Protein Structure Prediction

BMI/CS 776

www.biostat.wisc.edu/bmi776/

Spring 2024

Anthony Gitter

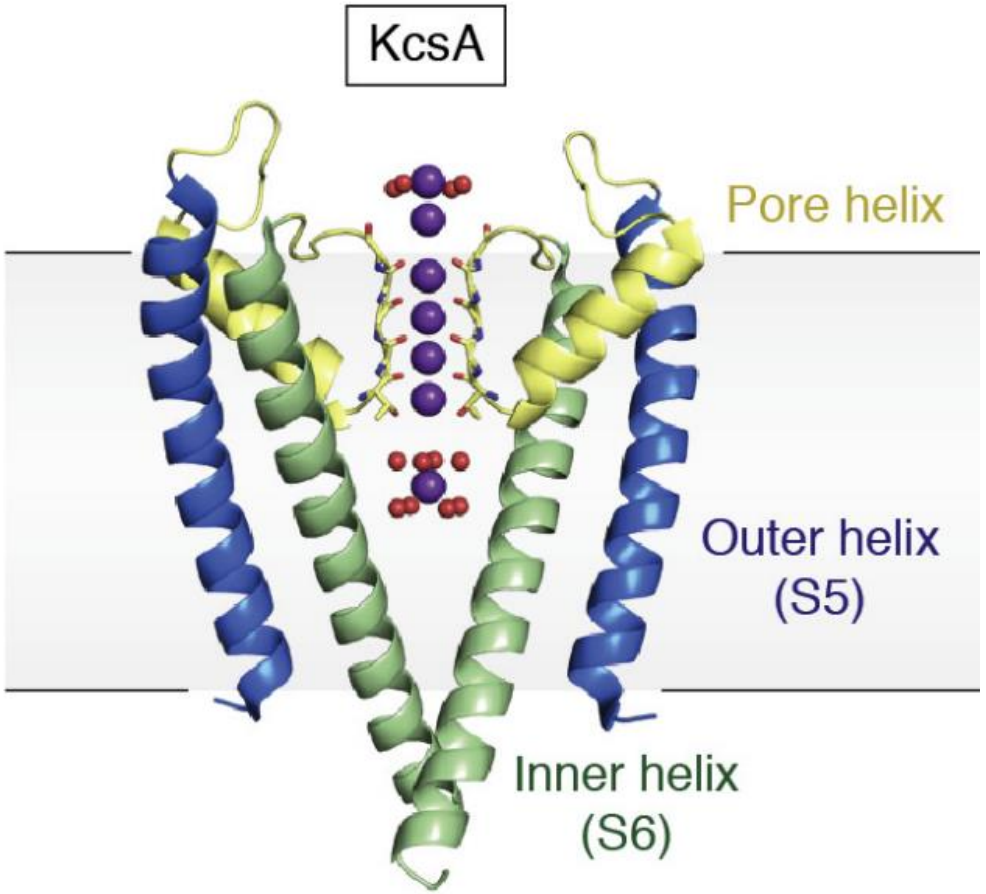
gitter@biostat.wisc.edu

Goals for lecture

- Why is protein structure important
- Elements of proteins
- AlphaFold2 and its impact
- What can we do with predicted structures
- What comes next for proteins

WHY IS PROTEIN STRUCTURE IMPORTANT?

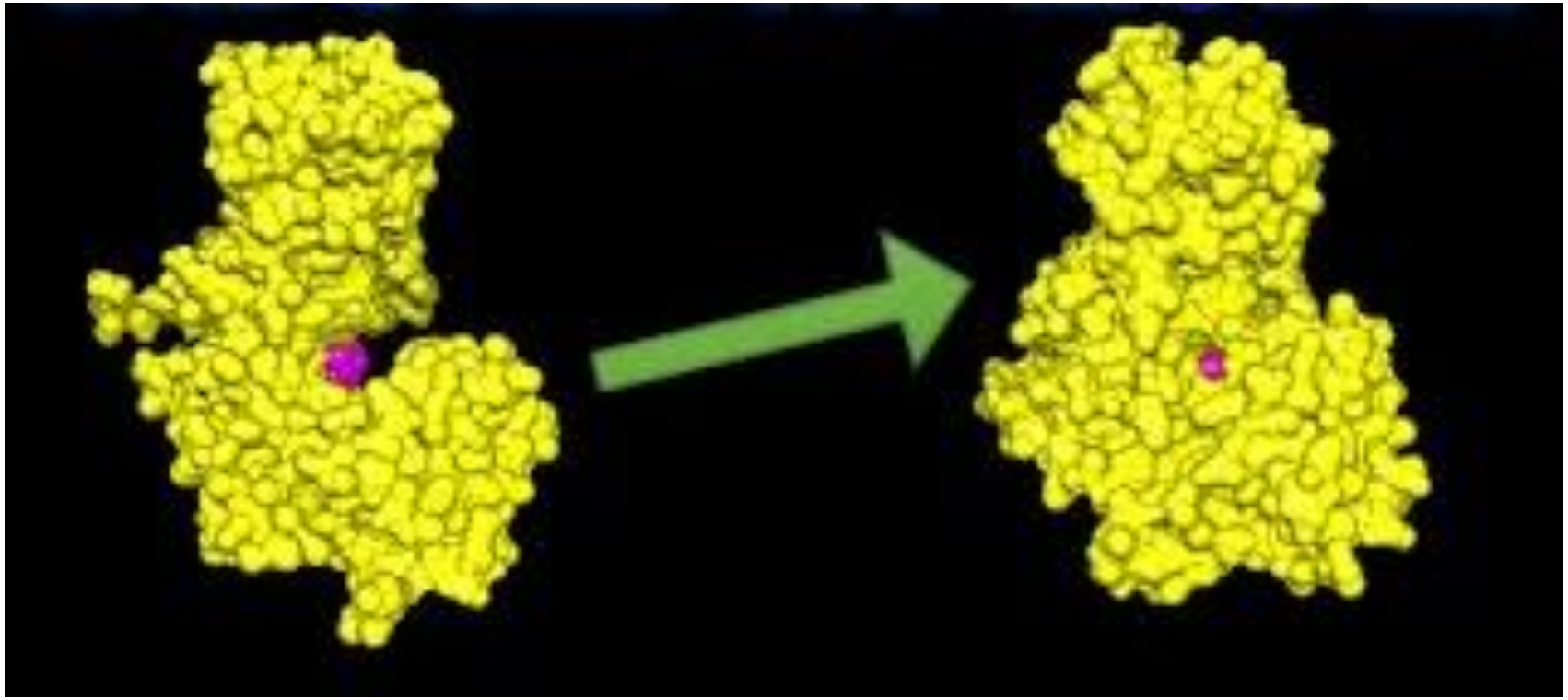
Protein structure determines function



Ion channel protein with K^+ ions at membrane

Li J Biol Chem 2021 doi:10.1016/j.jbc.2021.100557

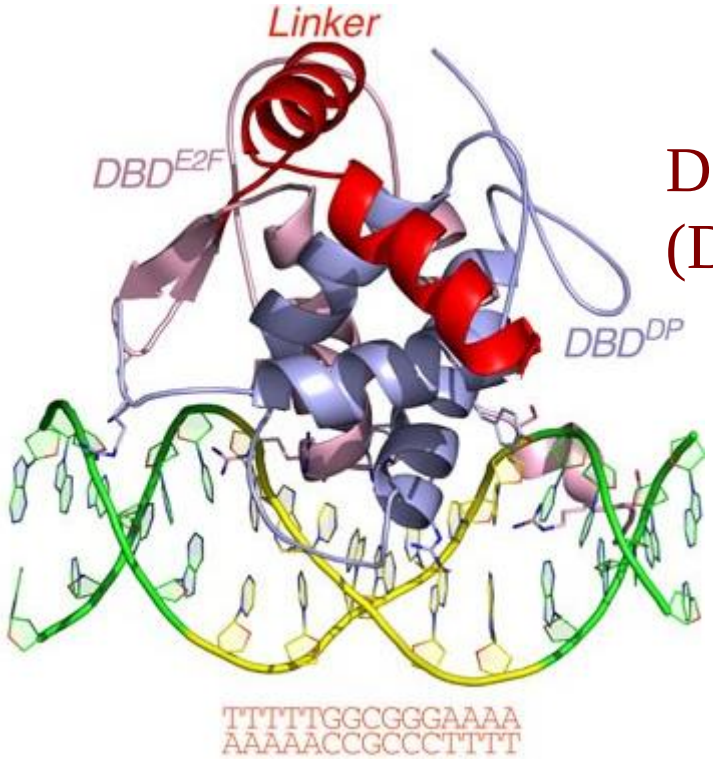
Protein structure determines function



Hexokinase binding glucose

Vishal Bhoir https://www.youtube.com/watch?v=U7GJWrt_VPA

Protein structure determines function



DNA-binding domain (DBD)

E2F8 protein bound to DNA

Morgunova Nat Commun 2015 doi:10.1038/ncomms10050

The Protein Folding Problem

- The function of a protein is determined in large part by its 3D shape (*fold*, *conformation*)
- Can we predict the 3D shape of a protein given only its 1D amino-acid sequence?

ELEMENTS OF PROTEINS

Protein Architecture

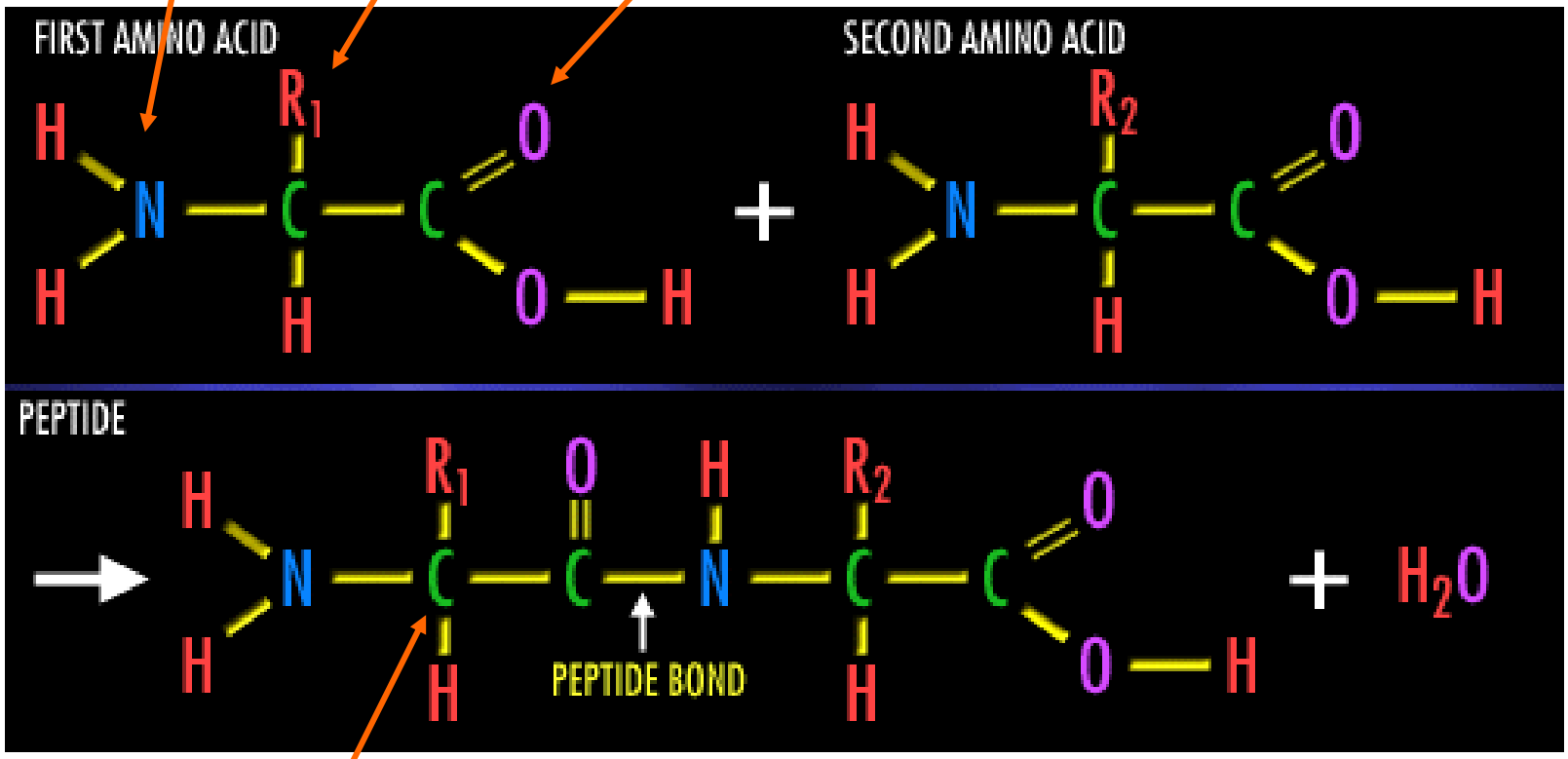
- Proteins are polymers consisting of amino acids linked by *peptide* bonds
- Each amino acid consists of
  - a central carbon atom (α-carbon)
  - an amino group, NH_2
  - a carboxyl group, COOH
  - a side chain (R)
- Differences in side chains distinguish 20 different amino acids

Amino Acids and Peptide Bonds

amino group

side chain

carboxyl group

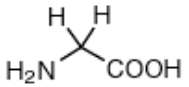


α carbon (common reference point for coordinates of a structure)

