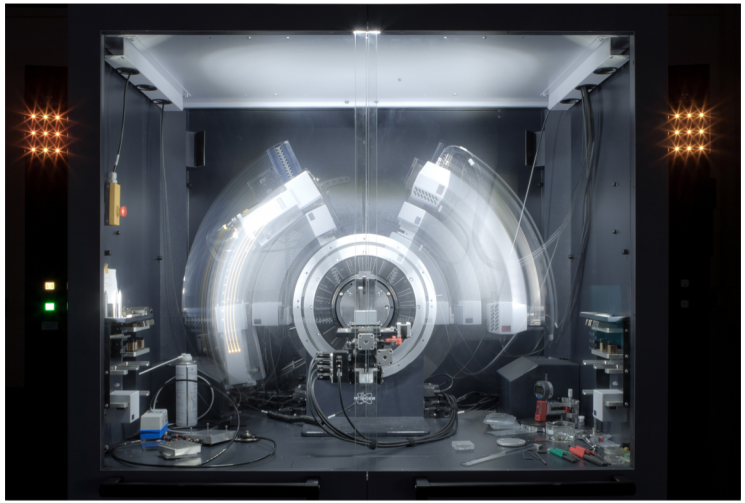
Methods of Protein Structure Prediction

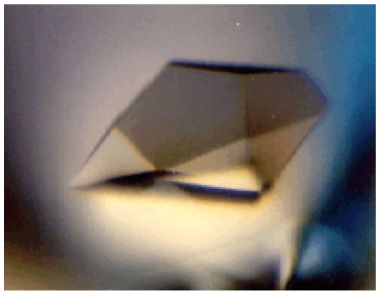
- **X-ray crystallography** – involves analyzing the diffraction pattern of X-rays by protein crystals, providing detailed atomic structure information.
- **NMR spectroscopy** – detects interactions between atomic nuclei in a magnetic field to determine protein structure in solution.
- Cyro-EM- is a cryomicroscopy technique applied on samples cooled to cryogenic temperatures, minimal amounts of biomolecules or complexes can be analyzed.
- **Homology modeling –** predicts protein structure by comparing target protein sequences to known structures, suitable for evolutionary related proteins.

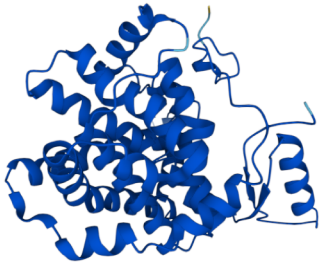


X-ray Crystallography

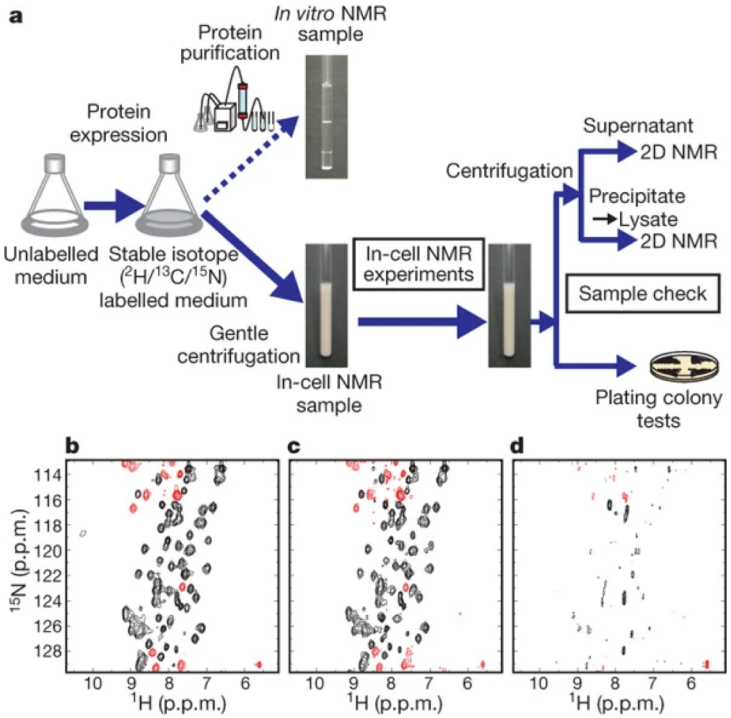
- X-ray diffraction involves exposing protein crystals to X-rays to produce diffraction patterns, which can be used to determine the atomic structure.
- The method has revolutionized structural biology by allowing scientists to obtain the 3D structures of proteins at atomic resolution.





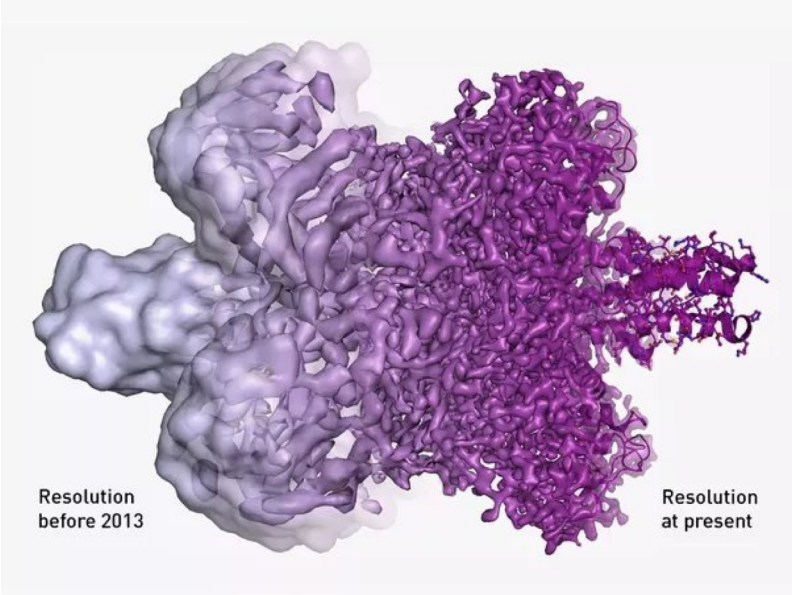


- NMR spectroscopy resolve protein structures in solution (buffer), allow to observe small changes in the conformation.
- Dedicated mainly to small proteins with a large number of flexible fragments.
- Limitations involve complexity of data interpretation and size restrictions.
- X-ray crystallography provides static protein structures but requires crystal formation, limiting analysis of dynamic behavior in solution.



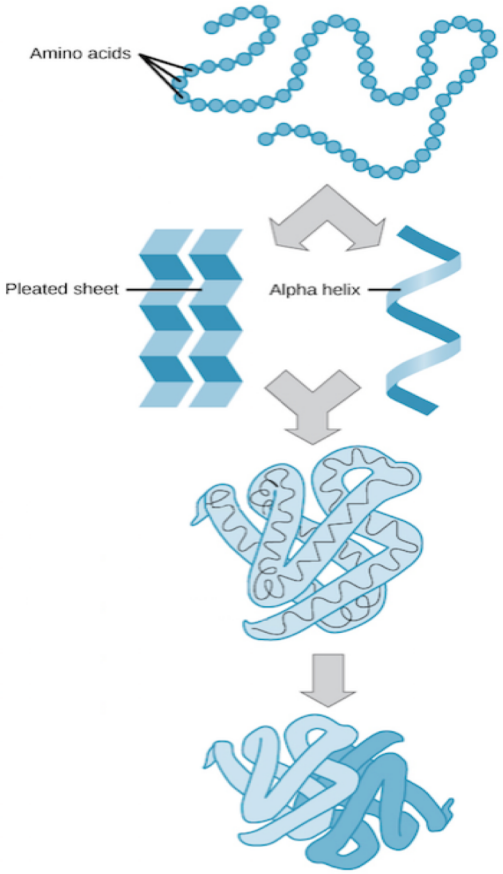
Limitations of Traditional Protein Structure Prediction Methods

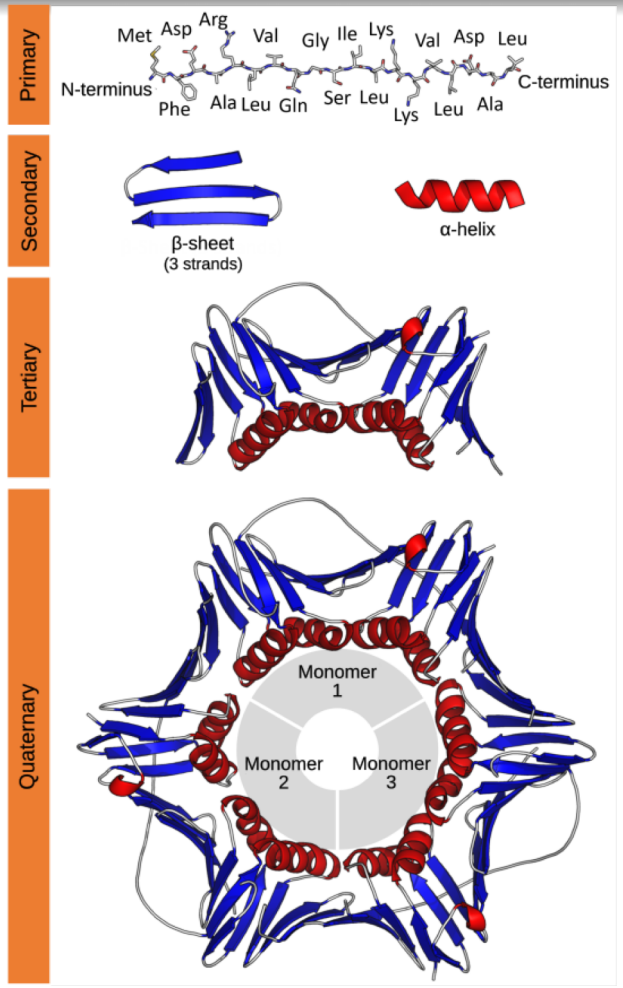
- Traditional prediction methods face challenges such as extensive time requirements, high costs, and limited experimentation feasibility due to protein complexity.
- The need for computational techniques arises to improve prediction accuracy, efficiency, and scalability, overcoming the limitations of traditional methods.



Protein - Hierarchical Nature of Protein Structures

Protein - Hierarchical Nature of Protein Structures

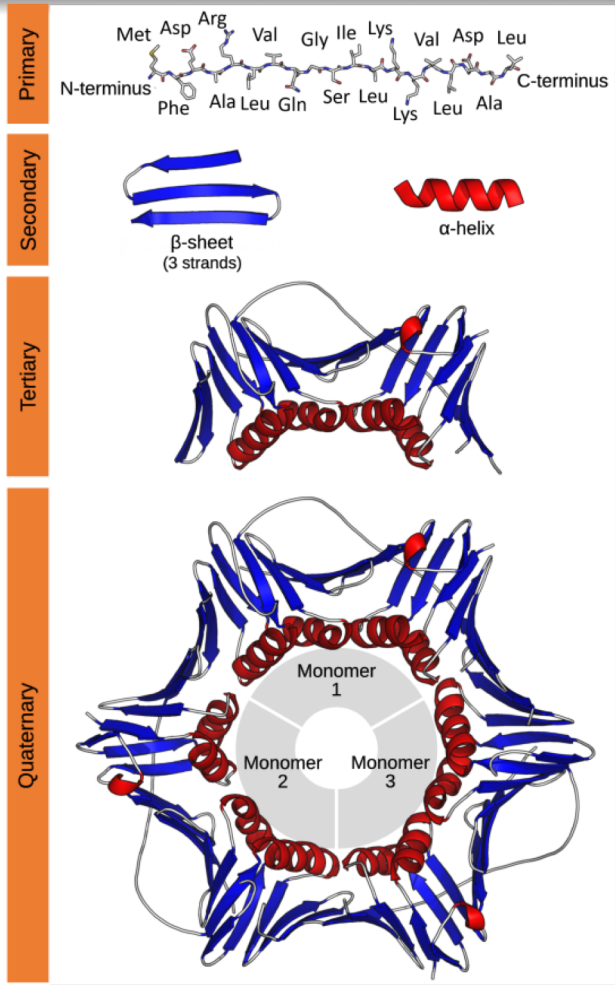




Credit:
https://en.wikipedia.org/wiki/Protein_structure

Protein - Hierarchical Nature of Protein Structures

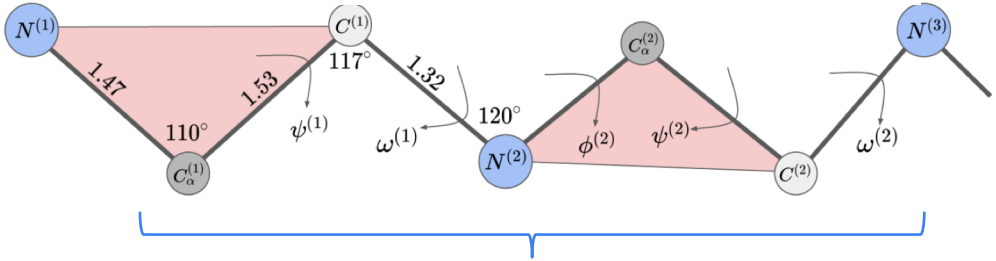
- Primary structure: Linear sequence of amino acids linked by peptide bonds. Determines overall protein function and characteristics.
- Secondary structure: Local folding patterns like alpha helices and beta sheets formed by hydrogen bonds between amino acids.
- Tertiary structure: 3D arrangement of secondary structures, held together by various bonds, crucial for protein function and specificity.
- Quaternary structure: Protein subunits come together to form a quaternary structure through non-covalent interactions and play specific roles in functionality.

