INSTITUT DU
DÉVELOPPEMENT ET DES
RESSOURCES EN
INFORMATIQUE
SCIENTIFIQUE

AI-DevTalk

AI for protein folding: focus on Alphafold

Thibaut Véry – IDRIS User support

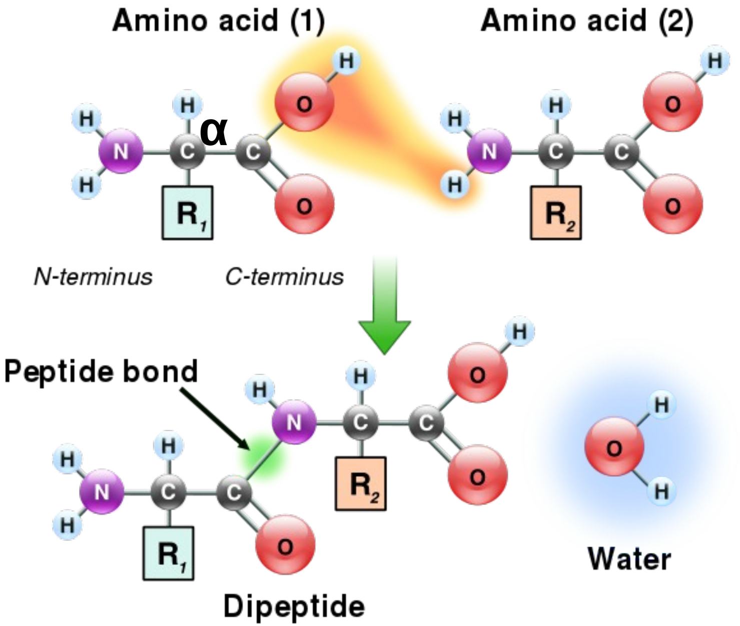
What is a protein?

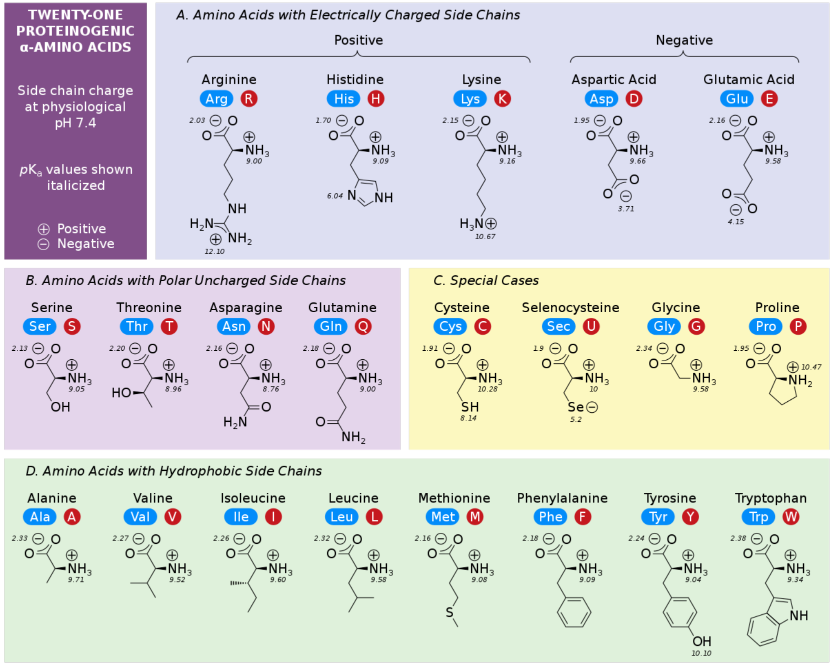
Sequence of aminoacids: FASTA format

```
>sp|P0DTC2|SPIKE_SARS2 Spike glycoprotein OS=Severe acute respiratory syndrome coronavirus 2 OX=2697049 GN=S PE=1 SV=1
MFVFLVLLPLVSSQCVNLTTRTQLPPAYTNSFTRGVYYPDKVFRSSVLHSTQDLFLPFFS
NVTWFHAIHVSGTNGTKRFDNPVLPFNDGVYFASTEKSNIIRGWIFGTTLDSKTQSLLIV
NNATNVVIKVCEFQFCNDPFLGVYYHKNNKSWMESEFRVYSSANNCTFEYVSQPFLMDLE
GKQGNFKNLREFVFKNIDGYFKIYSKHTPINLVRDLPQGFSALEPLVDLPIGINITRFQT
LLALHRSYLTPGDSSSGWTAGAAAYYVGYLQPRTFLLKYNENGTITDAVDCALDPLSETK
CTLKSFTVEKGIYQTSNFRVQPTESIVRFPNITNLCPFGEVFNATRFASVYAWNRKRISN
CVADYSVLYNSASFSTFKCYGVSPTKLNDLCFTNVYADSFVIRGDEVRQIAPGQTGKIAD
YNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPC
NGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVN
FNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP
GTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY
ECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI
SVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE
VFAQVKQIYKTPPIKDFGGFNFSQILPDPSKPSKRSFIEDLLFNKVTLADAGFIKQYGDC
LGDIAARDLICAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAM
QMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQDSLSSTASALGKLQDVVNQNAQALN
TLVKQLSSNFGAISSVLNDILSRLDKVEAEVQIDRLITGRLQSLQTYVTQQLIRAAEIRA
SANLAATKMSECVLGQSKRVDFCGKGYHLMSFPQSAPHGVVFLHVTYVPAQEKNFTTAPA
ICHDGKAHFPREGVFVSNGTHWFVTQRNFYEPQIITTDNTFVSGNCDVVIGIVNNTVYDP
LQPELDSFKEELDKYFKNHTSPDVDLGDISGINASVVNIQKEIDRLNEVAKNLNESLIDL
QELGKYEQYIKWPWYIWLGFIAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKFDEDD
SEPVLKGVKLHYT
```

What is a protein?