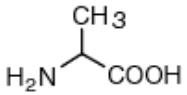
Amino Acid Side Chains

Side chains vary in
- shape
- size
- charge
- polarity

Small

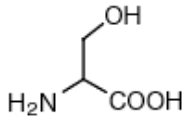


Glycine (Gly, G)
MW: 57.05

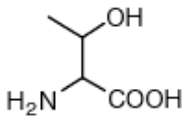


Alanine (Ala, A)
MW: 71.09

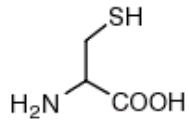
Nucleophilic



Serine (Ser, S)
MW: 87.08, pK_a ~ 16

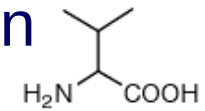


Threonine (Thr, T)
MW: 101.11, pK_a ~ 16

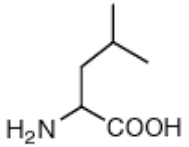


Cysteine (Cys, C)
MW: 103.15, pK_a = 8.35

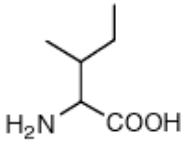
Hydrophobic



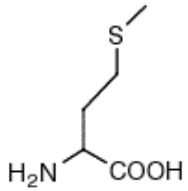
Valine (Val, V)
MW: 99.14



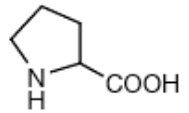
Leucine (Leu, L)
MW: 113.16



Isoleucine (Ile, I)
MW: 113.16

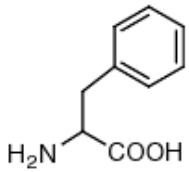


Methionine (Met, M)
MW: 131.19

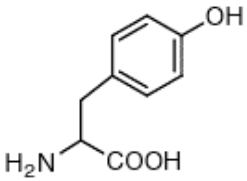


Proline (Pro, P)
MW: 97.12

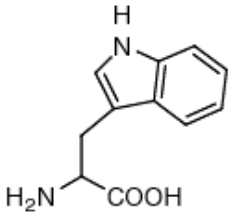
Aromatic



Phenylalanine (Phe, F)
MW: 147.18

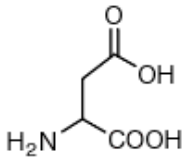


Tyrosine (Tyr, Y)
MW: 163.18

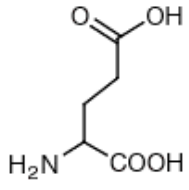


Tryptophan (Trp, W)
MW: 186.21

Acidic

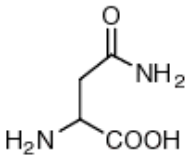


Aspartic Acid (Asp, D)
MW: 115.09, pK_a = 3.9

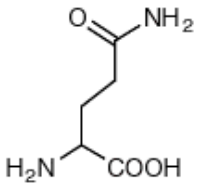


Glutamic Acid (Glu, E)
MW: 129.12, pK_a = 4.07

Amide

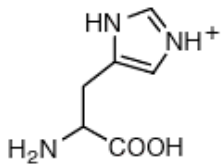


Asparagine (Asn, N)
MW: 114.11

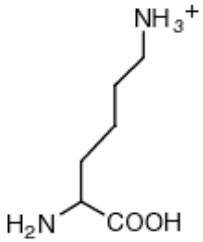


Glutamine (Gln, Q)
MW: 128.14

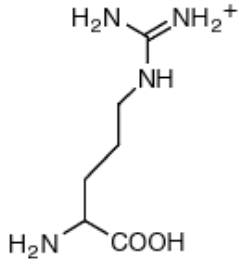
Basic



Histidine (His, H)
MW: 137.14, pK_a = 6.04

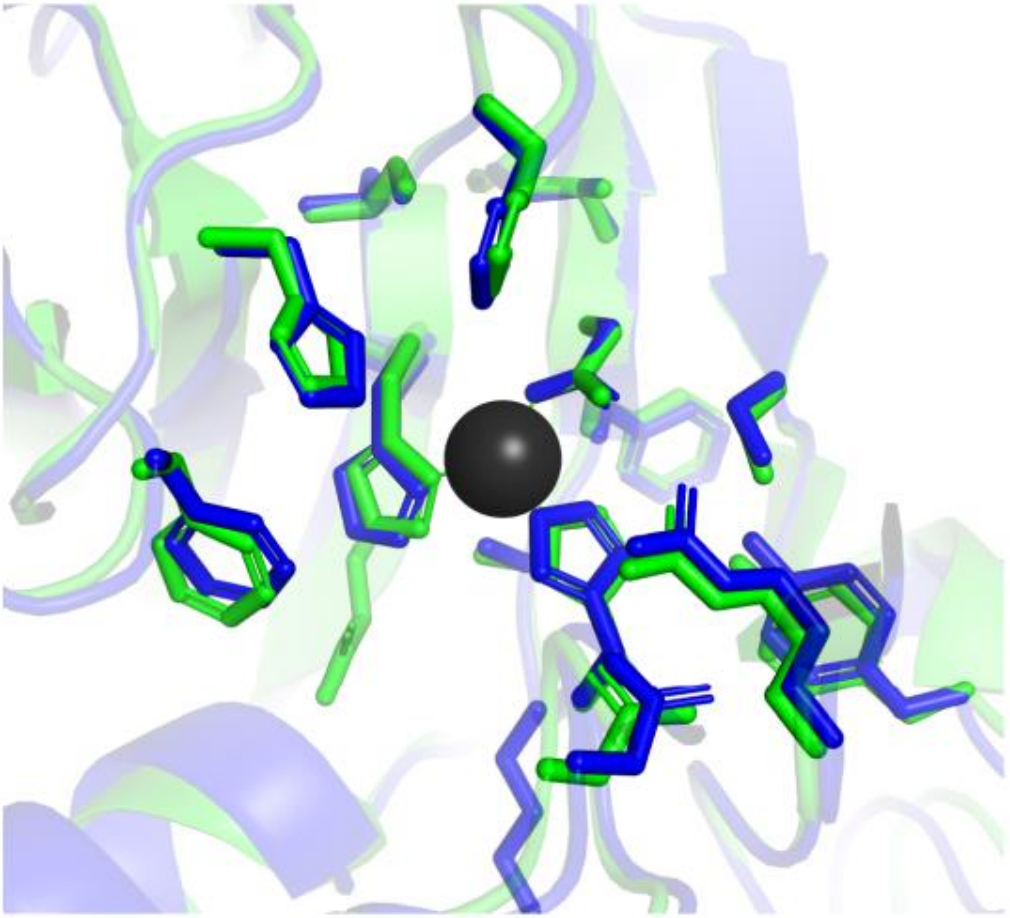


Lysine (Lys, K)
MW: 128.17, pK_a = 10.79



Arginine (Arg, R)
MW: 156.19, pK_a = 12.48

Predicting Side Chains is Hard but Important



AlphaFold Experiment

r.m.s.d. = 0.59 Å within 8 Å of Zn

Jumper Nature 2021 doi:10.1038/s41586-021-03819-2

What Determines Conformation?

- In general, the amino-acid sequence of a protein determines the 3D shape of a protein [Anfinsen et al., 1950s]

- But some qualifications
  - all proteins can be denatured
  - some proteins are inherently *disordered* (i.e. lack a regular structure)
  - some proteins get folding help from *chaperones*
  - there are various mechanisms through which the conformation of a protein can be changed in vivo
    - post-translational modifications such as *phosphorylation*
    - *prions*
    - etc.

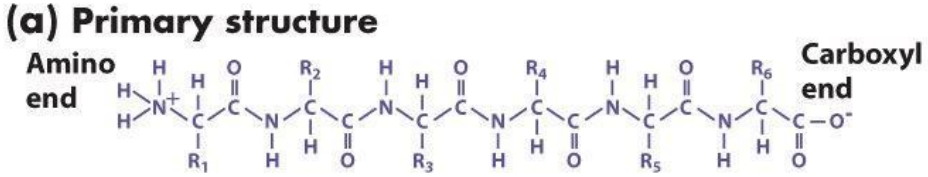
What Determines Conformation?

- Which physical properties of the protein determine its fold?
  - rigidity of the protein backbone
  - interactions among amino acids, including
    - electrostatic interactions
    - van der Waals forces
    - volume constraints
    - hydrogen, disulfide bonds
  - interactions of amino acids with water
    - hydrophobic and hydrophilic residues

Levels of Description

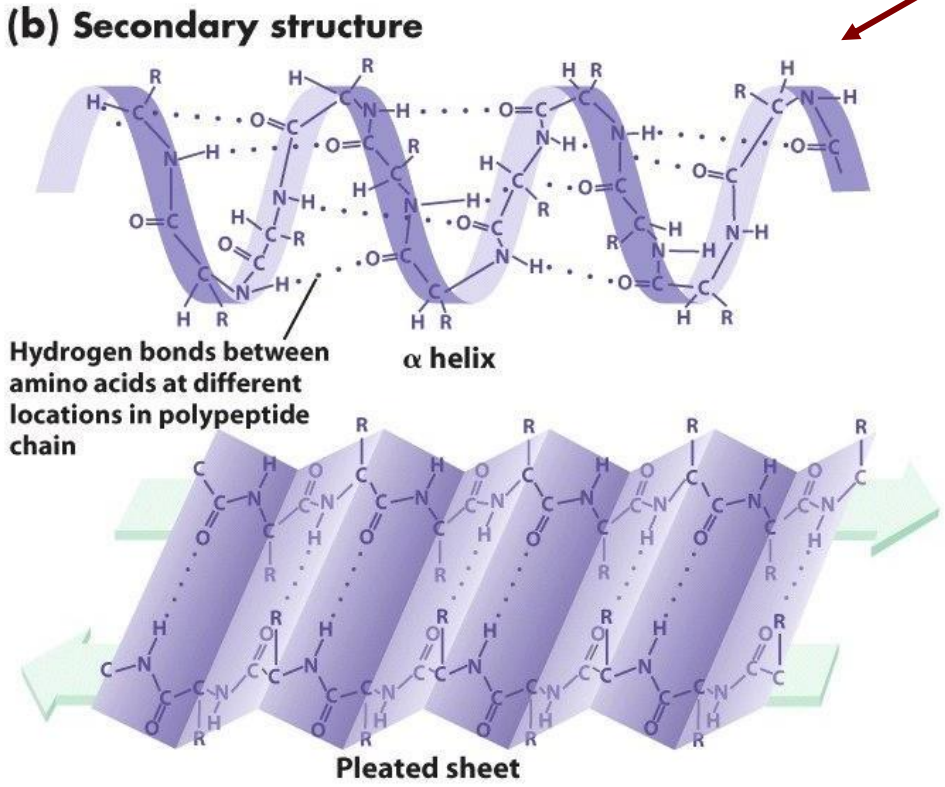
- Protein structure is often described at four different scales
  - primary structure
  - secondary structure
  - tertiary structure
  - quaternary structure

Levels of Description

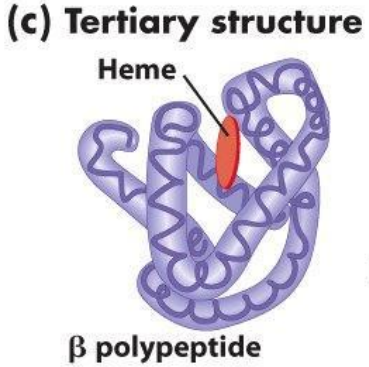


the amino acid sequence itself

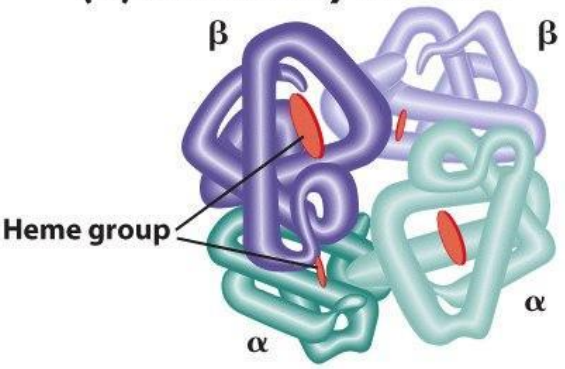
"local" description of structure: describes it in terms of certain common repeating elements



3D conformation of a polypeptide



3D conformation of a complex of polypeptides



Secondary Structure

- Secondary structure refers to certain common repeating structures
- It is a “local” description of structure
- Two common secondary structure

  α helices

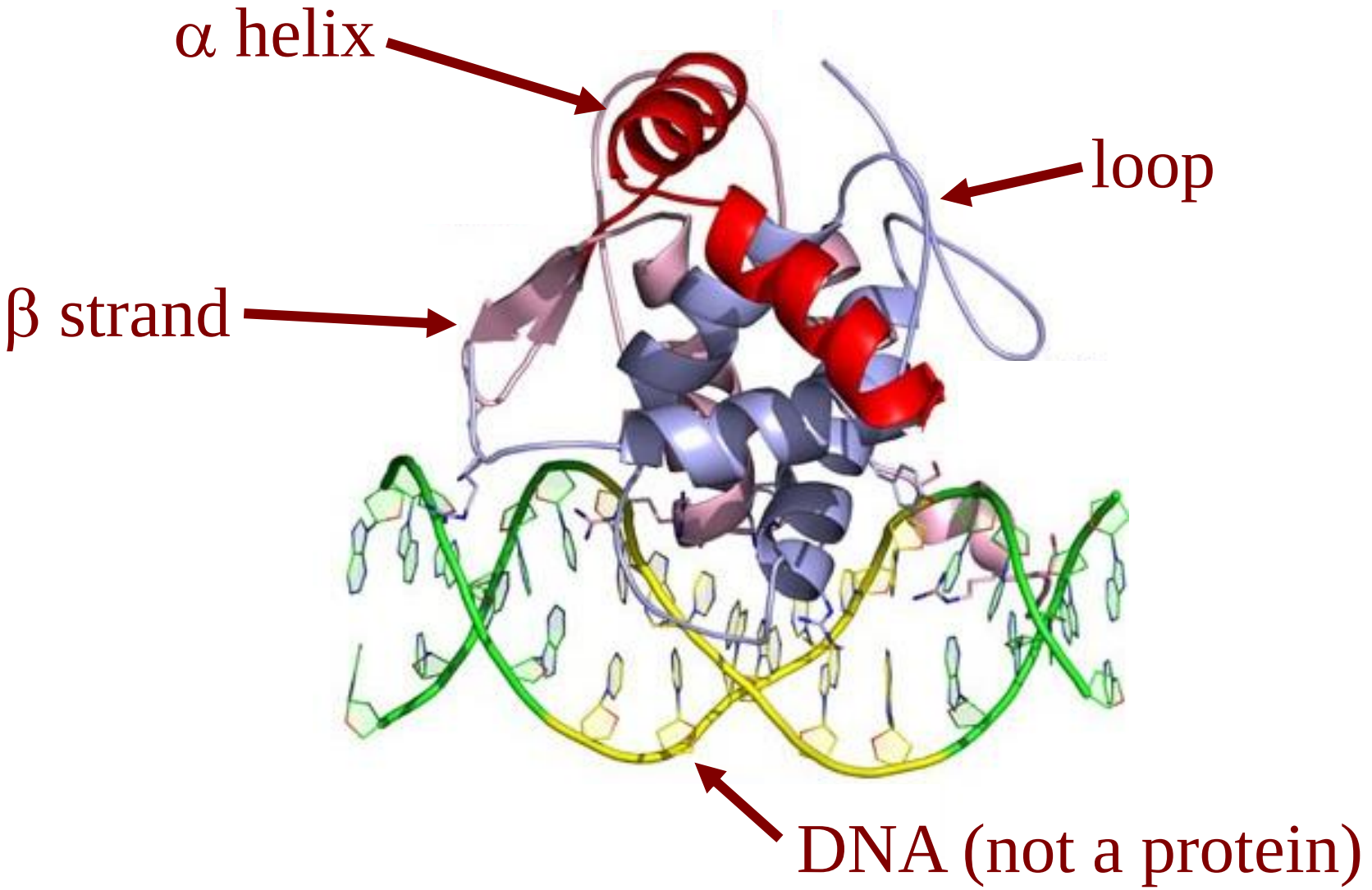
  β strands/sheets (pleated sheet on previous slide)
- A third category, called *coil* or *loop*, refers to everything else

Secondary Structure

“Is the neural network an essential tool for the most accurate secondary structure prediction?”

