Protein Structures

Primary Structure

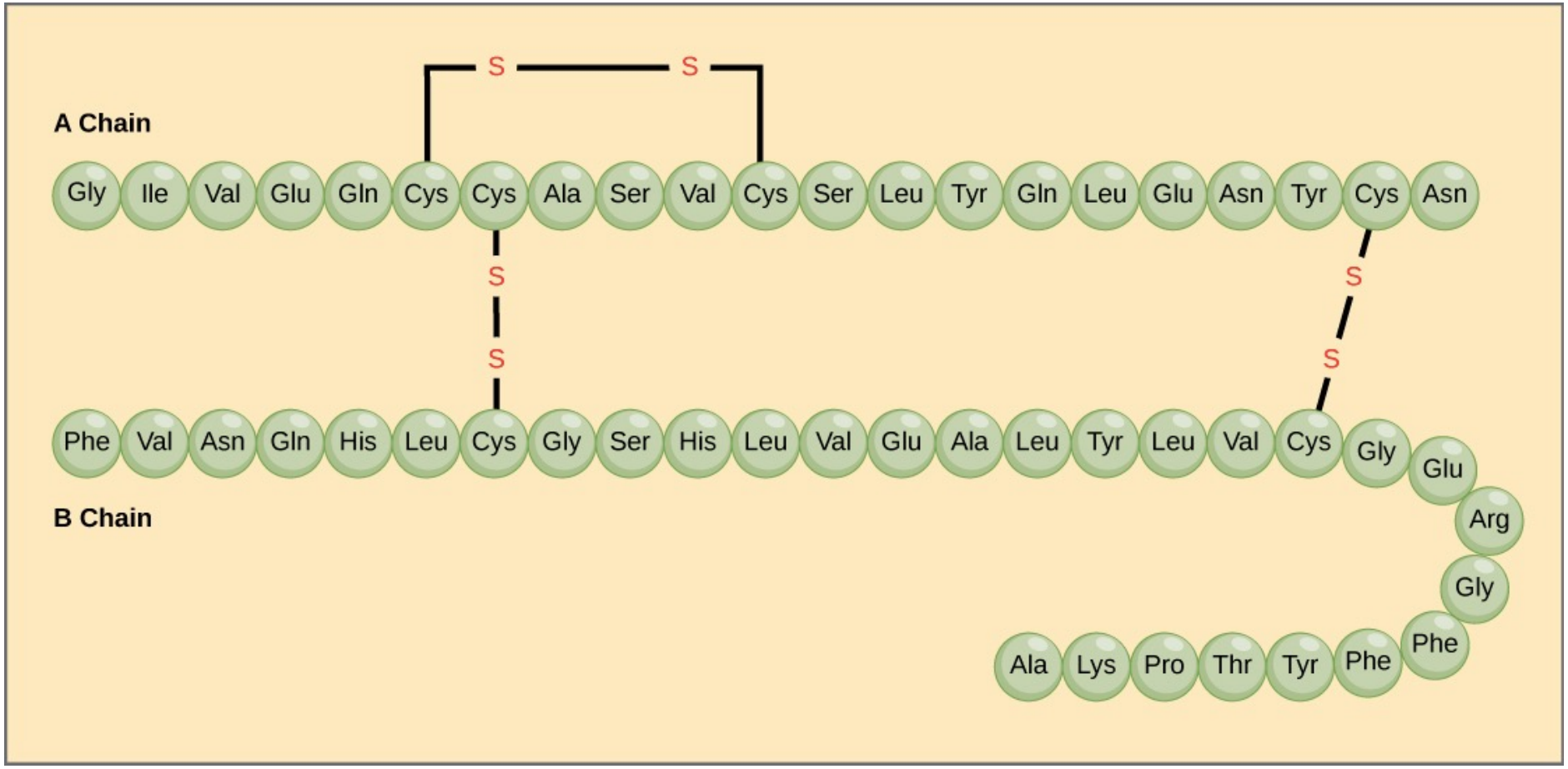
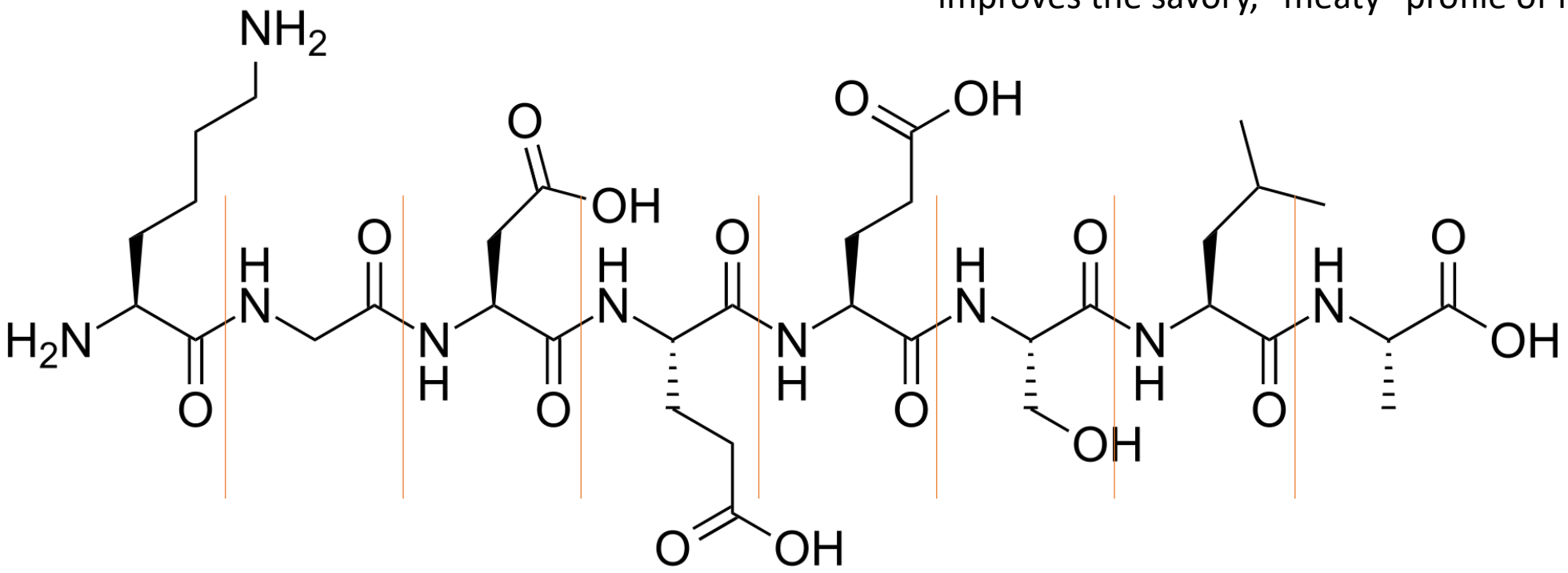


Image Credit: Khanacademy

Beefy meaty peptide (delicious peptide)

KGDEESLA

Originally isolated from beef soup by Yamasaki and Maekawa in 1978, it is recognized for its potential as a natural **flavor enhancer** that mimics or improves the savory, "meaty" profile of food.

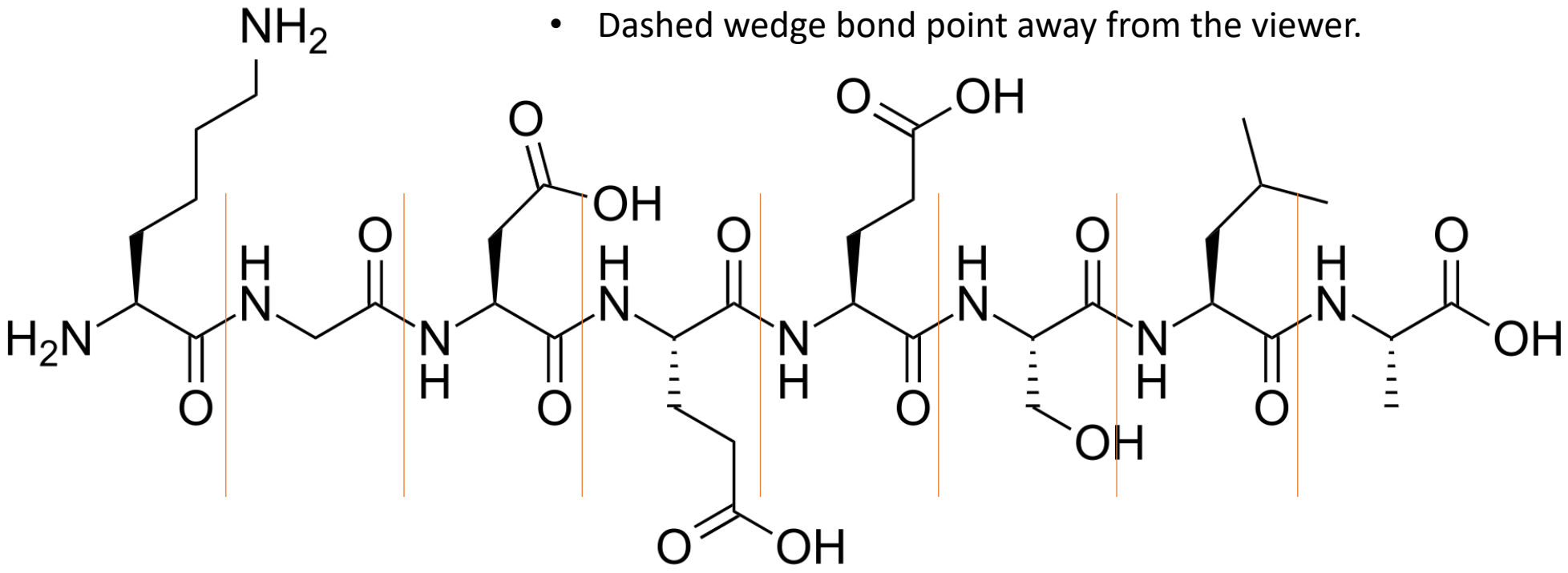


K G D E E S L A

Skeletal (line-angle) Structure

KGDEESLA

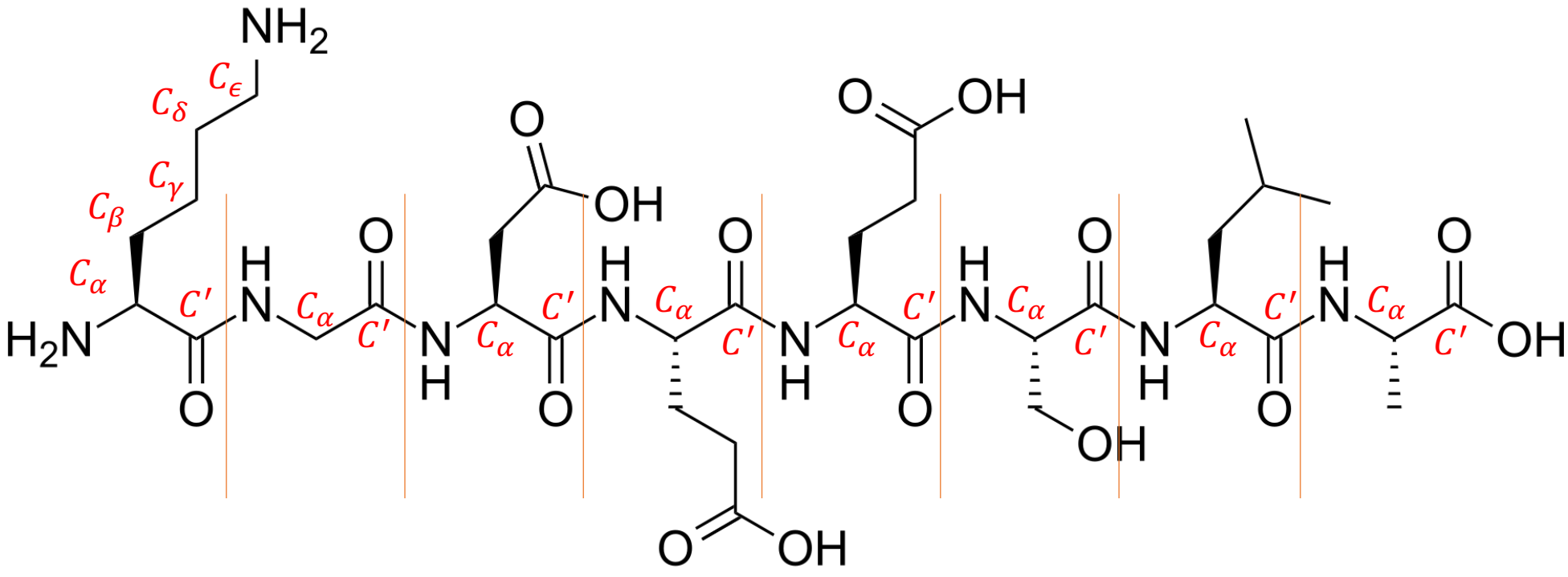
- C atom label is omitted, assume each node is a C atom.
- H atoms often omitted, can be inferred by the valence rule.
- Edges are covalent bonds.
- Solid wedge bonds come out of the plane (pointing towards the viewer).
- Dashed wedge bond point away from the viewer.



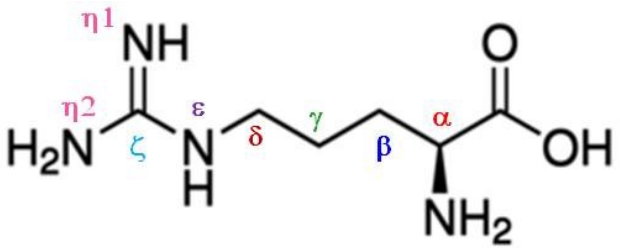
Skeletal (line-angle) Structure

KGDEESLA

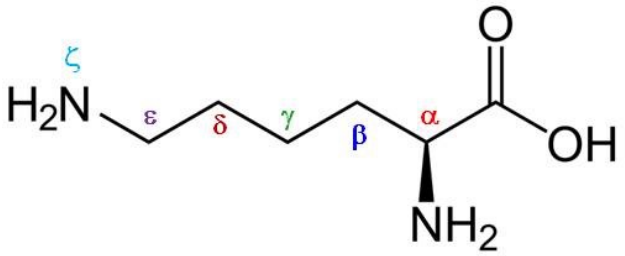
- Backbone is N-C_α-C'-N-C_α-C'- …
- Side chains attach to C_α
- C' is also called the carbonyl carbon and often just referred to as C.