<u>Primary structure</u>: sequence of aminoacids (called residues)

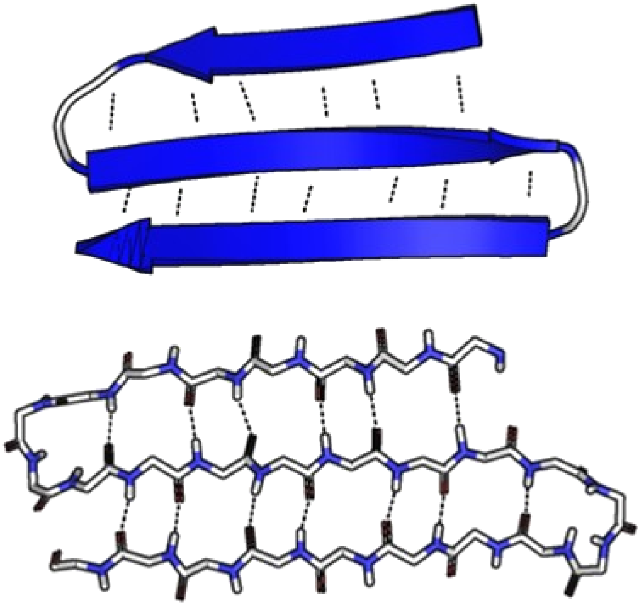



TWENTY-ONE PROTEINOGENIC α-AMINO ACIDS
Side chain charge at physiological pH 7.4
pK_a values shown italicized
⊕ Positive
⊖ Negative
A. Amino Acids with Electrically Charged Side Chains
Positive
Negative
Arginine Arg R 2.03 9.00 12.10
Histidine His H 1.70 9.09 6.04
Lysine Lys K 2.15 9.16 10.67
Aspartic Acid Asp D 1.95 9.66 3.71
Glutamic Acid Glu E 2.16 9.58 4.15
B. Amino Acids with Polar Uncharged Side Chains
Serine Ser S 2.13 9.05
Threonine Thr T 2.20 8.96
Asparagine Asn N 2.16 8.76
Glutamine Gln Q 2.18 9.00
C. Special Cases
Cysteine Cys C 1.91 10.28 8.14
Selenocysteine Sec U 1.9 10 5.2
Glycine Gly G 2.34 9.58
Proline Pro P 1.95 10.47
D. Amino Acids with Hydrophobic Side Chains
Alanine Ala A 2.33 9.71
Valine Val V 2.27 9.52
Isoleucine Ile I 2.26 9.60
Leucine Leu L 2.32 9.58
Methionine Met M 2.16 9.08
Phenylalanine Phe F 2.18 9.09
Tyrosine Tyr Y 2.24 9.04 10.10
Tryptophan Trp W 2.38 9.34

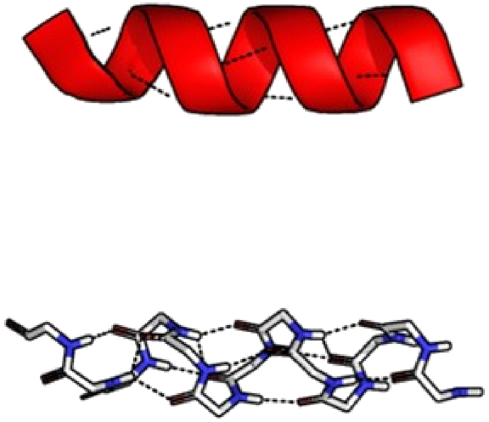
Number of residues: ~40 < n < ~35000 (average ~2000)

What is a protein?

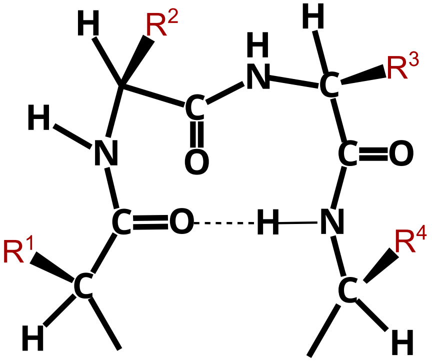
<u>Secondary structure</u>: Local structure thanks to inter-residue bonds (H-bond, ...)



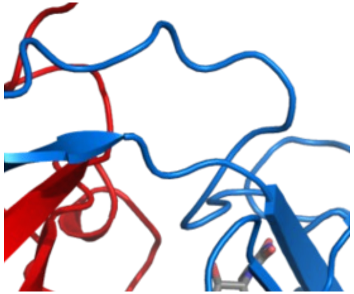
β-Sheet (3 strands)



α-helix



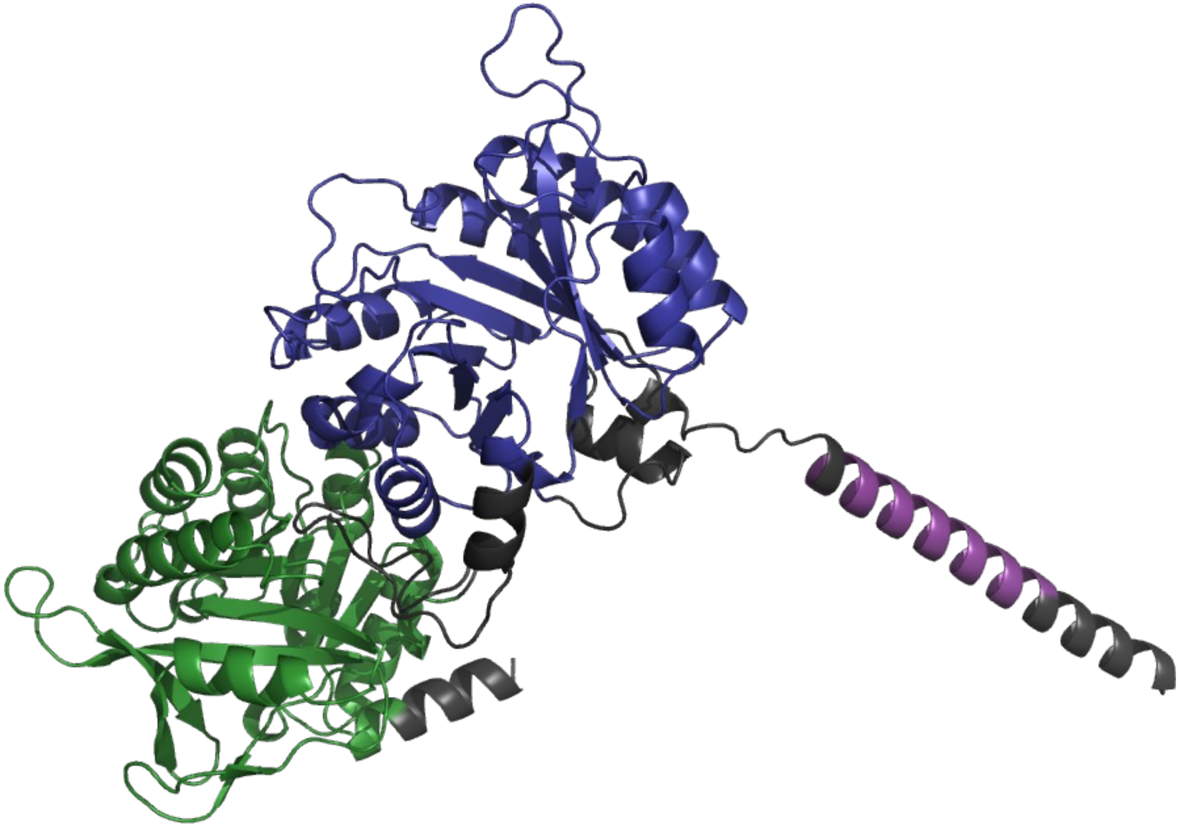
β turn: Type I



Loop

What is a protein?