- Burkhard Rost, 1998

Ribbon Diagram Showing Secondary Structures



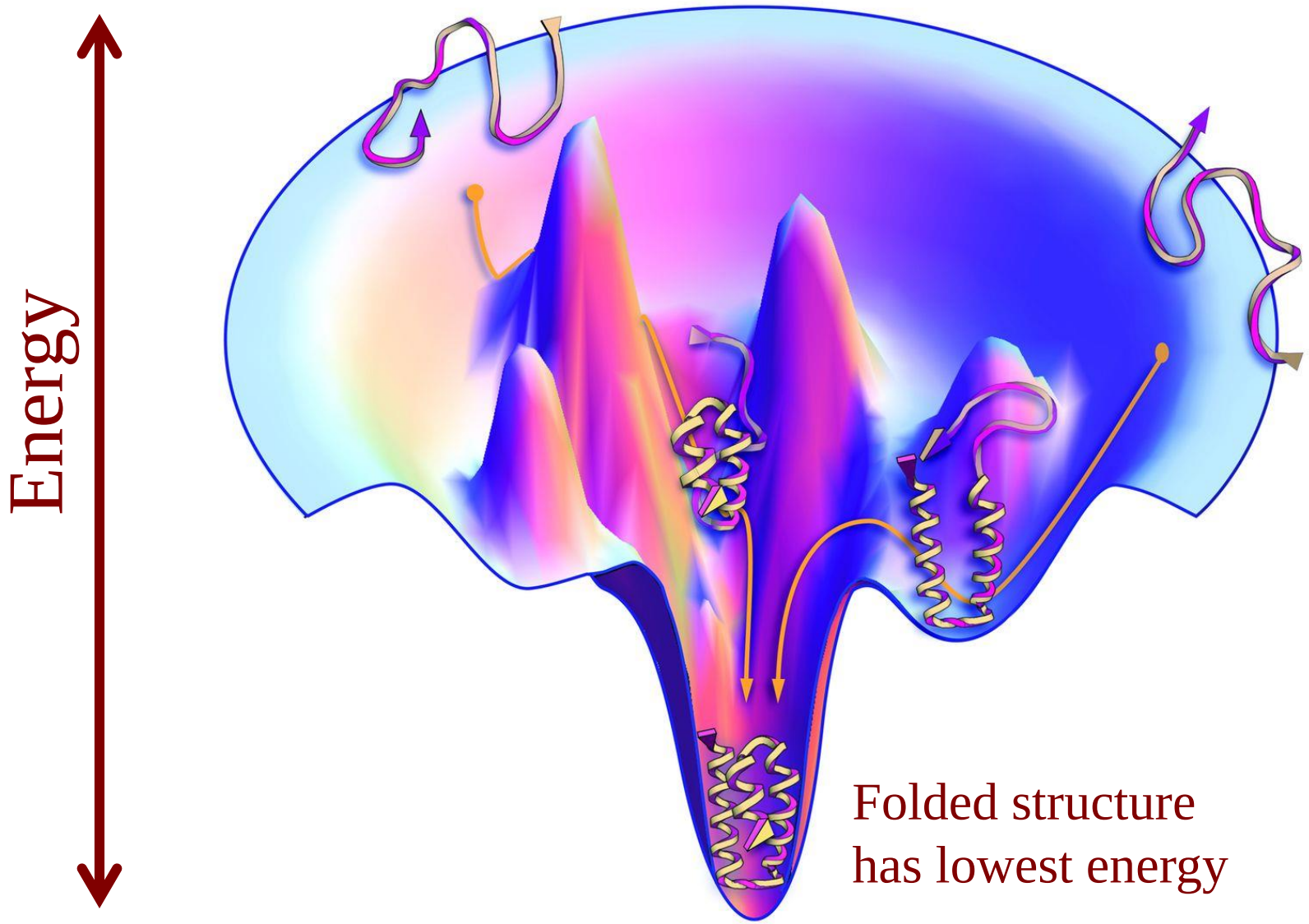
STRATEGIES FOR PREDICTING STRUCTURE

Determining Protein Structures

- Protein structures can be determined experimentally (in most cases) by
  - x-ray crystallography
  - nuclear magnetic resonance (NMR)
  - cryo-electron microscopy (cryo-EM)
- But this is very expensive and time-consuming
- There is a large sequence-structure gap
  - ≈ 1B protein sequences available
  - ≈ 100K protein structures in PDB database

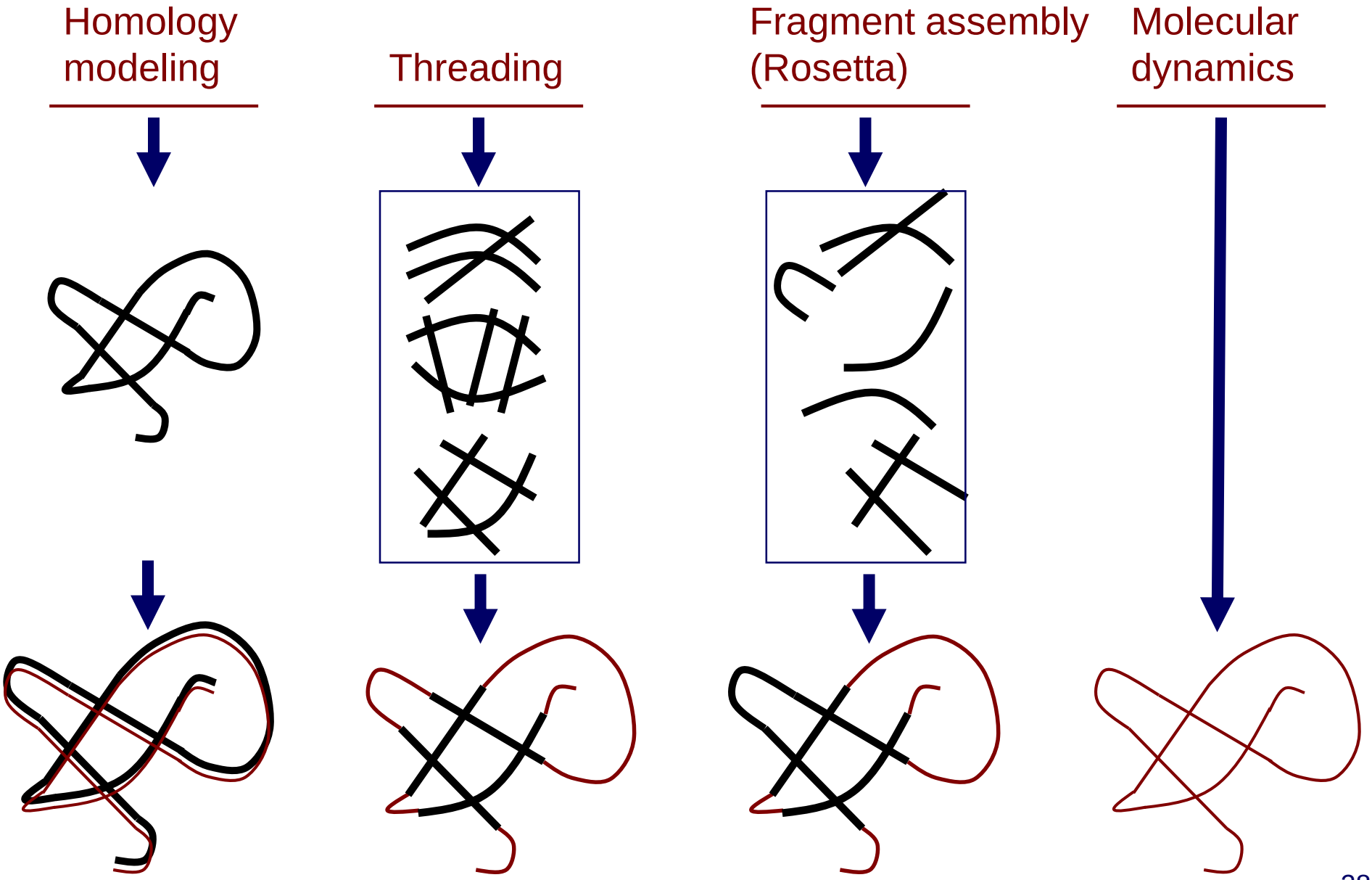
- Key question: can we predict structures by computational means instead?

Determining Protein Structures

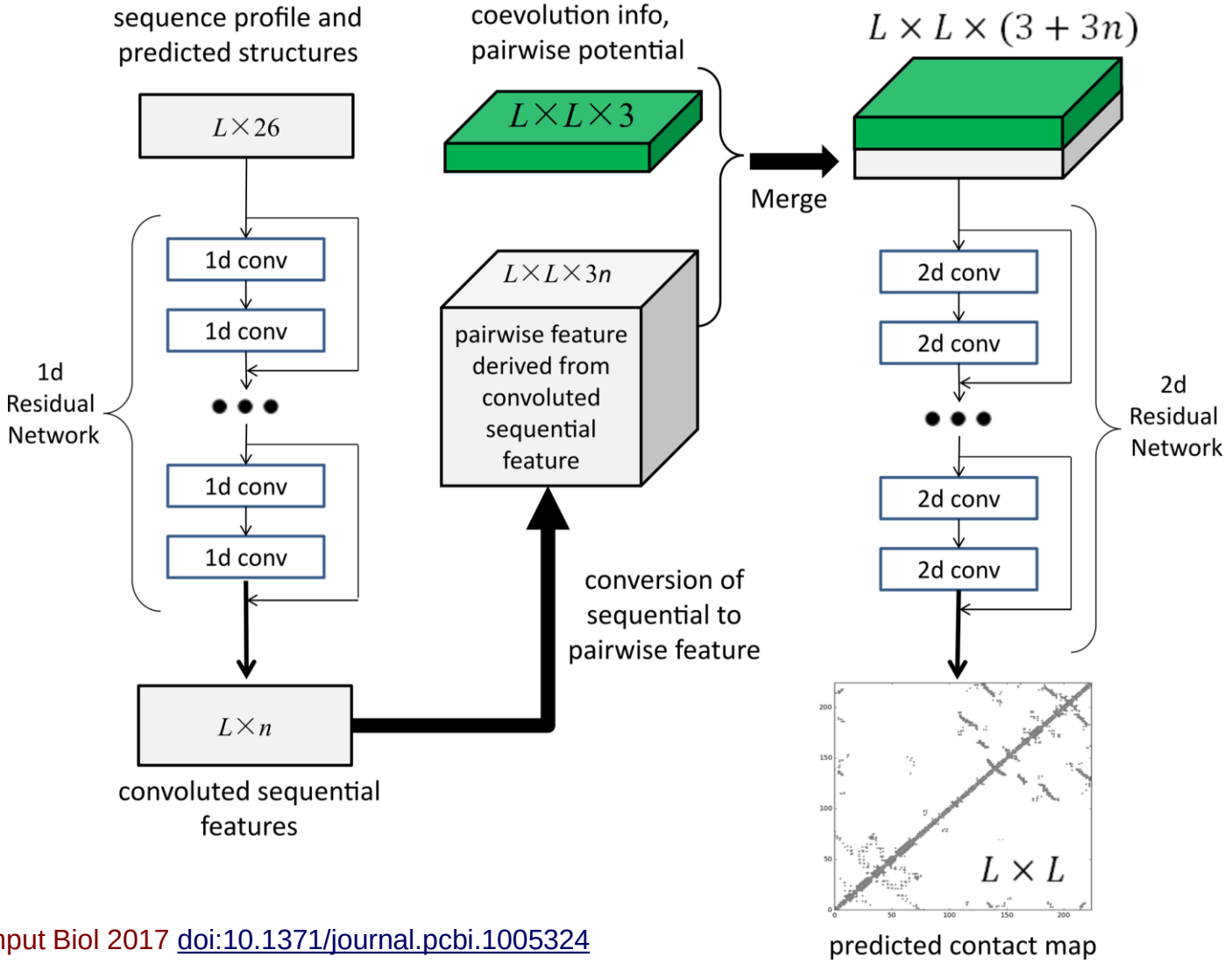


Dill Science 2012 doi:10.1126/science.1219021

Existing 3D structure prediction ideas



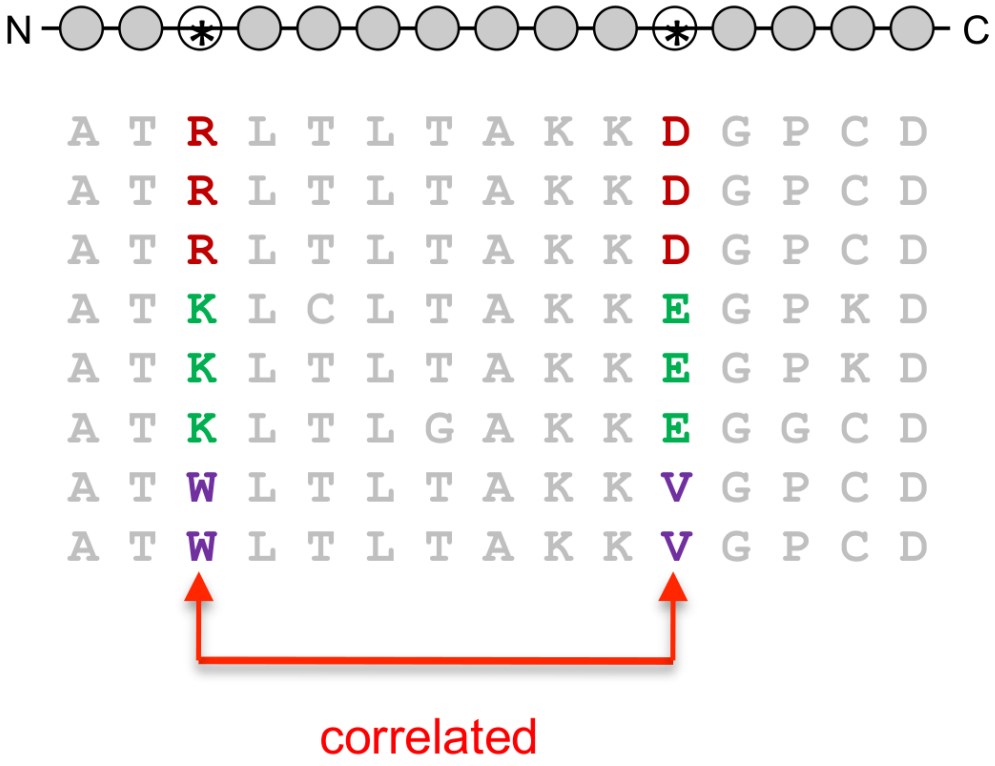
Deep learning contact map prediction

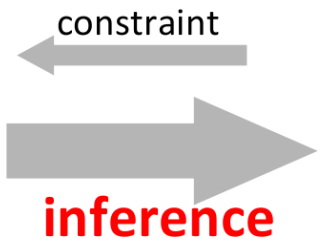


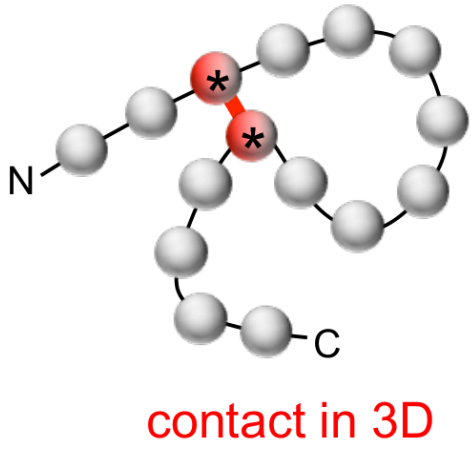
Wang PLoS Comput Biol 2017 doi:10.1371/journal.pcbi.1005324