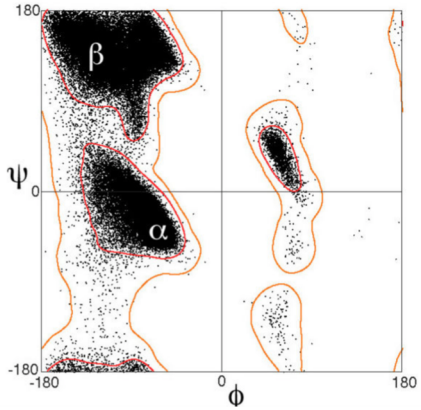


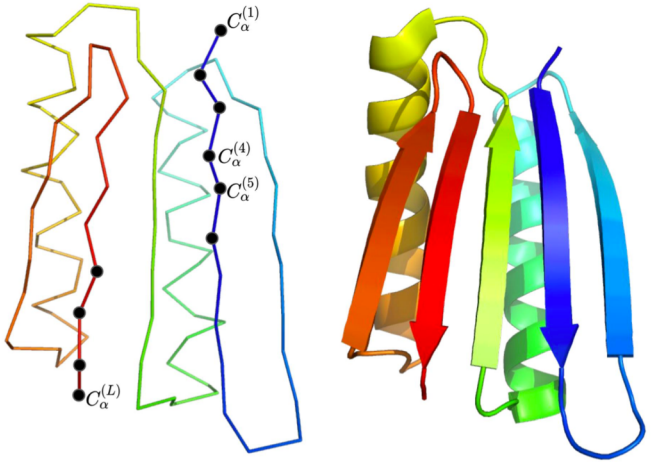
Geometric proteins close-up

- A protein backbone is a repeating sequence (linear chain) of 3 atoms: nitrogen, carbon, and another carbon …
- The Ramachandran plot shows the statistical distribution of the combinations of the backbone dihedral angles φ and ψ.

$$\underbrace{N^{(1)}, C_\alpha^{(1)}, C^{(1)}}, \underbrace{N^{(2)}, C_\alpha^{(2)}, C^{(2)}}, \ldots, \underbrace{N^{(L)}, C_\alpha^{(L)}, C^{(L)}}$$







https://dauparas.github.io/post/af2/

Protein – Folding/Structure based model (Go -model)

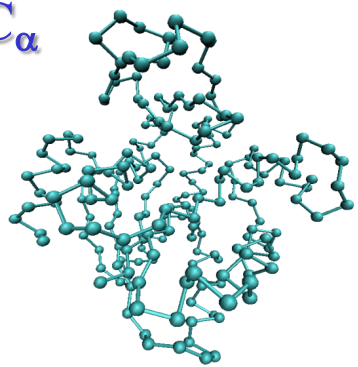
Covalent bonds

Contact map

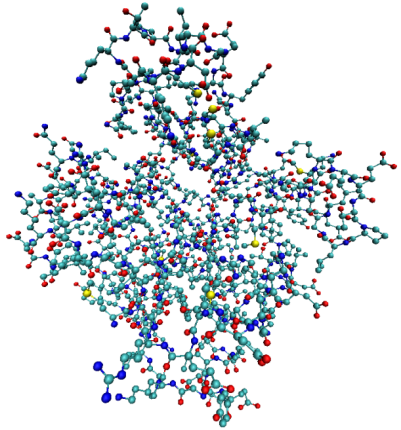
Native interactions

Non native interaction

C_α

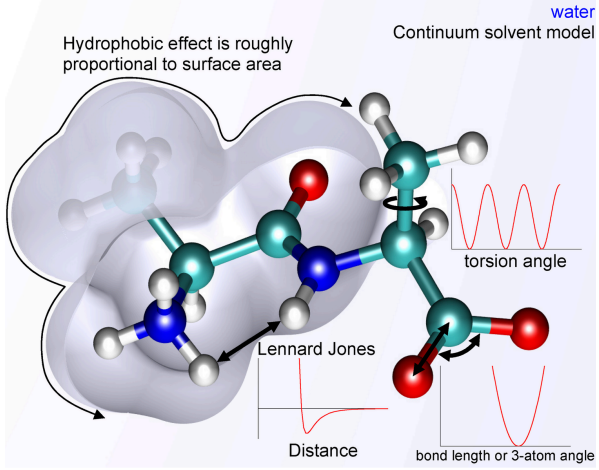


all atoms



$$E = \sum_{bonds} K_r (r - r_0)^2 + \sum_{angles} K_\theta (\theta - \theta_0)^2 + \sum_{dihedrals} K_\phi^{(n)} [1 - \cos(n \times (\phi - \phi_0))]$$

$$+ \sum_{i<j-3} \left\{ \varepsilon(i,j) \left[5 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 6 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{10} \right] + \varepsilon_2(i,j) \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} \right\}$$



Clementi C, Nymeyer H & Onuchic JN (2000), *J. Mol. Biol.* **298**, 937-953

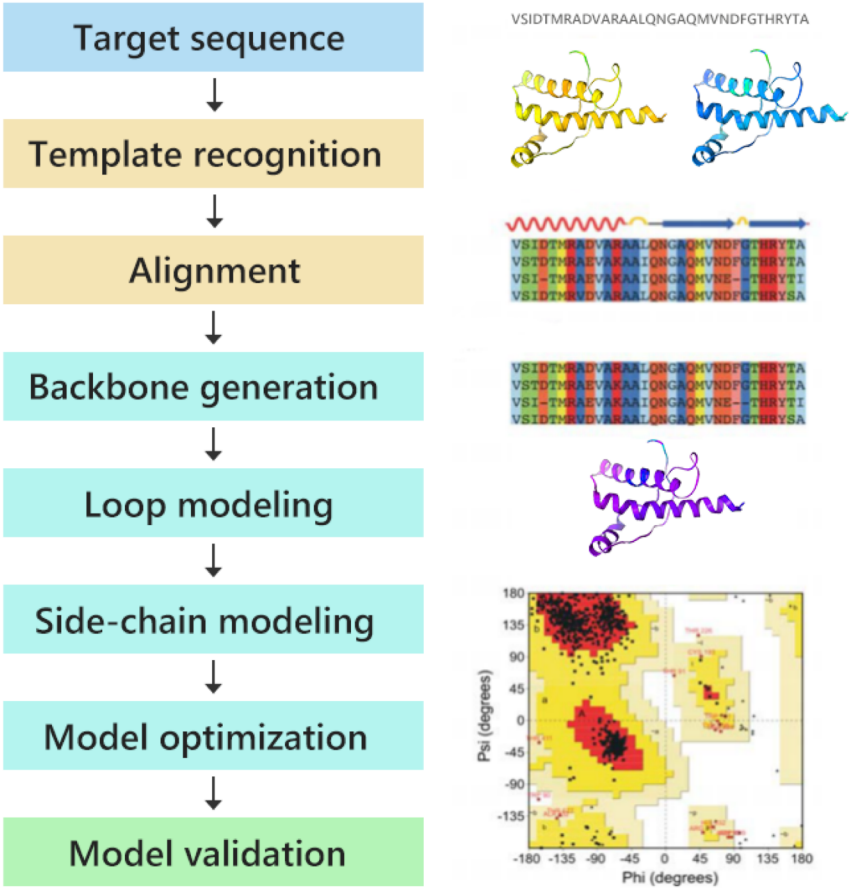
Sulkowska JI &Cieplak, (2008) Biophys J. 95, 3174-3191**95**

Whitford PC, Noel JK, Gosavi S, Schug A, Sanbonmatsu KY & Onuchic JN (2009), *Proteins*, 75, 430-441

Noel JN, Whitford PC, Sanbonmatsu KY & Onuchic JN, (2010) *Nucleic Acids Res.* Doi: 10.1093/nar/gkg498

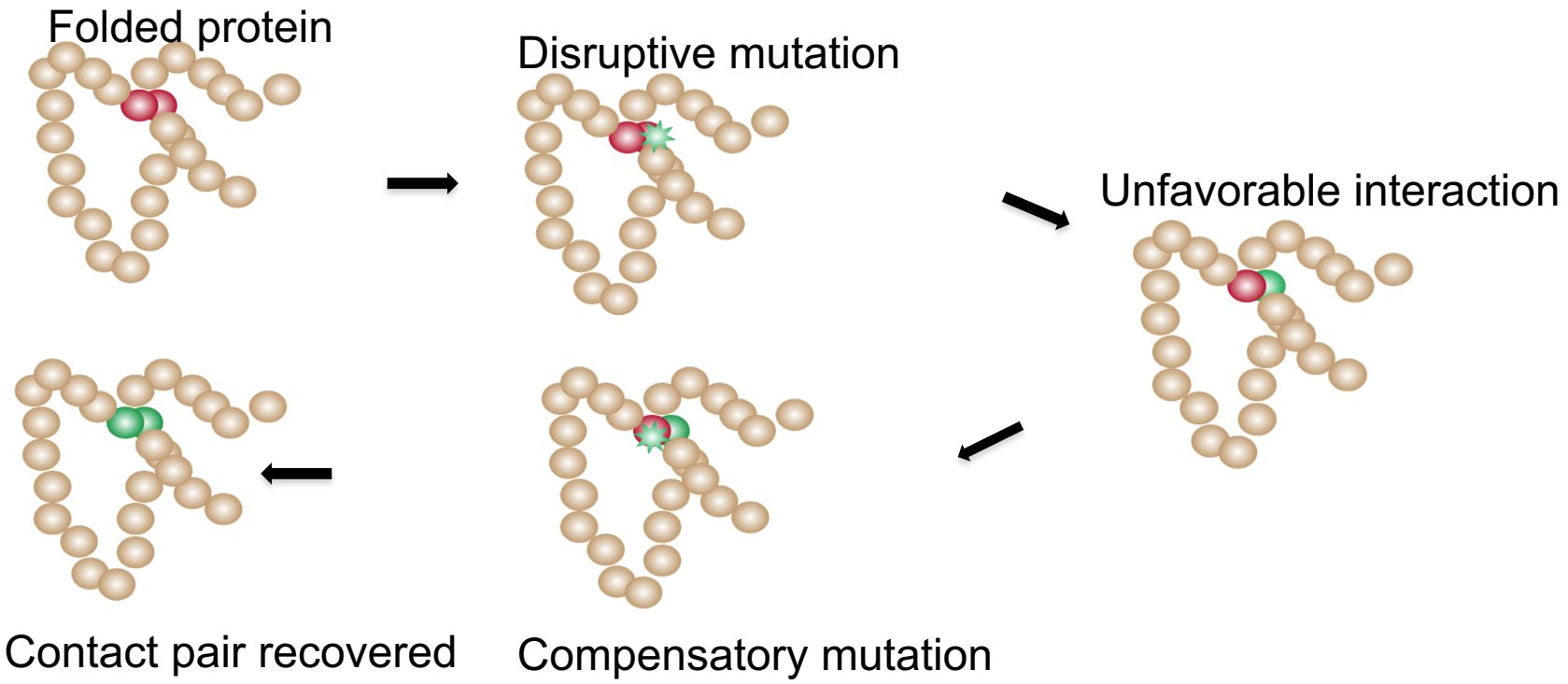
Methods of Protein Structure Prediction

- Homology modeling is a predictive method using evolutionary relationships to infer the structure of a target protein.
- Template structures play a crucial role in homology modeling by serving as a 3D framework for predicting the target protein's structure.
- The quality and similarity of the chosen template structures significantly impact the accuracy of the final homology model.



Multiple Sequence Alignment (MSA) – co-evolution approach (2012)

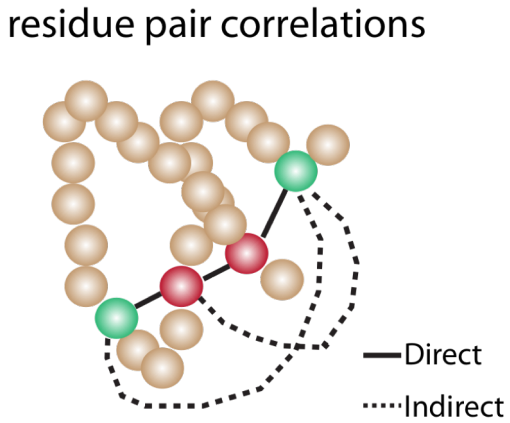
- Sequence alignment of a significant number of sequences
- MSA is used in AlphaFold to generate feature maps -> Provide information how to construct 3D structure



- Sequence alignment of a significant number of sequences
- MSA is used in AlphaFold to generate feature maps
- Provide information how to construct 3D structure

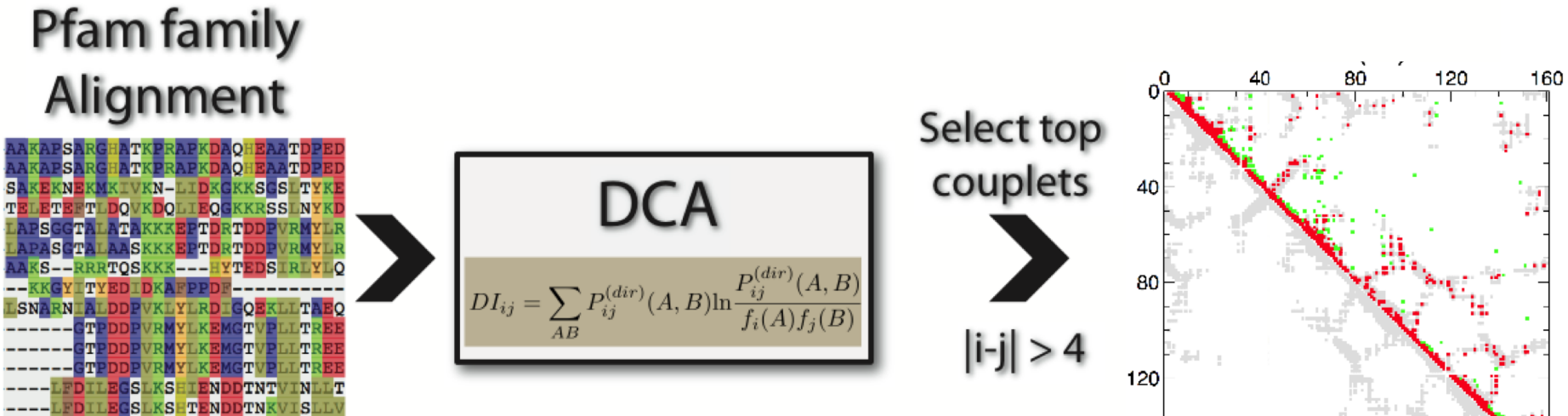
Analysis of Correlated Residue Pairs

- The **problem of directly coupled residue pairs** has been a long standing challenge due to several factors:
  1. Correlations can be **direct** (e.g. physical contacts) or **indirect** (chain effects)
  2. **Data quality:** Spurious correlations induced by **noisy or incomplete data**
  3. The use of **local statistical models** which assume independence between pairs of residue positions



- **"Recently" the panorama has changed:**
  1. A significant **increase of sequence information** due to improvements in technology
  2. The concept of a **protein family** has been **formalized and improved with better statistical models** (e.g. Pfam)
  3. **Global statistical** models (traditionally intractable) can be attacked with **novel approximate solutions**

To try to disentangle **direct** from **indirect** correlations we developed a statistical inference method called Direct Coupling Analysis (**DCA**).