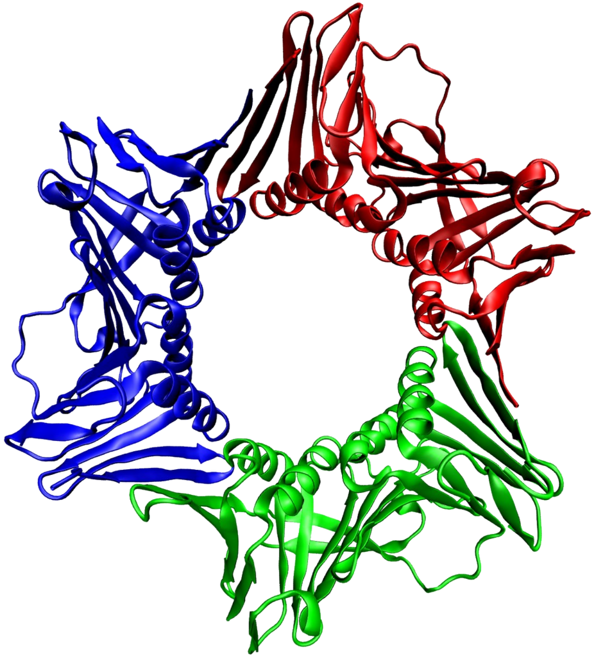
Tertiary structure: Global folding of the protein



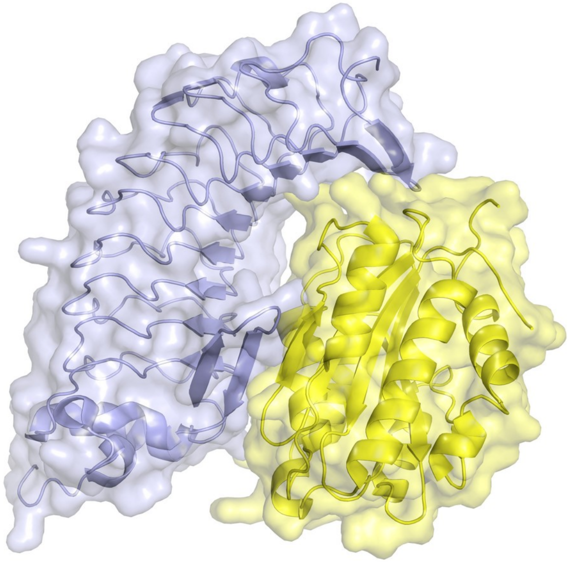
What is a protein?

<u>Quaternary structure</u>: Assembly of several protein units

Homo-n-mer (here trimer)

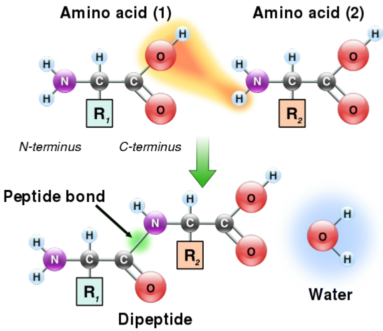


Hetero-n-mer (here dimer)

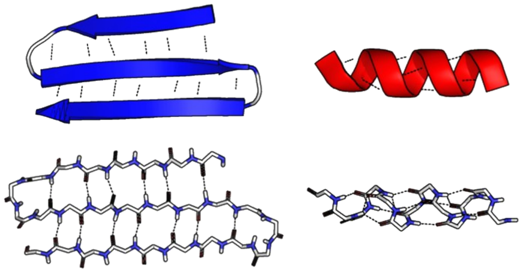


What is a protein?

Summary



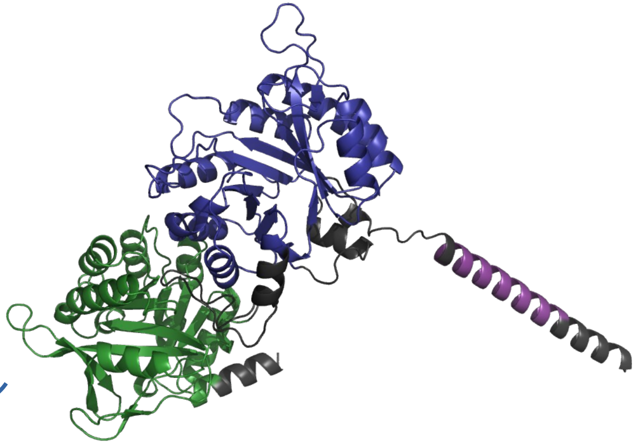
Interaction of AA gives secondary structures



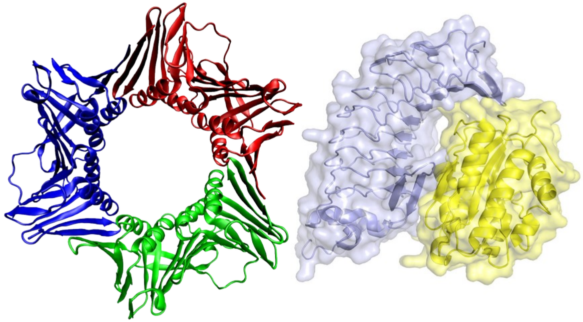
β-Sheet (3 strands)

α-helix

Folding of the secondary structure



Association of monomers (same or different)

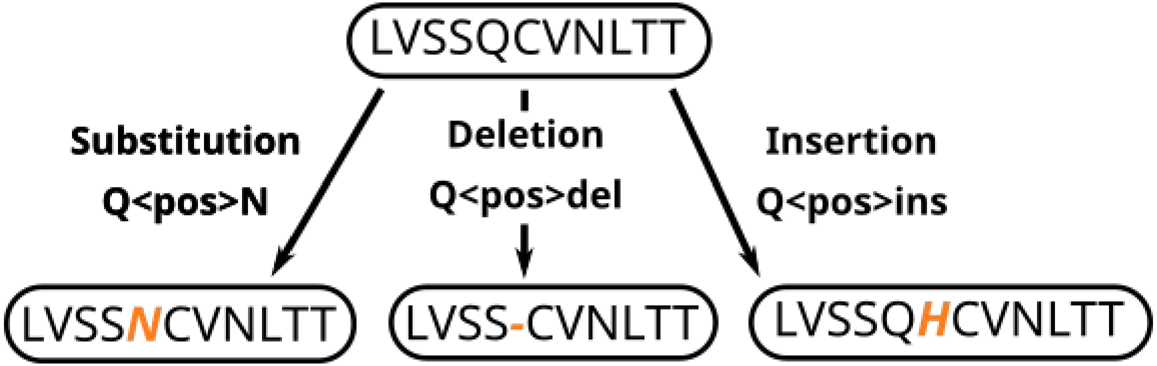


Protein structure/function

- The structure of the protein gives its function.
- Mutations can occur in the primary sequence as long as the structure does not change too much as to break the function.
- Some amino acids are important for chemical reactions in active sites and might be difficult to replace without breaking the function.