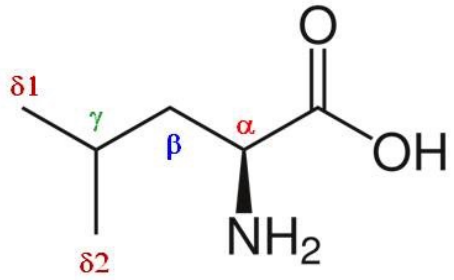
Names of the Side Chain Carbon Atoms



Arginine



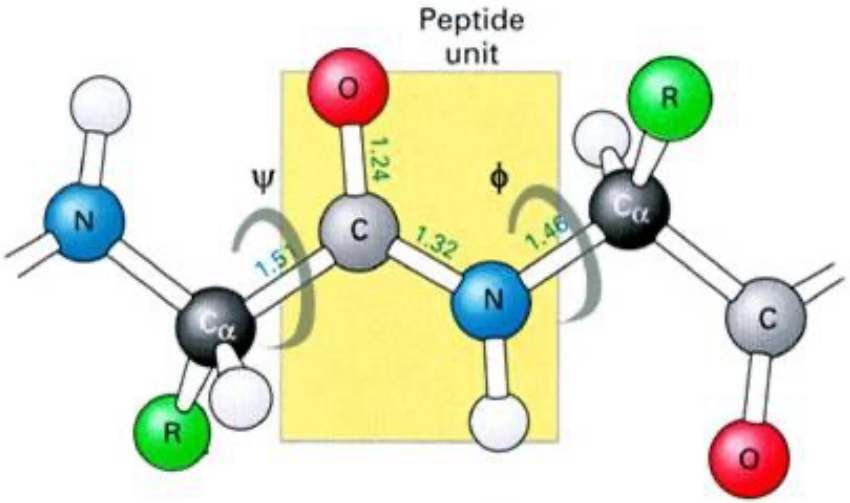
Lysine

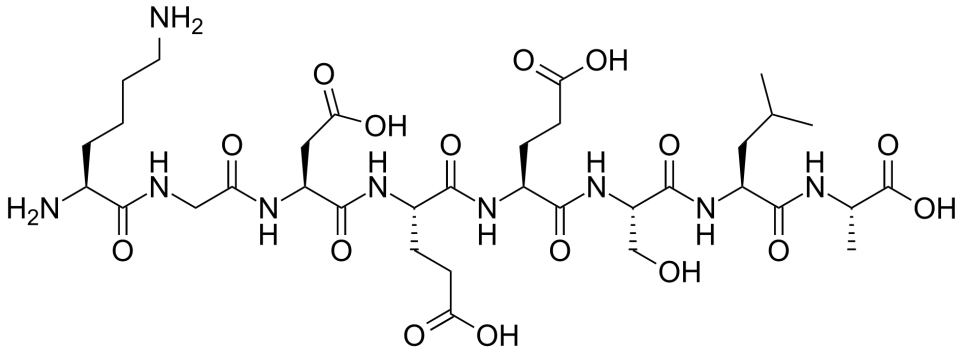


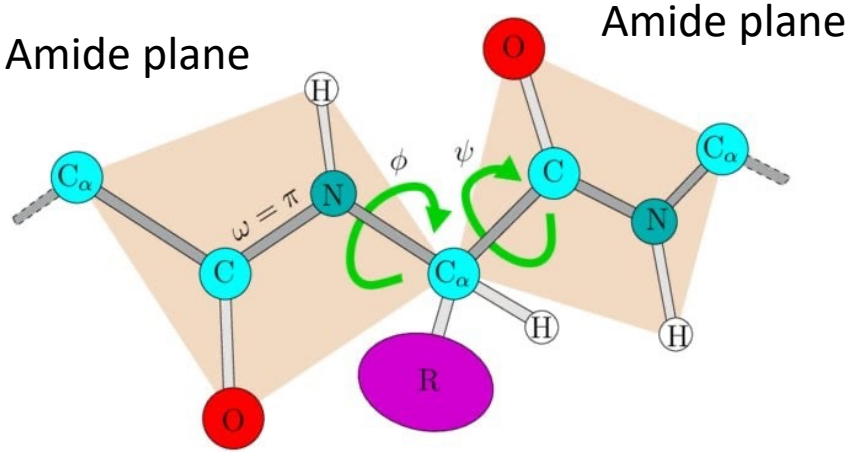
Leucine

Torsion Angles

- Torsion angles provide the spatial flexibility of the protein backbone structure.
- Two adjacent C_α, and the C, N in between usually form a rigit plane, named amide plane.
- The ϕ and ψ angles are flexible.
- Length unit is ångström (Å) = 0.1 nm







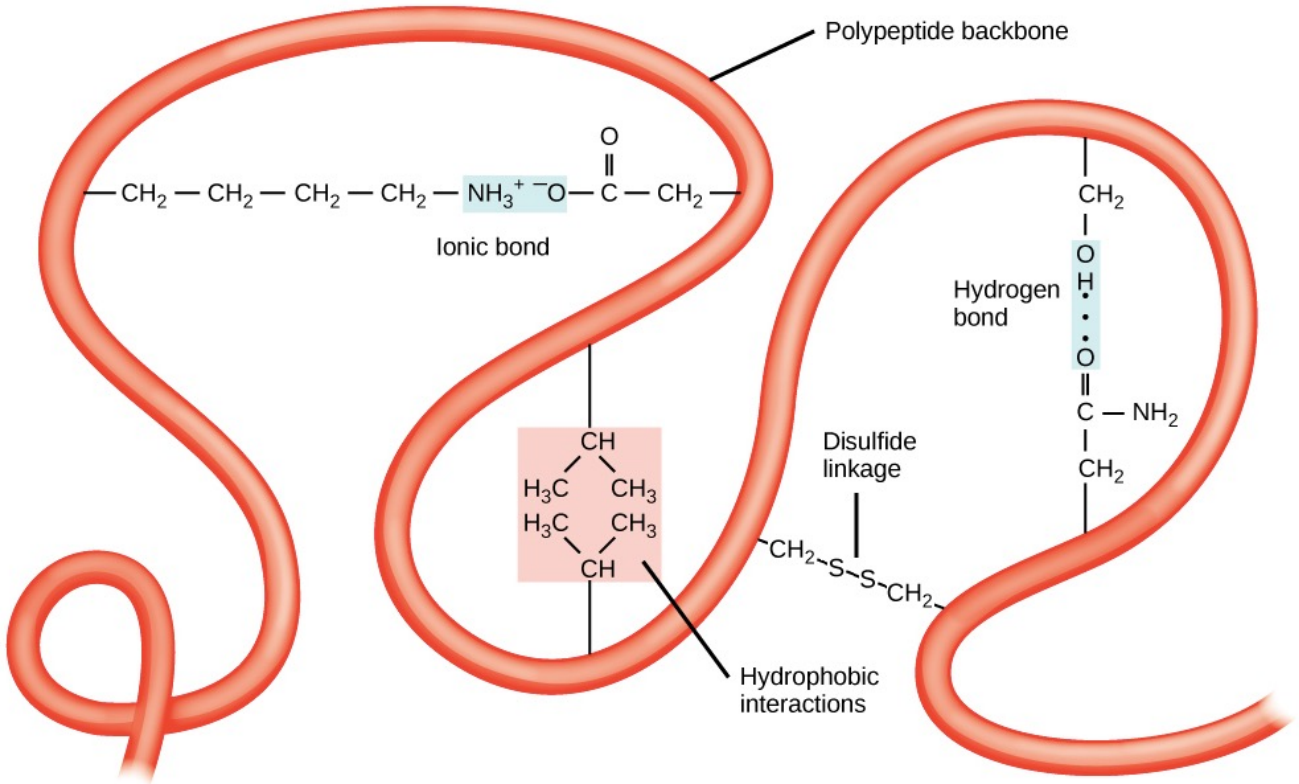
Forces Stablizing Protein Structure

- Protein structure is stabilized by non-covalent interactions and covalent bonds.
- Hydrophobic interaction
  - Nonpolar side chains avoid water.
  - They cluster inside the protein to form a hydrophobic core.
- Hydrogen bonds
  - C=O ⋯ H–N
- Ionic interaction (salt bridges)
  - D^- --- K^+
  - E^- --- R^+
- van der Waals interactions
  - Weak attractive forces between atoms that are **very close together**.
- Disulfide bonds
  - Cys–SH + HS–Cys → Cys–S–S–Cys

Gibbs Free Energy

- G (Gibbs free energy) is a thermodynamic state function that represents the energy available to do useful work in a system at constant temperature and pressure.
- The change from an unfolded structure to the folded structure will form the aforementioned iteractions, reducing the G.
- Or $\Delta G < 0$. This will cause the protein to fold.

Forces Stablizing the Structure



- Certain interactions reduces the “free energy”.

Anfinsen's Dogma

20 July 1973, Volume 181, Number 4096 SCIENCE

Principles that Govern the Folding of Protein Chains

Christian B. Anfinsen

The telegram that I received from the Swedish Royal Academy of Sciences specifically cites ". . . studies on ribonuclease, in particular the relationship between the amino acid sequence and the biologically active conformation. . . ." The work that my colleagues and I have carried out on the nature of the process that controls the folding of polypeptide chains into the unique three-dimensional structures of proteins was, indeed, strongly influenced by observations on the ribonuclease molecule. Many others, including Anson and Mirsky (*1*) in the 1930's and Lumry and Eyring (*2*) in the 1950's, had observed and discussed the reversibility of denaturation of proteins. However, the true elegance of this consequence of natural selection was dramatized by the ribonuclease work, since the refolding of this molecule, after full denaturation by reductive cleavage of its four disulfide bonds (Fig. 1), required that only 1 of the 105 possible pairings of eight sulfhydryl groups to form four disulfide linkages take place. The original observations that led to this conclusion were made together with my colleagues Michael Sela and Fred White in 1956–1957 (*3*). These were, in actuality, the beginnings of a long series of studies that rather vaguely aimed at the eventual total synthesis of the protein. As we all know, Gutte and Merrifield (*4*) at the Rockefeller Institute, and Ralph Hirschman and his colleagues at the Merck Research Institute (*5*), have now accomplished this monumental task.

The studies on the renaturation of fully denatured ribonuclease required many supporting investigations (*6–8*) to establish, finally, the generality which we have occasionally called (*9*) the "thermodynamic hypothesis." This hypothesis states that the three-dimensional structure of a native protein in its normal physiological milieu (solvent, *pH*, ionic strength, presence of other components such as metal ions or prosthetic groups, temperature, and other) is the one in which the Gibbs free energy of the whole system is lowest; that is, that the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment. In terms of natural selection through the "design" of macromolecules during evolution, this idea emphasized the fact that a protein molecule only makes stable, structural sense when it exists under conditions similar to those for which it was selected—the so-called physiological state.

After several years of study on the ribonuclease molecule it became clear to us, and to many others in the field of protein conformation, that proteins devoid of restrictive disulfide bonds or other covalent cross-linkages would make more convenient models for the study of the thermodynamic and kinetic aspects of the nucleation, and subsequent pathways, of polypeptide chain folding. Much of what I will review deals with studies on the flexible and convenient staphylococcal nuclease molecule, but I will first summarize some of the older background experiments on bovine pancreatic ribonuclease itself.

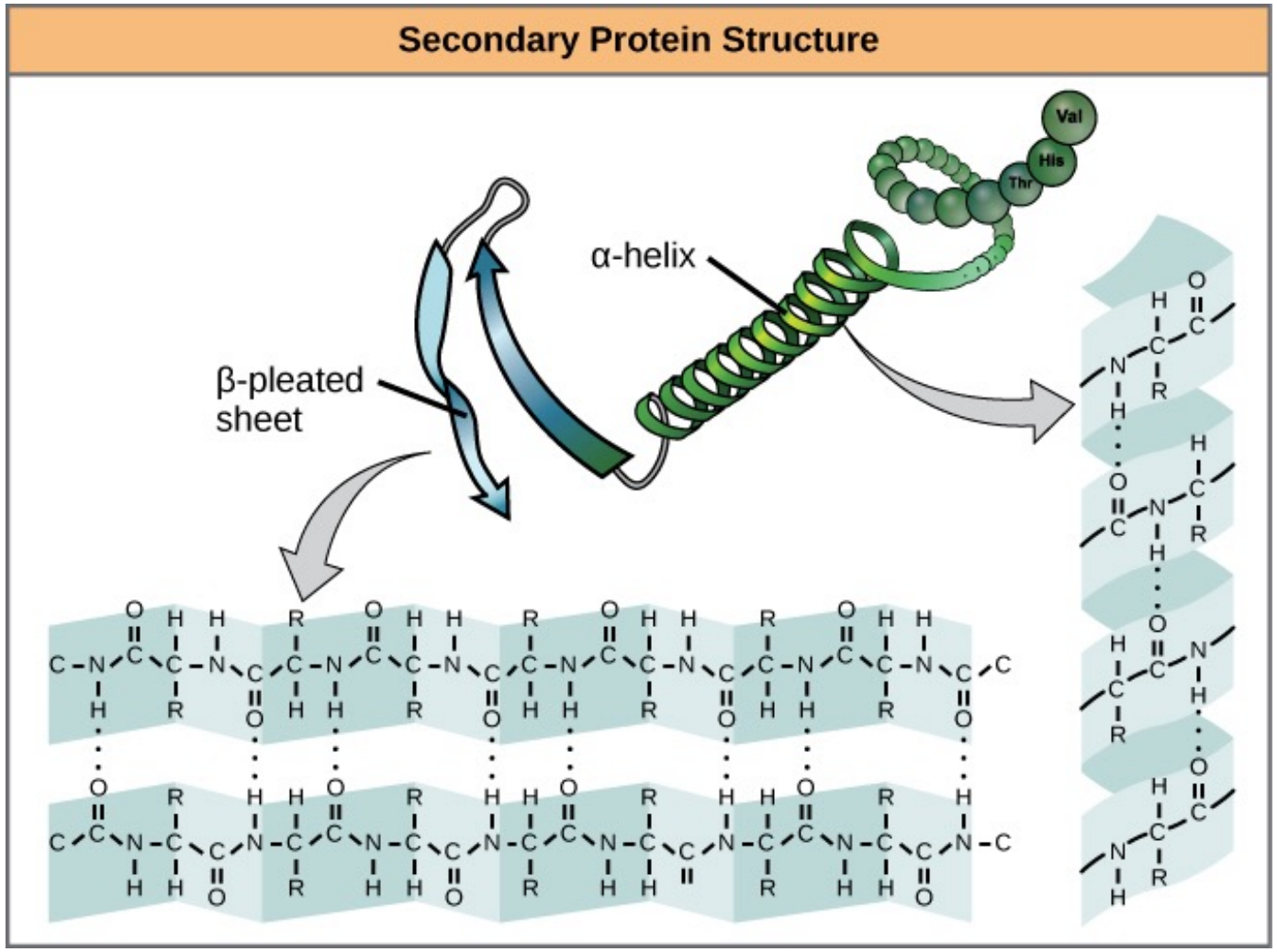
Support for the "Thermodynamic Hypothesis"

An experiment that gave us a particular satisfaction in connection with the translation of information in the linear amino acid sequence into native conformation involved the rearrangement of so-called "scrambled" ribonuclease (*8*). When the fully reduced protein, with eight SH groups, is allowed to reoxidize under denaturing conditions such as exist in a solution of 8 molar urea, a mixture of products is obtained containing many or all of the possible 105 isomeric disulfide bonded forms (schematically shown at the bottom right of Fig. 2). This mixture is essentially inactive—having on the order of 1 percent the activity of the native enzyme. If the urea is removed and the "scrambled" protein is exposed to a small amount of a sulfhydryl group–containing reagent such as mercaptoethanol, disulfide interchange takes place, and the mixture eventually is converted into a homogeneous product, indistinguishable from native ribonuclease. This process is driven entirely by the free energy of conformation that is gained in going to the stable, native

The author is chief of the Laboratory of Chemical Biology, National Institute of Arthritis, Metabolic and Digestive Diseases, National Institutes of Health, Bethesda, Maryland 20014. This article is the lecture he delivered in Stockholm, Sweden, on 11 December 1972 when he received the Nobel Prize for Chemistry, a prize he shared with Stanford Moore and William H. Stein. It is published here with the permission of the Nobel Foundation and will also be included in the complete volume of *Les Prix Nobel en 1972* as well as in the series Nobel Lectures (in English) published by the Elsevier Publishing Company, Amsterdam and New York. Dr. Moore's and Dr. Stein's combined lecture appeared as a single article in the 4 May issue of *Science*, page 458.