Evolutionary conservation

- Multiple sequence alignments provide information about 3D structure







Marks PLoS One 2011 doi:10.1371/journal.pone.0028766

ALPHAFOLD2

Cannot understate the leap AlphaFold2 made

- Critical Assessment of Structure Prediction (CASP) is community challenge to predict new held out structures
- Run since 1994
- DeepMind competed with AlphaFold in CASP13 in 2018
- Then AlphaFold2 in 2020…

https://predictioncenter.org/index.cgi

Cannot understate the leap AlphaFold2 made

- Image circulating Twitter before CASP14 conference started

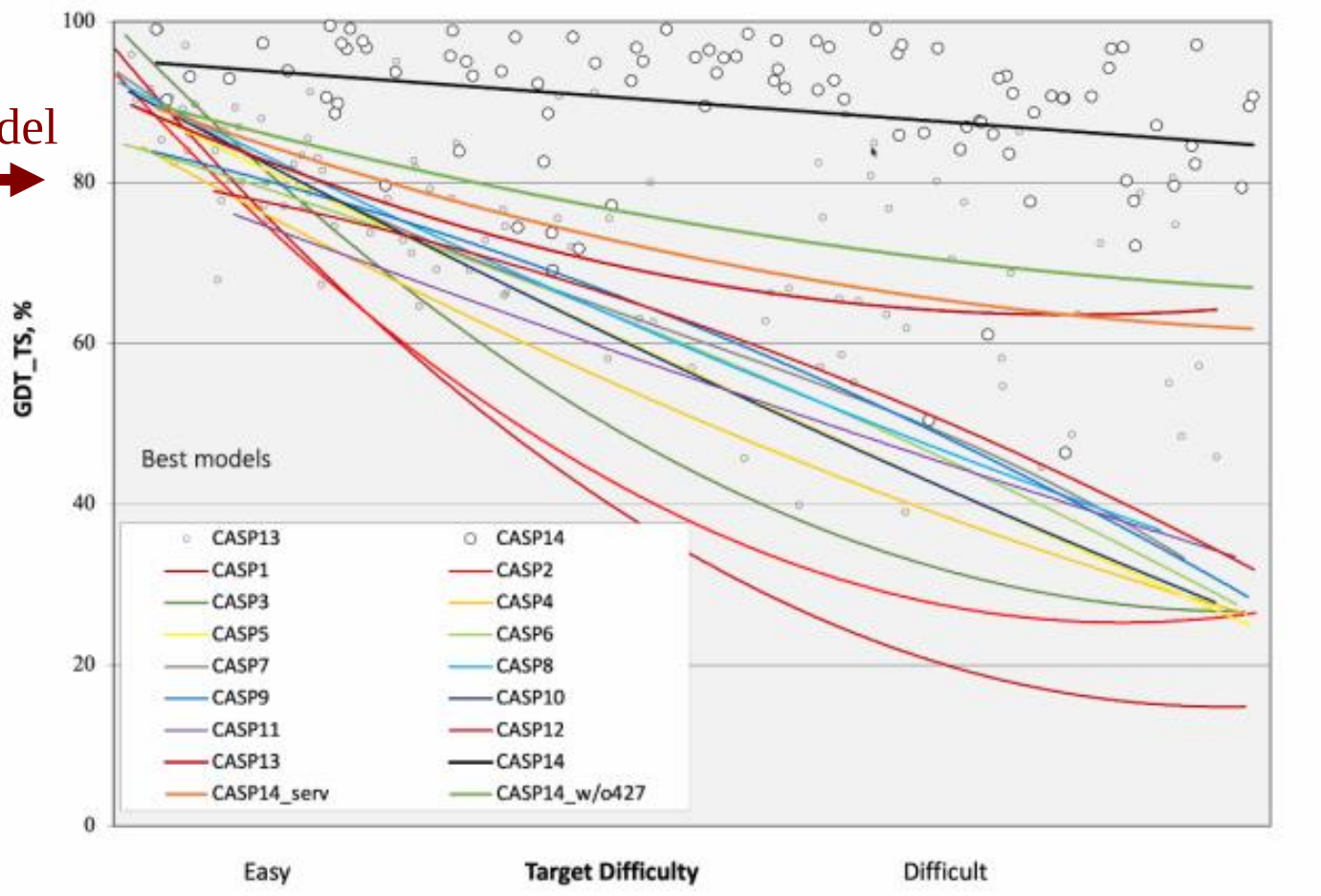
https://predictioncenter.org/casp14/zscores_final.cgi

https://www.blopig.com/blog/2020/12/casp14-what-google-deepminds-alphafold-2-really-achieved-and-what-it-means-for-protein-folding-biology-and-bioinformatics/

Cannot understate the leap AlphaFold2 made

- CASP performance over the years

Side chains resemble model



https://www.blopig.com/blog/2020/12/casp14-what-google-deepminds-alphafold-2-really-achieved-and-what-it-means-for-protein-folding-biology-and-bioinformatics/

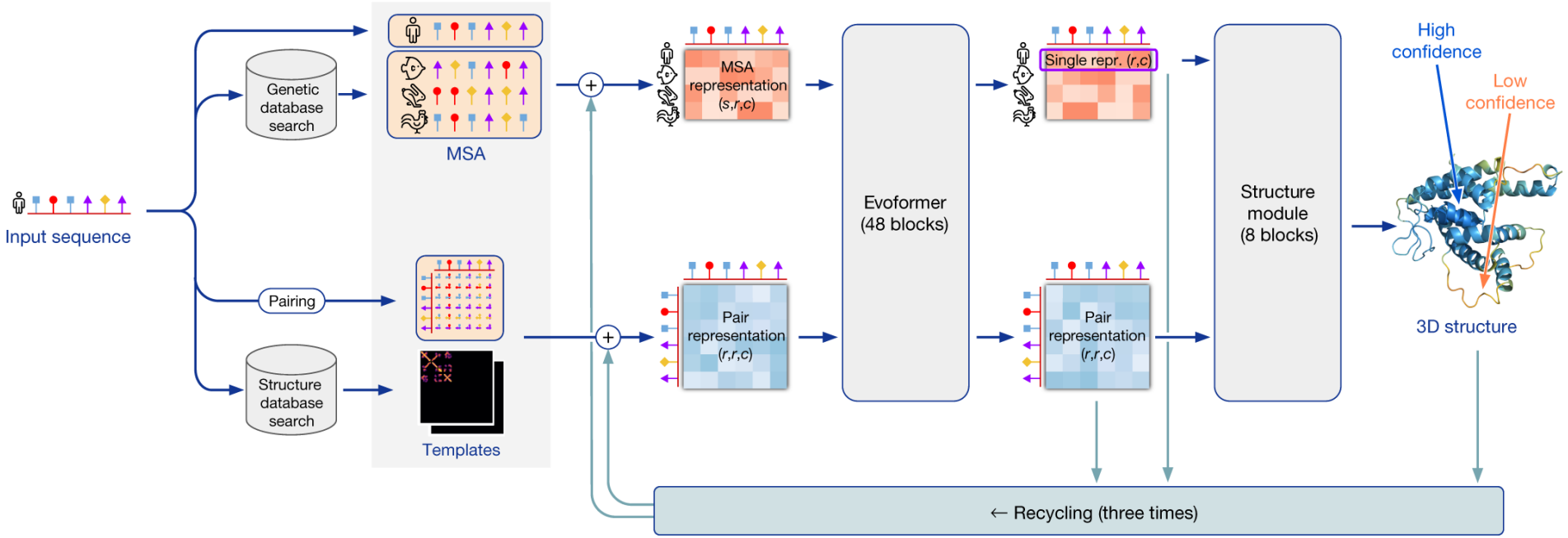
Cannot understate the leap AlphaFold2 made

- Zhavoronkov speculates in *Forbes* about AlphaFold winning Nobel prize
- Mohammed AlQuraishi's famous blog post
  - "my expectation… not until the late 2020s would we see >90 GDT_TS for most targets." (AlphaFold2 median 92.4)
  - "The core field has been blown to pieces; there's just no sugar-coating it."
  - "This was captured poignantly by a panelist at the very last session of the conference who remarked that CASP14 feels a bit like when one's child leaves home for the very first time."

https://www.forbes.com/sites/alexzhavoronkov/2023/11/02/will-there-be-a-nobel-prize-for-ai/

https://moalquraishi.wordpress.com/2020/12/08/alphafold2-casp14-it-feels-like-ones-child-has-left-home/

The AlphaFold2 model



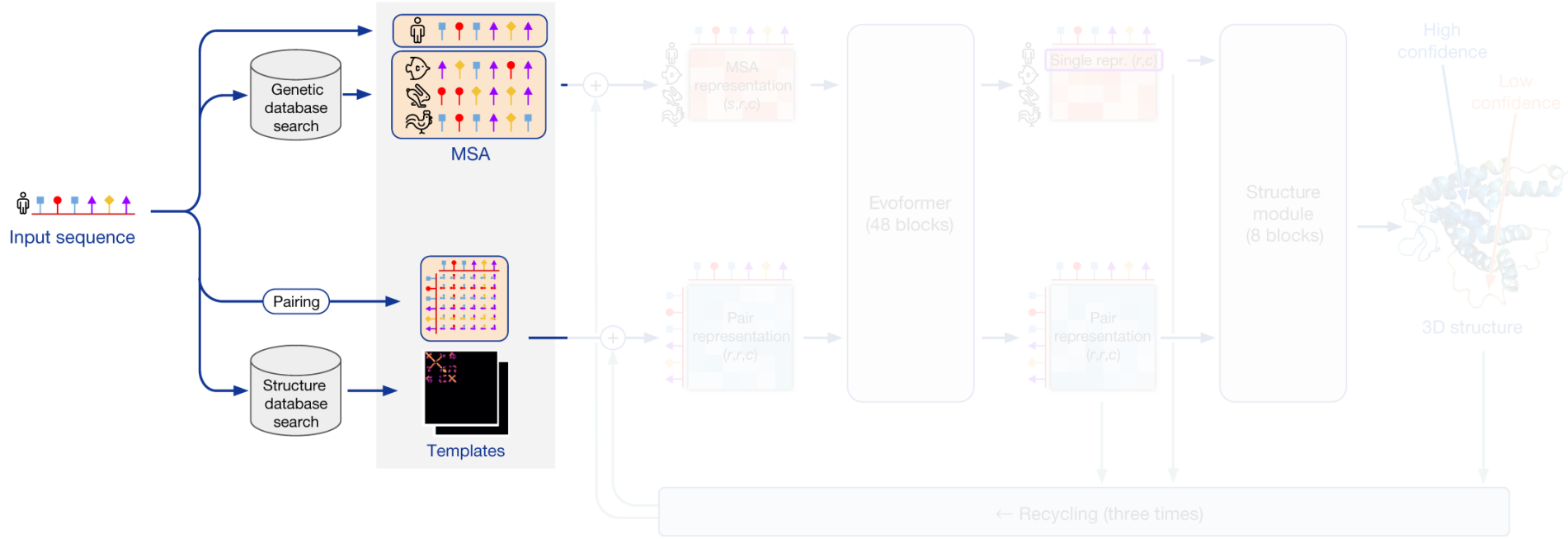
Input: amino acid sequence

Primary output: 3D coordinates for atoms

Jumper Nature 2021 doi:10.1038/s41586-021-03819-2

The AlphaFold2 model

Use input seq to search huge sequence database
Build multiple sequence alignment (MSA)

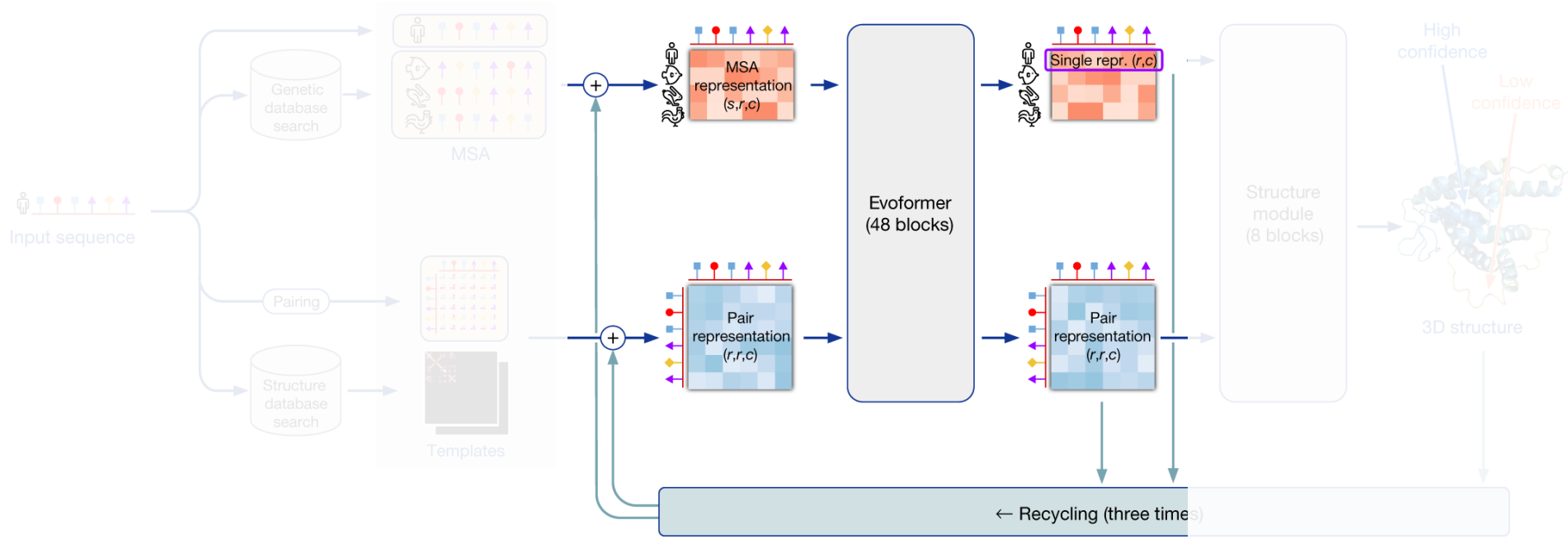


Use input seq to structure database
Find similar structure templates

Jumper Nature 2021 doi:10.1038/s41586-021-03819-2

The AlphaFold2 model

Iteratively improve embedding of MSA

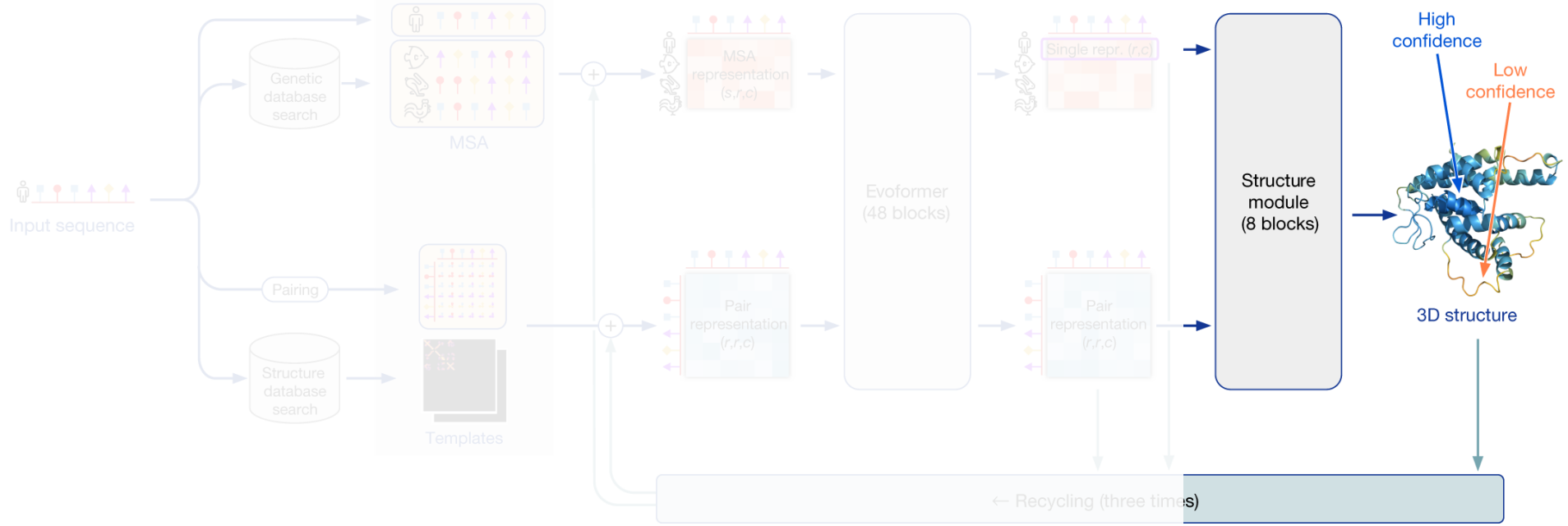


Iteratively improve embedding of residue pairs
Share information across these embeddings