Define frequency counts as probabilities

$$P_i(A_i) = \sum_{\{A_k | k \neq i\}} P(A_1, ..., A_L) \equiv f_i(A_i)$$

$$P_{ij}(A_i, A_j) = \sum_{\{A_k | k \neq i,j\}} P(A_1, ..., A_L) \equiv f_{ij}(A_i, A_j)$$

Using maximum entropy principle to model the joint probability distribution

$$P(A_1, \ldots, A_L) = \frac{1}{Z} \exp\{\sum_{i<j} e_{ij}(A_i, A_j) + \sum_i h_i(A_i)\}$$

Direct Probabilities are defined as:

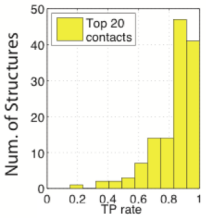
$$P_{ij}^{(dir)}(A,B) = \frac{1}{Z}\exp\{e_{ij}(A,B) + \hat{h}_i(A) + \hat{h}_i(B)\}$$

The probabilities for residue couplets are ranked using Direct Information

$$DI_{ij} = \sum_{A,B=1}^{q} P_{ij}^{(dir)}(A,B)\ln\frac{P_{ij}^{(dir)}(A,B)}{f_i(A)\,f_j(B)}$$

Select top couplets

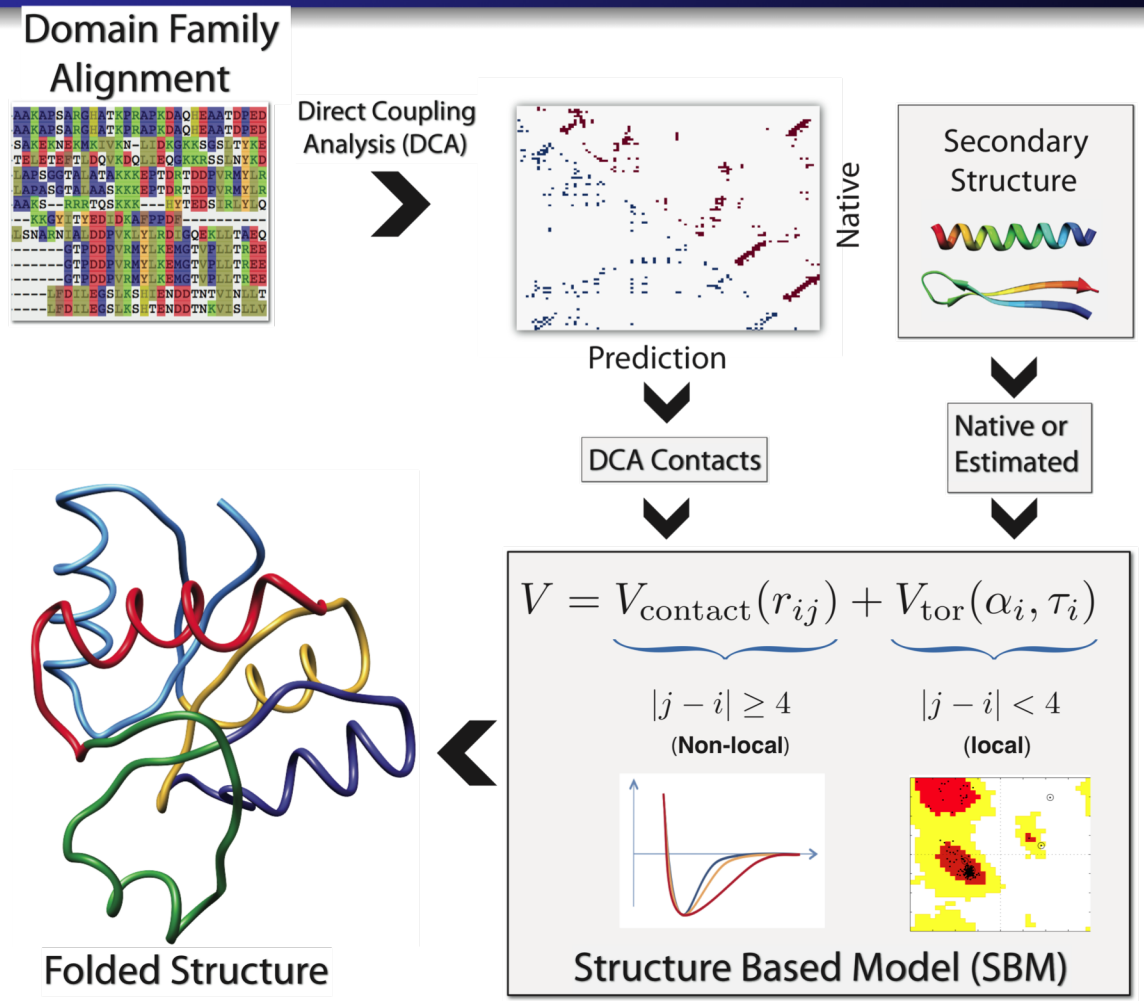
|i-j| > 4



True Positive (TP) rates

True positive contacts (<8A) are evaluated from top couplets

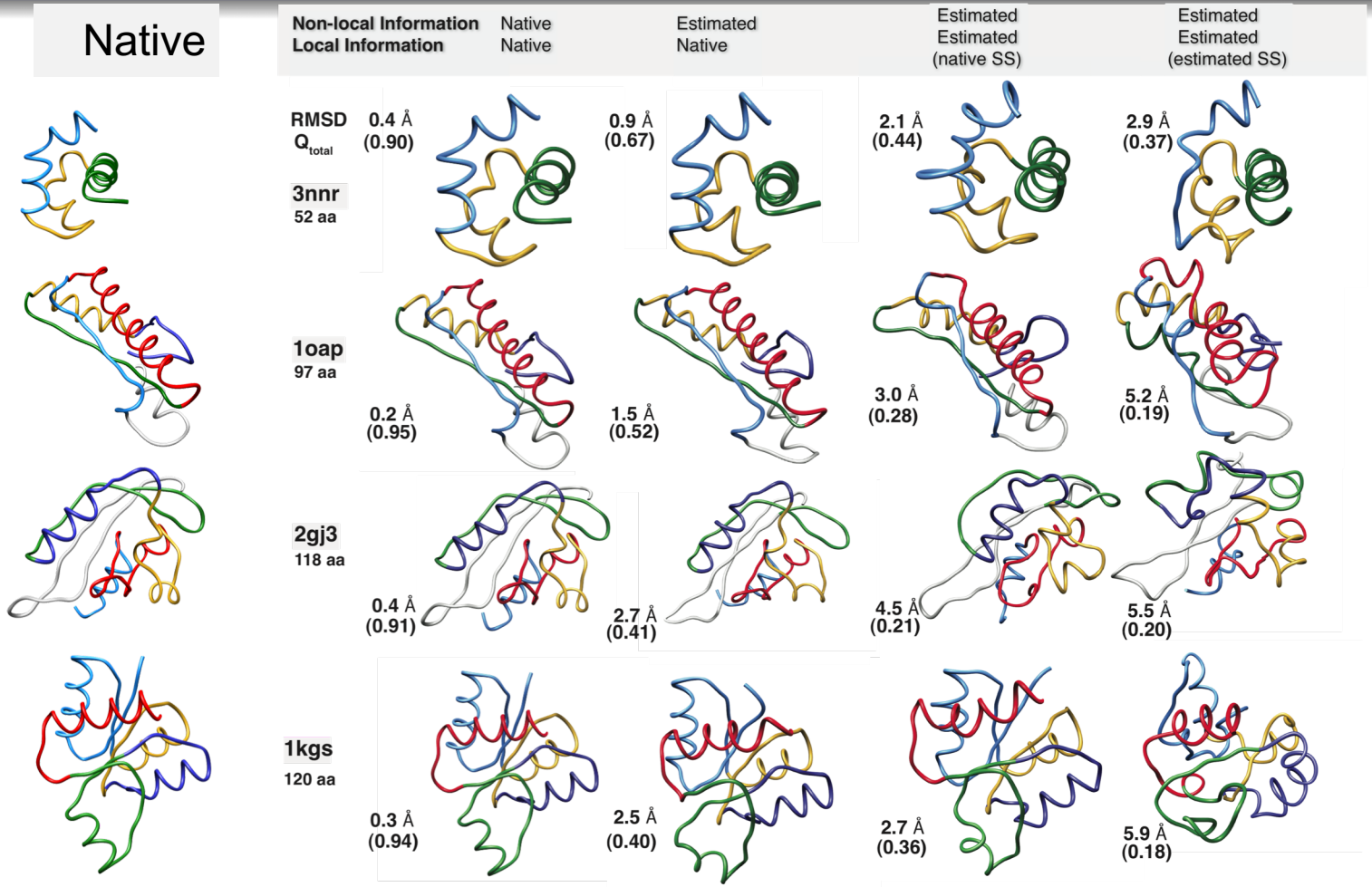
Genomics-Aided Structure Prediction (DCAfold)



[Sulkowska, Morcos et al. PNAS 2012]

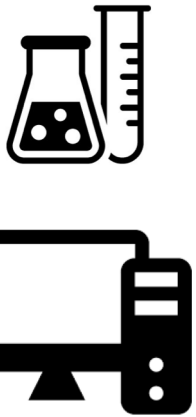
Review: Marks et al. *Nature Biotechnology* 2012

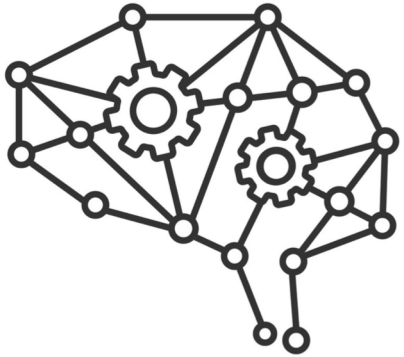
Genomics-Aided Structure Prediction (DCAfold)



Theoretical approach to protein structure prediction

- Machine learning enables faster and more accurate protein structure prediction by learning patterns from large datasets.
- Molecular dynamics simulations provide insights into protein dynamics, aiding in understanding structure-function relationships.
- Protein threading offers a valuable alternative to predict protein folds, potentially overcoming challenges faced by traditional methods.





AlphaFold v1: Improved protein structure prediction using potentials from deep learning

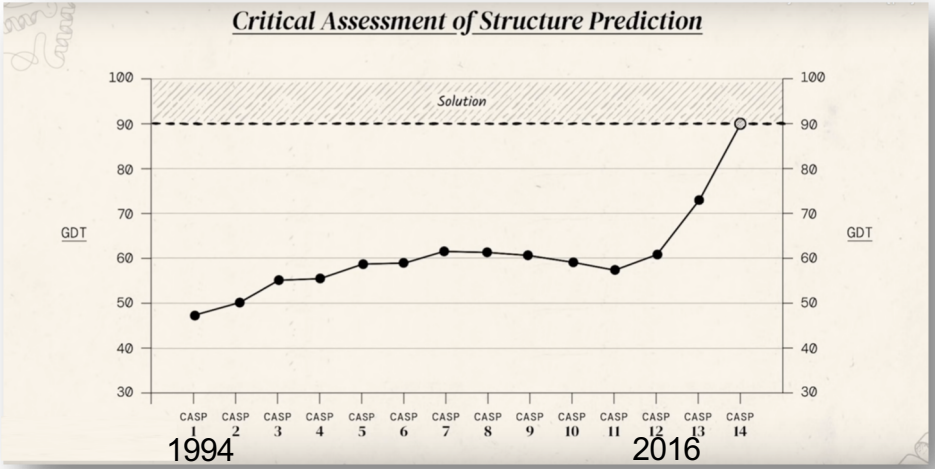
One CNN-supported Protein Folding Model

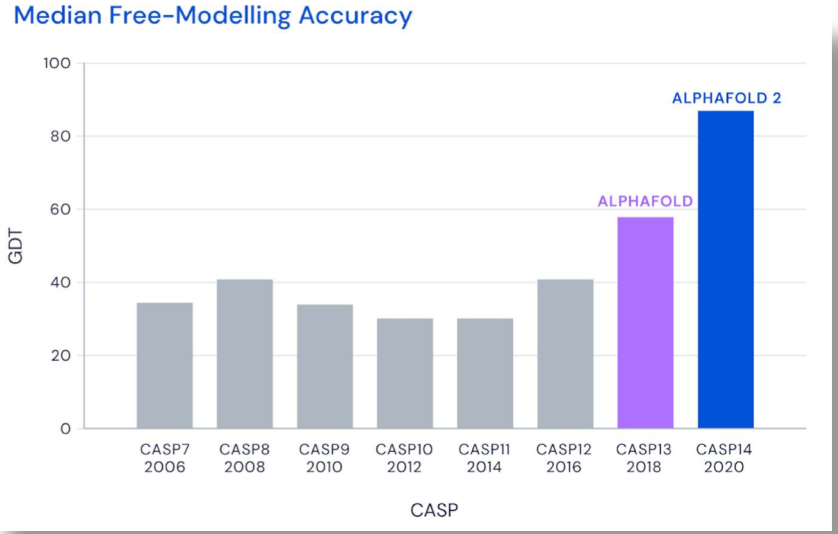
- Introduction to AlphaFold: history, key components.
- Discussion of the groundbreaking results achieved by AlphaFold in the CASP competitions.
- Overview of the AlphaFold algorithm: attention mechanism, residual networks, and transformer architecture.
- Demonstration of how AlphaFold predicts protein structures.
- Case studies showcasing AlphaFold's accuracy and utility in protein structure prediction.