"thermodynamic hypothesis." This hypothesis states that the three-dimensional structure of a native protein in its normal physiological milieu (solvent, *pH*, ionic strength, presence of other components such as metal ions or prosthetic groups, temperature, and other) is the one in which the Gibbs free energy of the whole system is lowest; that is, that the native conformation is determined by the totality of interatomic interactions and hence by the amino acid sequence, in a given environment. In terms of natural selection through the "design" of macromolecules during evolution, this idea emphasized the fact that a protein molecule only makes stable, structural sense when it

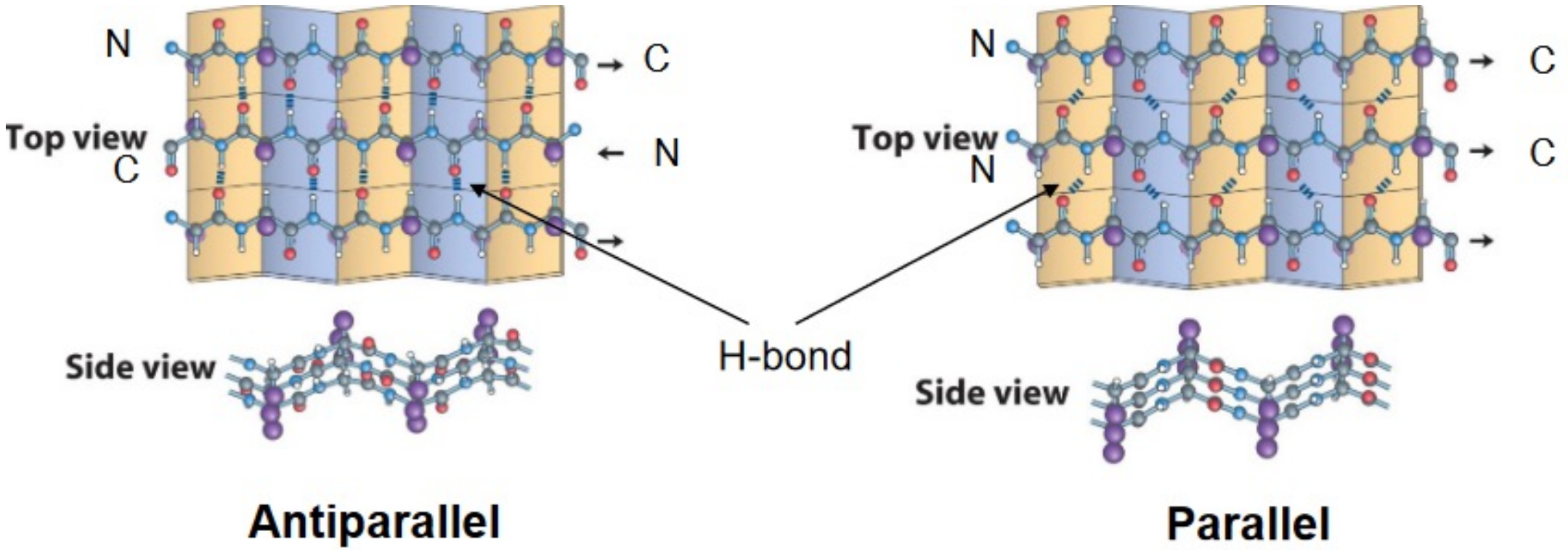
Secondary Structure



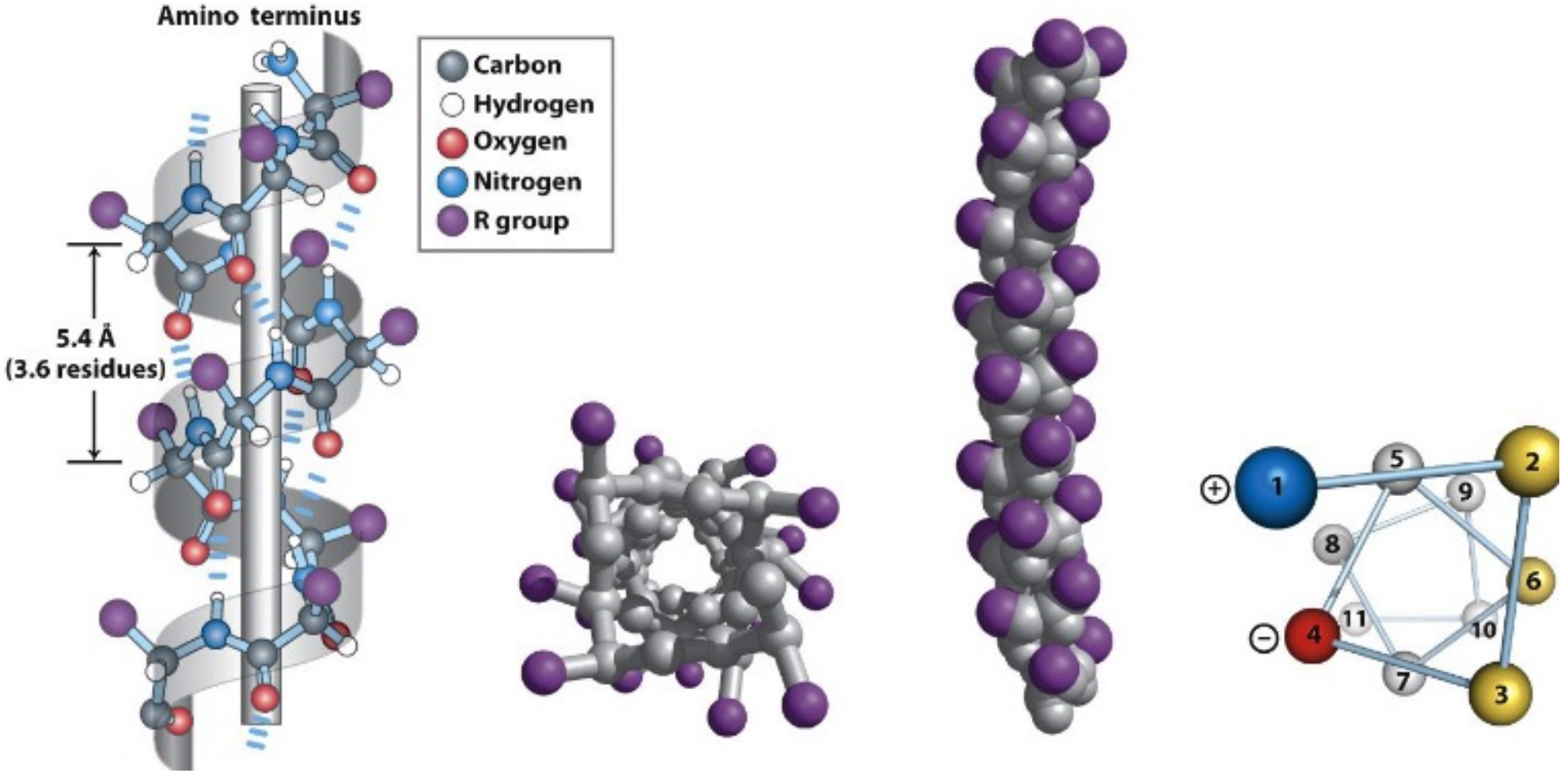
- Secondary structures are local, regular arrangements of polypeptide backbones.
- Stablized by hydrogen bonds.
- Alpha helix: 30-40%
- Beta sheet: 20-30%
- Loops/turns/irregular regions: 30-50%.

Image credit: OpenStax Biology.

Beta Sheet


N
C
Top view
C
N
Side view
H-bond
N
C
Top view
N
C
Side view
Antiparallel
Parallel

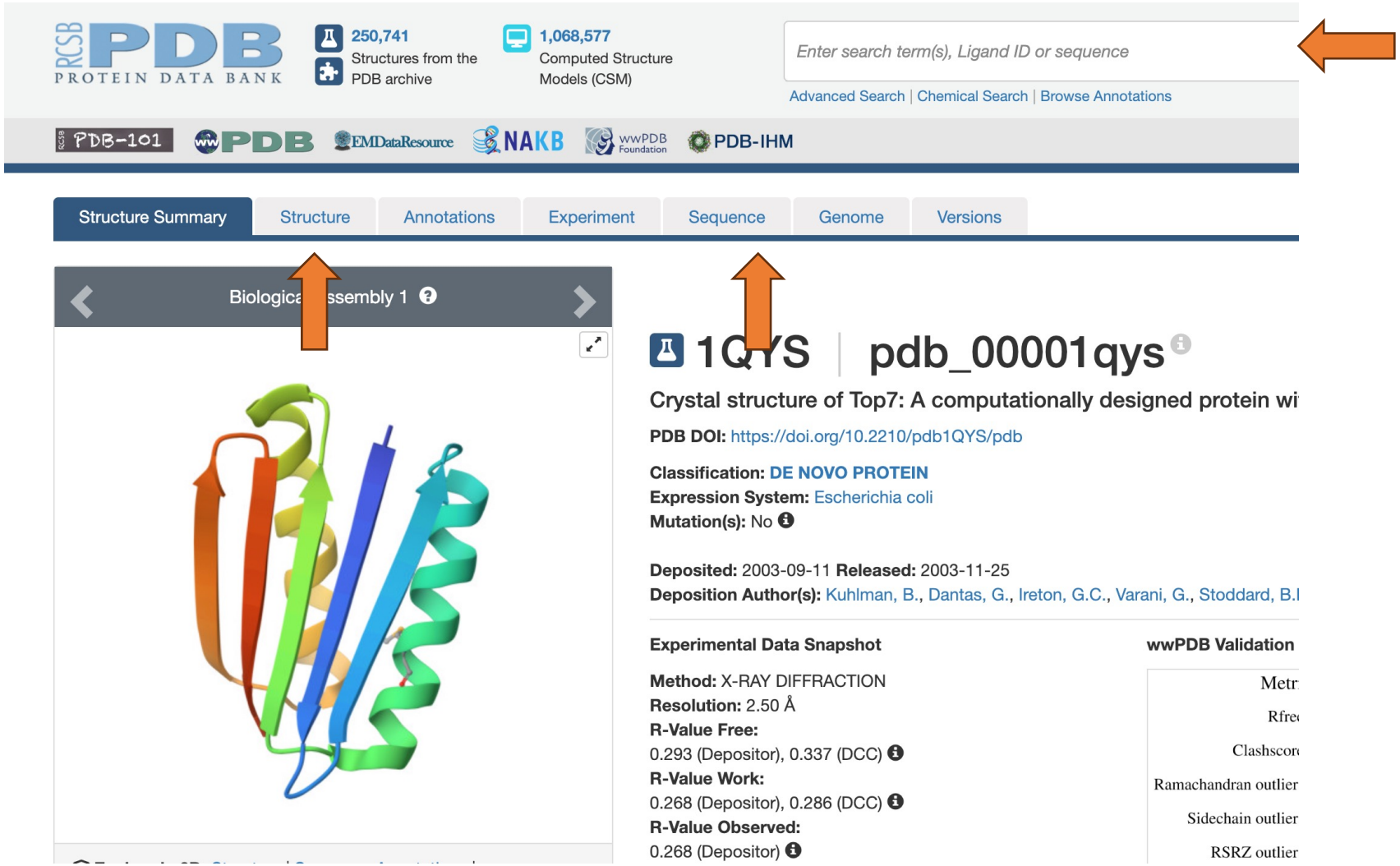
Alpha Helix


Amino terminus
Carbon
Hydrogen
Oxygen
Nitrogen
R group
5.4 Å
(3.6 residues)
1
2
3
4
5
6
7
8
9
10
11

PDB Database

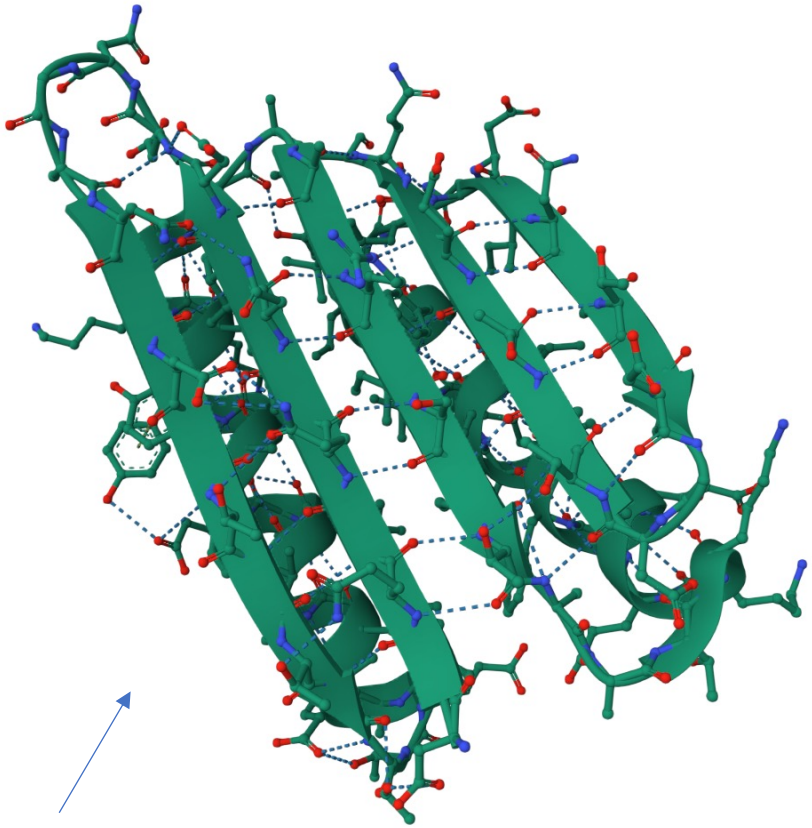
- https://www.rcsb.org/
- RCSB Protein Data Bank - US data center for the global Protein Data Bank (PDB) archive of 3D structure data for large biological molecules (proteins, DNA, and RNA).
- RCSB: Research Collaboratory for Structural Bioinformatics
- PDB was established in 1971 at Brookhaven National Laboratory under the leadership of Walter Hamilton and originally contained 7 structures. Today (Mar. 2026) it has over 250K experimentally determined structures and over 1 million computed structures.
- Critical resource for studying structures and structure prediction.