Jumper Nature 2021 doi:10.1038/s41586-021-03819-2

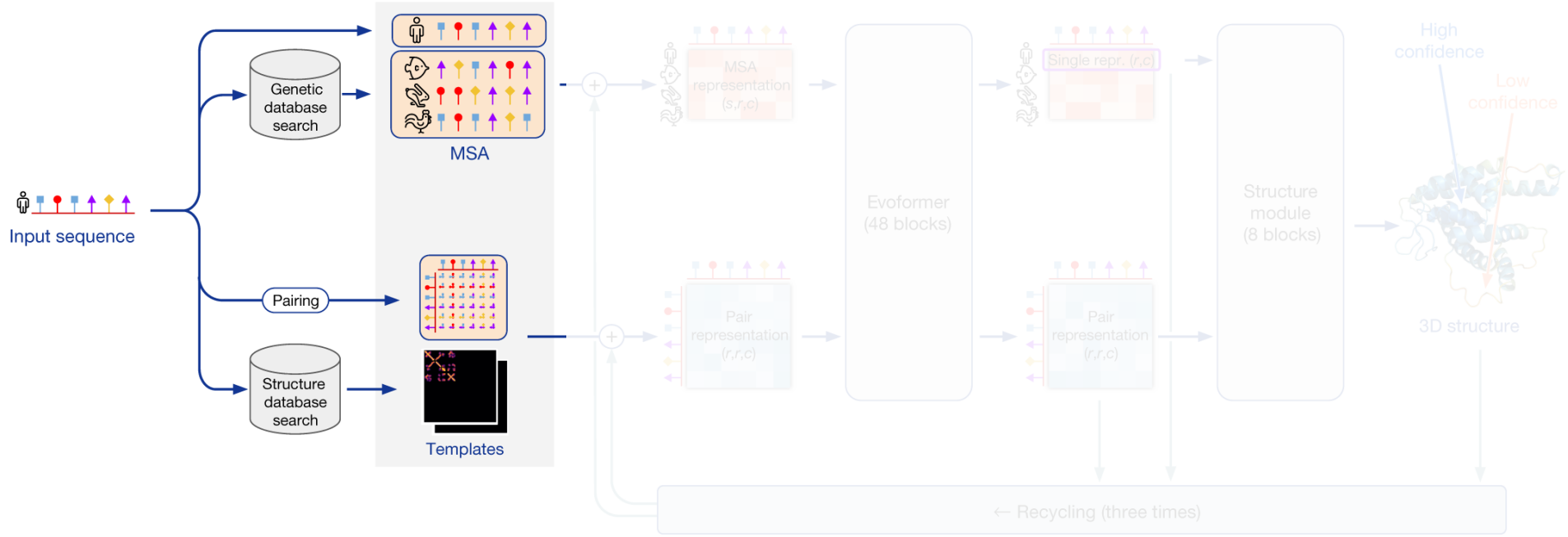
The AlphaFold2 model

Convert abstract sequence representation to 3D coordinates



Jumper Nature 2021 doi:10.1038/s41586-021-03819-2

AlphaFold2 input



Jumper Nature 2021 doi:10.1038/s41586-021-03819-2

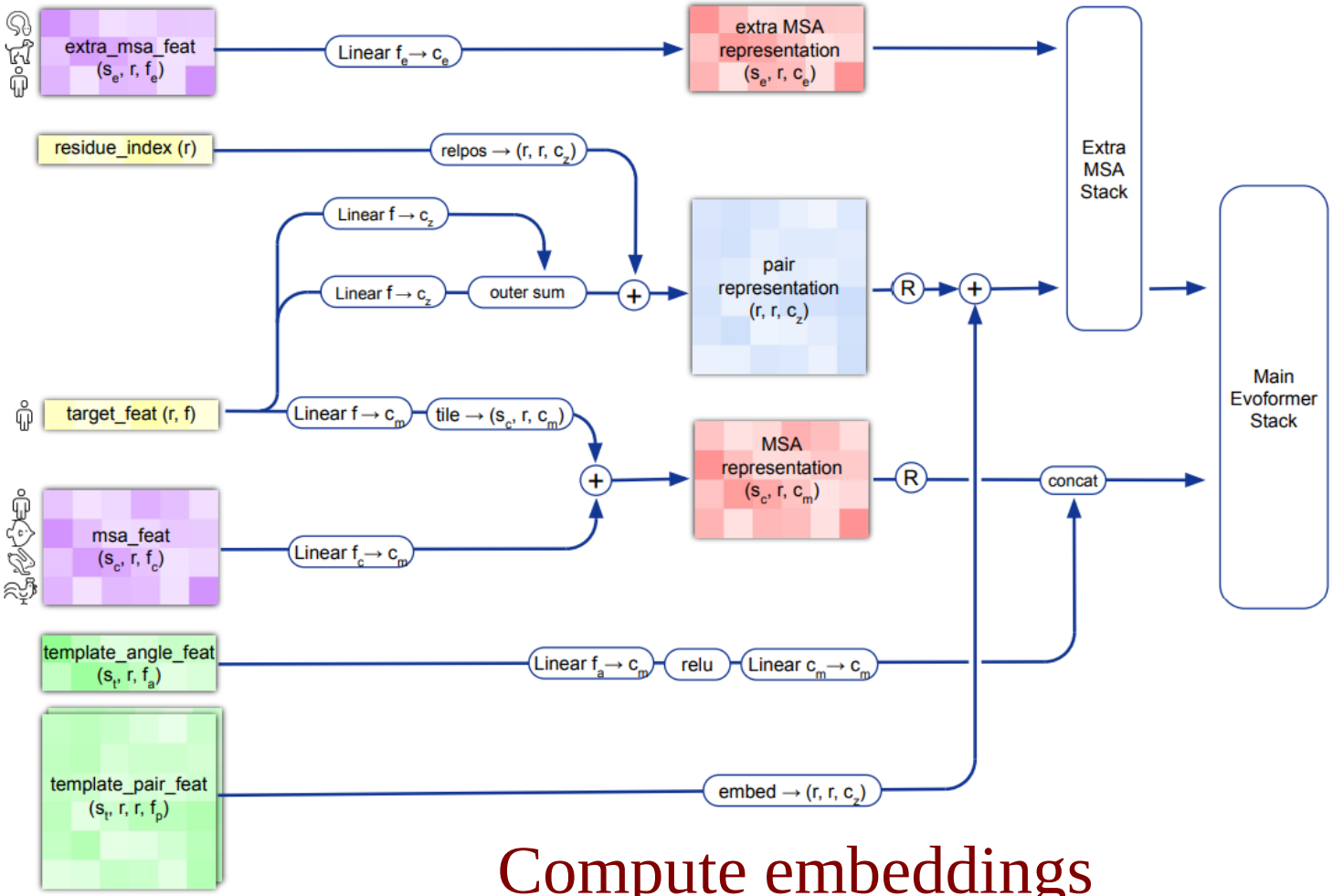
AlphaFold2 input

Extra MSA seqs

Input sequence

Clustered MSA seqs

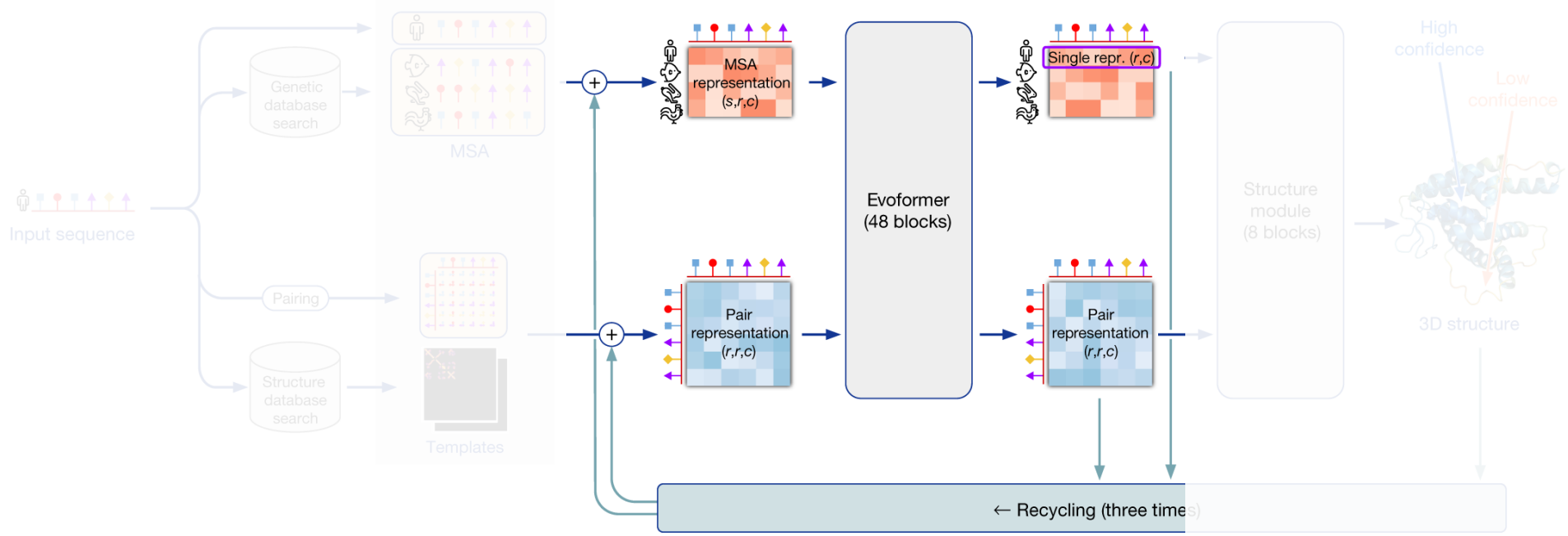
Template inputs



Compute embeddings

Jumper Nature 2021 doi:10.1038/s41586-021-03819-2

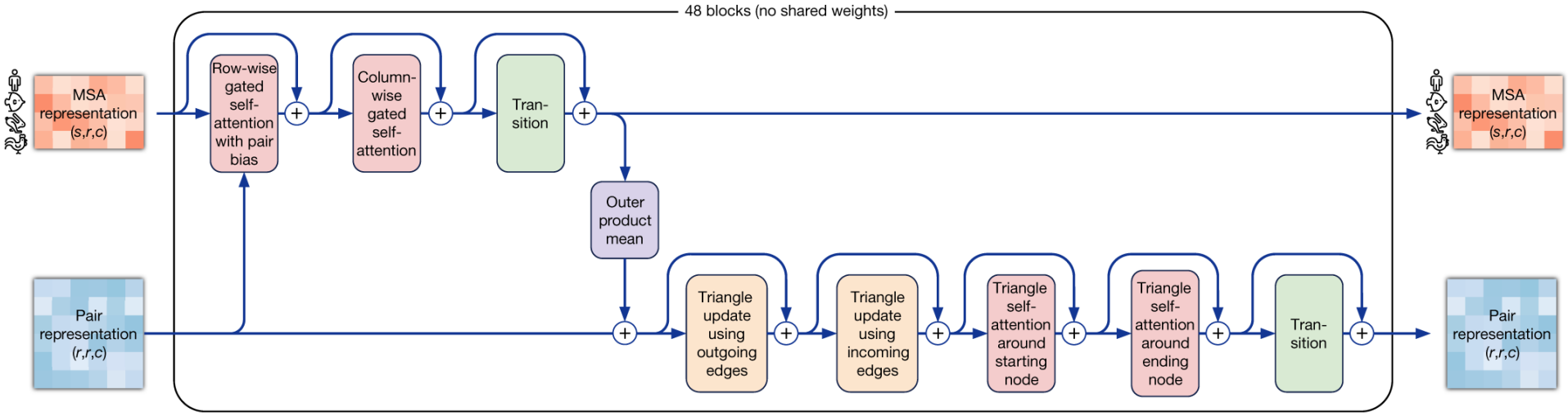
AlphaFold2 Evoformer



Jumper Nature 2021 doi:10.1038/s41586-021-03819-2

AlphaFold2 Evoformer

Iteratively improve embedding of MSA



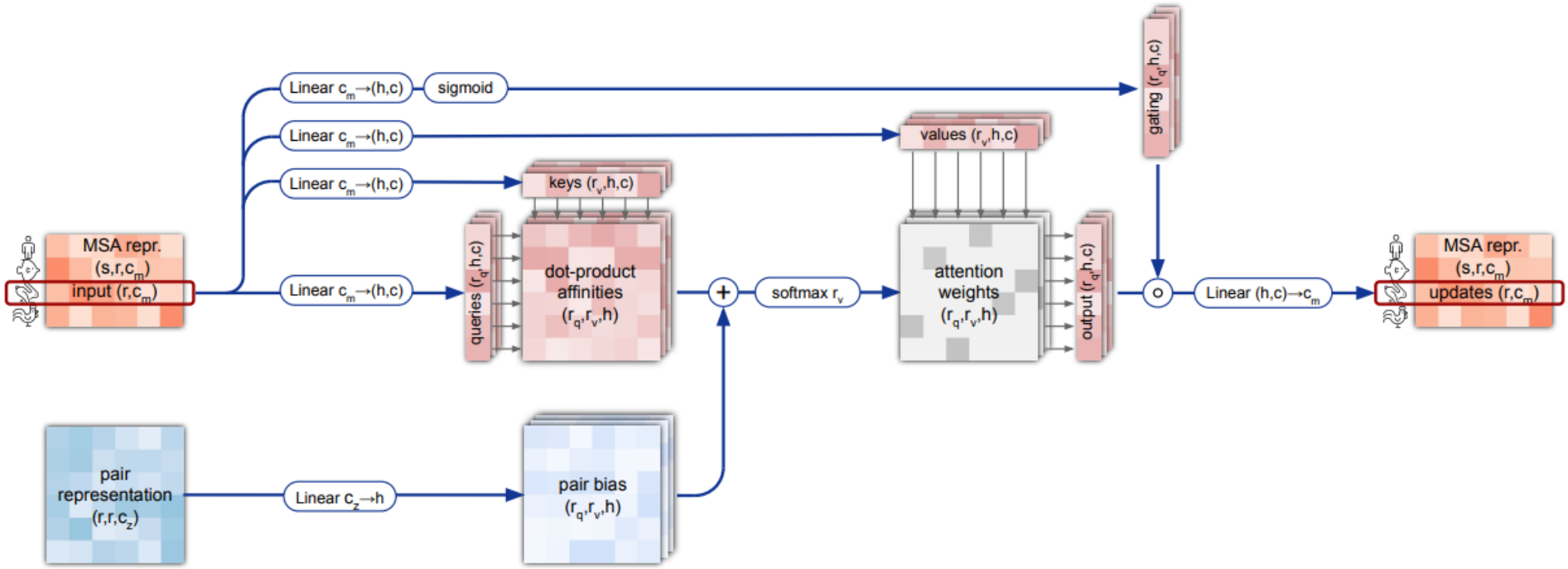
Iteratively improve embedding of residue pairs