- AlphaFold: Improved protein structure prediction using potentials from deep learning (Published on Nature, Jan 2020)
- AlphaFold2: Highly accurate protein structure prediction with AlphaFold (Published on Nature, July 2021)

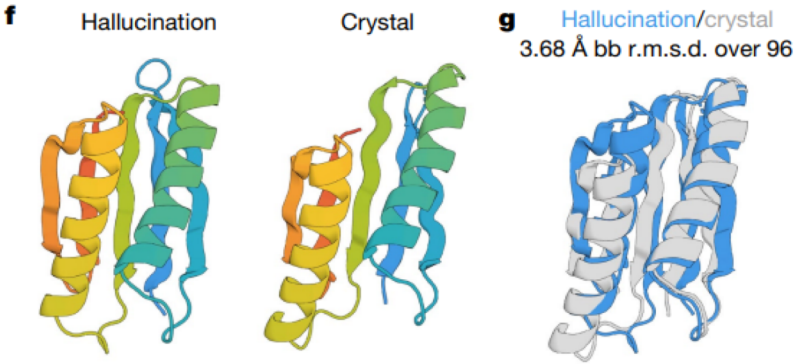
AlphaFold v1 – Artificial Intelligence model for protein structure prediction

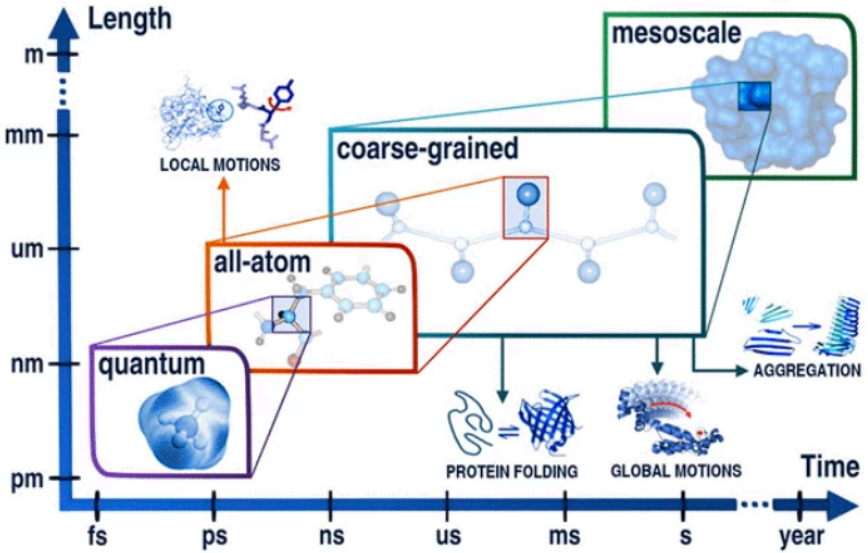
- Domination of the AlphaFold model in CASP13/14
- Database with predictions from the UniProt Database: AlphaFold Database
- Deep Hallucinations: predicting novel proteins and folds



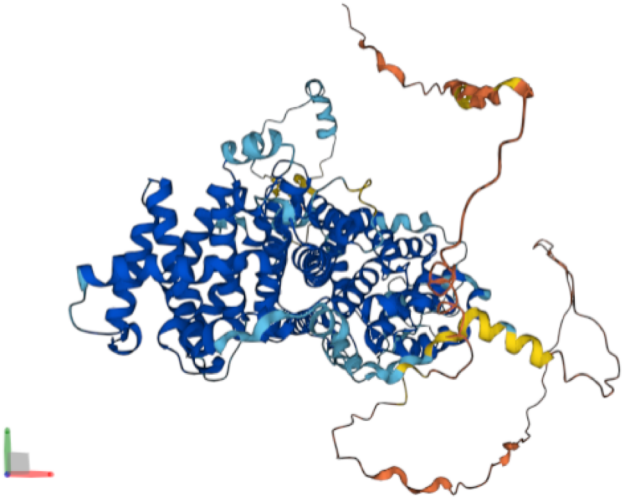


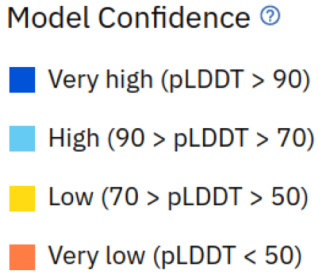
- Domination of the AlphaFold model in CASP13/14
- Database with predictions from the UniProt Database: AlphaFold Database
- Deep Hallucinations: predicting novel proteins and folds

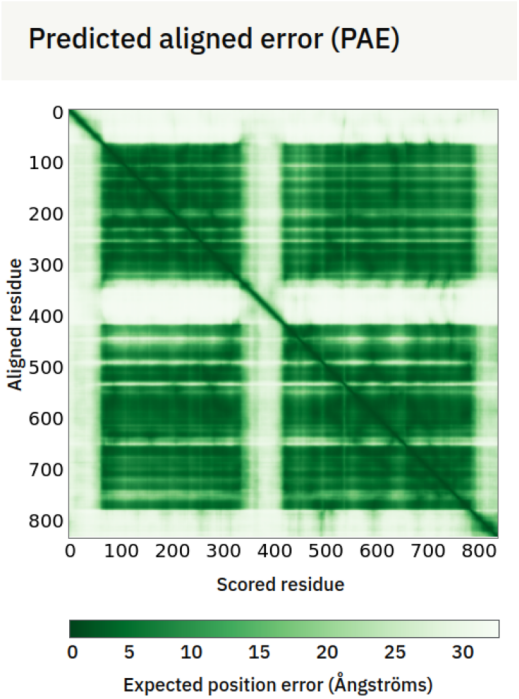




- Each prediction comes with a per-residue confidence score, known as pLDDT (predicted Local Distance Difference Test).
- Scores range from 0 to 100, with higher scores indicating greater confidence in the accuracy of the predicted local structure.
- PAE matrix for long distance relation evaluation



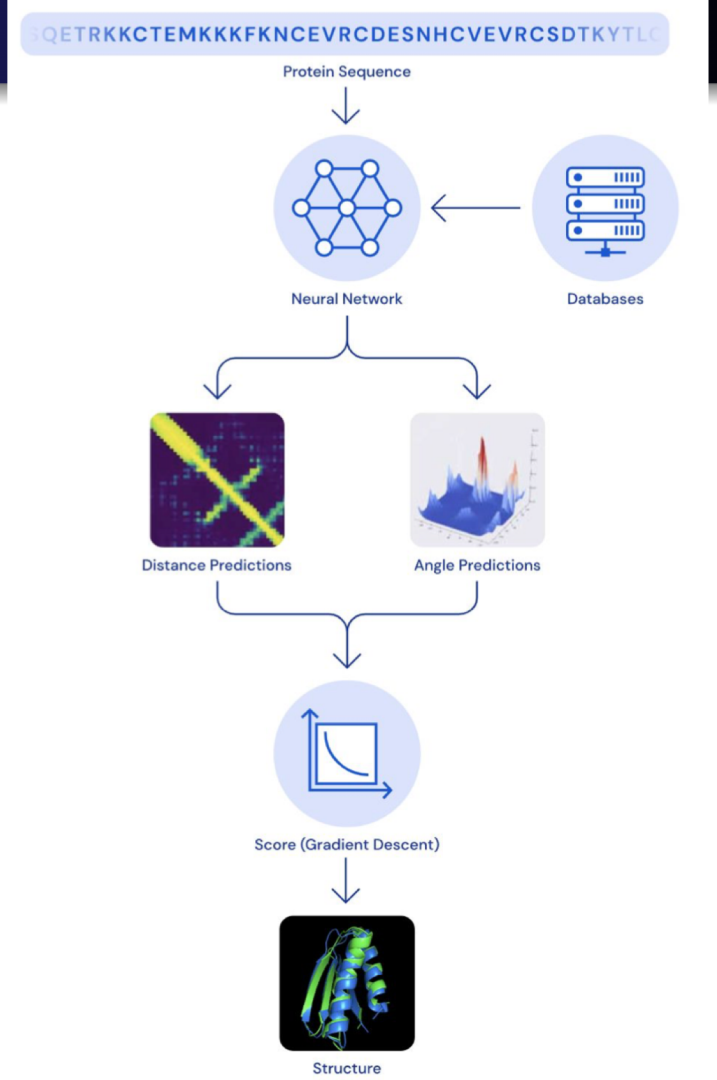




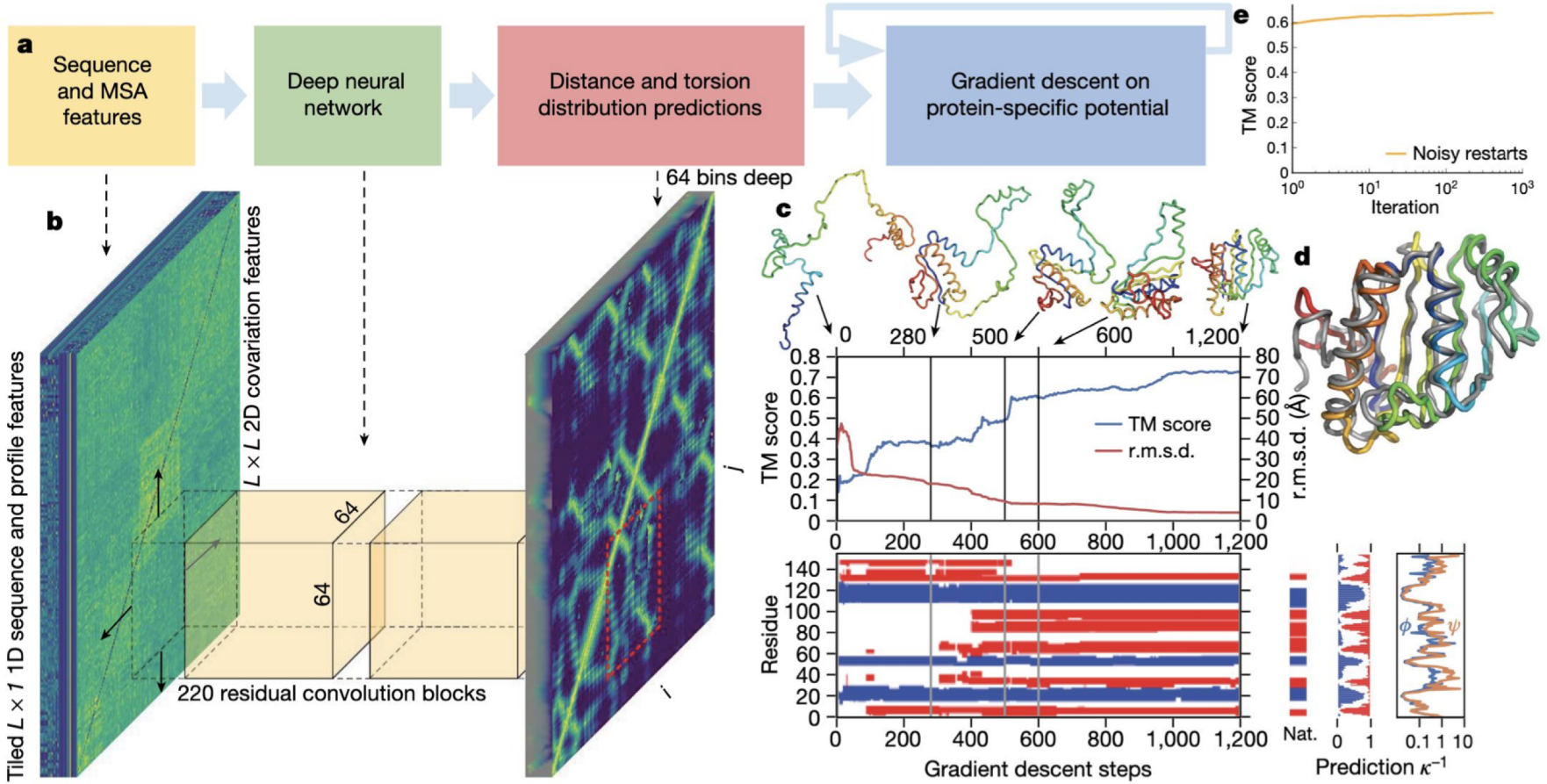
AlphaFold v1: Schematic Architecture

- Residual CNN as core model to predict distance and angle to create final structure output