Mutations

- Several types of mutations appear



LVSSQCVNLTT

Substitution

Q<pos>N

Deletion

Q<pos>del

Insertion

Q<pos>ins

LVSSNCVNLTT

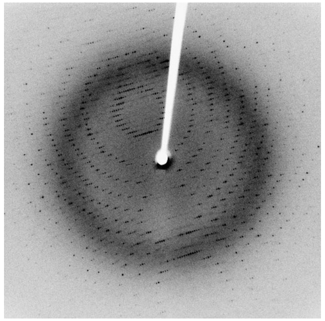
LVSS-CVNLTT

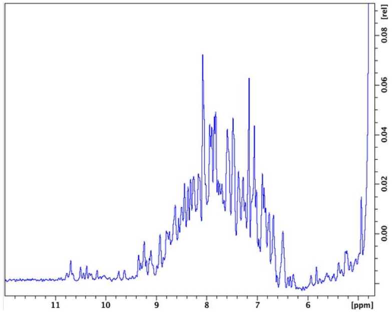
LVSSQHCVNLTT

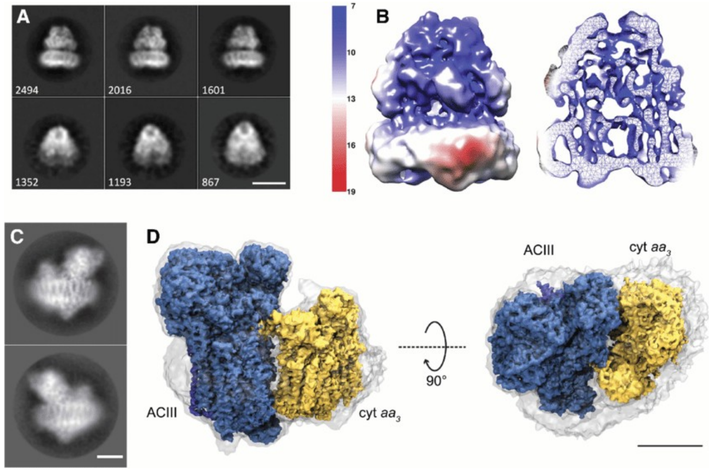
Institut du Développement et des Ressources en Informatique Scientifique

Finding protein structures: experimental methods

- Protein sequencing (no structure)
- X-ray crystallography
- NRM
- Cryo-electron microscopy

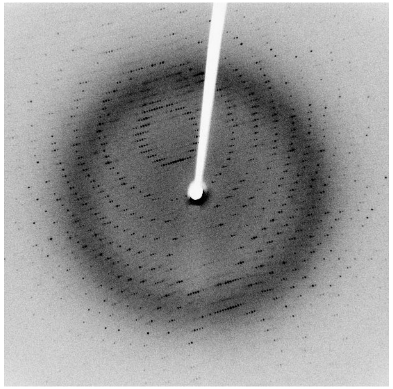


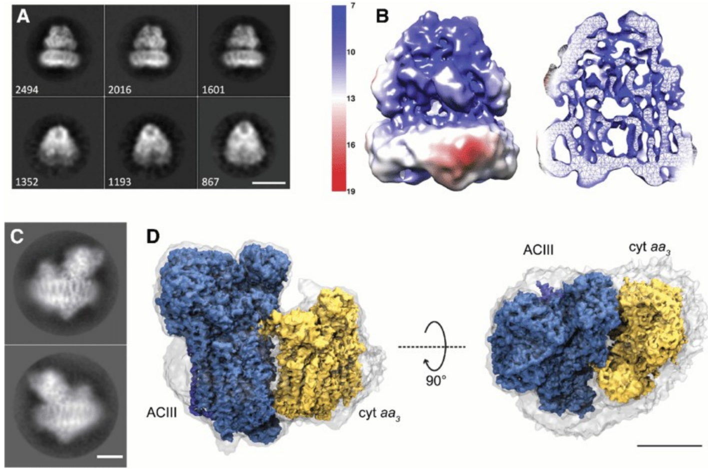


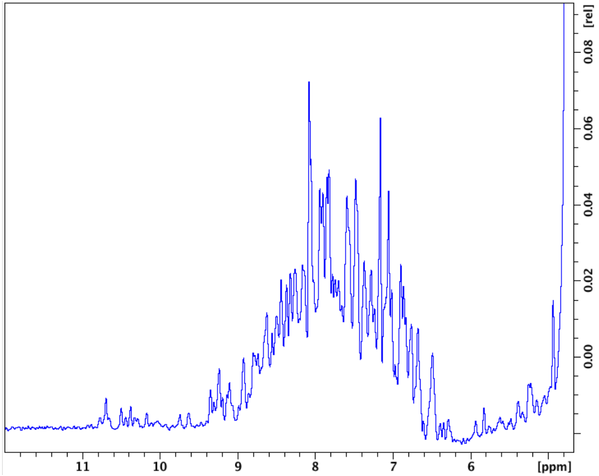


Finding protein structures: experimental methods

- Each method has strength and drawbacks.
- It might require a long time (and money) to get the structure.







Numerical methods

- Different groups of methods are available
  - Molecular dynamics
  - Conformational sampling
  - Comparative modeling
  - Fold recognition and threading
  - ...