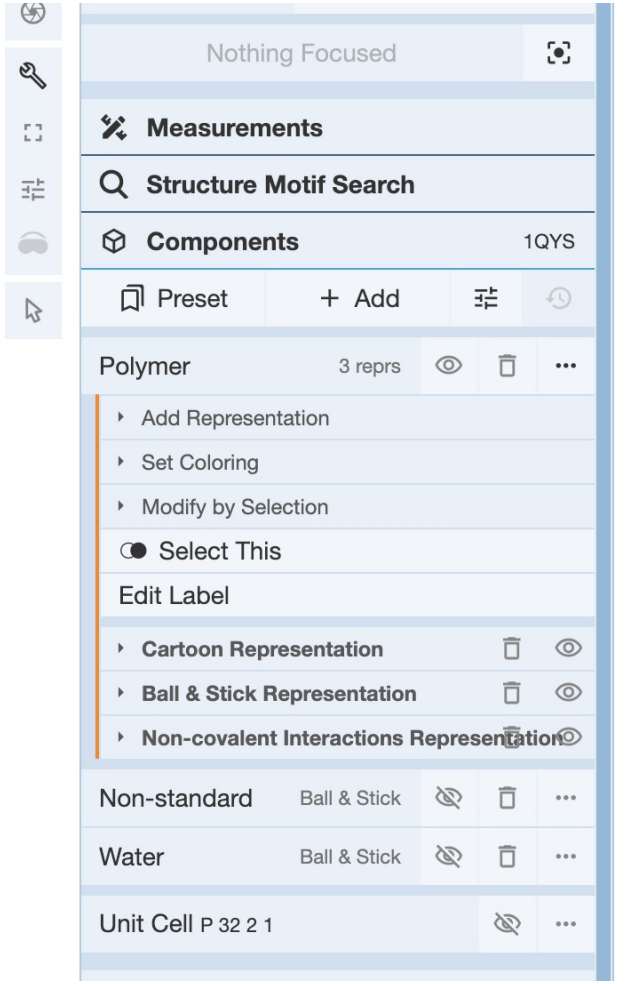
Example: Top7 Protein



Example: 1QYS (Top7)

toolbar on/off





Show/hide/edit components

Add views/representations

Show/hide/delete representations

Try: Cartoon, Ball & Stick, Non-covalent Interacton, Molecular surface (adjust opacity)

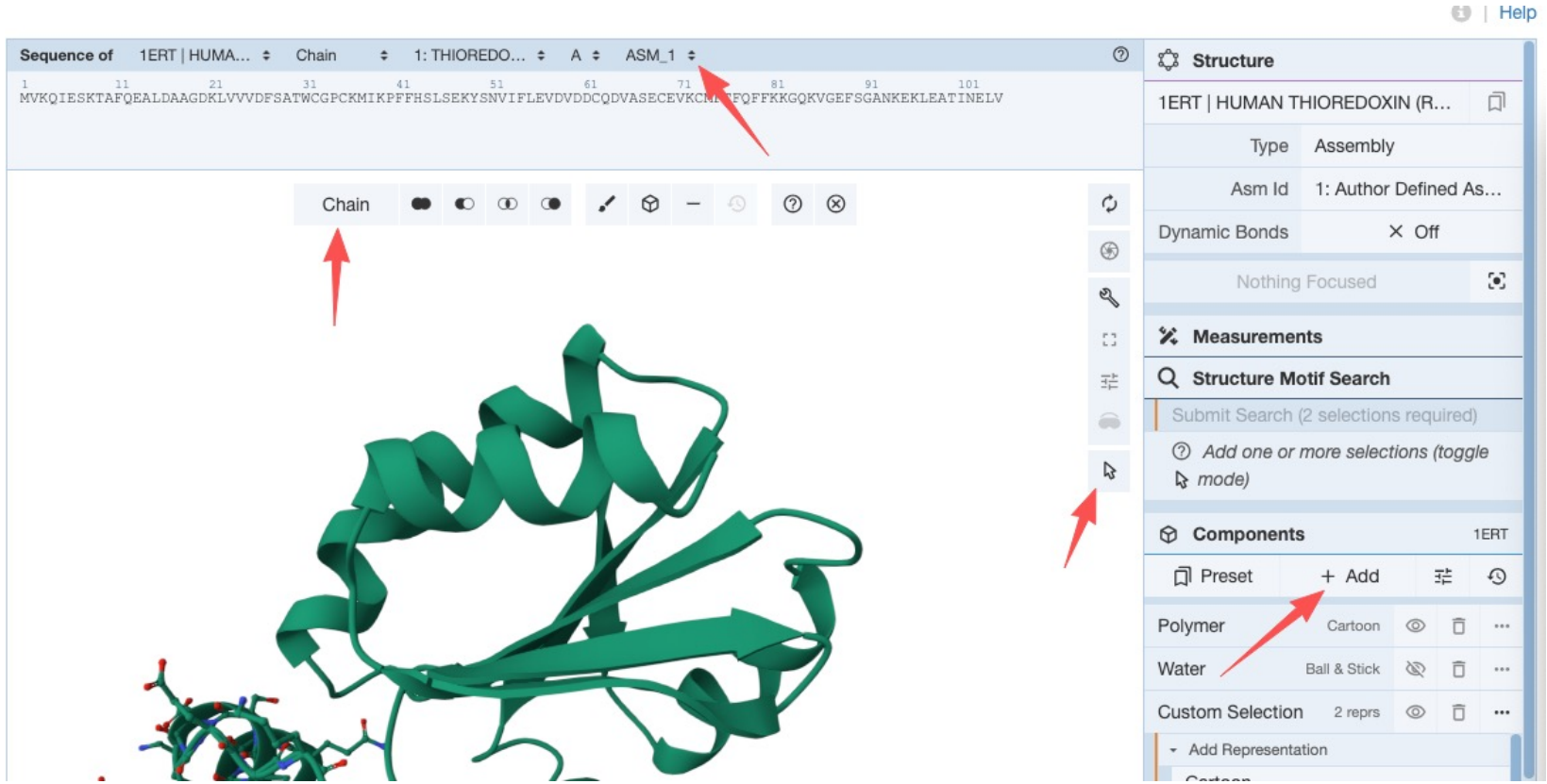
- Rotate and move the structure.
- Examine the hydrogen bonds stablizing α helix and β sheet.

Example: 1ERT

- Human Thioredoxin (Trx1) is an antioxidant. It reduces disulphide bonds, maintaining the balance.
- The CGPC motif is its active site.
- In crystal structure (necessary for X-ray structure determination) it forms a homodimer. In physiological state it is believed to be monomer, or both monomer and dimer exist.
- Check its structure.

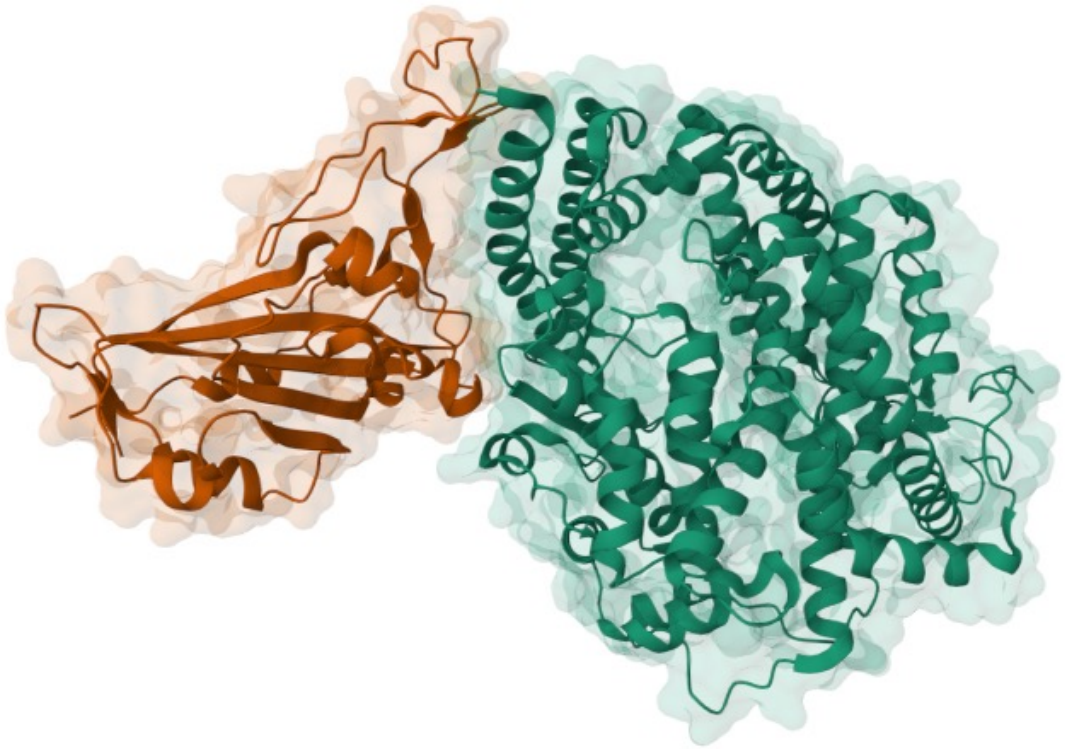
Example: 1ERT

- Check 1ERT structure.
- Identify the active site CGPC. Notice it’s buried in the dimer structure.
- Use select tool to select a chain and add as a component.
- Hide a chain so you can focus on the monomer.
- Now check the CGPC site again. It is exposed.



Example: 6M0J

- SARS-CoV-2 spike receptor-binding domain (RBD) bound with ACE2.
- Check the interaction between RBD and ACE2.



Structure Prediction

Levinhal's Paradox

- Levinhal's paradox is a thought experiment proposed by Cyrus Levinthal in 1969.
- Each additional amino acid gives two additional torsion angles.
- If each torsion angle has c configurations, protein has length L, then degree of freedom: $c^{2(L-1)}$.
- This is an astronomical number of possible combinations.
- Yet biology can find the right folding in microseconds to seconds.

Counter Argument

- This paper argues that if Gibbs free energy can be accuately computed, then random search bias against the locally unfavorable configurations will find the right structure quickly.

Proc. Natl. Acad. Sci. USA
Vol. 89, pp. 20–22, January 1992
Biophysics

Levinthal's paradox

ROBERT ZWANZIG, ATTILA SZABO, AND BIMAN BAGCHI*

Laboratory of Chemical Physics, National Institute of Diabetes and Digestive and Kidney Diseases, Building 2, National Institutes of Health, Bethesda, MD 20892

Contributed by Robert Zwanzig, October 7, 1991

ABSTRACT **Levinthal's paradox is that finding the native folded state of a protein by a random search among all possible configurations can take an enormously long time. Yet proteins can fold in seconds or less. Mathematical analysis of a simple model shows that a small and physically reasonable energy bias against locally unfavorable configurations, of the order of a few *kT*, can reduce Levinthal's time to a biologically significant size.**

time to the fully correct state can be very much shorter. In fact, this time can become biologically significant.

Model and Results

Since the goal is not to understand the folding of any particular protein, but only to present an elementary resolution of Levinthal's paradox, precise details of the protein structure will be ignored. Consequently, the model to be

Protein Structure Prediction

- **Energy fuction + search algorithm**.
- Energy function can be approximate or scoring function.
- Search algorithm can be monte carlo or heuristics.

NOBELPRISET I KEMI 2024
THE NOBEL PRIZE IN CHEMISTRY 2024

KUNGL. VETENSKAPS-AKADEMIEN
THE ROYAL SWEDISH ACADEMY OF SCIENCES


Photo: University of Washington

David Baker
University of Washington
USA

”för datorbaserad proteindesign”

“for computational protein design”


Photo: The Royal Society

Demis Hassabis
Google DeepMind
United Kingdom


Photo: BBVA Foundation

John M. Jumper
Google DeepMind
United Kingdom

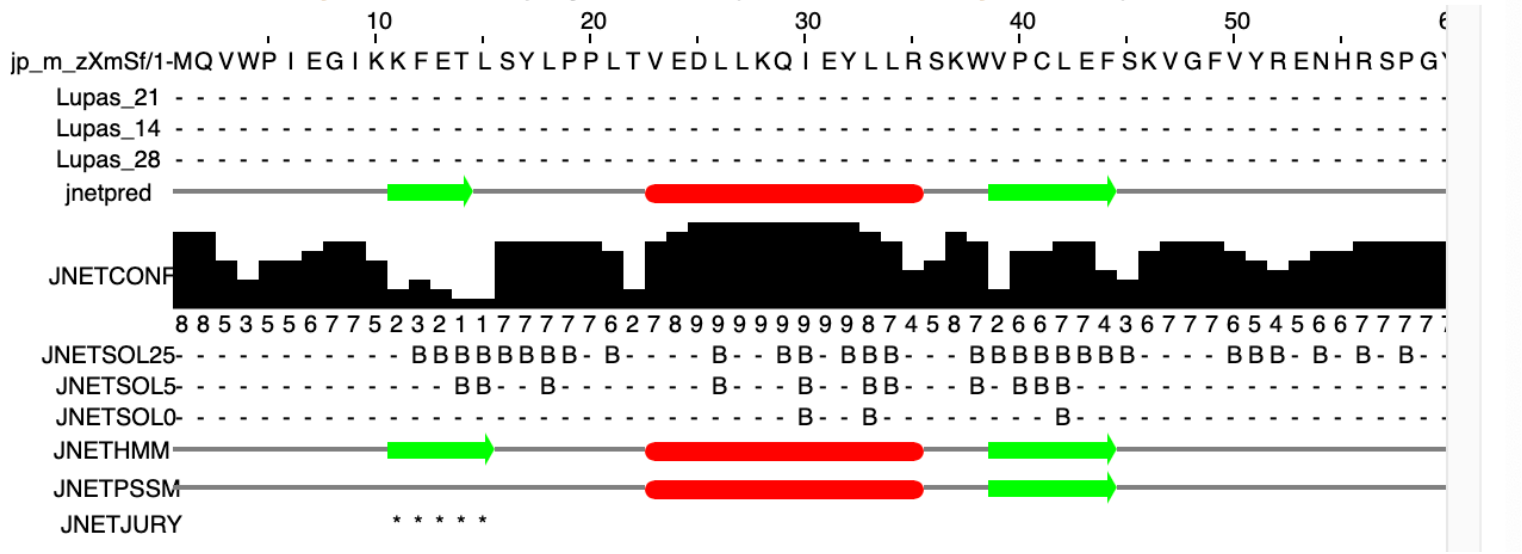
”för proteinstrukturprediktion”

“for protein structure prediction”

Secondary Structure Prediction

- Historically, the secondary structure is predicted first before predicting the tertiary structure.
- Jpred is one of the existing tools for secondary structure prediction from sequence
- https://www.compbio.dundee.ac.uk/jpred/

Red: α-helix.
Blue: β-sheet.
Gray: coil.



Secondary Structure Prediction

- Possible ways for secondary structure prediction include:
- Hidden Markov Model
- Neural network, e.g. sliding window, RNN, LSTM, GRN, transformer.
- Homology based.
- Multiple sequence alignment first – structure is more conserved than sequence.
- Today, very often the tertiary structure is predicted first, and then secondary structure is recognized afterward.

Protein Structure Prediction Methods

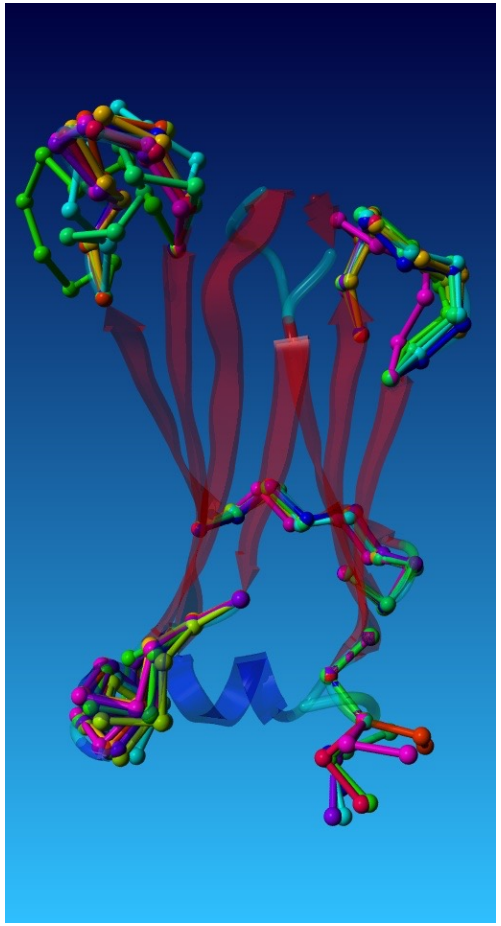
- Method 1. Homology modelling
  - Find an existing structure with a simliar sequence
  - Align the sequence with the template
  - Refine the structure
  - Typical tool: MODELLER
- Method 2. Threading
  - Regardless of the sequence similarity, just align the sequence with each template
  - Find the best fit
  - Refine the structure
  - Typical tool: RaptorX
- These two can all be regarded as template based.

Protein Structure Prediction Methods

- Method 3. Ab initio (de novo) prediction
  - Explore many possible conformation
  - Evaluate using an energy function
  - Choose the structure with the lowest energy
  - Typical tool: Rosetta
- Method 4. AI-based prediction
  - Typical tools: AlphaFold, OpenFold, RoseTTAFold
  - Belongs to the Ab initio category.
- These two can all be regarded as ab initio.