Share information across these embeddings

Jumper Nature 2021 doi:10.1038/s41586-021-03819-2

AlphaFold2 Evoformer: MSA

MSA row-wise gated self-attention with pair bias

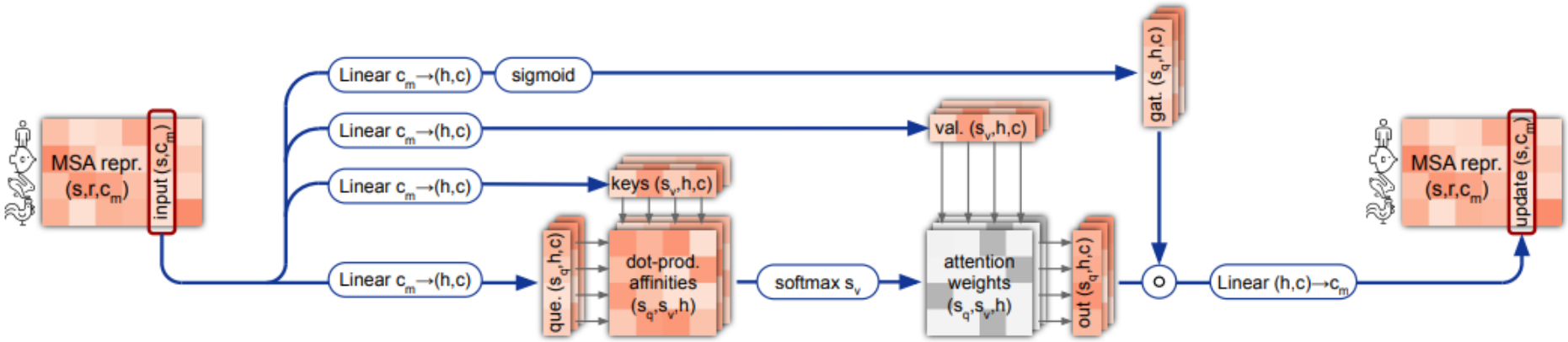


Pair representation influences the attention calculations

Jumper Nature 2021 doi:10.1038/s41586-021-03819-2

AlphaFold2 Evoformer: MSA

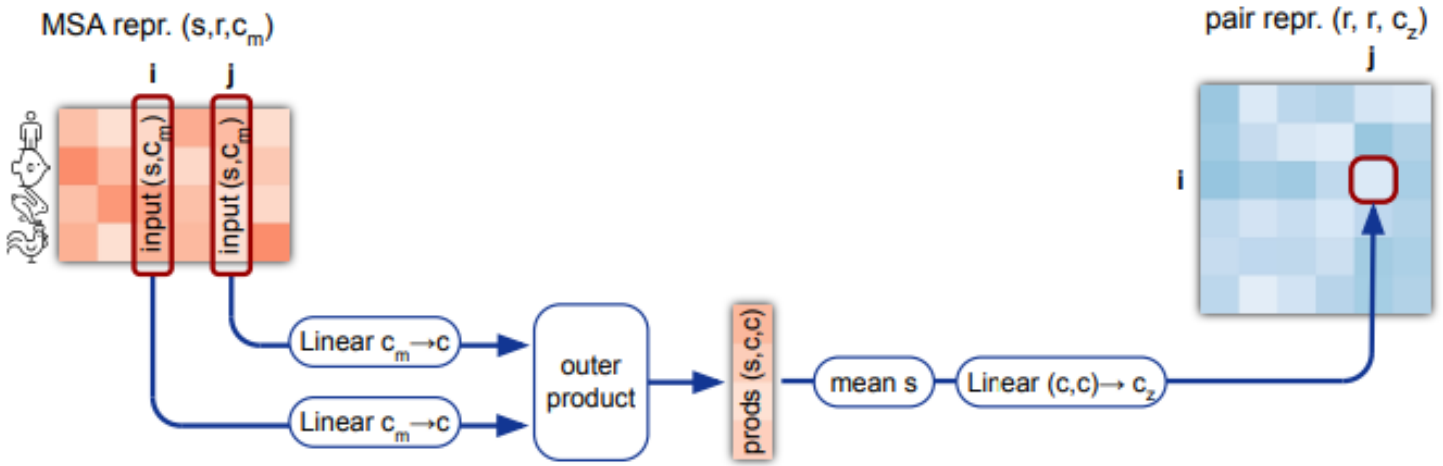
MSA column-wise gated self-attention



Jumper Nature 2021 doi:10.1038/s41586-021-03819-2

AlphaFold2 Evoformer: pairs

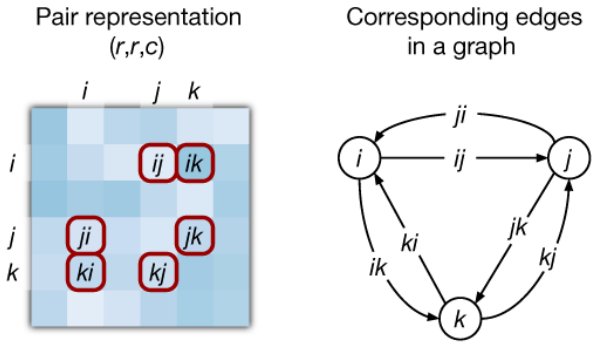
Outer product mean

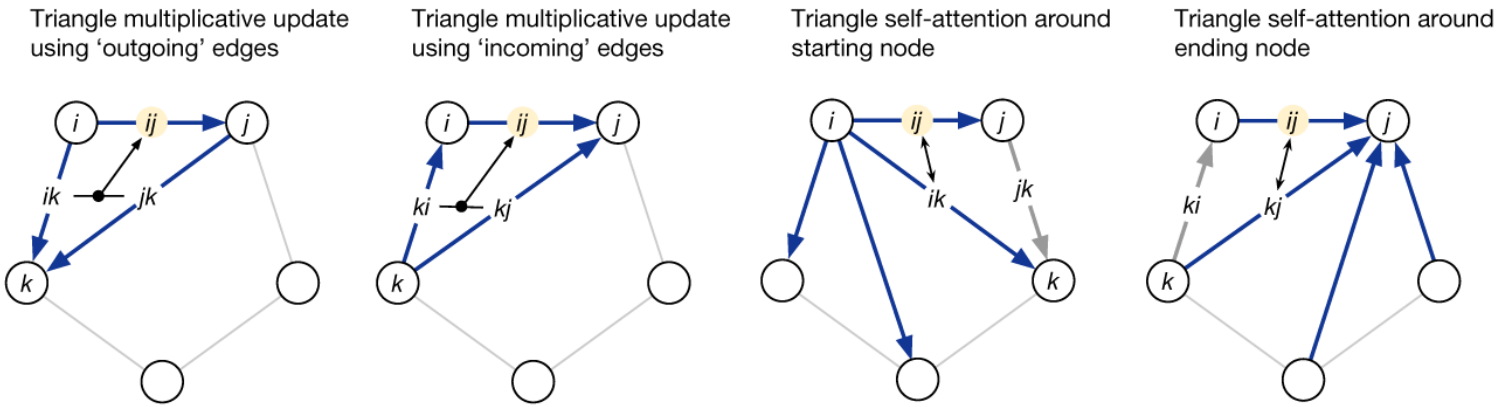


Extract pairwise information from updated MSA

AlphaFold2 Evoformer: pairs

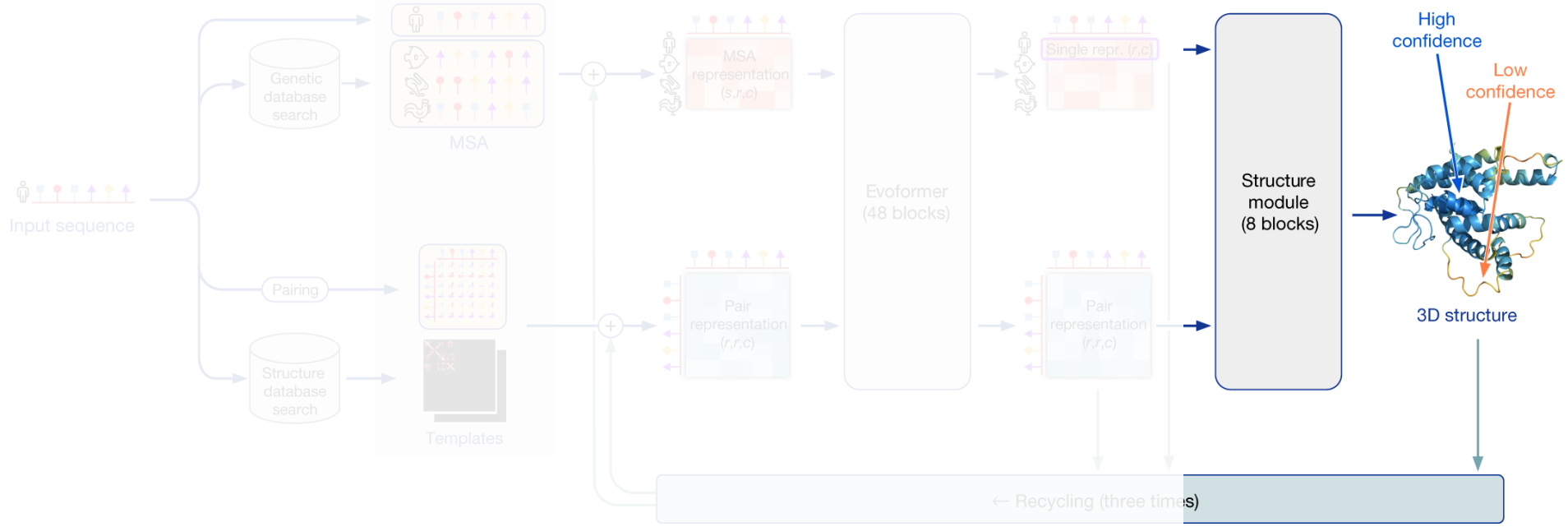
Triangle multiplicative updates and self-attention





Jumper Nature 2021 doi:10.1038/s41586-021-03819-2

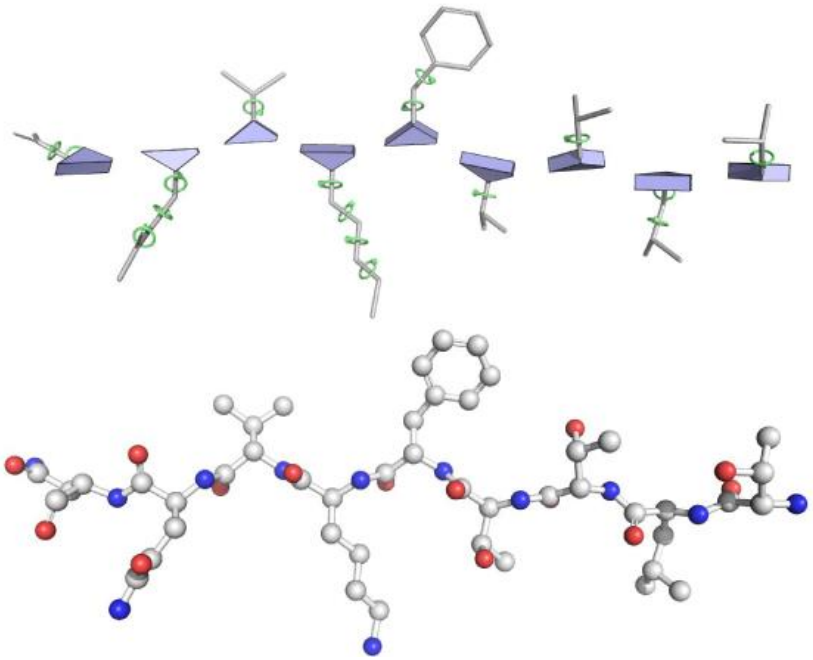
AlphaFold2 structure module

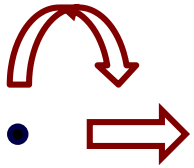


Jumper Nature 2021 doi:10.1038/s41586-021-03819-2

AlphaFold2 structure module

Residue modeled as a triangle
of three backbone atoms
Learn the side chain angles



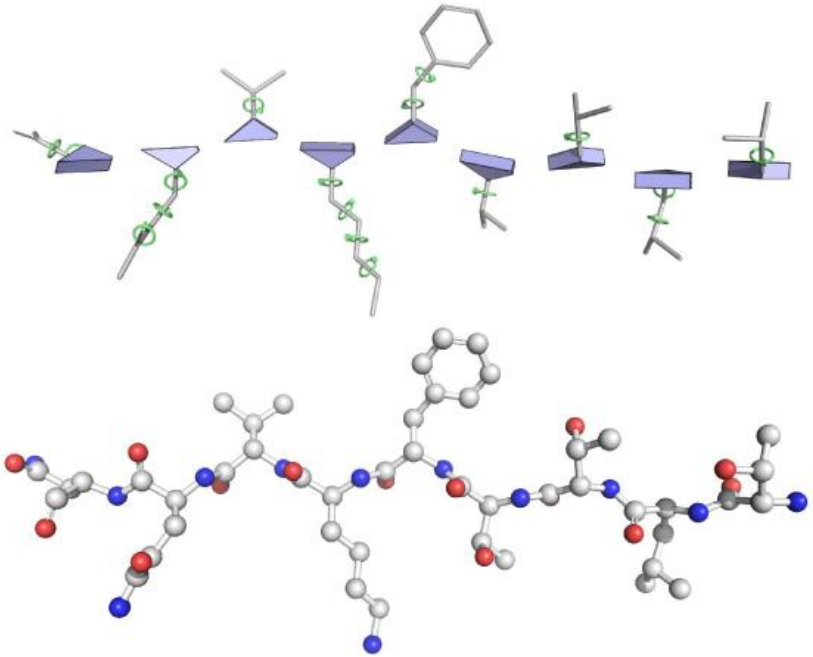


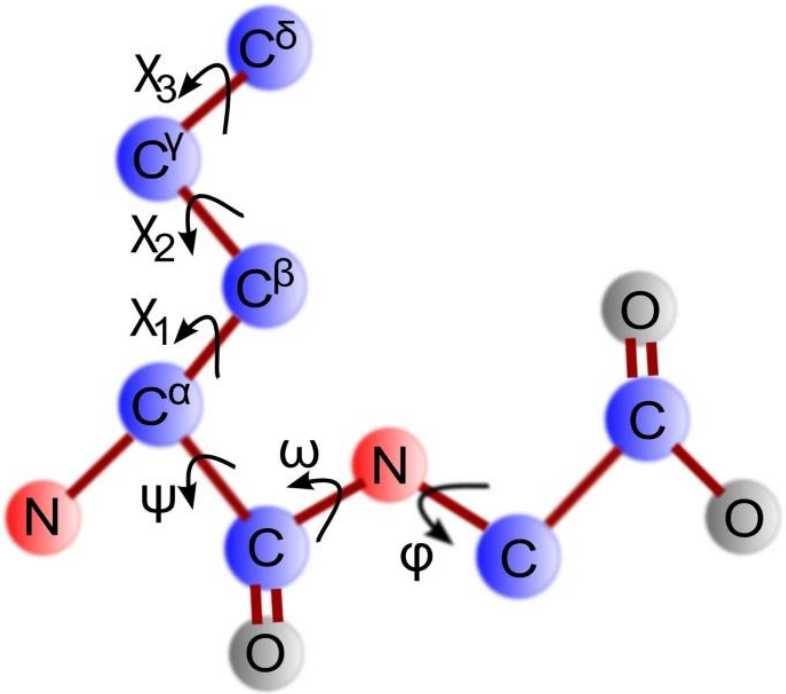
Learn a rotation and
translation for each residue
in the sequence

Jumper Nature 2021 doi:10.1038/s41586-021-03819-2

AlphaFold2 structure module

Residue modeled as a triangle
of three backbone atoms
Learn the side chain angles

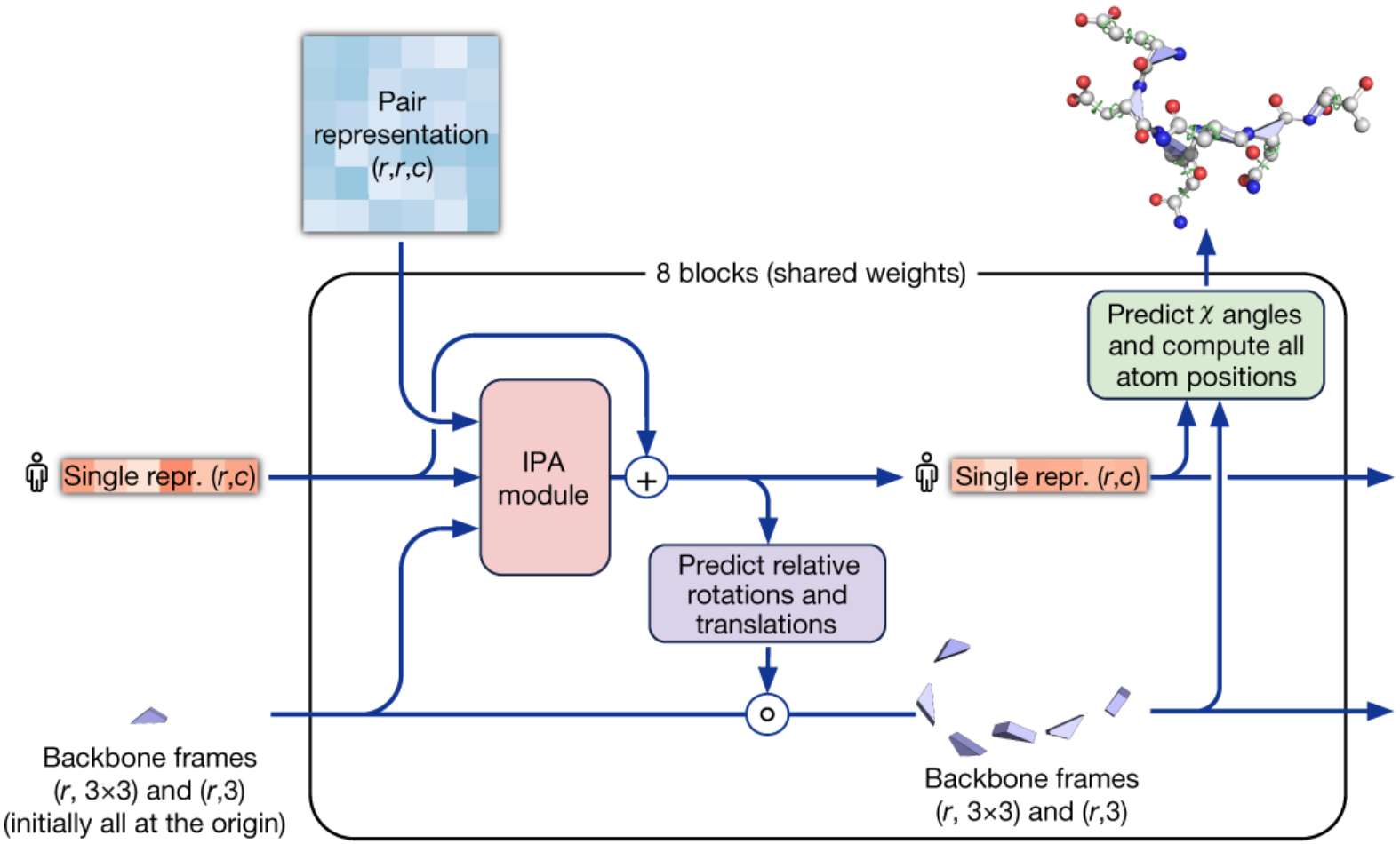




Torsion angles of residue

Jumper Nature 2021 doi:10.1038/s41586-021-03819-2 BMC Bioinformatics 2011 doi:10.1186/1471-2105-12-S14-S10