- Using Multiple Sequence Alignment (MSA) from databases for feature generation

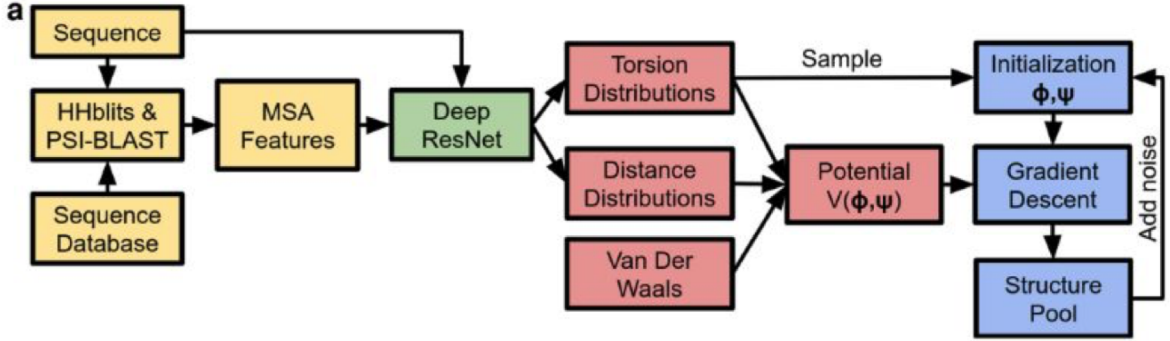


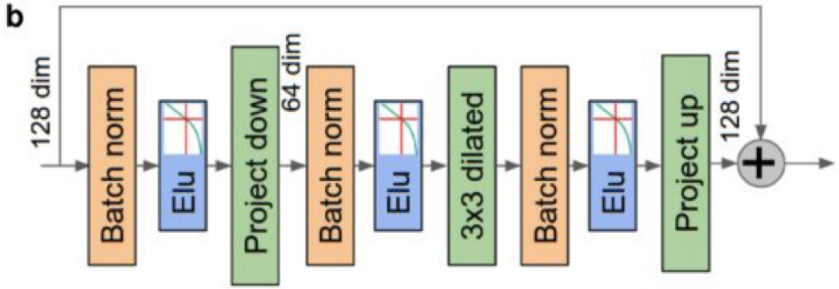
AlphaFold v1: Model Overview

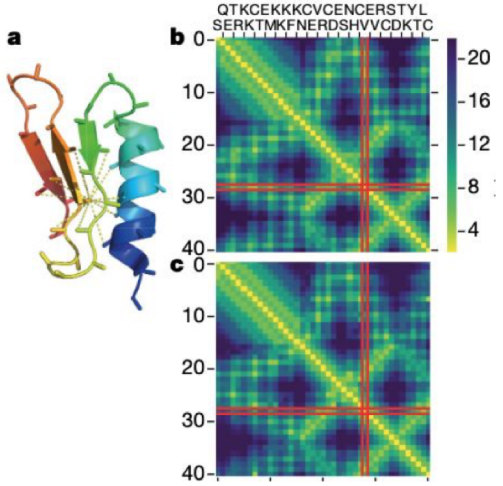


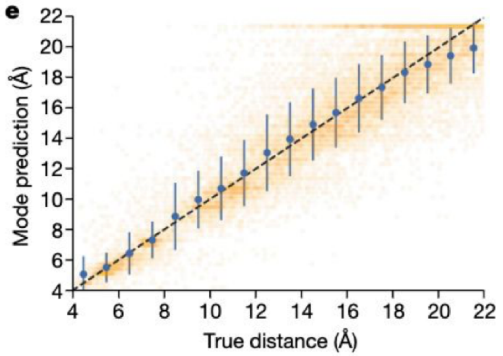
AlphaFold v1: Model Details

UCLA
Engineer Change.

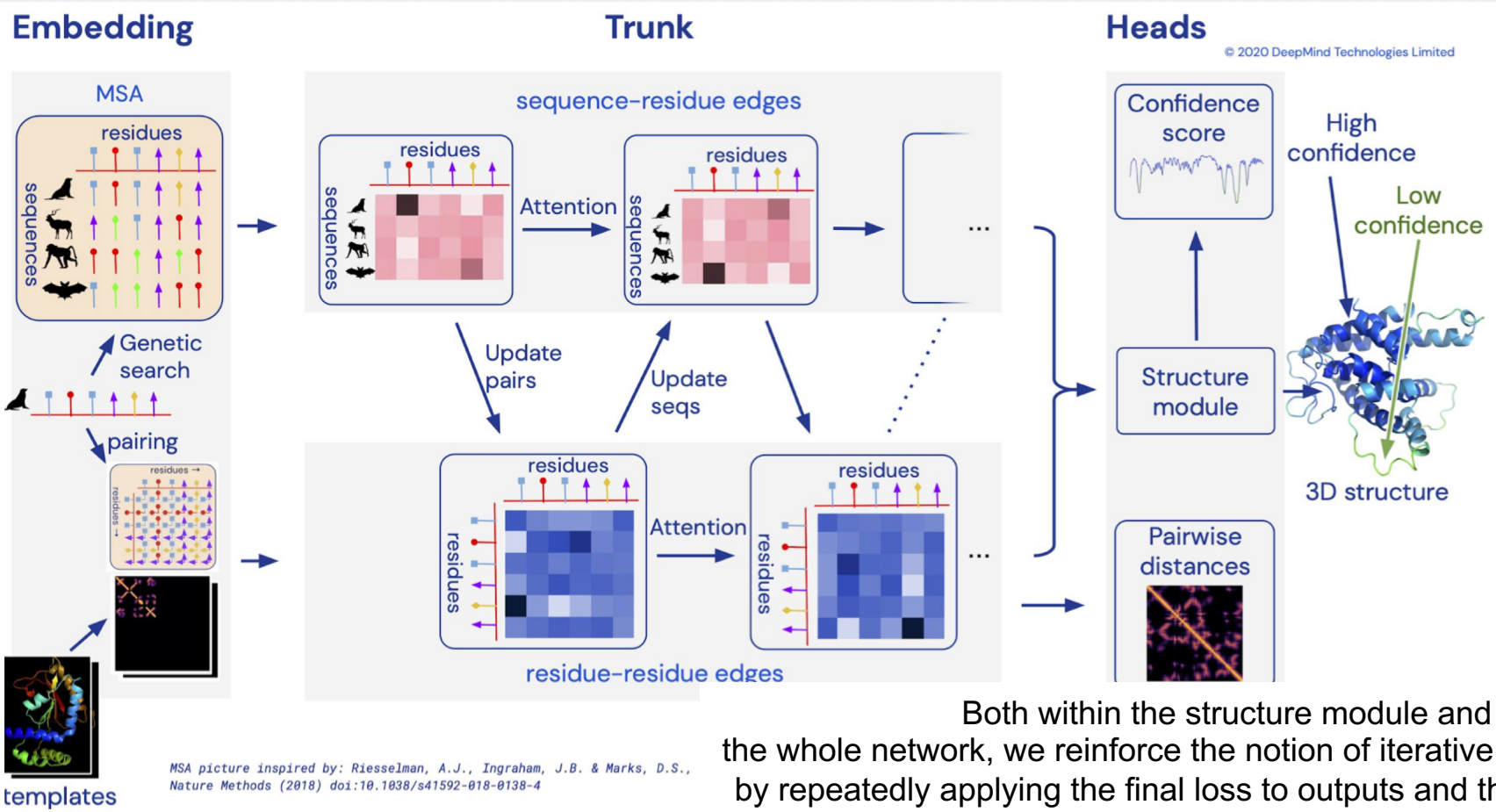








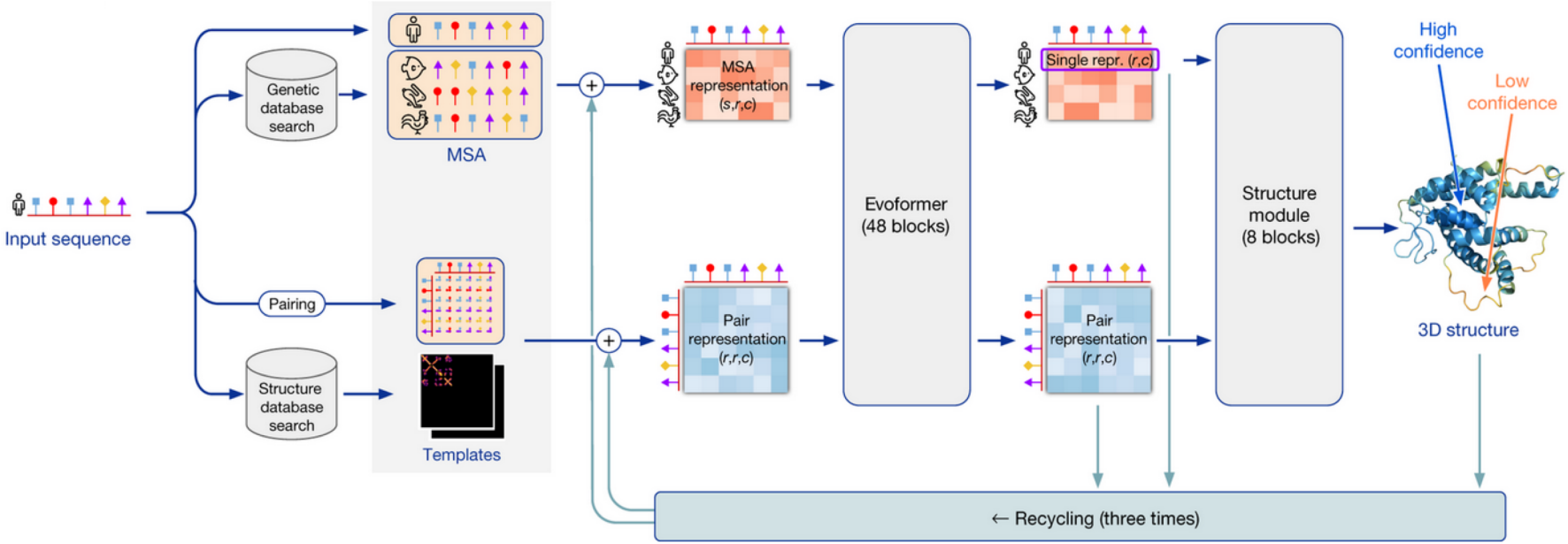
AlphaFold 2 - model architecture



Both within the structure module and throughout the whole network, we reinforce the notion of iterative refinement by repeatedly applying the final loss to outputs and then feeding the outputs recursively into the same modules.

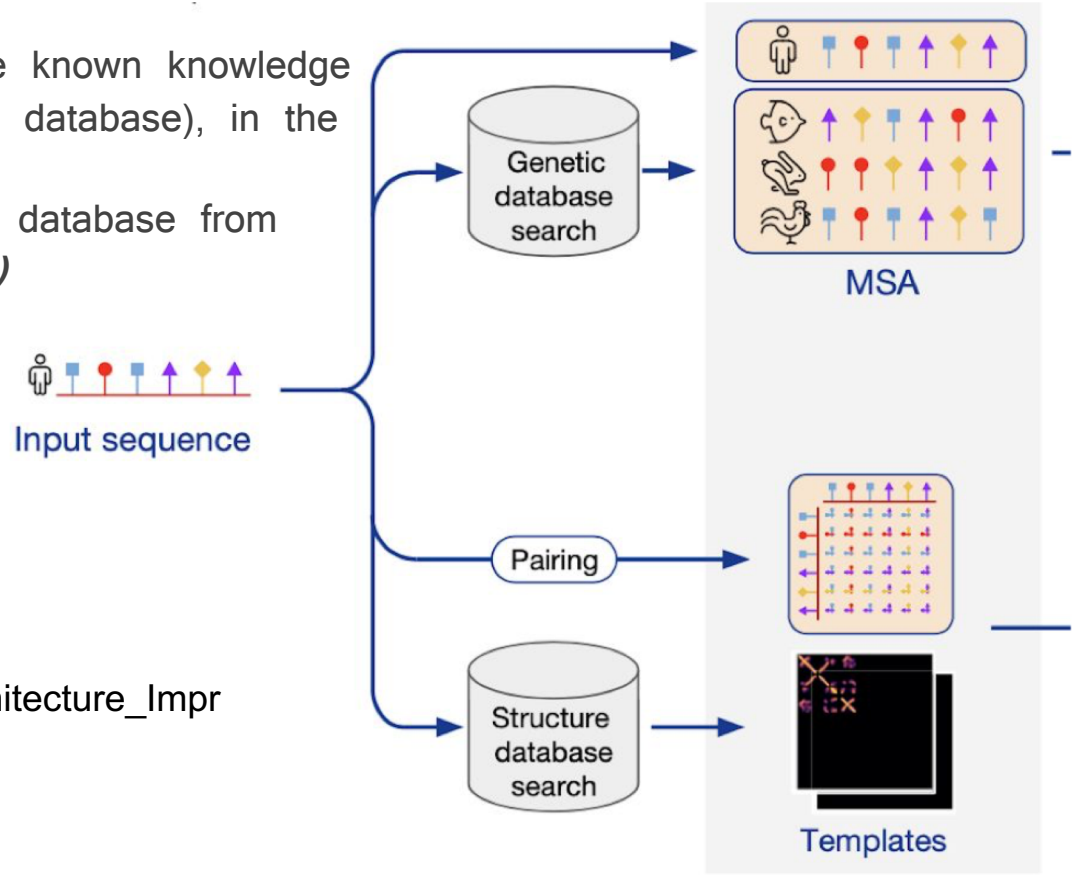
AlphaFold 2 - model architecture

- Converts the amino acid sequence and associated multiple sequence alignments (MSAs) into an initial numeric representation.
- Data processing through Evoformer Blocks and Structure Module
- Recycling: Iteratively refines the output structure multiple times based on feedback from the previous iterations.
- Converts the predicted distances (distogram) and angles into 3D coordinates.
- Each residue's position is scored with a predicted Local Distance Difference Test (pLDDT)



AlphaFold 2 - embeding/input

- Not significantly different from AlphaFold v1, or even other models
- Input sequence, and leveraging some known knowledge
- MSA (**sequence-residue** from genetic database), in the shapeof ***(sequence,residue,column)***
- Templates (**residue-residue**, structure database from known proteins), in the shape of ***(r,r,c)***