Overall Workflow of Traditional Ab initio Method

1. Secondary structure prediction
2. Backbone construction / fold assembly
3. Loop modeling
4. Side-chain packing



Loop modeling

Rosetta

- Software developed by David Baker (U. Washington) and collaborators (since late 1990s)
  - Software at: https://www.rosettacommons.org/software
  - Structure prediction server: http://robetta.bakerlab.org/
- Why use Rosetta as an example?
  - Among the better ab initio modeling packages (for some years it was the best)
  - Approach is similar to that of many ab initio modeling packages
  - Rosetta provides a common framework that has become very popular for a wide range of molecular prediction and design tasks.
- Knowledge based energy function + Knowledge based search

Rosetta Energy Function

TABLE I
COMPONENTS OF ROSETTA ENERGY FUNCTION[a]

Name	Description (putative physical origin)	Functional form	Parameters (values)
env[b]	Residue environment (solvation)	$\sum_i -\ln[P(\mathrm{aa}_i\|\mathrm{nb}_i)]$	i = residue index aa = amino acid type nb = number of neighboring residues[c] (0, 1, 2. . . 30, >30)
pair[b]	Residue pair interactions (electrostatics, disulfides)	$\sum_i \sum_{j>i} -\ln\left[\frac{P(\mathrm{aa}_i, \mathrm{aa}_j\|s_{ij}d_{ij})}{P(\mathrm{aa}_i\|s_{ij}d_{ij})P(\mathrm{aa}_i\|s_{ij}d_{ij})}\right]$	i, j = residue indices aa = amino acid type d = centroid–centroid distance (10–12, 7.5–10, 5–7.5, <5 Å) s = sequence separation (>8 residues)
SS[d]	Strand pairing (hydrogen bonding)	SchemeA : $\mathrm{SS}_{\phi,\theta} + \mathrm{SS}_{hb} + \mathrm{SS}_d$ SchemeB : $\mathrm{SS}_{\phi,\theta} + \mathrm{SS}_{hb} + \mathrm{SS}_{d\sigma}$ where $\mathrm{SS}_{\phi,\theta} = \sum_m \sum_{n>m} -\ln[P(\phi_{mn}, \theta_{mn}\|d_{mn}, \mathrm{sp}_{mn}, s_{mn})]$ $\mathrm{SS}_{hb} = \sum_m \sum_{n>m} -\ln[P(\mathrm{hb}_{mn}\|d_{mn}, s_{mn})]$ $\mathrm{SS}_d = \sum_m \sum_{n>m} -\ln[P(d_{mn}\|s_{mn})]$ $\mathrm{SS}_{d\sigma} = \sum_m \sum_{n>m} -\ln[\mathrm{P}(d_{mn}\sigma_{mn}\|\rho_m, \rho_n)]$	m, n = strand dimer indices; dimer is two consecutive strand residues V = vector between first N atom and last C atom of dimer m = unit vector between $\hat{V}_m$ and $\hat{V}_n$ midpoints x = unit vector along carbon–oxygen bond of first dimer residue y = unit vector along oxygen–carbon bond of second dimer residue ϕ, θ = polar angles between $\hat{V}_m$ and $\hat{V}_n$ (36° bins) hb = dimer twist, $\sum_{k=m,n} 0.5(\lvert\hat{\boldsymbol{m}}\cdot\hat{\boldsymbol{x}}_k\rvert + \lvert\hat{\boldsymbol{m}}\cdot\hat{\boldsymbol{y}}_k\rvert)$ (< 0.33, 0.33–0.66, 0.66–1.0, 1.0–1.33, 1.33–1.6, 1.6–1.8, 1.8–2.0) d = distance between $\hat{V}_m$ and $\hat{V}_n$ midpoints (< 6.5 Å) σ = angle between $\hat{V}_m$ and $\hat{M}$ (18° bins) sp = sequence separation between dimer-containing strands (< 2, 2–10, > 10 residues) s = sequence separation between dimers (>5 or >10) ρ = mean angle between vectors $\hat{m}, \hat{x}$ and $\hat{m}, \hat{y}$ (180° bins)

From Rohl et al., *Methods in Enzymology* 2004

- No need to remember.
- A lot of the values are based on statistics (e.g. log likelihood) instead of physical measurements.

Updated Rosetta Energy Function

sheet[e]	Strand arrangement into sheets	$-\ln\left[P(n_{\text{sheets}} n_{\text{lonestrands}} \mid n_{\text{strands}})\right]$	n_{sheets} = number of sheets $n_{\text{lone strands}}$ = number of unpaired strands n_{strands} = total number of strands
HS	Helix–strand packing	$\sum_m \sum_n -\ln\left[P(\phi_{mn}, \psi_{mn} \mid sp_{mn} d_{mn})\right]$	m = strand dimer index; dimer is two consecutive strand residues n = helix dimer index; dimer is central two residues of four consecutive helical residues $\hat{V}$ = vector between first N atom and last C atom of dimer ϕ, θ = polar angles between $\hat{V}_m$ and $\hat{V}_n$ (36° bins) sp = sequence separation between dimer-containing helix and strand (binned < 2, 2–10, >10 residues) d = distance between $\hat{V}_m$ and $\hat{V}_n$ midpoints (< 12 Å)
rg	Radius of gyration (vdw attraction; solvation)	$\sqrt{\langle d_{ij}^2 \rangle}$	i, j = residue indices d = distance between residue centroids
cbeta	Cβ density (solvation; correction for excluded volume effect introduced by simulation)	$\sum_i \sum_{sh} -\ln\left[\frac{P_{\text{compact}}(\text{nb}_{i,sh})}{P_{\text{random}}(\text{nb}_{i,sh})}\right]$	i = residue index sh = shell radius (6, 12 Å) nb = number of neighboring residues within shell[f] P_{compact} = probability in compact structures assembled from fragments P_{random} = probability in structures assembled randomly from fragments
vdw[g]	Steric repulsion	$\sum_i \sum_{j>i} \frac{\left(r_{ij}^2 - d_{ij}^2\right)^2}{r_{ij}};\ d_{ij} < r_{ij}$	i, j = residue (or centroid) indices d = interatomic distance r = summed van der Waals radii[h]

No need to remember.

Alford et al., Journal of Chemical Theory and Computation, 2017

Rosetta’s Search Stategy

- Step 1: Coarse search many times.
- Step 2: Refinement on each corse search result.

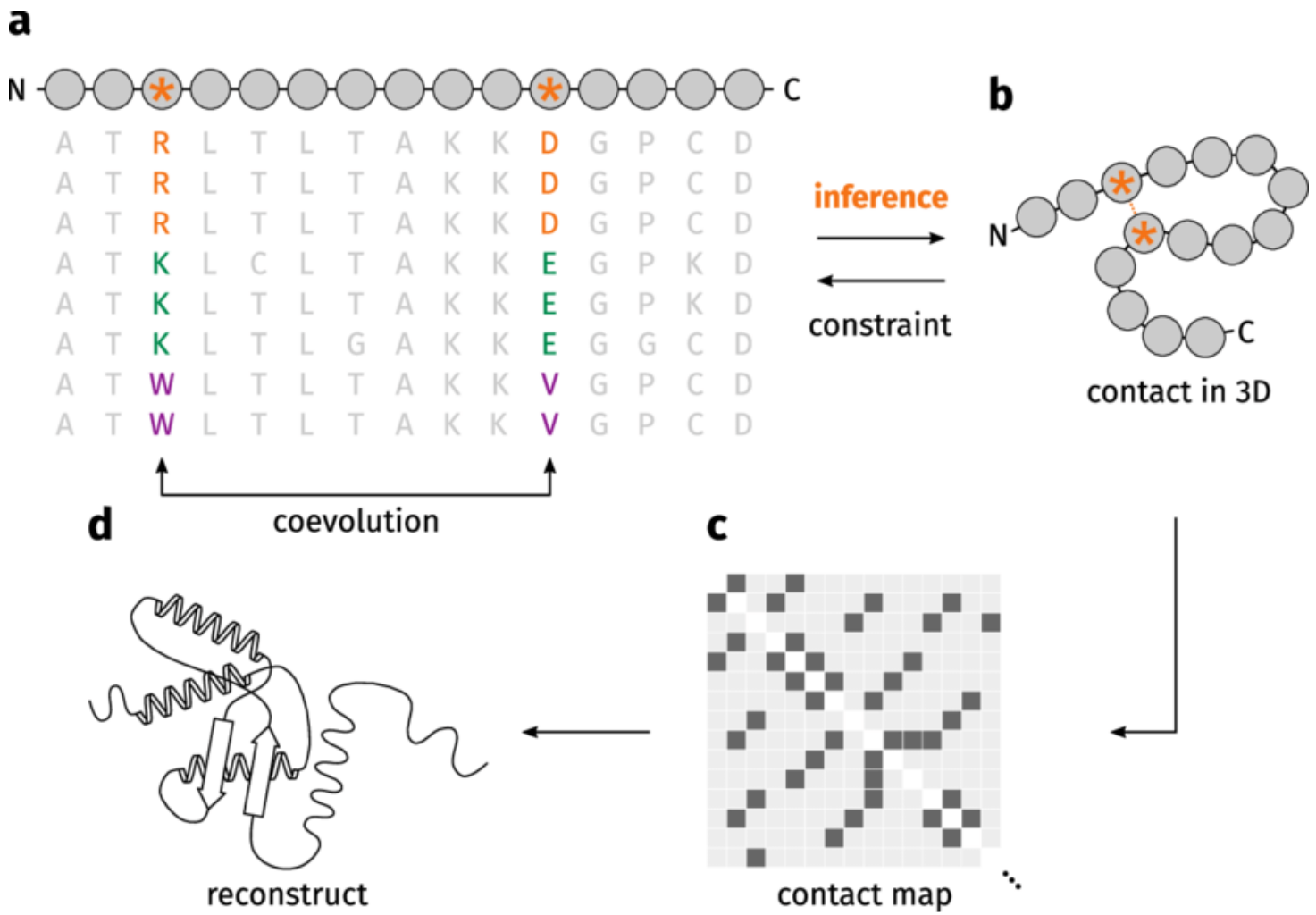
Coarse Search: Fragment Assembly

- Uses a large database of 3-residue and 9-residue fragments, taken from structures in the PDB
- Monte Carlo sampling algorithm proceeds as follows:
  1. Start with the protein in an extended conformation
  2. Randomly select a 3-residue or 9-residue section
  3. Find a fragment in the library whose sequence resembles it
  4. Consider a move in which the torsion angles of the selected section are replaced by those of the fragment
  5. Calculate the effect on the entire protein structure
  6. Evaluate the Rosetta energy function before and after the move
  7. Use the Metropolis criterion to accept or reject the move
  8. Return to step 2

Refinement

- Refinement is performed using the Rosetta all-atom energy function, after building in side chains
- Refinement involves a combination of Monte Carlo moves and energy minimization
- The Monte Carlo moves are designed to perturb the structure much more gently than those used in the coarse search
  - May still involve the use of fragments

Multiple Alignment Commonly Used in Prediction

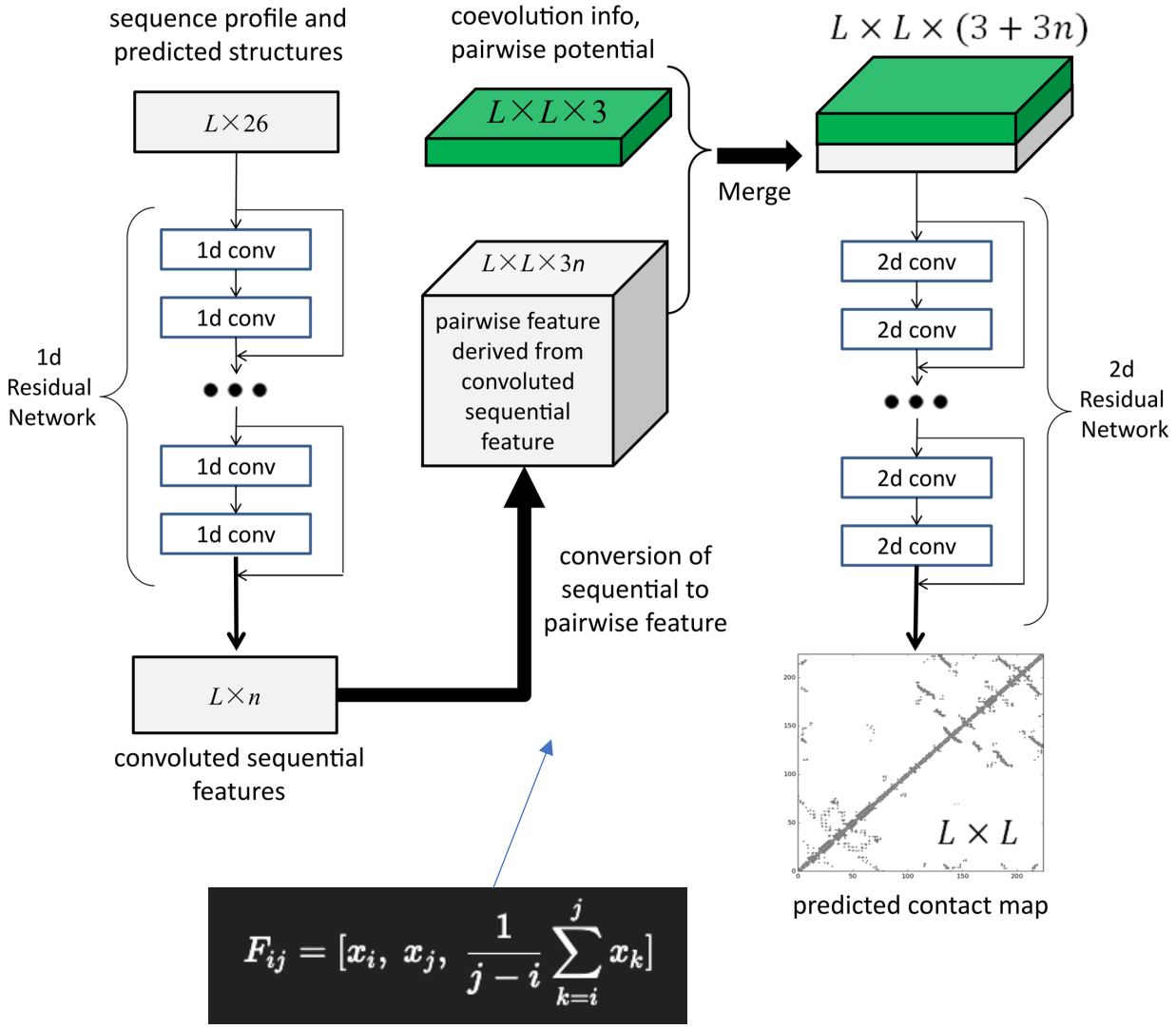


- Coevolution often suggests contact pairs.
- Contact map limits the possible structure configurations, serving as a constraint for search algorithm.
- Very helpful for determining the structure (e.g. fold assembly after secondary structure prediction).

Image credit: DOI:10.1038/s41598-019-55047-4

Raptor X Contact Map

- Use ResNet to predict the contact map.
- Use some traditional approach to predict structure assisted by contact map.
- Wang S, Sun S, Li Z, Zhang R, Xu J (2017) Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. PLoS Comput Biol 13(1): e1005324. doi:10.1371/journal.pcbi.1005324



AlphaFold

- Contact mapping is used to calculate a "potential function", where variables are torsion angles.
- Gradient descient used for the search.
- Senior et al. Nature 577. 2020.