Comparative modeling

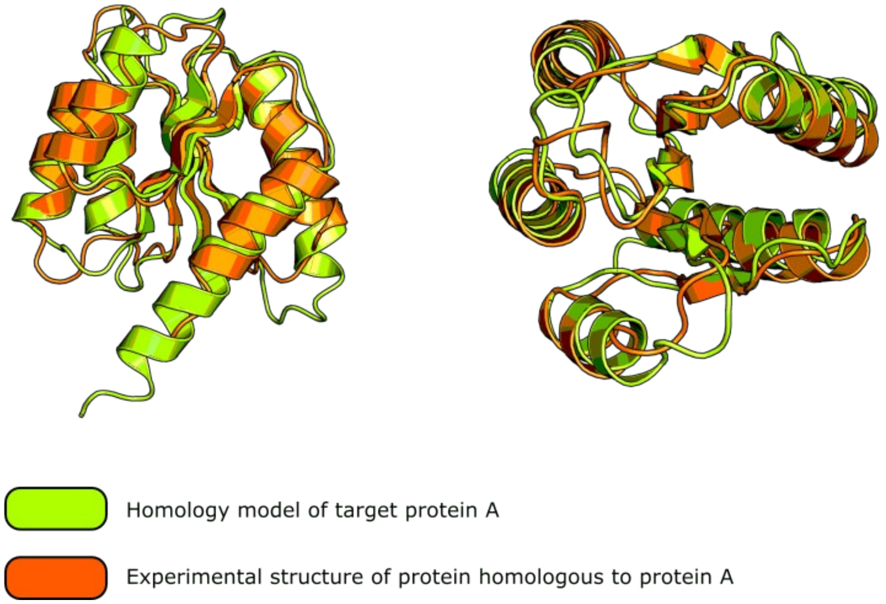
- Search homologous proteins (template): eg different species. The structure of templates are known
- Align the sequences to get information about:
  - Conserved secondary structures
  - Aminoacids that are mandatory to keep the function
  - ...

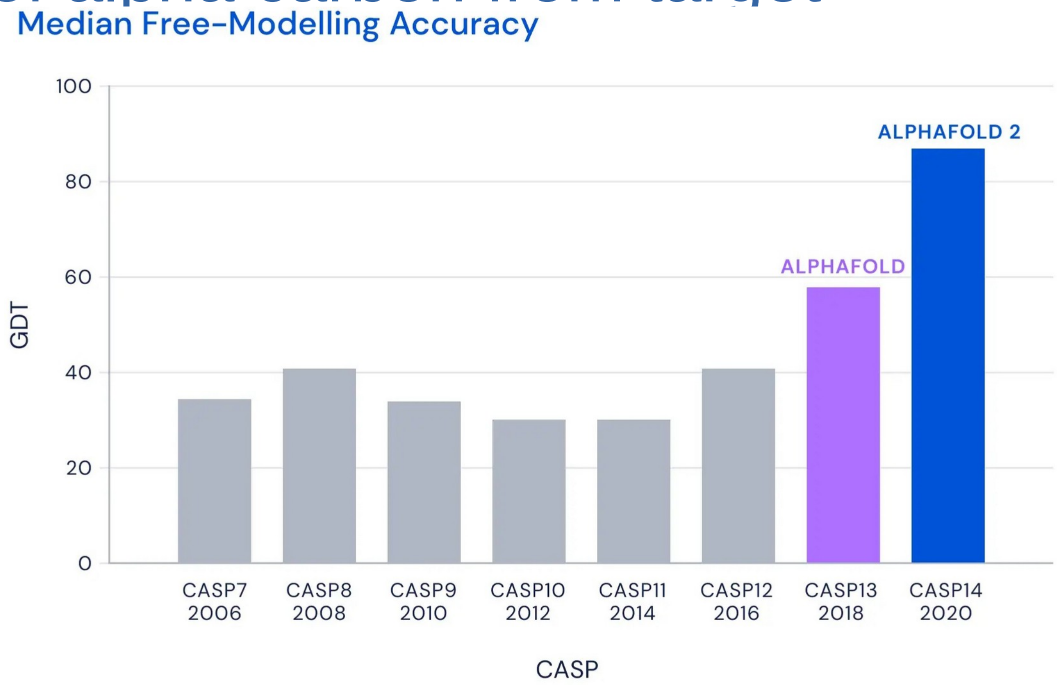
CASP competition

- Critical Assessment of protein Structure Prediction
- Every 2 years since 1994
- Unknown protein structures resolved experimentally then compared to numerical models

Performance evaluation

- Global Distance Test (GDT)
  - Derived from distance of alpha carbon from target


🟩 Homology model of target protein A
🟧 Experimental structure of protein homologous to protein A



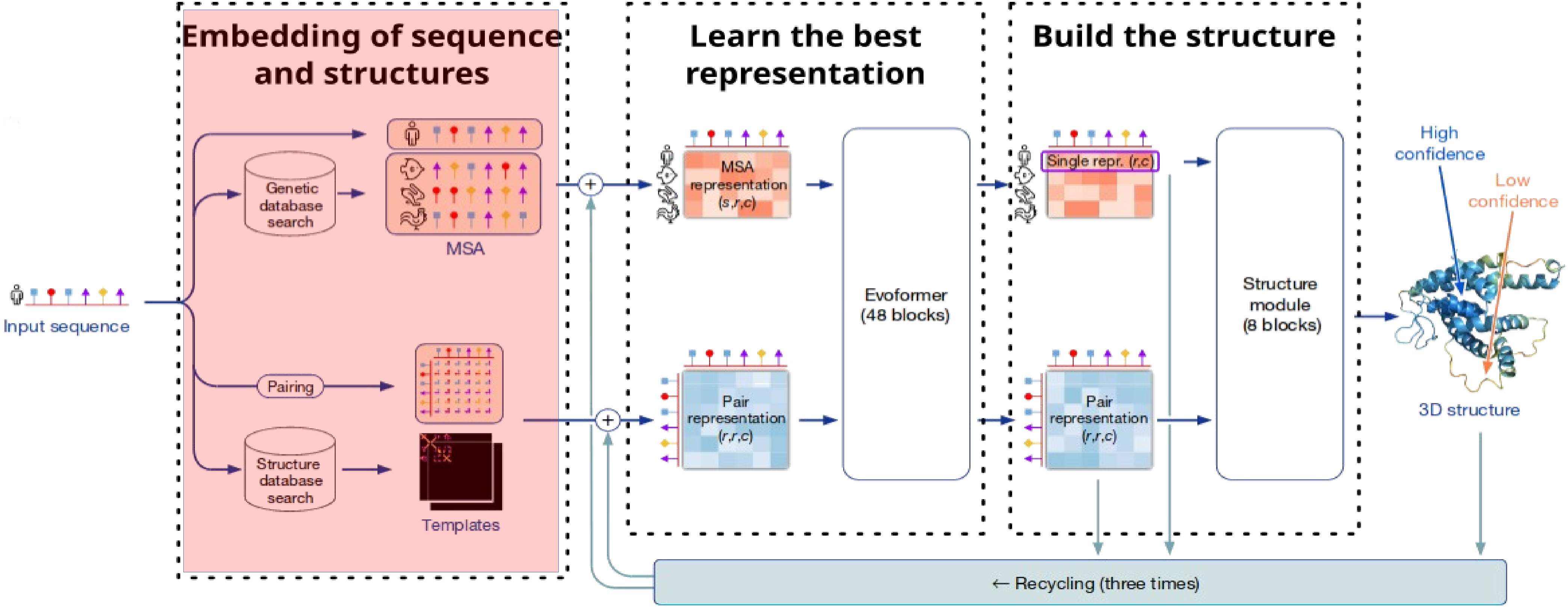
Alphafold[a] “hardware”

- Written with Tensorflow 2 + JAX
- Runs on GPU, TPU, CPU
- Depends on tools for sequence alignment
  - HH-suite (hhblits, hhsearch, ...)
  - hmmer-suite (jackhmmer)
- Part of the dataset was self-distilled[b] (noisy student)

[a] Jumper et al., “Highly accurate protein structure prediction with AlphaFold,” Nature, vol. 596, no. 7873, Art. no. 7873, 2021

[b] Xie et al. Self-training with noisy student improves imagenet classification. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10687–10698, 2020.