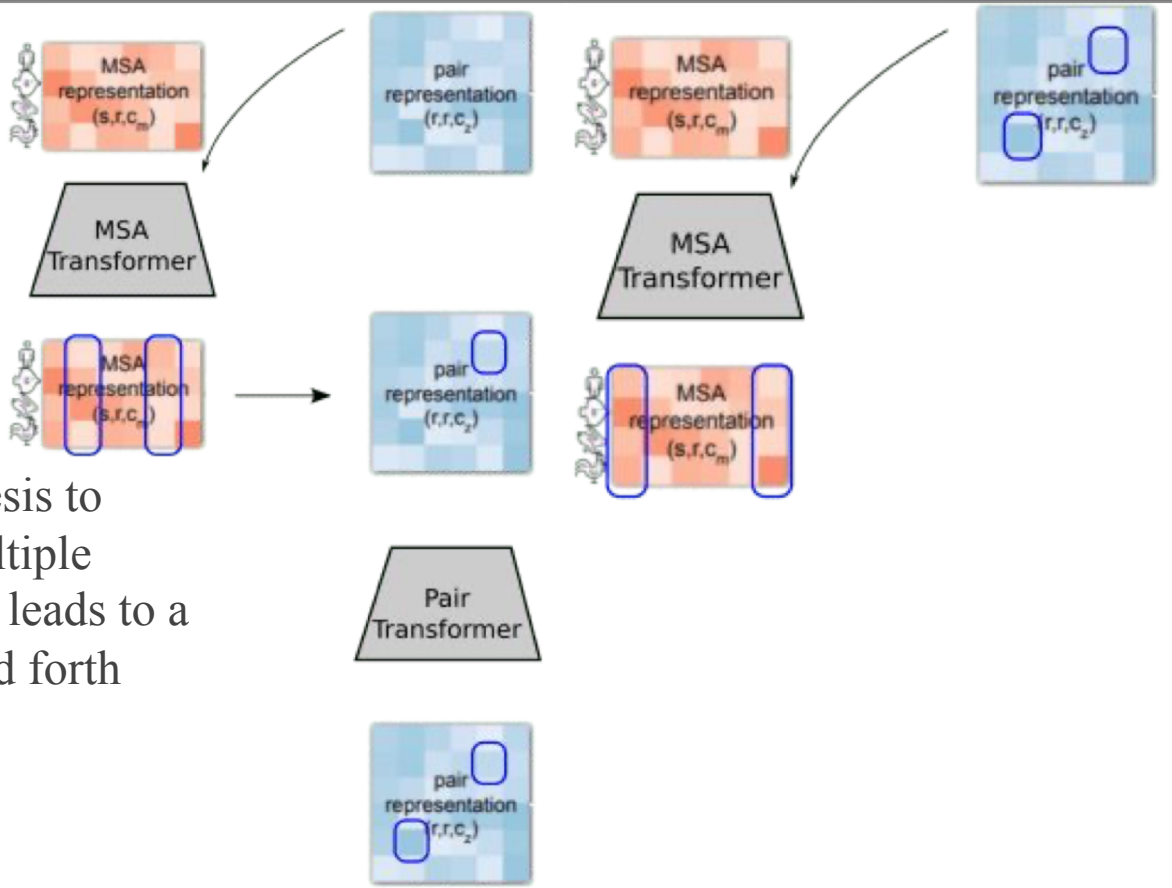


Very good description:

https://piip.co.kr/en/blog/AlphaFold2_Architecture_Improvements


Folded protein
Disruptive mutation
AAKAPSARGHATKPRAPKDAQHEAA
AAKAPSARGHATKPRAPKDAQHEAA
Unfavorable interaction
SAKEKNEKMKIVKN-LIDKGKKSGS
TELETKFTLDQVKDQLEEQGKKRSS
LAPSGNTALATAKKKEITDRTDDPV
TELETKFTLDQVKPRAEKDGKKRSS
Contact pair recovered
Compensatory mutation

AlphaFold 2 – Evoformer (evolutional transformer)



- Central idea: AlphaFold2 uses the current structural hypothesis to improve the assessment of the multiple sequence alignment, which in turns leads to a new structural hypothesis, back and forth at every cycle.

- Two transformers (a "two-tower architecture"), with one clear communication channel.

https://www.haojunheng.com/files/talks/alphafold.pdf

AlphaFold v2 - MSA transformer

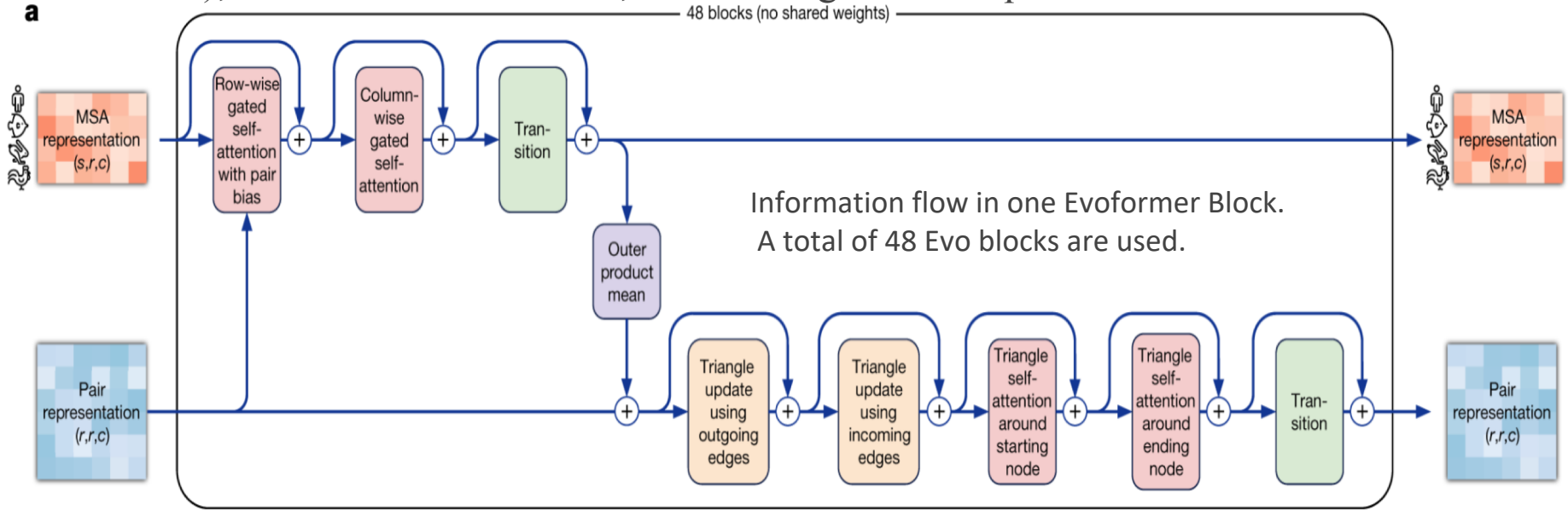
Core component that iteratively processes sequence information using attention mechanisms to model interactions between residues

- The attention is “factorized” in “row-wise” and “column-wise” components.
- MSA Transformer first computes attention in the horizontal direction, allowing the network to identify which pairs of amino acids are more related; and then in the vertical direction, determining which sequences are more informative.
- MSA Transformer’s row-wise (horizontal) attention mechanism incorporates information from the “pair representation”.
- Gated attention applied.

https://www.haojunheng.com/files/talks/alphafold.pdf

Core component that iteratively processes sequence information using attention mechanisms to model interactions between residues

- The attention is "factorized" in "row-wise" and "column-wise" components.
- MSA Transformer: attention in the horizontal direction (which pairs of amino acids are more related); in the vertical direction, determining which sequences are more informative.



AlphaFold v2 - voformer Stack: Algorithm Workflow

Algorithm 6 Evoformer stack

def EvoformerStack($\{\mathbf{m}_{si}\}, \{\mathbf{z}_{ij}\}, N_{\text{block}} = 48, c_s = 384$) :

for all $l \in [1, \ldots, N_{\text{block}}]$ **do**

 # MSA stack

$\{\mathbf{m}_{si}\}$ += DropoutRowwise$_{0.15}$(MSARowAttentionWithPairBias($\{\mathbf{m}_{si}\}, \{\mathbf{z}_{ij}\}$))

$\{\mathbf{m}_{si}\}$ += MSAColumnAttention($\{\mathbf{m}_{si}\}$)

$\{\mathbf{m}_{si}\}$ += MSATransition($\{\mathbf{m}_{si}\}$)

 # Communication

$\{\mathbf{z}_{ij}\}$ += OuterProductMean($\{\mathbf{m}_{si}\}$)

 # Pair stack

$\{\mathbf{z}_{ij}\}$ += DropoutRowwise$_{0.25}$(TriangleMultiplicationOutgoing($\{\mathbf{z}_{ij}\}$))

$\{\mathbf{z}_{ij}\}$ += DropoutRowwise$_{0.25}$(TriangleMultiplicationIncoming($\{\mathbf{z}_{ij}\}$))

$\{\mathbf{z}_{ij}\}$ += DropoutRowwise$_{0.25}$(TriangleAttentionStartingNode($\{\mathbf{z}_{ij}\}$))

$\{\mathbf{z}_{ij}\}$ += DropoutColumnwise$_{0.25}$(TriangleAttentionEndingNode($\{\mathbf{z}_{ij}\}$))

$\{\mathbf{z}_{ij}\}$ += PairTransition($\{\mathbf{z}_{ij}\}$)

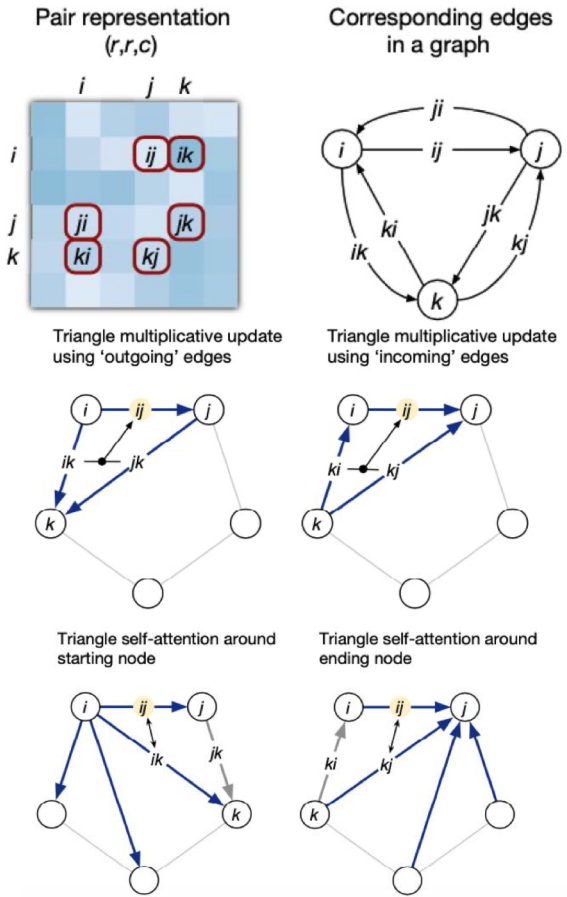
end for

 # Extract the single representation

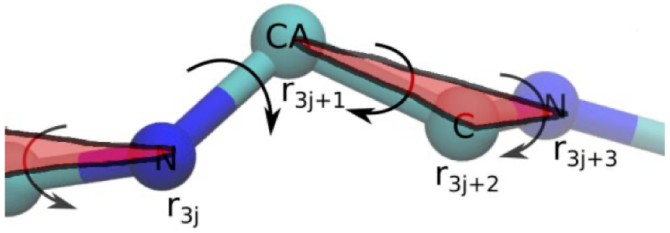
$\mathbf{s}_i = \text{Linear}(\mathbf{m}_{1i})$ $\quad \mathbf{s}_i \in \mathbb{R}^{c_s}$

return $\{\mathbf{m}_{si}\}, \{\mathbf{z}_{ij}\}, \{\mathbf{s}_i\}$

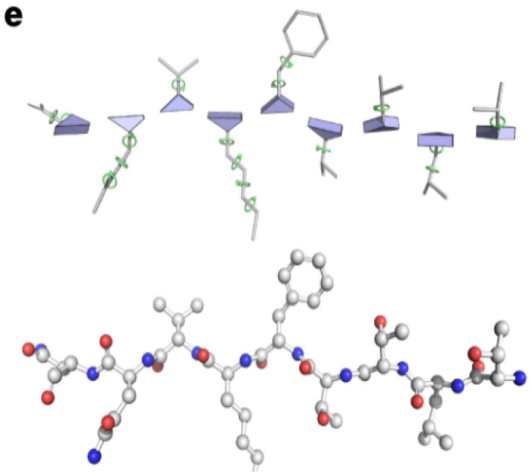
Attention is arranged in terms of triangles of residues.

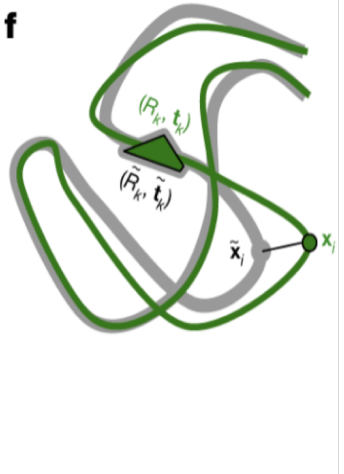


- The structure module treats the protein as a “residue gas”, **a floating backbone**.
- Every amino acid is modelled as a triangle, representing the three atoms of the backbone.
- The triangles float around in space and are moved by the network to form the structure.
- These transformations are parametrized as “affine matrices”.
- At every step of the iterative process, AlphaFold 2 produces a set of affine matrices that displace and rotate the residues in space.