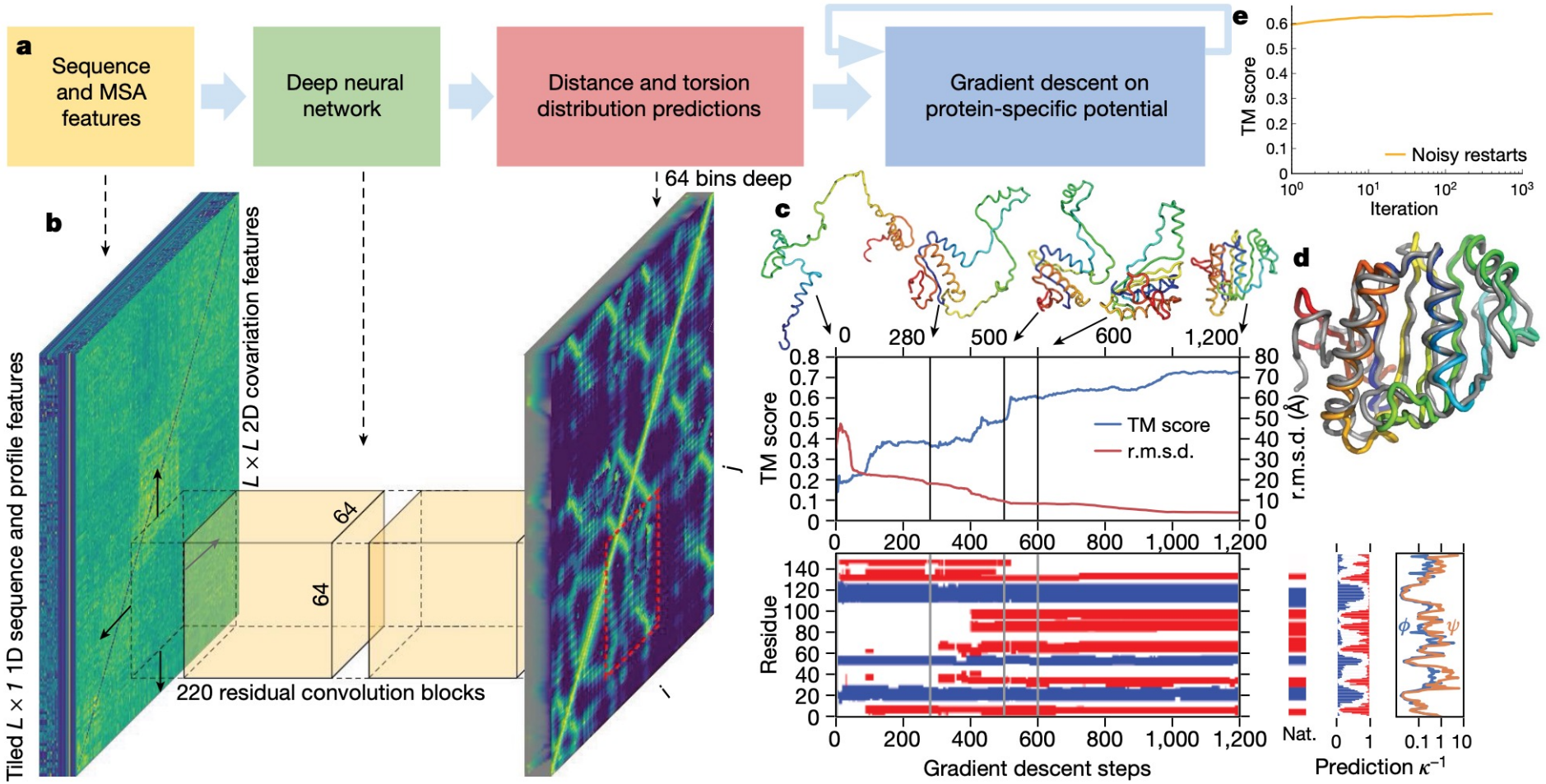
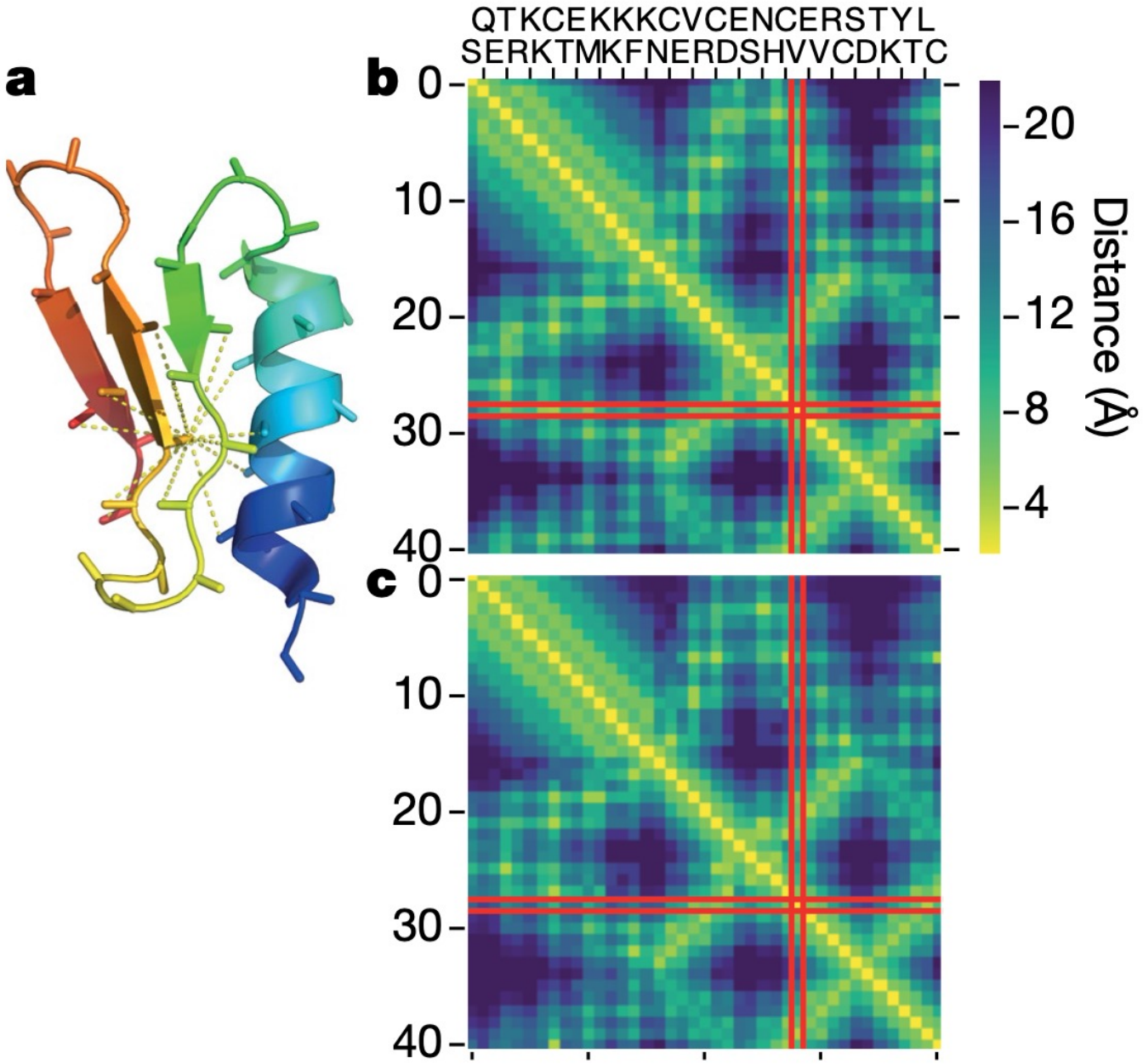


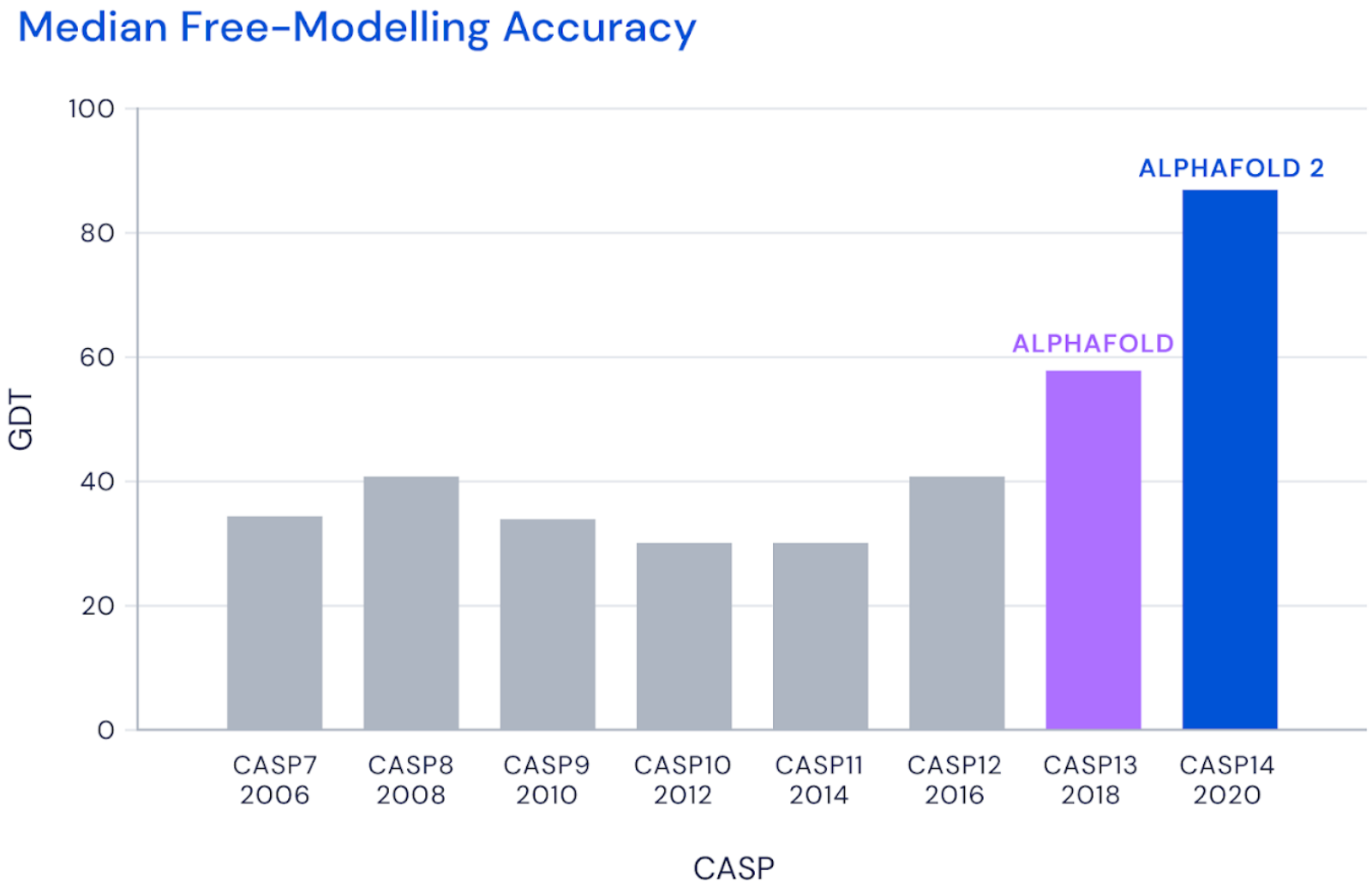
Fig. 2 | The folding process illustrated for CASP13 target T0986s2. CASP target T0986s2, $L = 155$, PDB: 6N9V. **a**, Steps of structure prediction. **b**, The neural network predicts the entire $L \times L$ distogram based on MSA features, accumulating separate predictions for 64×64-residue regions. **c**, One iteration of gradient descent (1,200 steps) is shown, with the TM score and root mean square deviation (r.m.s.d.) plotted against step number with five snapshots of the structure. The secondary structure (from SST[33]) is also shown (helix in blue, strand in red) along with the native secondary structure (Nat.), the secondary structure prediction probabilities of the network and the uncertainty in torsion angle predictions (as κ^{-1} of the von Mises distributions fitted to the predictions for φ and ψ). While each step of gradient descent greedily lowers the potential, large global conformation changes are effected, resulting in a well-packed chain. **d**, The final first submission overlaid on the native structure (in grey). **e**, The average (across the test set, $n = 377$) TM score of the lowest-potential structure against the number of repeats of gradient descent per target (log scale).

AlphaFold Example

(b) Real pair distance
(c) Predicted pair distance



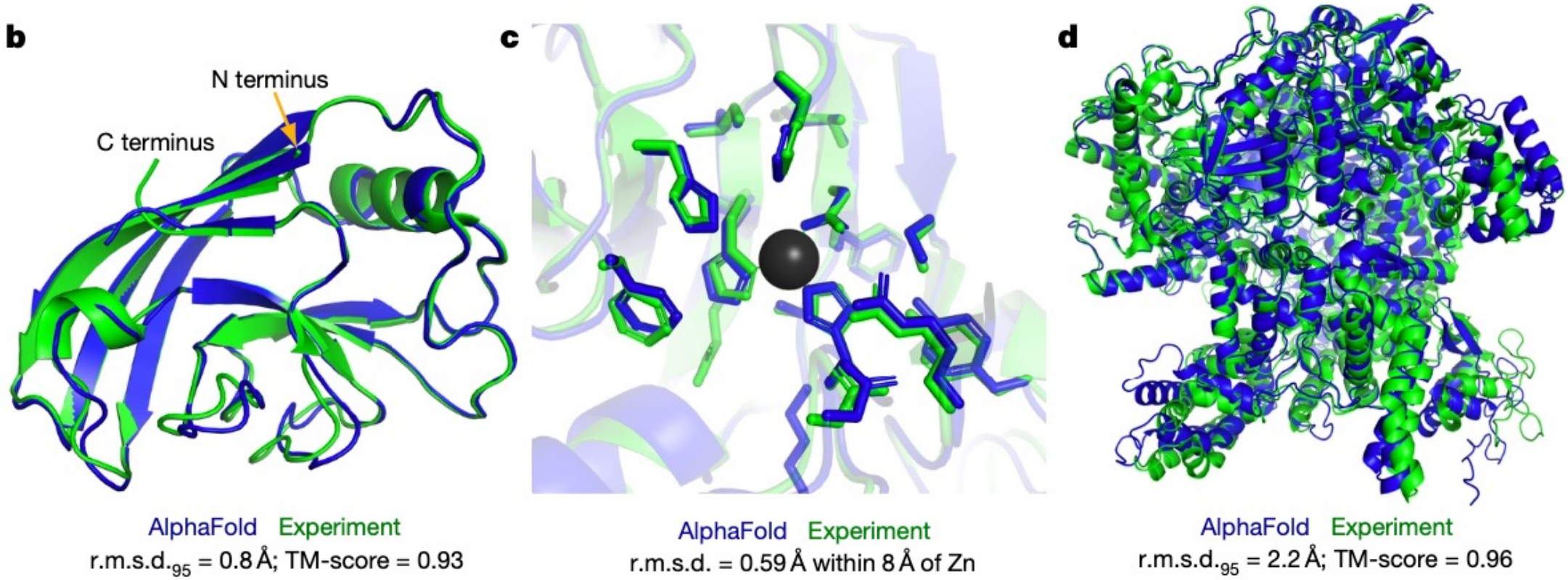
CASP: Critical Assessment of Protein Structure Prediction



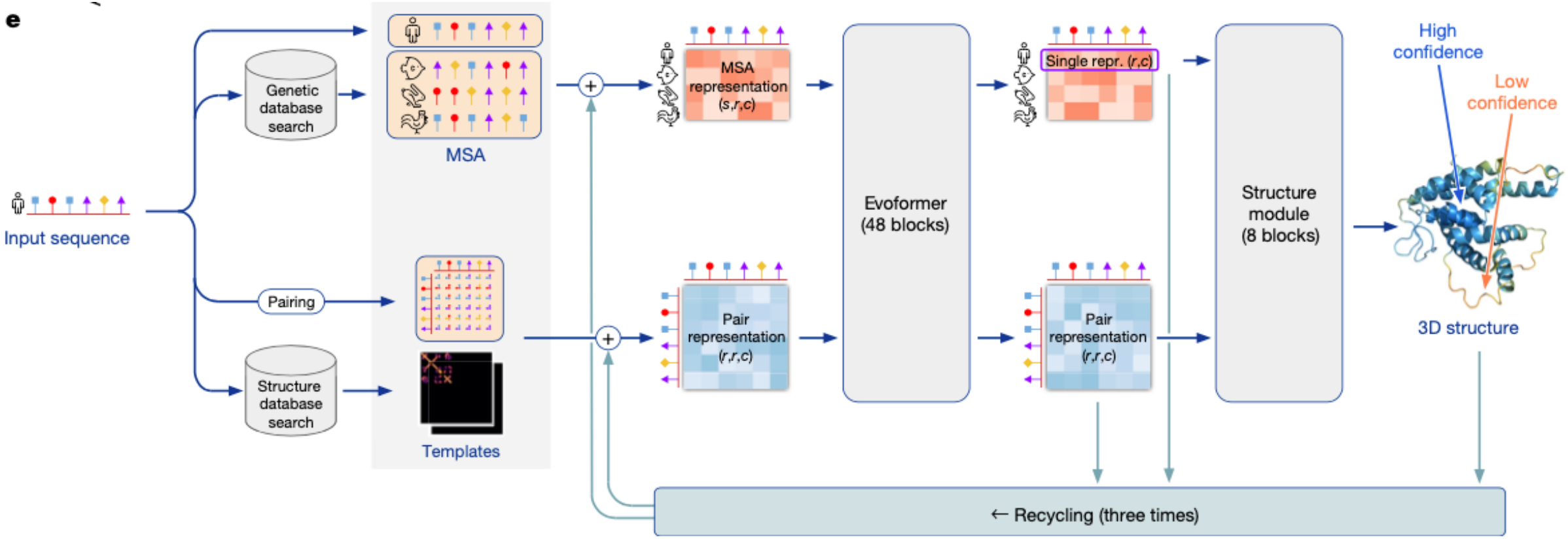
AlphaFold 2

- Jumper et al. Highly accurate protein structure prediction with AlphaFold. Nature 596. 2021.
- New transformer-based prediction of pair contact features.
- New NN based prediction of torsion angel and structure.