Alphafold model

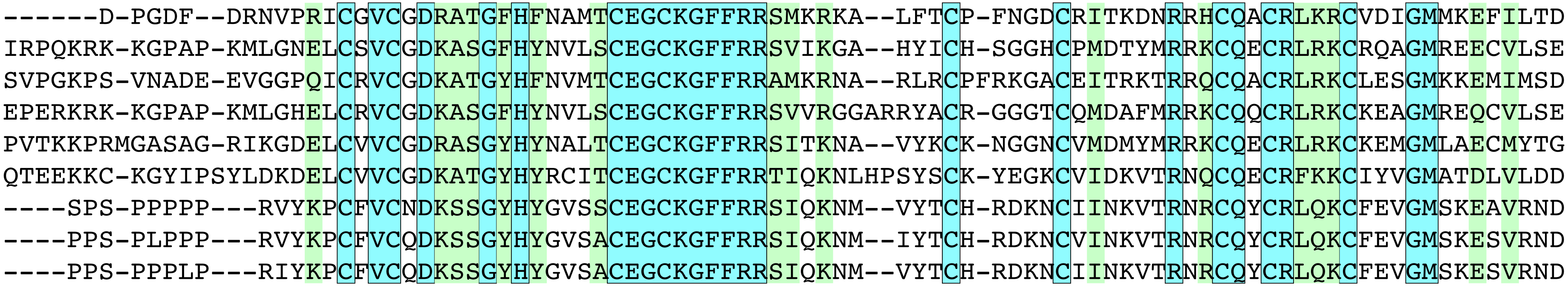

Embedding of sequence and structures
Learn the best representation
Build the structure
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

MSA representation

- Search genetic databases for sequences
  - MGnify (metagenomics)
  - UniRef90 (protein clusters from UniProt)
  - Uniclust30 + BFD (protein clusters from various databases by Soeding lab)
- Now database with around 214M proteins → 23TB and 600M files

Sequence Alignment

- Comparison of several related proteins (eg. different species)



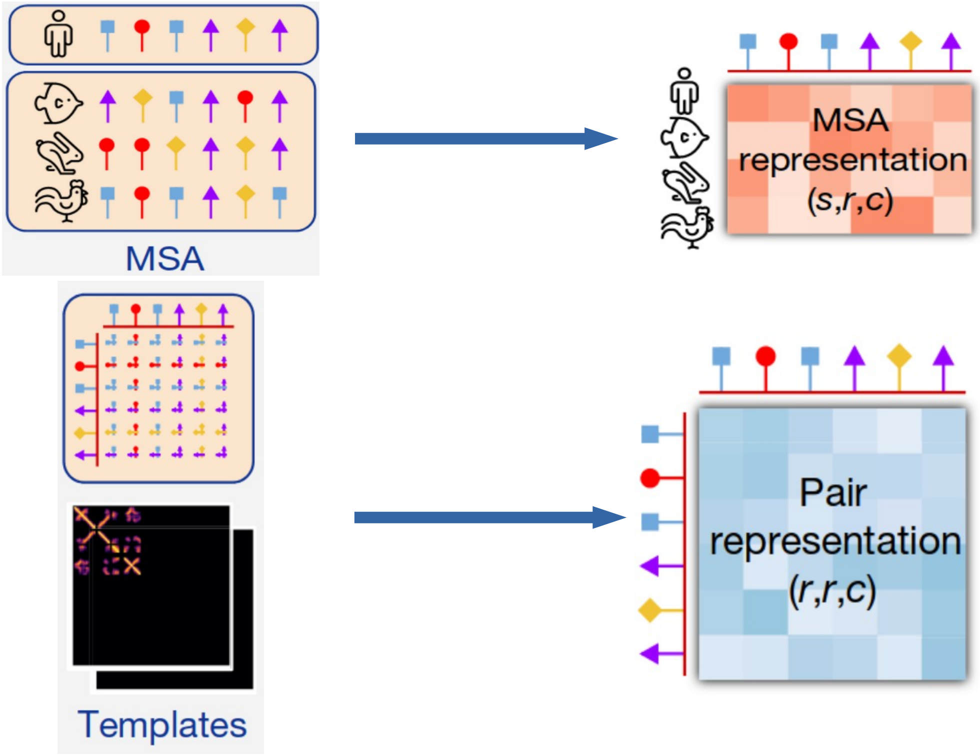
```
------D-PGDF--DRNVPRICGVCGDRATGFHFNAMTCEGCKGFFRRSMKRKA--LFTCP-FNGDCRITKDNRRHCQACRLKRCVDIGMMKEFILTD
IRPQKRK-KGPAP-KMLGNELCSVCGDKASGFHYNVLSCEGCKGFFRRSVIKGA--HYICH-SGGHCPMDTYMRRKCQECRLRKCRQAGMREECVLSE
SVPGKPS-VNADE-EVGGPQICRVCGDKATGYHFNVMTCEGCKGFFRRAMKRNA--RLRCPFRKGACEITRKTRRQCQACRLRKCLESGMKKEMIMSD
EPERKRK-KGPAP-KMLGHELCRVCGDKASGFHYNVLSCEGCKGFFRRSVVRGGARRYACR-GGGTCQMDAFMRRKCQQCRLRKCKEAGMREQCVLSE
PVTKKPRMGASAG-RIKGDELCVVCGDRASGYHYNALTCEGCKGFFRRSITKNA--VYKCK-NGGNCVMDMYMRRKCQECRLRKCKEMGMLAECMYTG
QTEEKKC-KGYIPSYLDKDELCVVCGDKATGYHYRCITCEGCKGFFRRTIQKNLHPSYSCK-YEGKCVIDKVTRNQCQECRFKKCIYVGMATDLVLDD
----SPS-PPPPP---RVYKPCFVCNDKSSGYHYGVSSCEGCKGFFRRSIQKNM--VYTCH-RDKNCIINKVTRNRCQYCRLQKCFEVGMSKEAVRND
----PPS-PLPPP---RVYKPCFVCQDKSSGYHYGVSACEGCKGFFRRSIQKNM--IYTCH-RDKNCVINKVTRNRCQYCRLQKCFEVGMSKESVRND
----PPS-PPPLP---RIYKPCFVCQDKSSGYHYGVSACEGCKGFFRRSIQKNM--VYTCH-RDKNCIINKVTRNRCQYCRLQKCFEVGMSKESVRND
```