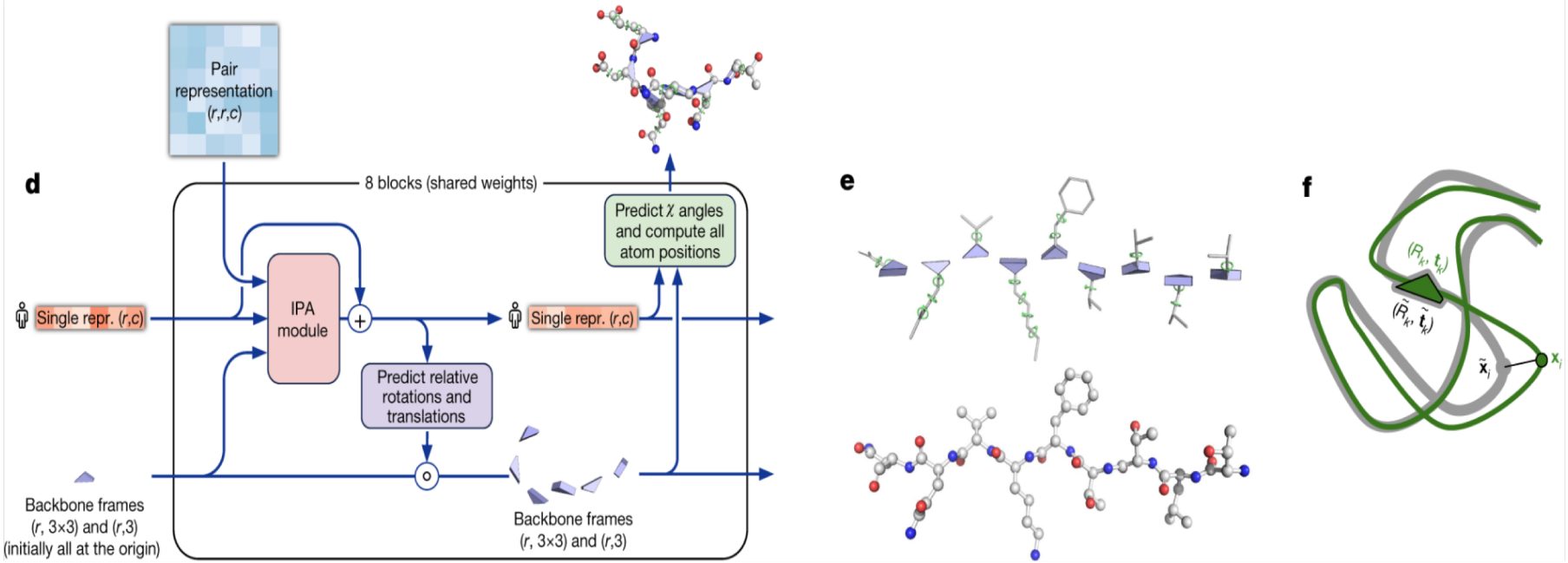
$$\mathbf{M} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$



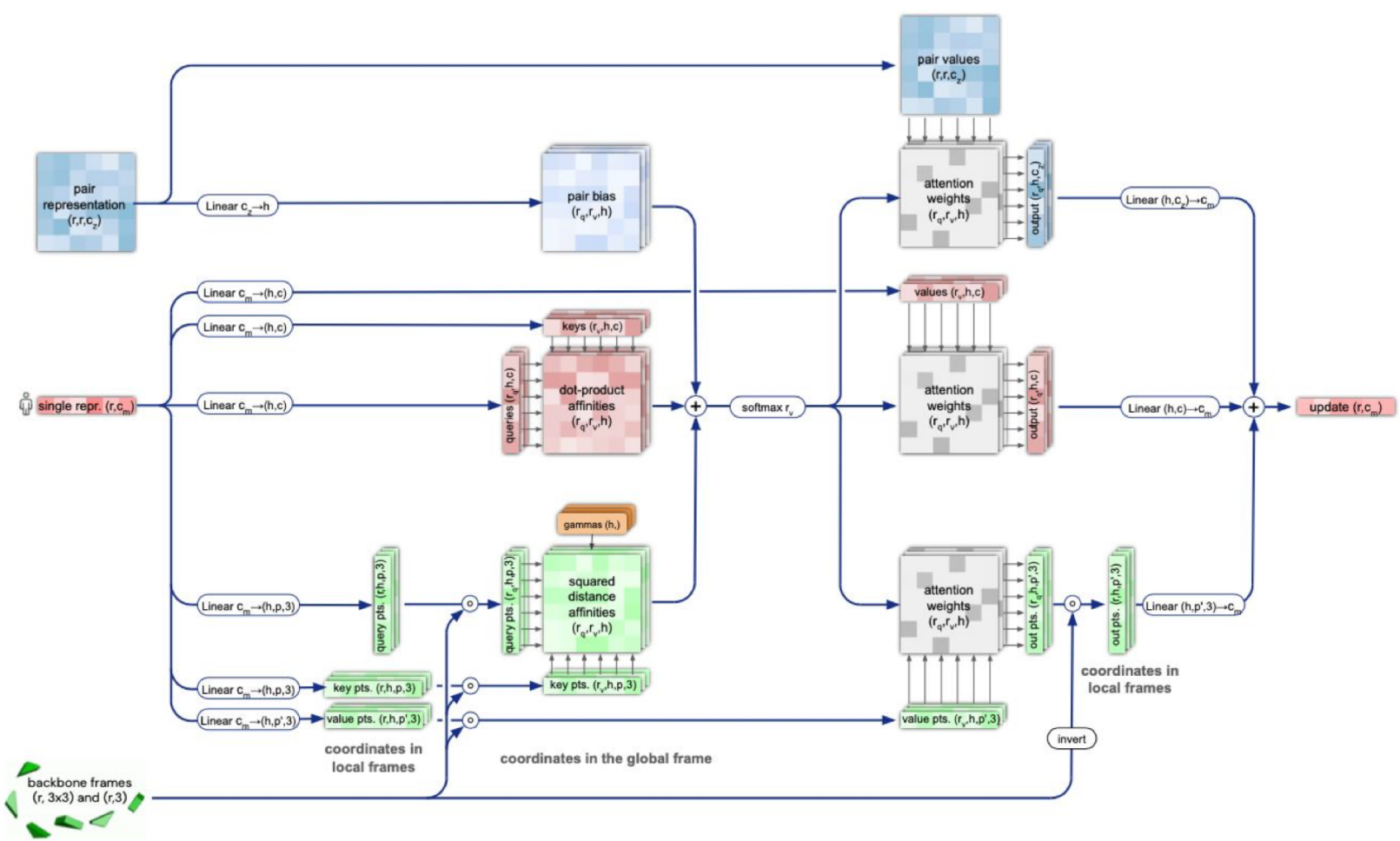


AlphaFold v2 – structure module

- Takes processed sequence features from Evoformer and predicts distances and angles between residues, forming a rough 3D structure.
- Utilizes a graph-based neural network that models the relationships between amino acids.
- Amino acid approximated as floating/rigid triangles



AlphaFold v2 – structure module/IPA Module



AlphaFold versions

- OpenFold – memory efficient, open-source version of AlphaFold
- ColabFold – version of AlphaFold for Google Collab
- AlphaFold multimer – AlphaFold version for protein and protein-RNA complex modelling
- OpenComplex – Open-source version of AlphaFold multimer, based on OpenFold

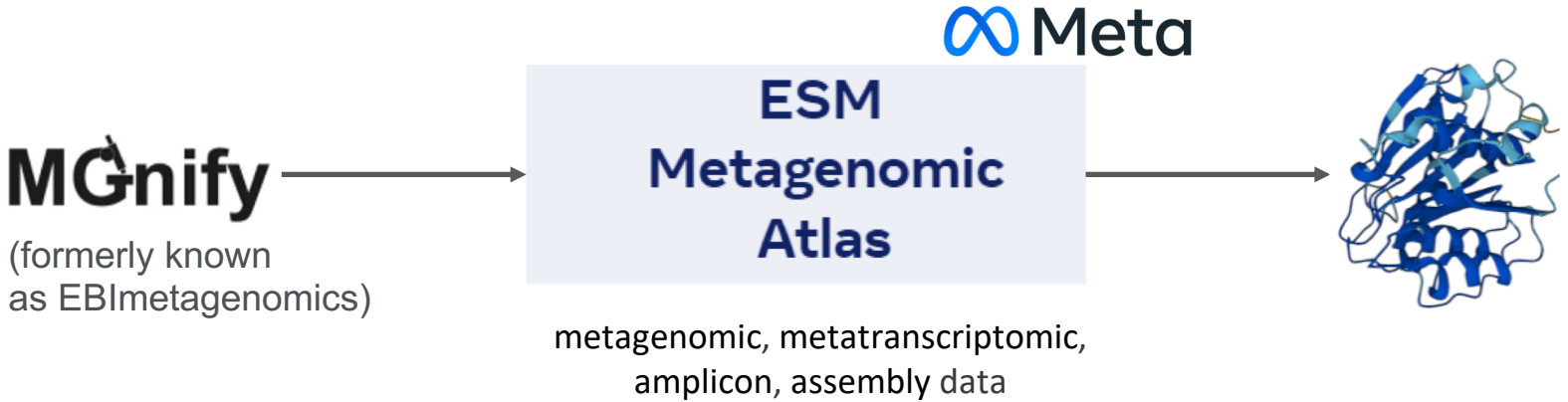
OpenComplex

ESMFold: Advancements in Protein Structure Prediction

- Introduction to ESMFold: Evolutionary Scale Modeling (ESM) approach.
- Overview of the ESMFold algorithm and its key features.
- Explanation of how ESMFold integrates evolutionary information with deep learning.
- Comparison between AlphaFold and ESMFold.
- Discussion of the potential advantages and limitations of ESMFold.
- Case studies demonstrating the effectiveness of ESMFold in predicting protein structures.

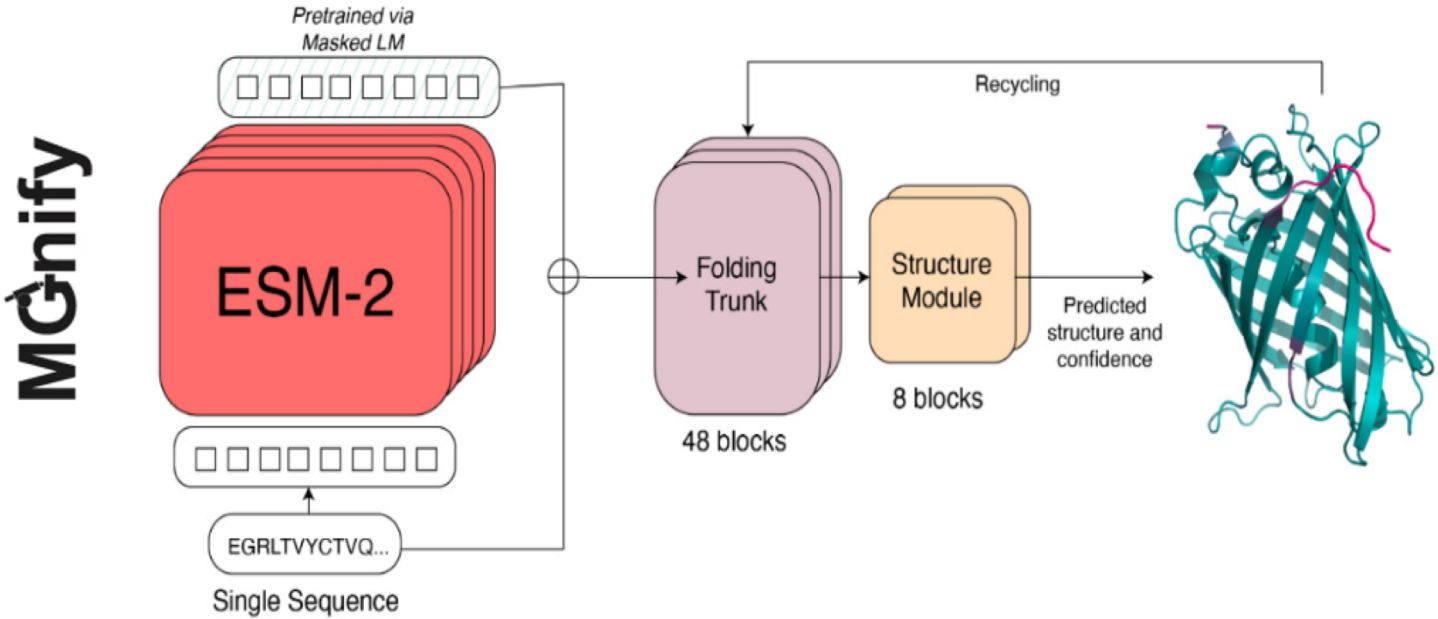
ESM Fold

- ESM Fold is a protein structure prediction model developed by researchers at Meta AI.
- Utilizes transformer-based machine learning architecture similar to those used in natural language processing, adapted to understand biological sequences.
  - a BERT-like architecture, large language model that utilizes stacked, transformer encoder layers.
- Demonstrates high accuracy in predicting protein structures
- ESM Fold is available as an open-source tool, making it accessible for researchers.