AlphaFold 2 Prediction Examples



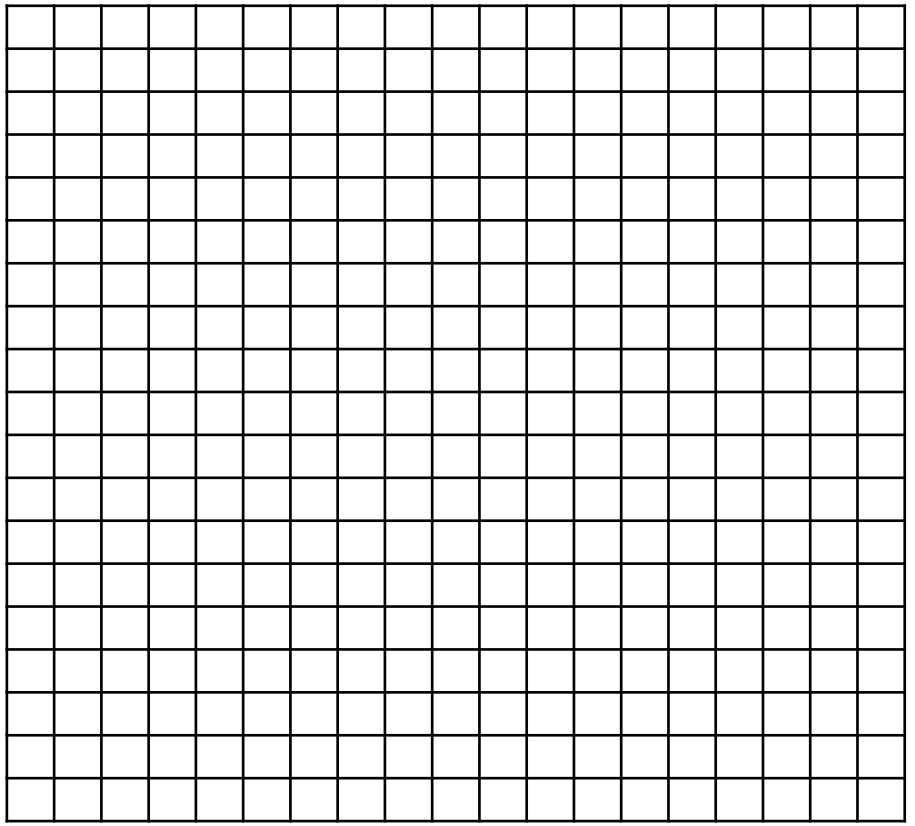
AlphaFold 2 Overall Architecture



MSA and Pair Representations

MSA representation

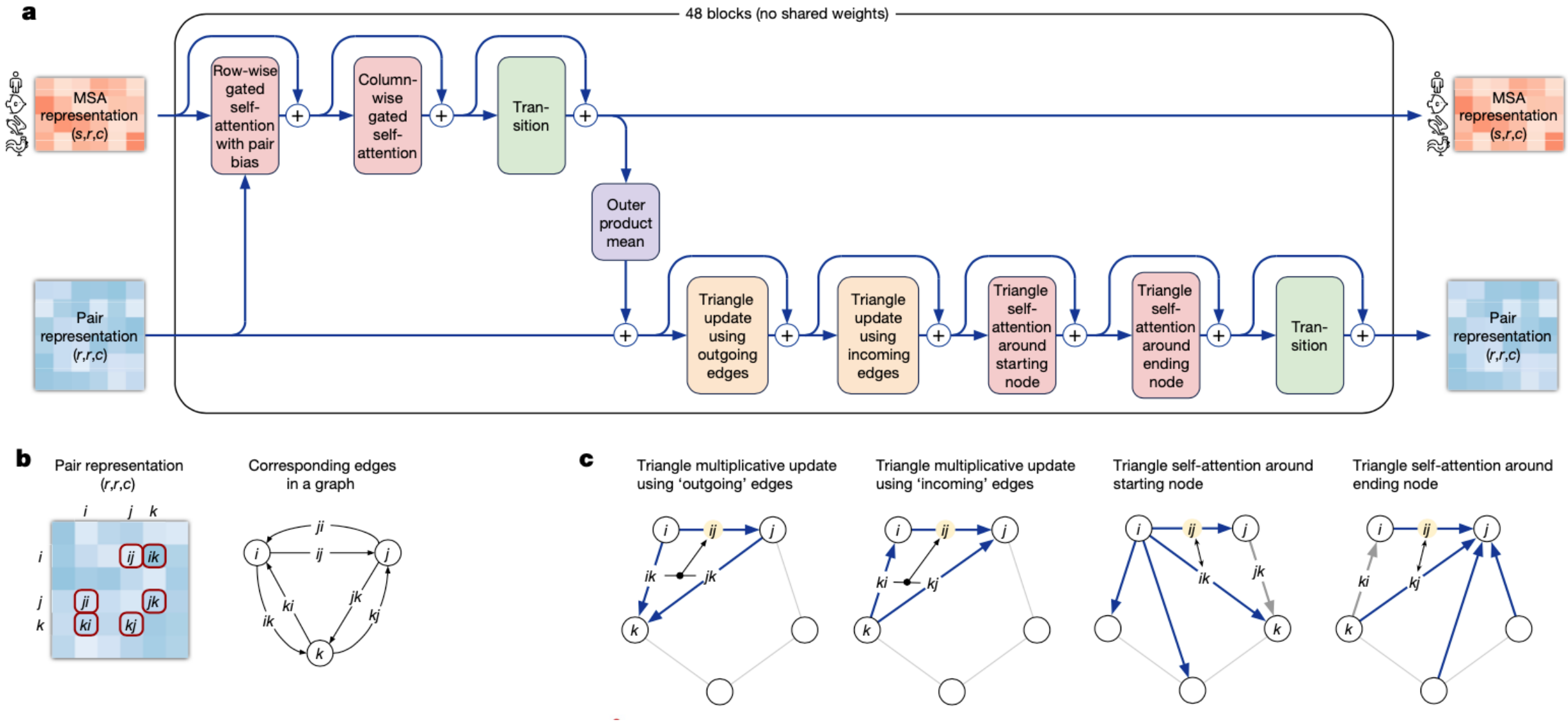
- Dimension (s, r, c).
- s is the number of seqeucnes.
- r is the number of residues in input sequence.
- c is the model's dimension.
- Represent each residue position in MSA.



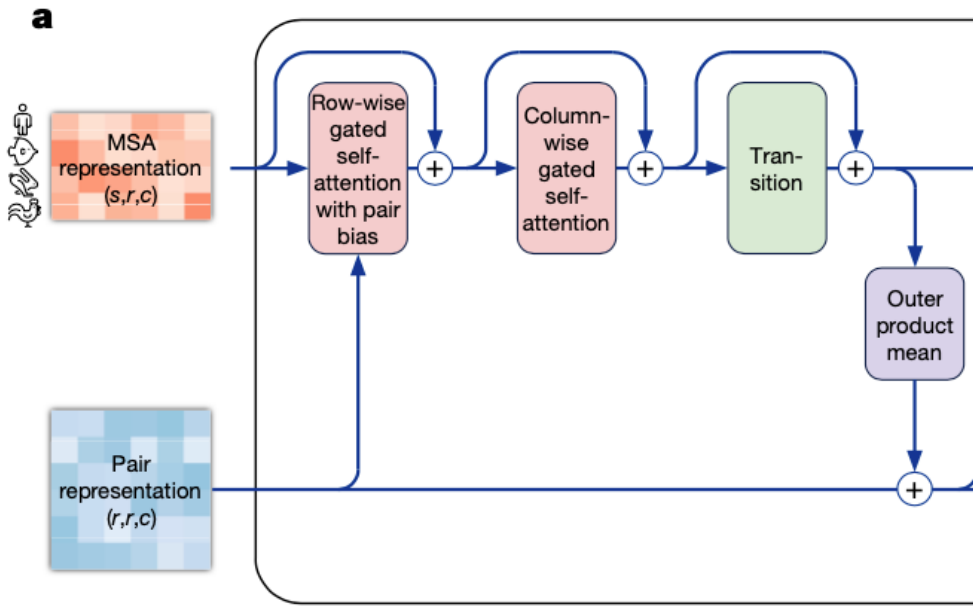
Pair representation

- Dimension (r, r, c).
- r is the number of residues in input sequence.
- c is the model's dimension.
- Represent the residue to residue relation.

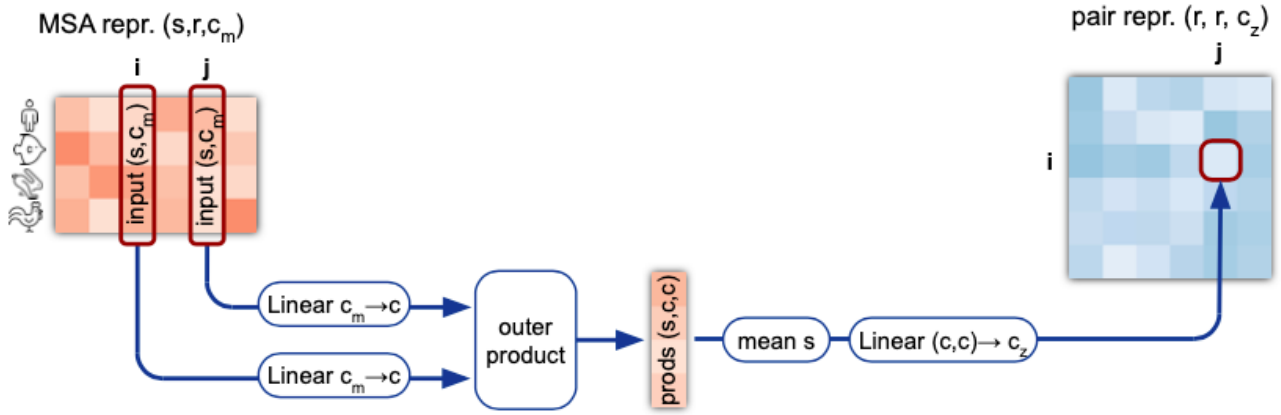
Evoformer


a
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)
b
Pair representation (r,r,c)
Corresponding edges in a graph
c
Triangle multiplicative update using 'outgoing' edges
Triangle multiplicative update using 'incoming' edges
Triangle self-attention around starting node
Triangle self-attention around ending node

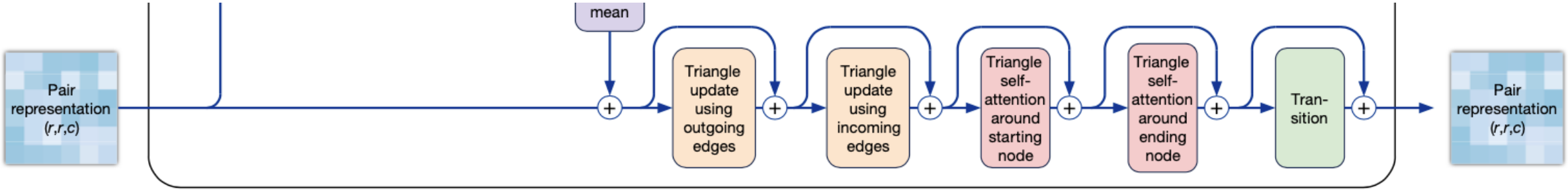
Update MSA

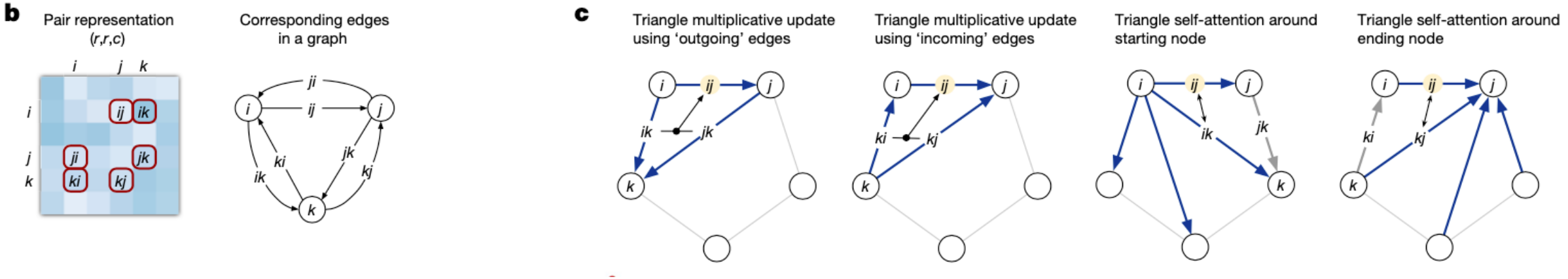


- Row-wise gated attention: for each seqeunce, self-attention between its residues. Add a bias computed from pair(i,j) to attention weight between j and j.
- Column-wise attention: for each column of MSA, self-attention between residues at this column.
- Transition: same role as the FFN in transformer.