Conserved	The sequence is identical
Semi-Cons.	Some mutations are possible

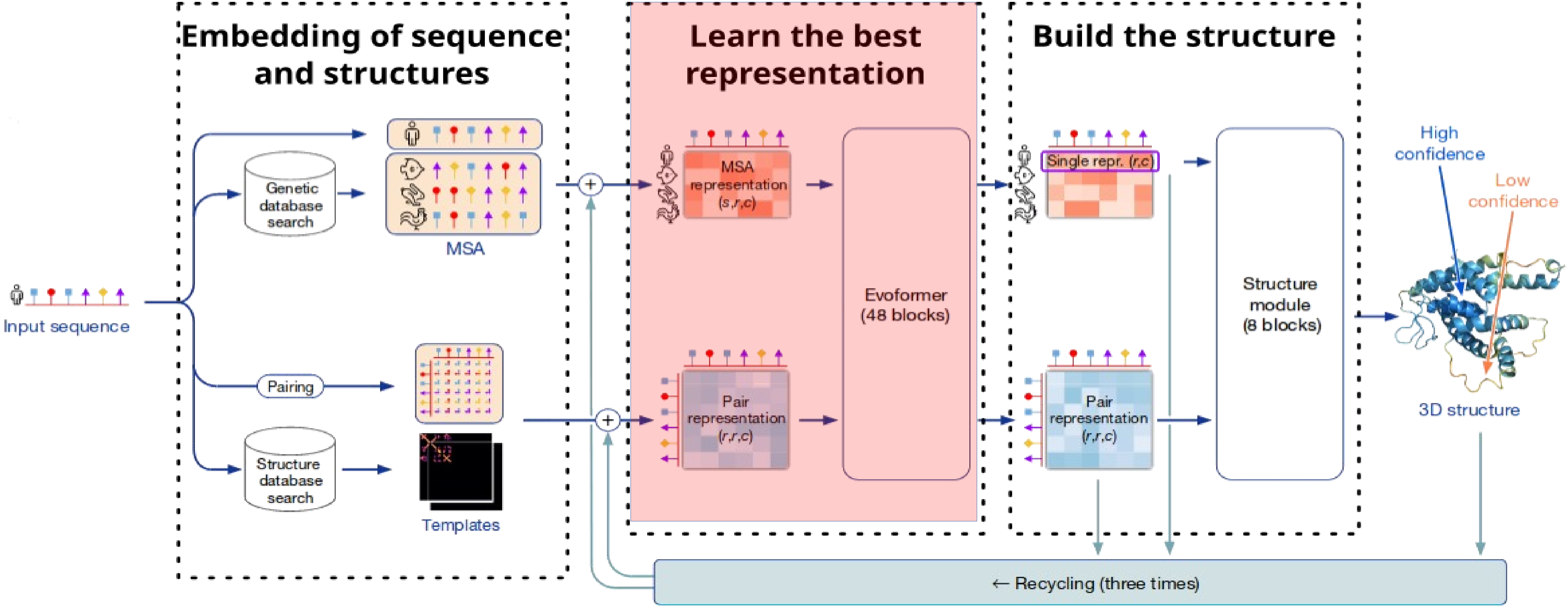
MSA: MultiSequence Alignement

Ingredients

- We need 2 ingredients
  - The similar sequences aligned with the input
  - Some structures close enough to serve as template
- Input for the Evoformer blocks

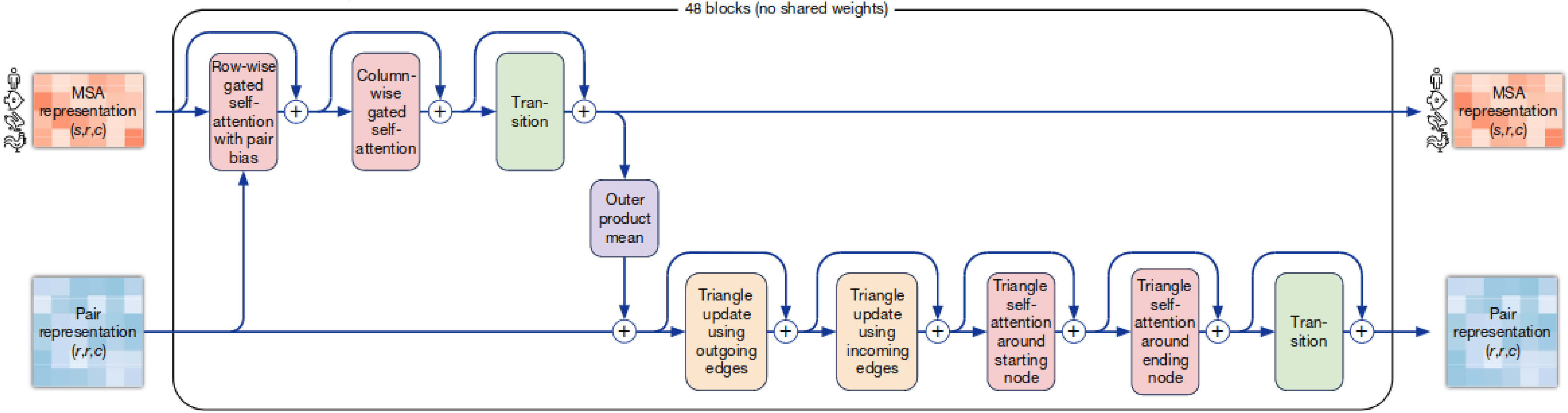


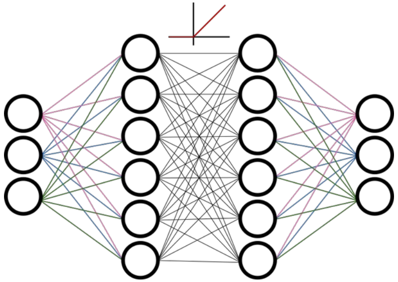
Alphafold model


Embedding of sequence and structures
Learn the best representation
Build the structure
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