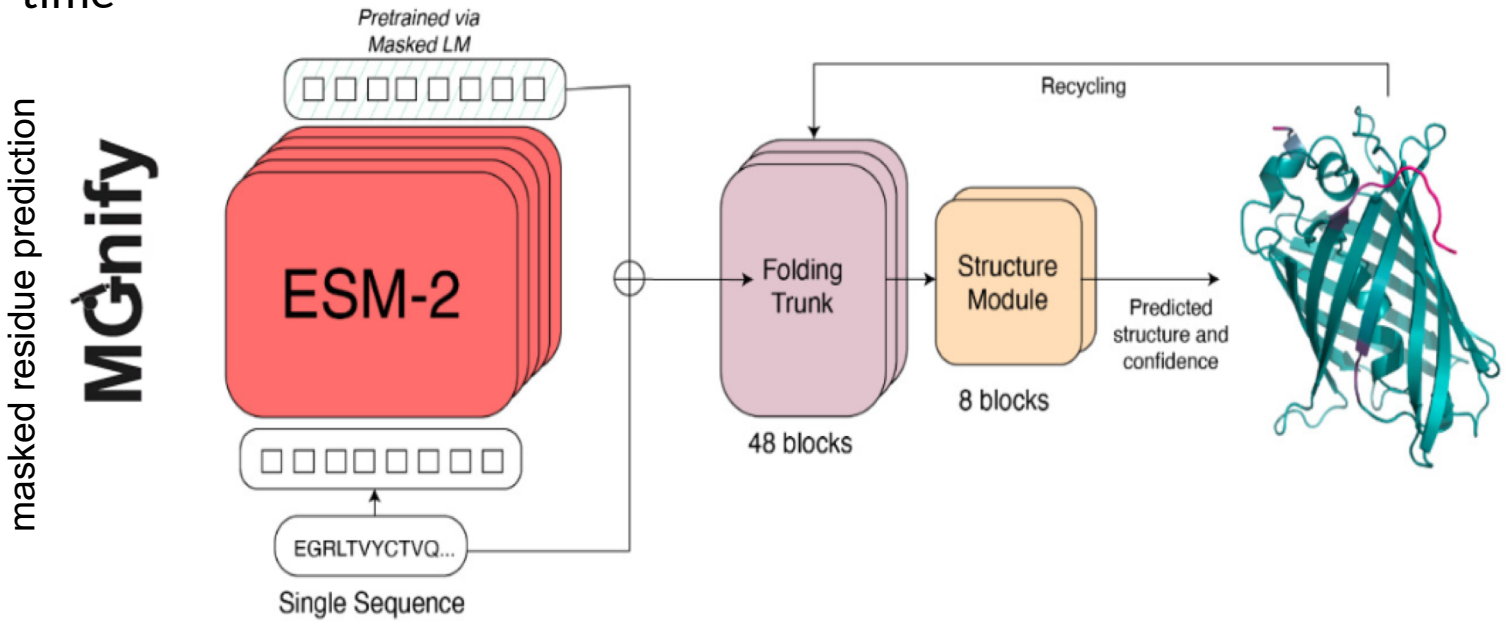
ESM model architecture

- ESM model works very similarly to AlphaFold model (Folding Trunk & Structure module, recycling described earlier)
- Difference lies in the ESM-2 module, which is a complex, pre-trained language model trained on the evolutionary focused database MGnify



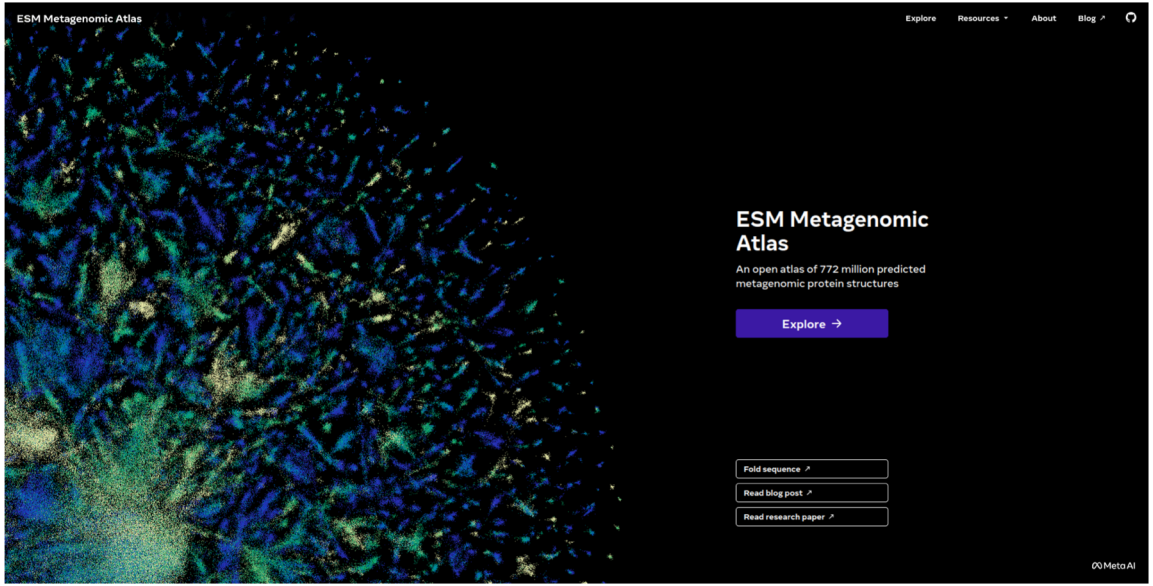
Compared to AlphaFold:

- ESMFold relies on the token embeddings from the large pre-trained protein language model stem
- ESMFold does not perform a multiple sequence alignment (MSA) step at inference time



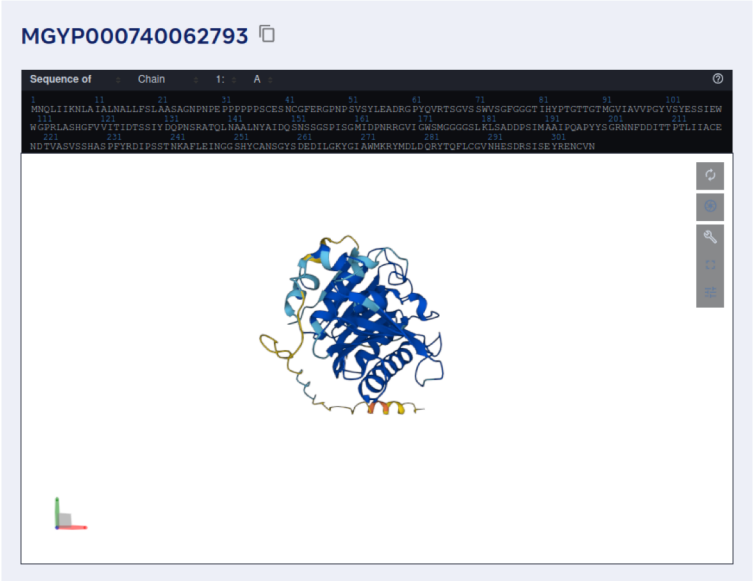
ESM Atlas

- ESM Atlas - Database with protein structure predictions
- Protein structure source: MGnify Database
- Available additional tools for quick 3D protein prediction



Prediction evaluation

- Structures can be evaluated using pLDDT value.
- Values range between 0 and 1 with higher scores representing higher model confidence.



RosettaFold and Other Modeling Approaches

- Introduction to other models like RosettaFold, explanations of the modelling approaches.
- Comparison to AlphaFold, and ESMFold.
- Discussion of the strengths and weaknesses of each approach.
- Case studies showcasing the utility of other algorithms.

How RosettaFold Works

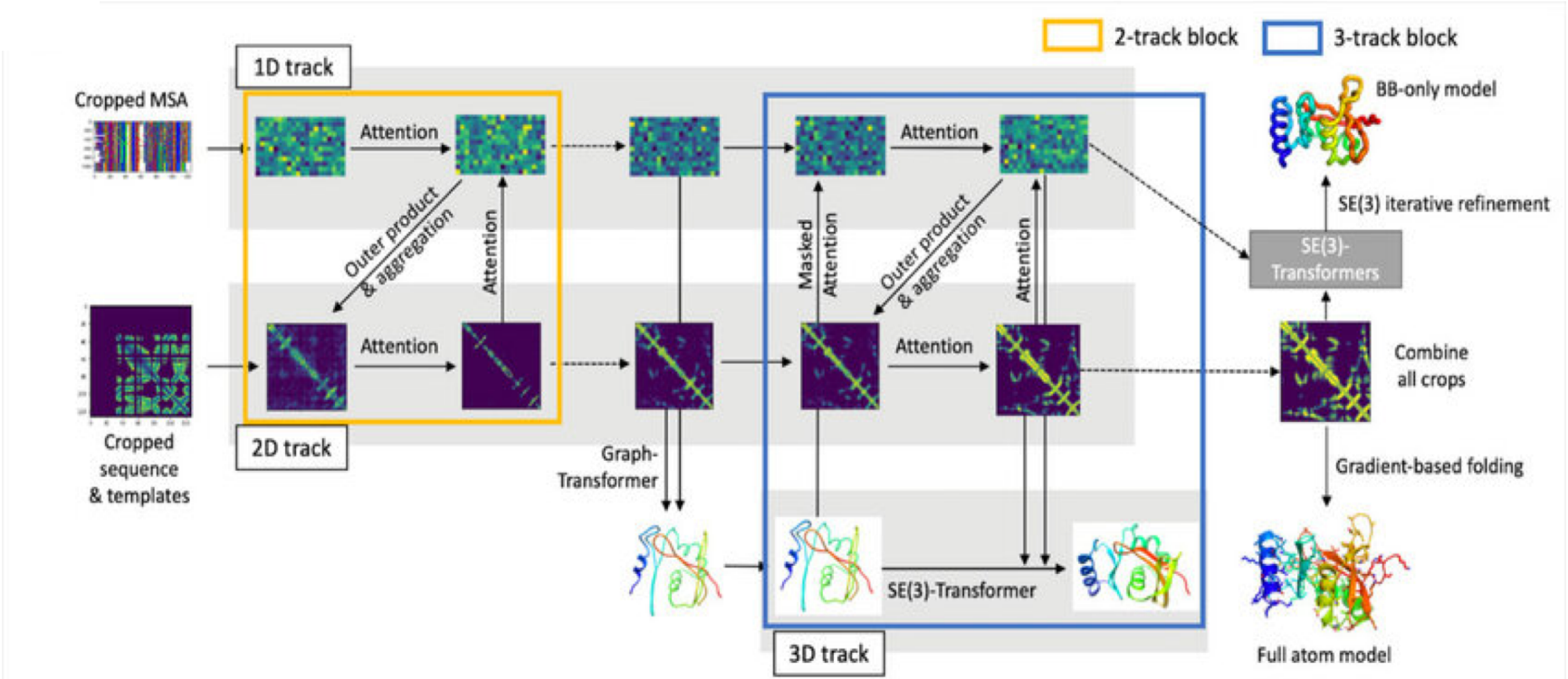
- Combines deep learning models with physical and biological insights.
- Uses multiple neural network components to process sequences and predict structures.

„RoseTTAFold is a “three-track” neural network, meaning it simultaneously considers patterns in protein sequences, how a protein’s amino acids interact with one another, and a protein’s possible three-dimensional structure.”
„…. allowing the network to collectively reason about the relationship between a protein’s chemical parts and its folded structure.”

- Integrates evolutionary information about protein families.

Baek M, *et al.* *Accurate prediction of protein structures and interactions using a 3-track network*. **Science**. July 2021

- RoseTTAFold architecture with 1D, 2D, and 3D attention tracks
- Multiple connections between tracks
- Relationships within and between sequences, distances, and coordinates

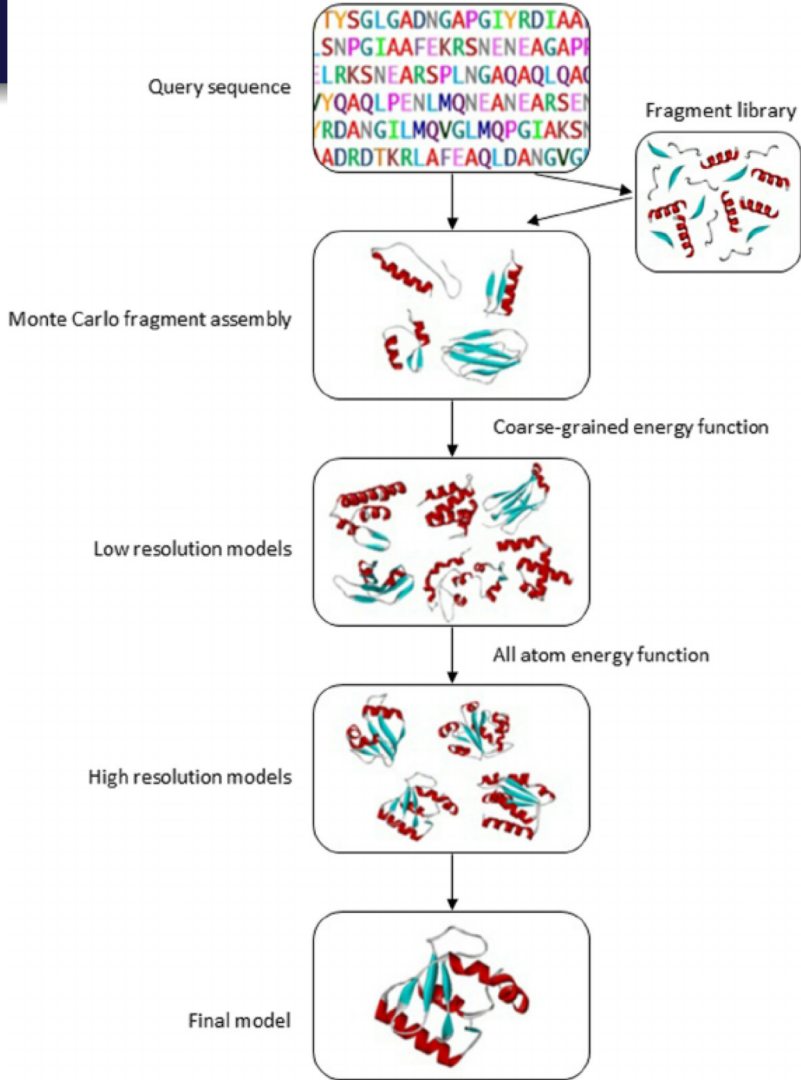


How RosettaFold Works

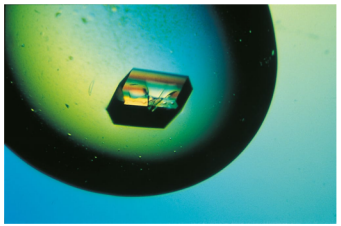
- RosettaFold architecture consists of fragment assembly, Monte Carlo sampling, and energy minimization steps to predict protein structures.

- Fragment assembly allows building plausible structures from short sequences, Monte Carlo refines conformational space, and energy minimization optimizes models.

RoseTTAFold code is available through the GitHub website



Thank you for your attention!



AlphaFold
Protein Structure
Database

RoseTTAFold

ESM
Metagenomic
Atlas

References

- https://www.nature.com/articles/nature07814
- https://www.researchgate.net/publication/281540998_General_overview_on_structure_prediction_of_twilight-zone_proteins
- https://www.ipd.uw.edu/2021/07/rosettafold-accurate-protein-structure-prediction-accessible-to-all/
- https://www.researchgate.net/publication/353282026_Accurate_prediction_of_protein_structures_and_interactions_using_a_three-track_neural_network
- https://news.yale.edu/2023/07/12/genetic-screen-reveals-protein-primed-stop-covid-19-virus
- https://www.nature.com/articles/s41586-021-03819-2
- https://www.frontiersin.org/articles/10.3389/frai.2022.875587/full

https://www.haojunheng.com/files/talks/alphafold.pdf