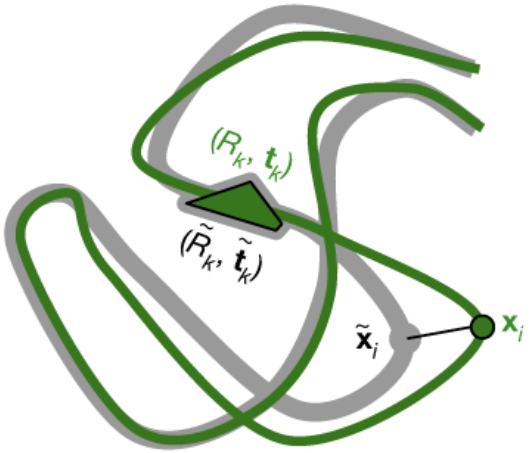
AlphaFold2 structure module



Jumper Nature 2021 doi:10.1038/s41586-021-03819-2

AlphaFold2 structure module

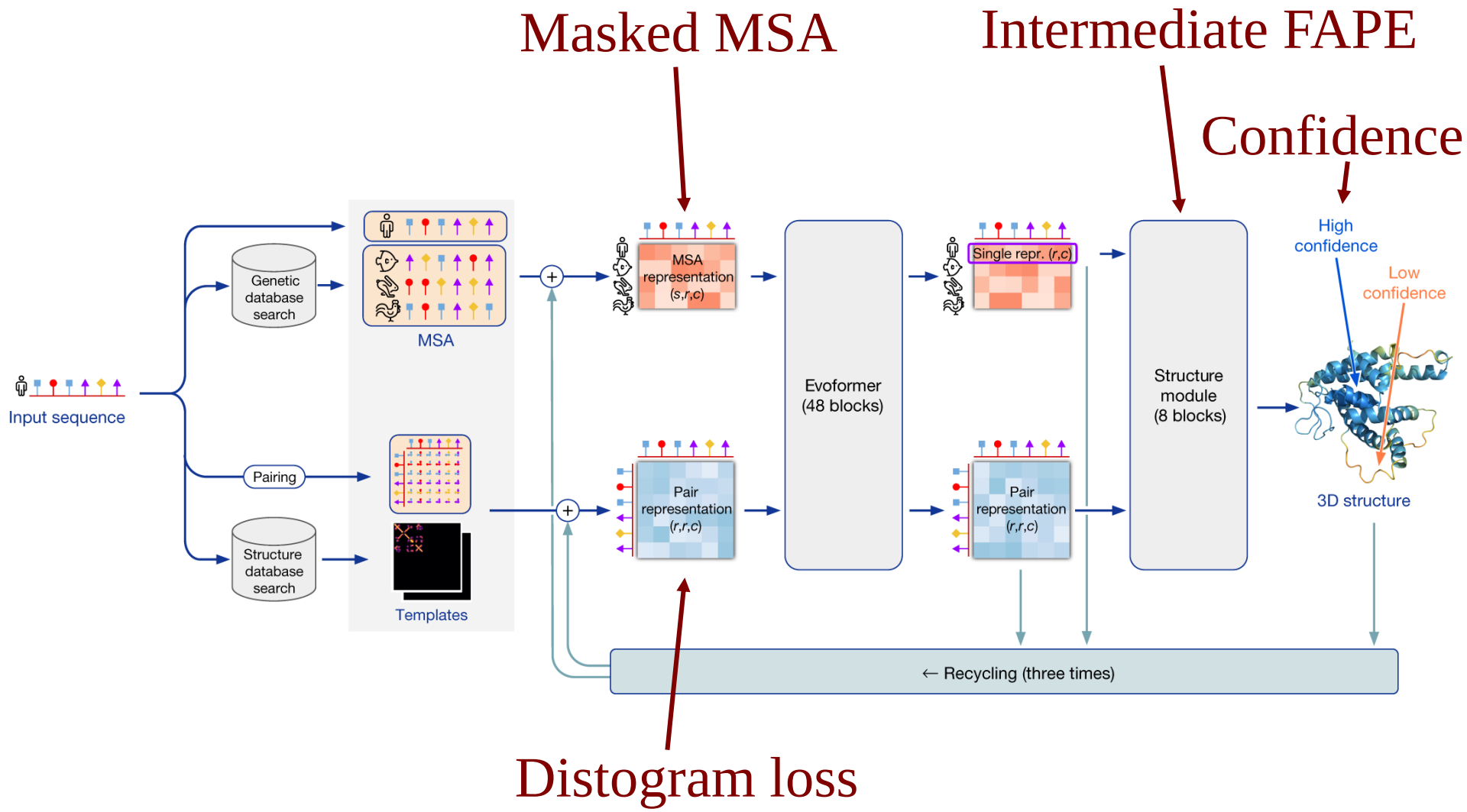
Main loss: frame aligned point error (FAPE)



Considers all atoms, must get side chains and chirality correct

Jumper Nature 2021 doi:10.1038/s41586-021-03819-2

Many auxiliary losses



Jumper Nature 2021 doi:10.1038/s41586-021-03819-2

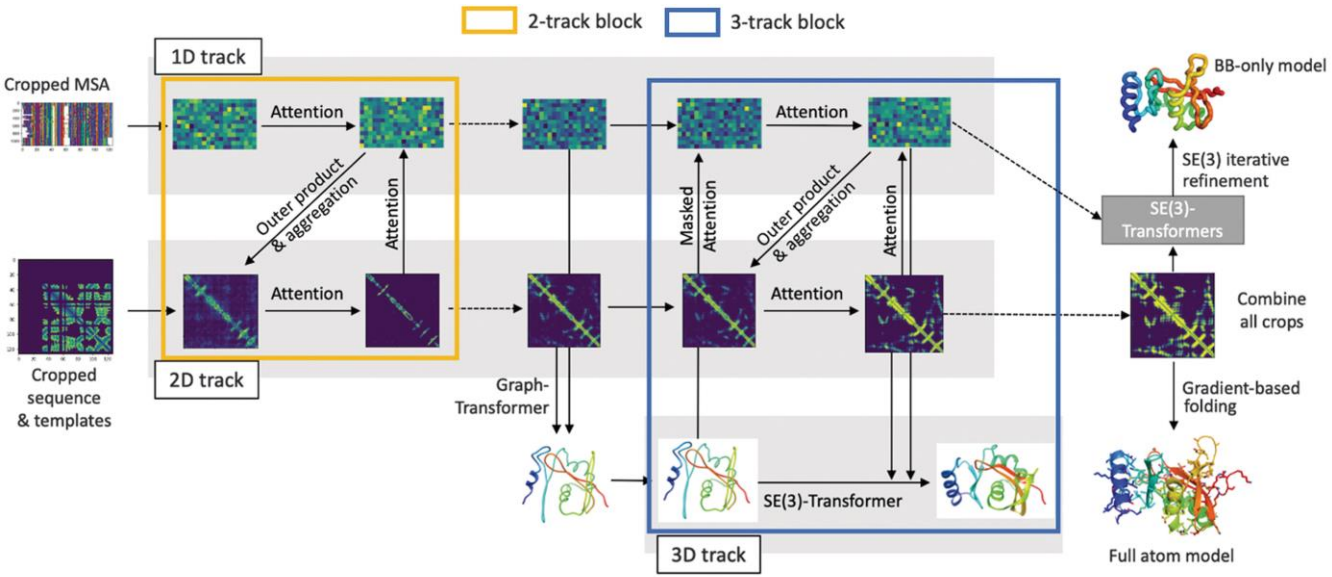
Many other important details

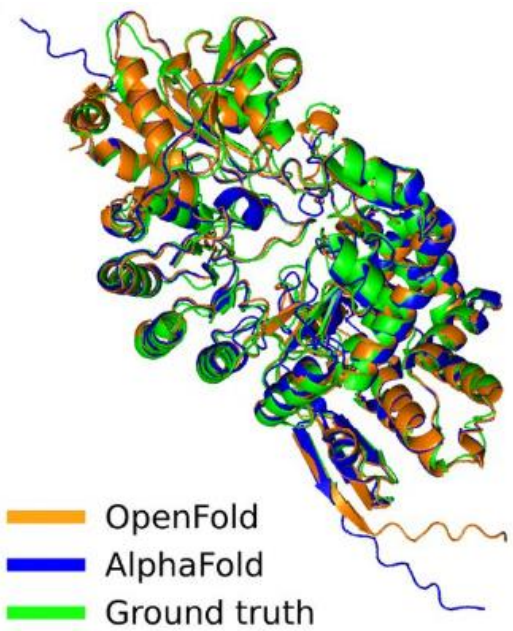
- Ensemble 5 models
- Relax structures with OpenMM
- Recycle the predictions multiple times
- Train on large number of predicted structures (self-distillation)
- Predict per-residue accuracy (predicted Cα local-distance difference test, pLDDT)

POST-ALPHAFOLD2

Academic inspiration

RoseTTAFold inspired by ideas from CASP14





OpenFold reproduces AlphaFold2 including training, improves efficiency

Baek Science 2021 doi:10.1126/science.abj8754

Ahdritz bioRxiv doi:10.1101/2022.11.20.517210

Access to AlphaFold2 predictions

- After source code released, still challenging to run
- Requires and a GPU or TPU
- Requires over 2.5 TB of data
  - 5 GB models
  - 238 GB structures
  - Everything else sequences
- Need to run MSA and template search preprocessing before model inference

Access to AlphaFold2 predictions

ColabFold v1.5.5: AlphaFold2 using MMseqs2

Easy to use protein structure and complex prediction using AlphaFold2 and Alphafold2-multimer. Sequence alignments/templates are generated through MMseqs2 and HHsearch. For more details, see bottom of the notebook, checkout the ColabFold GitHub and read our manuscript. Old versions: v1.4, v1.5.1, v1.5.2, v1.5.3-patch

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: Making protein folding accessible to all. *Nature Methods*, 2022

Input protein sequence(s), then hit `Runtime` -> `Run all`

query_sequence: SKGEELFTGVVPILVELDGDVNGHKFRVSGEGEGDATYGKLTLKFICTTGKLPVPWPTLVTTLSYGVQCFSRYPDHMKQHDFFKSAMPEGYVQERTIFFKDDGNYKTRAEV

- Use `:` to specify inter-protein chainbreaks for **modeling complexes** (supports homo- and hetro-oligomers). For example **PI...SK:PI...SK** for a homodimer

jobname: GFP_O10

num_relax: 1

- specify how many of the top ranked structures to relax using amber

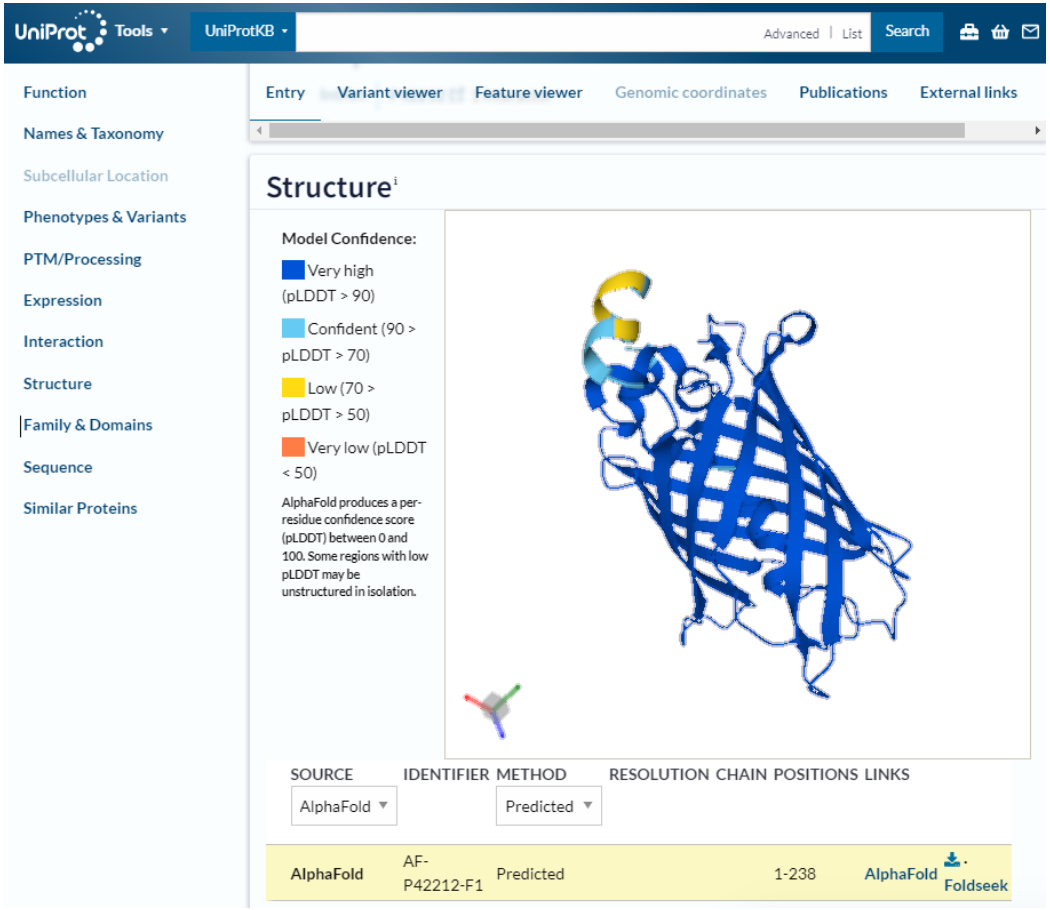
template_mode: none

- `none` = no template information is used. `pdb100` = detect templates in pdb100 (see notes). `custom` - upload and search own templates (PDB or mmCIF format, see notes)

ColabFold makes it trivial to generate and visualize a single structure prediction

Mirdita Nat Methods 2022 doi:10.1038/s41592-022-01488-1

Access to AlphaFold2 predictions



AlphaFold Protein Structure Database contains over 200M predicted structures, integrated into UniProt