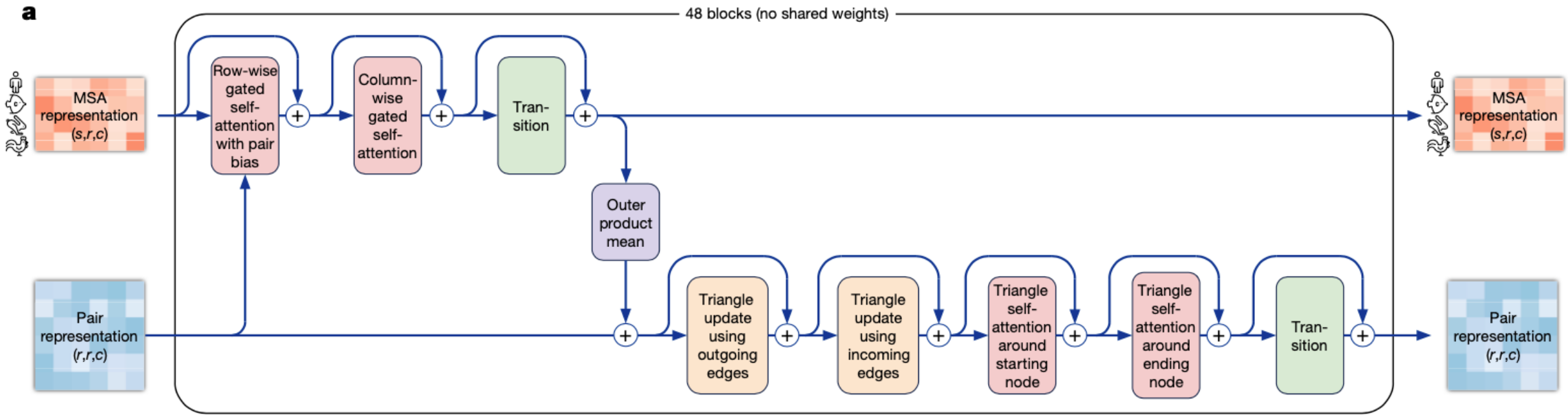
Evoformer





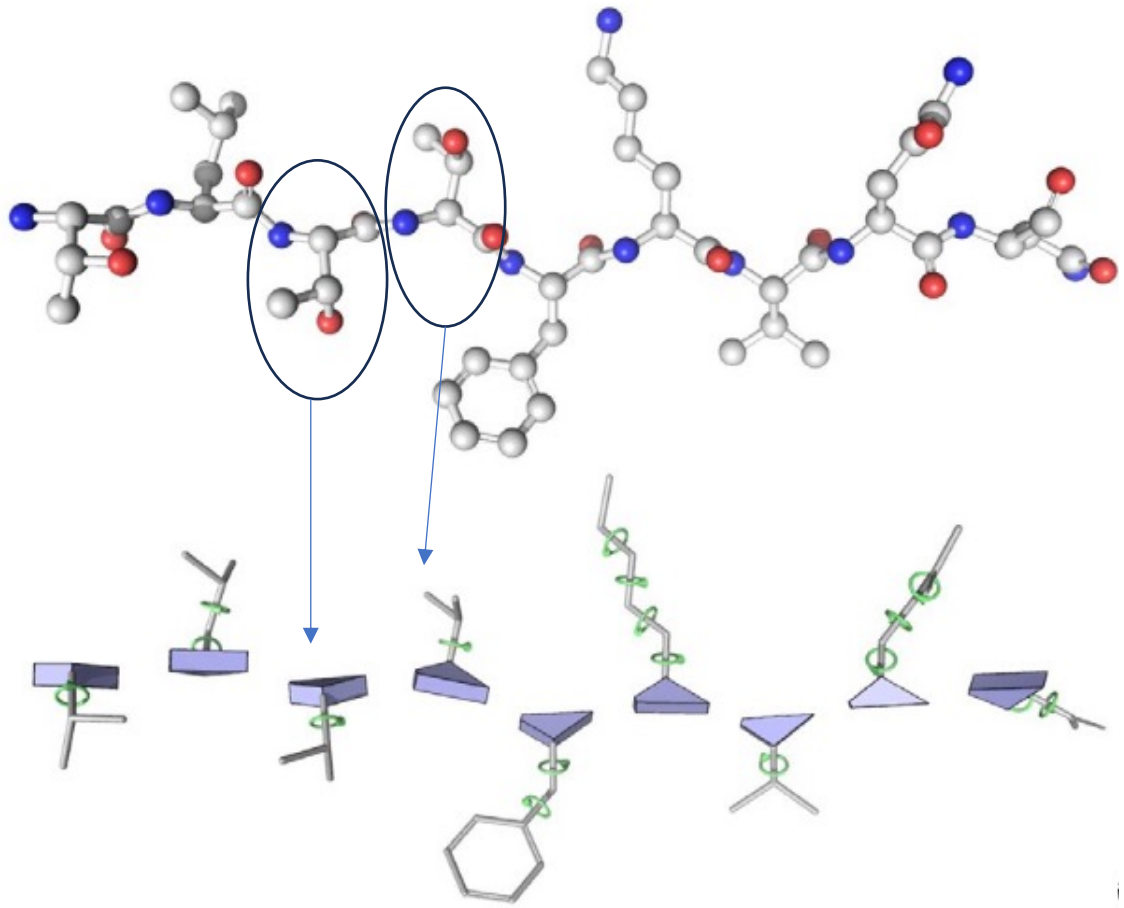
- The triangle update reflects the triangular inequality constraint among there residues.
- Triangule self-attention is similar to row and col wise self attentions.
- Triangle multiplicative update uses values computed from two edges of a triangle to update the third edge.

Evoformer

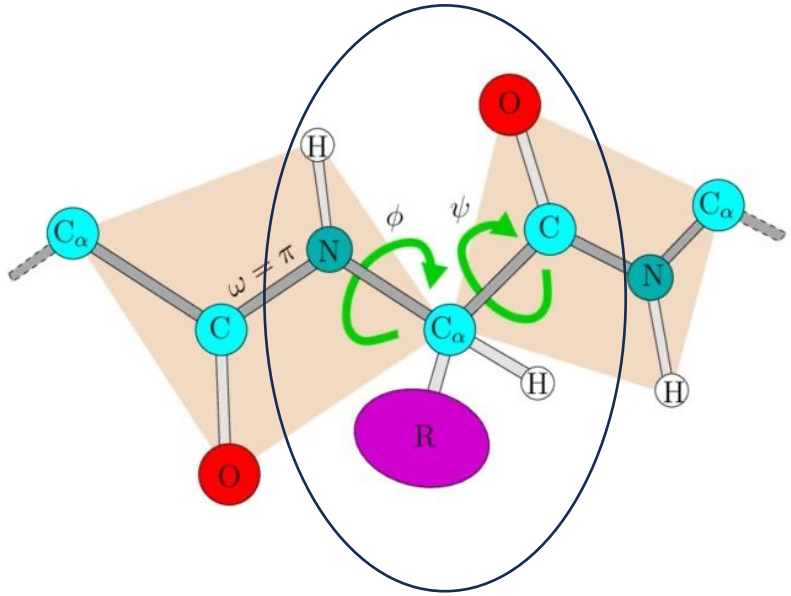


- Each block is called an evoformer block, it updates MSA and pair representations.
- AlphaFold 2 stacks 48 blocks together. The output is then sent to the structure module.

Residue Gas

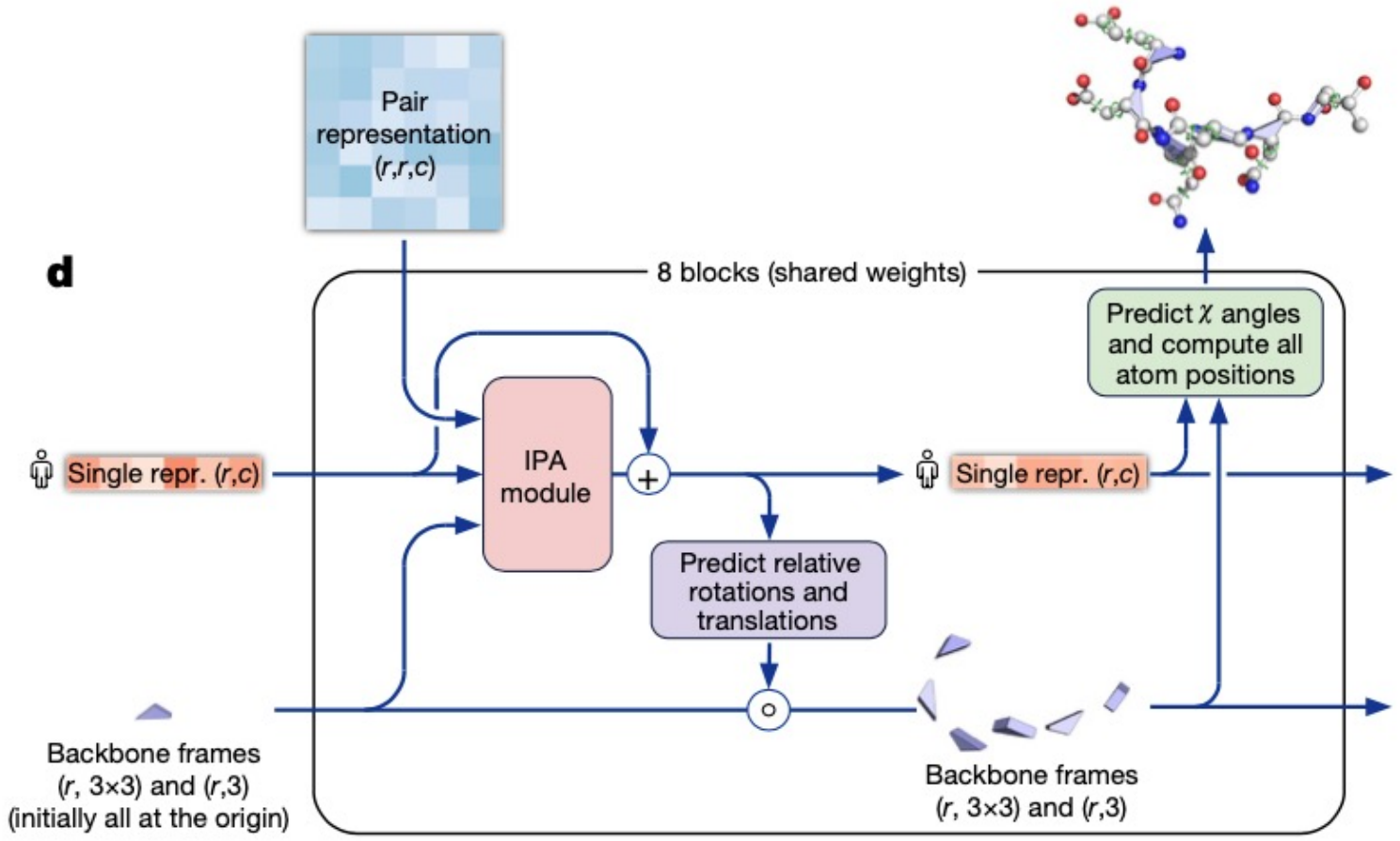


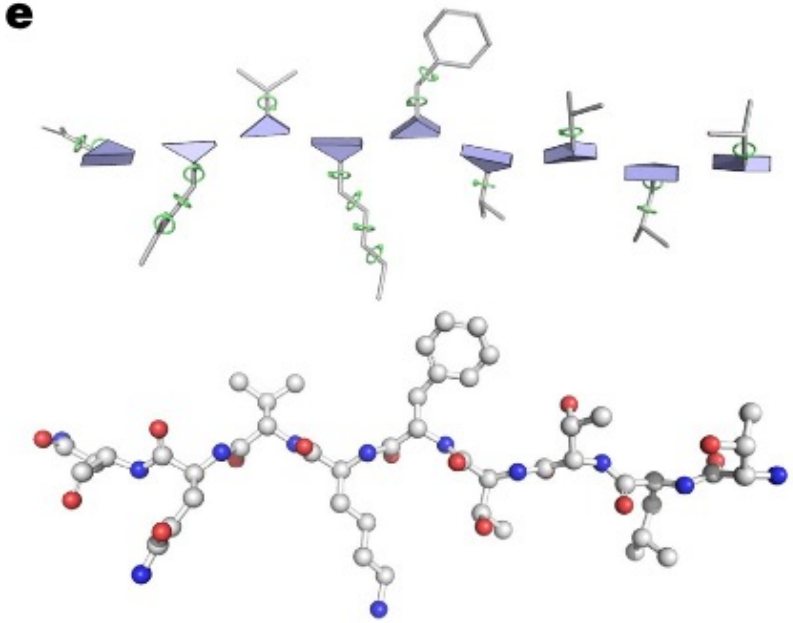
residue



- Each residue's N-C_α-C determines a local frame, which is represented by a rotation (3x3 matrix) and a translation (length-3 vector).
- The structural module tries to predict the local frame for each residue, and a representation for each residue (named single representation).
- The representation is supposed to encode structural information relative to the local frame.

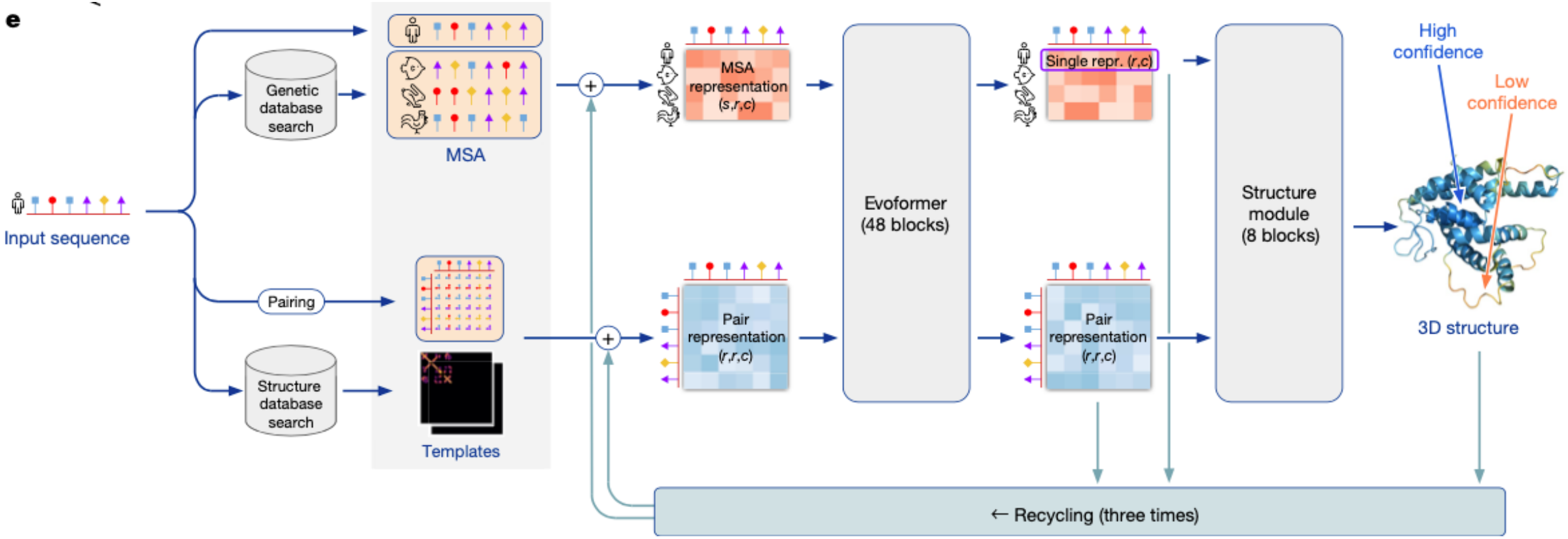
Invariant Point Attention





- Single representation is initialized to be the row of the input sequence of MSA representation. Backbone frames are initialized to be at the origin.
- Self attention on single representation is invariant to global rotations and translations (IPA).
- Backbone frames are updated in each block as well.
- 8 weight-sharing blocks are used. The final output of single representations is used to predict side chain.

AlphaFold 2 Overall Architecture



Summary

- Raptor X (Jingbo Xu)
  - Use CNN to predict contact map. Then uses traditional algorithms to find structure to fit the contact map.
- AlphaFold
  - Use CNN to predict contact map. Define an energy function with contact map. Then use gradient descent search strategy to minimize the energy.
- AlphaFold2
  - Predict the contact map (2D representation) with transformer.
  - Predict residue gas with invariant point attention.

Structure Prediction Software

- Rosetta: David Baker, University of Washington
- RaptorX: Jinbo Xu, Toyota Technological Institute at Chicago
- Alpha Fold: DeepMind
- Many others..

Protein Structure Prediction & Overlay (Class Activity)

- 1. Get sequences from UniProt: Myoglobin (P02144) and Neuroglobin (Q9NPG2). http://uniprot.org
- 2. Run AlphaFold prediction for each protein: https://alphafoldserver.com/
- 3. Download results
- 4. View the structures by drag and drop to Mol*: https://molstar.org/viewer/
- 5. Overlay structures: Select two chains and then click superpose.

Myoglobin vs Nuroglobin