Varadi Nucleic Acids Res 2022 doi:10.1093/nar/gkab1061

Subjectively sensed a shift in the community

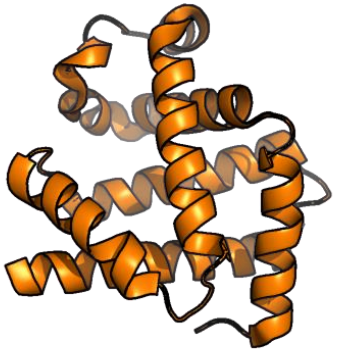
- Curiosity to see what AlphaFold2 predictions could bring to one's problem
  - "A structural biology community assessment of AlphaFold2 applications" Akdel Nat Struct Mol Biol doi:10.1038/s41594-022-00849-w
- Algorithm developers could assume (model of) protein structure available
- Structure models help interpret experimental data

Structure clustering and exploration

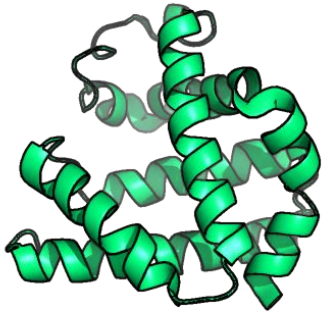
Proteins can have conserved structure without conserved sequence

Myoglobin proteins with sequence identity to human

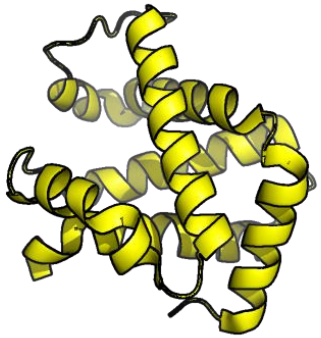
Human



African elephant (80%)



Pigeon (25%)

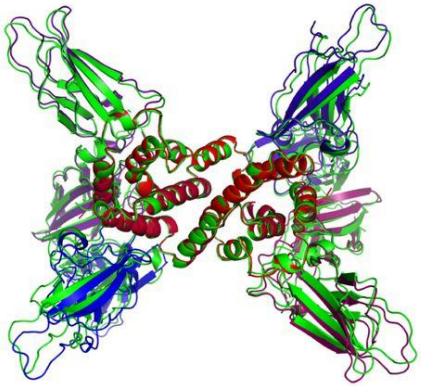


Black fin tuna (45%)

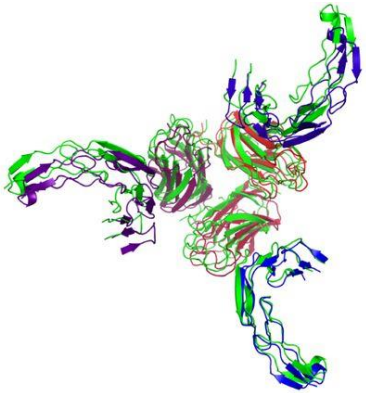
Cluster and analyze similarity within AlphaFold Protein Structure Database

Barrio-Hernandez Nature 2023 doi:10.1038/s41586-023-06510-w
https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/

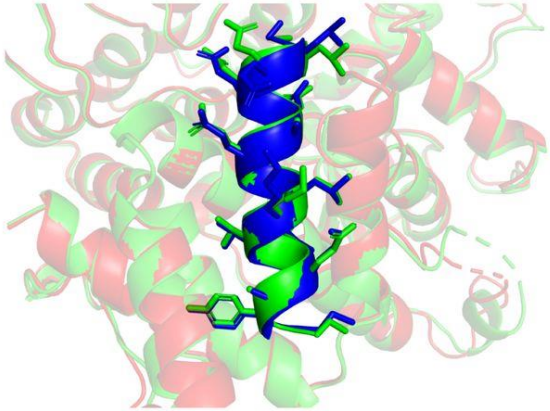
AlphaFold-Multimer



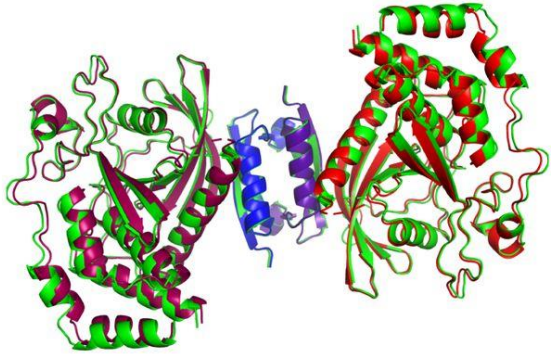
(a) A2B2C2 heteromer
TM-score = 97.4, N_{res} = 1,246, PDB ID = 6E3K



(b) A3B3 heteromer
TM-score = 85.4, N_{res} = 795, PDB ID = 7KHD



(c) Protein-peptide complex
TM-score = 96.6, DockQ = 0.954,
N_{res} = 385, PDB ID = 6JMT

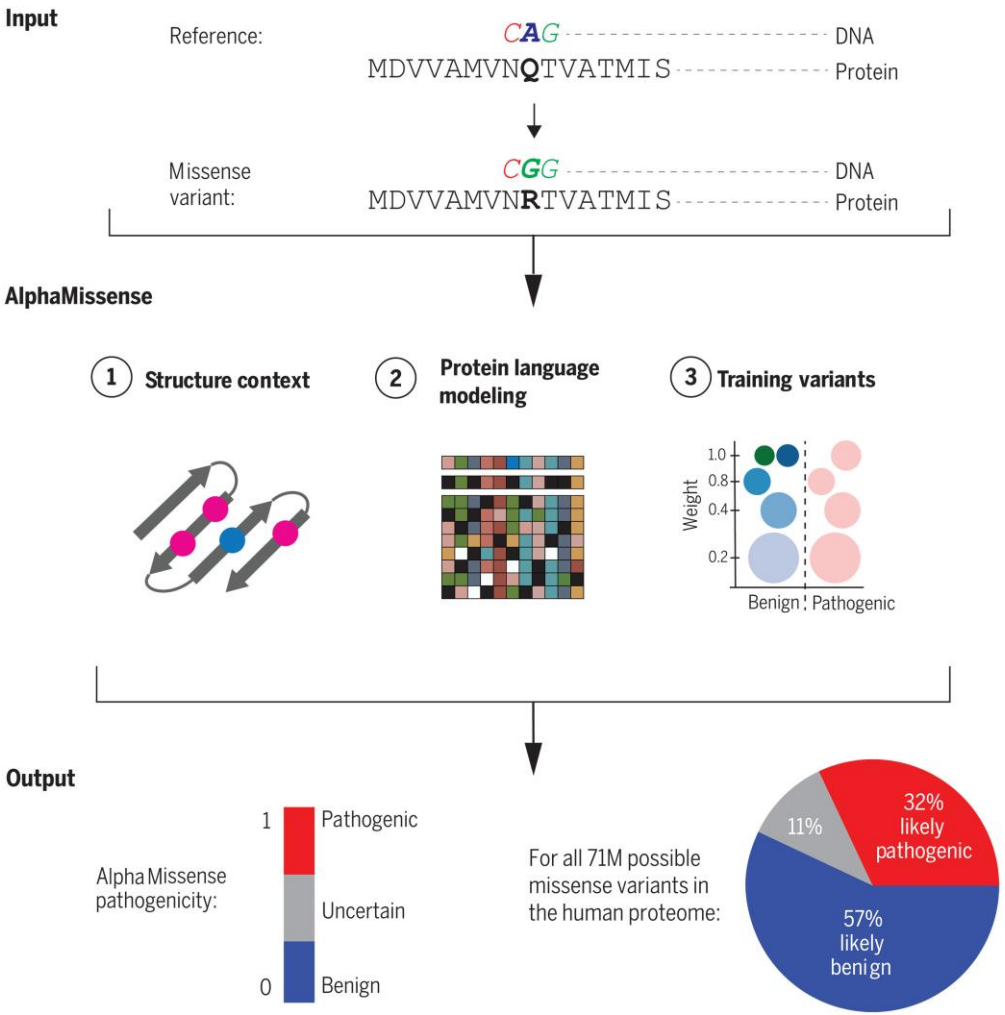


(d) A2B2 heteromer
TM-score = 98.5, N_{res} = 716, PDB ID = 6IWD

Extension to protein complexes

Evans bioRxiv doi:2021.10.04.463034

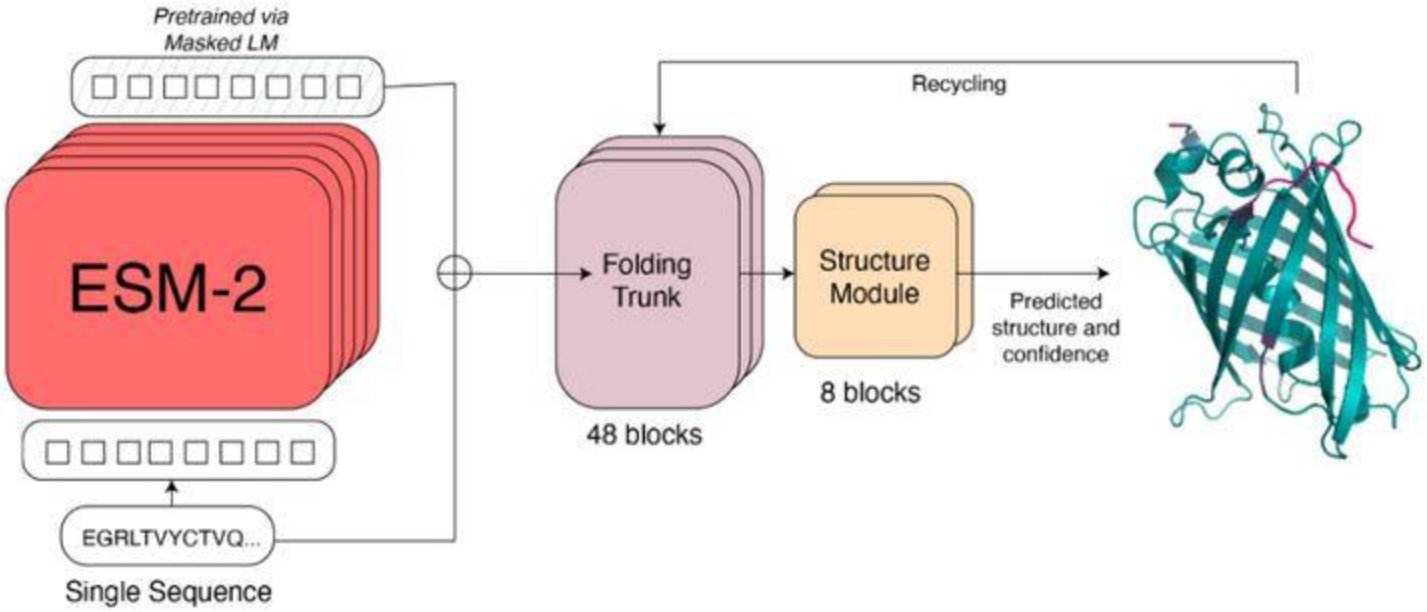
AlphaMissense



Predict pathogenicity of genetic variants

Cheng Science 2023 doi:10.1126/science.adg7492

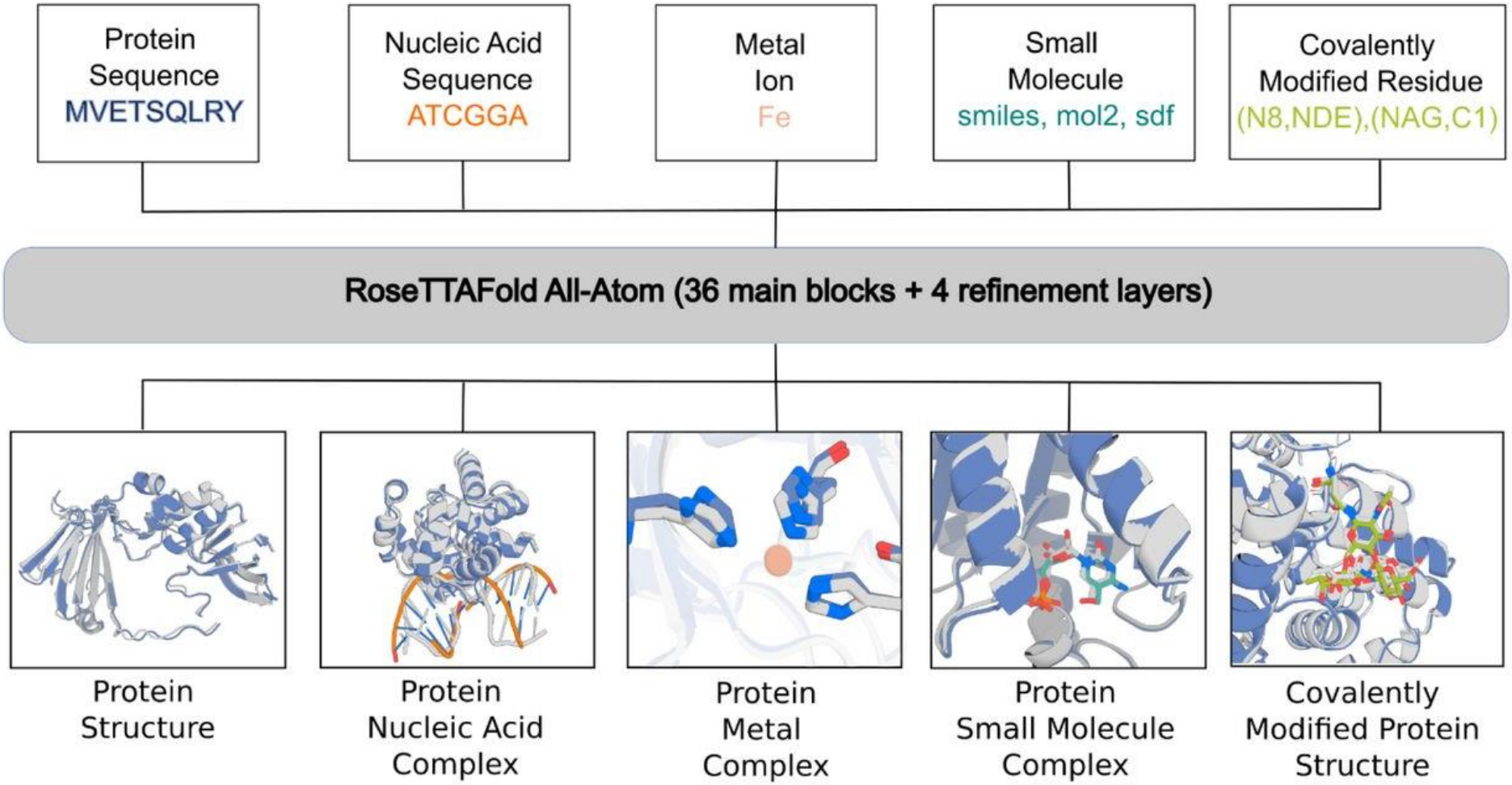
Single-sequence prediction



Train a language model on all natural protein sequences
Use instead of MSA

Lin Science 2023 doi:10.1126/science.ade2574

RoseTTAFold All-Atom



Protein structures in diverse complexes

Krishna Science 2024 doi:10.1126/science.adl2528

Conclusions

- Protein structure prediction from sequence was an open problem for over 50 years
- One version of it is now largely solved by AlphaFold2
- AlphaFold2 combines expert modeling of MSAs, templates, and protein geometry; professional deep learning engineering; large sequence and structure databases
- Protein machine learning and computational structural biology have flourished in the wake of AlphaFold2

Resources

- https://www.blopig.com/blog/2020/12/casp14-what-google-deepminds-alphafold-2-really-achieved-and-what-it-means-for-protein-folding-biology-and-bioinformatics/
- https://www.blopig.com/blog/2021/07/alphafold-2-is-here-whats-behind-the-structure-prediction-miracle/
- https://moalquraishi.wordpress.com/2020/12/08/alphafold2-casp14-it-feels-like-ones-child-has-left-home/
- https://bio.libretexts.org/Bookshelves/Biochemistry/Book%3A_Biochemistry_Free_For_All_(Ahern_Rajagopal_and_Tan)/02%3A_Structure_and_Function/203%3A_Structure__Function_-_Proteins_I
- https://predictioncenter.org/index.cgi