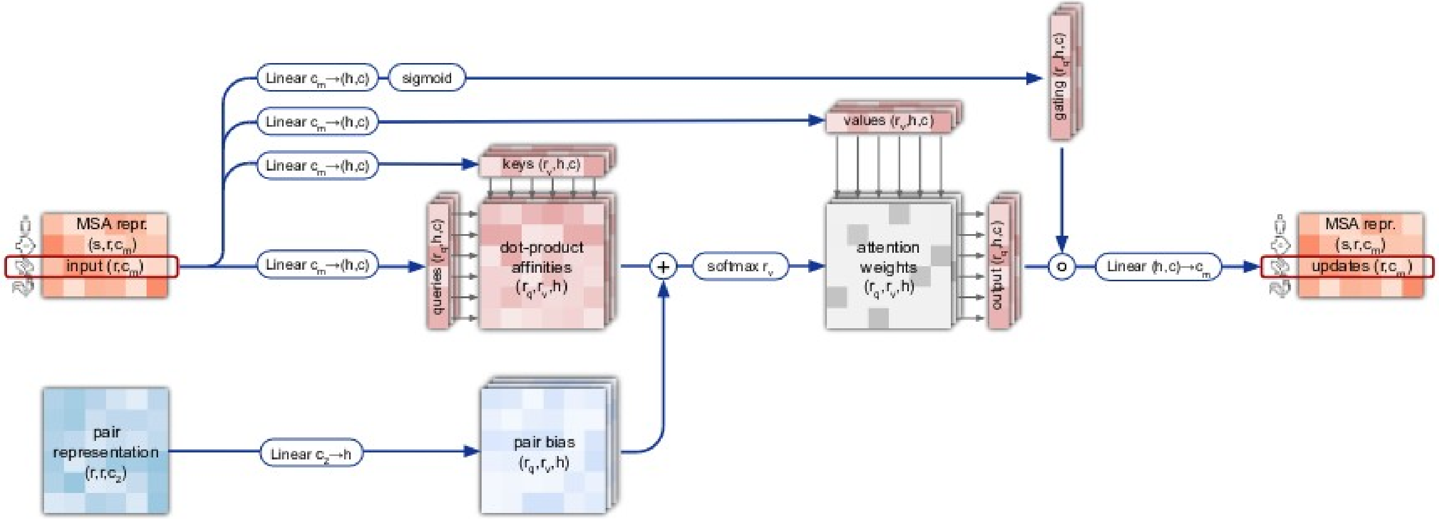
Evoformer

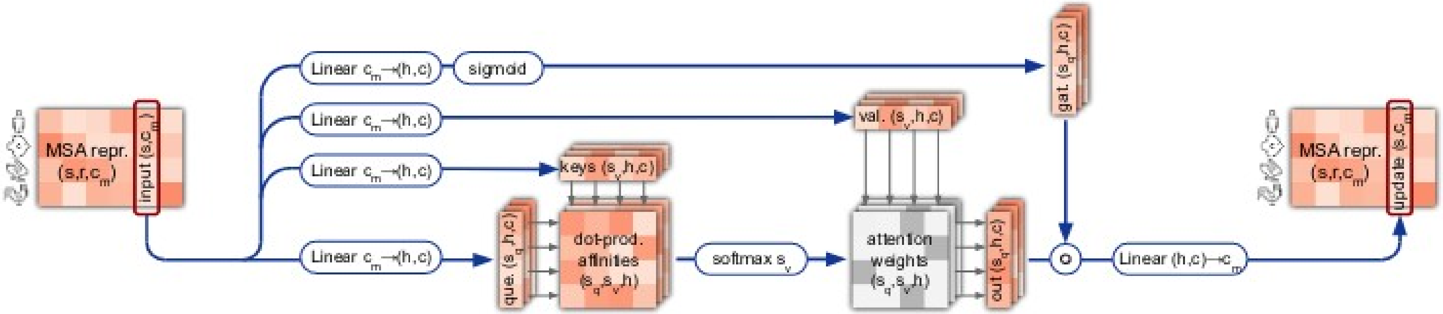
- Evolutionary Transformer


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)

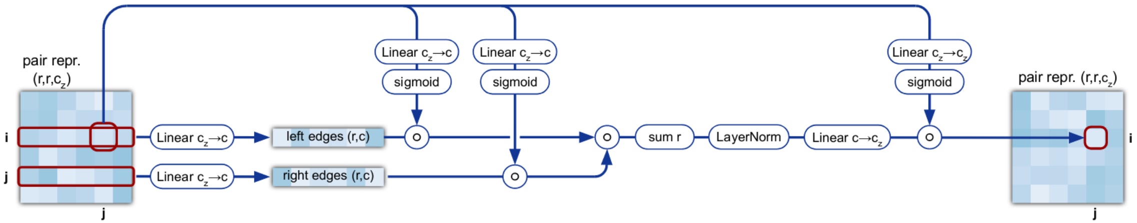
Transition = 

MSA attention


Linear $c_m \rightarrow (h,c)$
sigmoid
gating (r_q,h,c)
Linear $c_m \rightarrow (h,c)$
values (r_v,h,c)
Linear $c_m \rightarrow (h,c)$
keys (r_v,h,c)
MSA repr. (s,r,c_m)
input (r,c_m)
Linear $c_m \rightarrow (h,c)$
queries (r_q,h,c)
dot-product affinities (r_q,r_v,h)
+
softmax r_v
attention weights (r_q,r_v,h)
output (r_q,h,c)
o
Linear $(h,c) \rightarrow c_m$
MSA repr. (s,r,c_m)
updates (r,c_m)
pair representation (r,r,c_z)
Linear $c_z \rightarrow h$
pair bias (r_q,r_v,h)

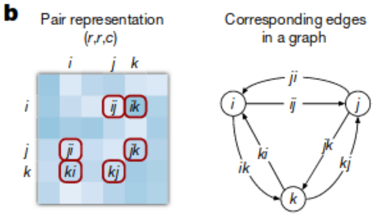

Linear $c_m \rightarrow (h,c)$
sigmoid
gat. (s_q,h,c)
Linear $c_m \rightarrow (h,c)$
val. (s_v,h,c)
Linear $c_m \rightarrow (h,c)$
keys (s_v,h,c)
MSA repr. (s,r,c_m)
input (s,c_m)
Linear $c_m \rightarrow (h,c)$
que. (s_q,h,c)
dot-prod. affinities (s_q,s_v,h)
softmax s_v
attention weights (s_q,s_v,h)
out (s_q,h,c)
o
Linear $(h,c) \rightarrow c_m$
MSA repr. (s,r,c_m)
update (s,c_m)

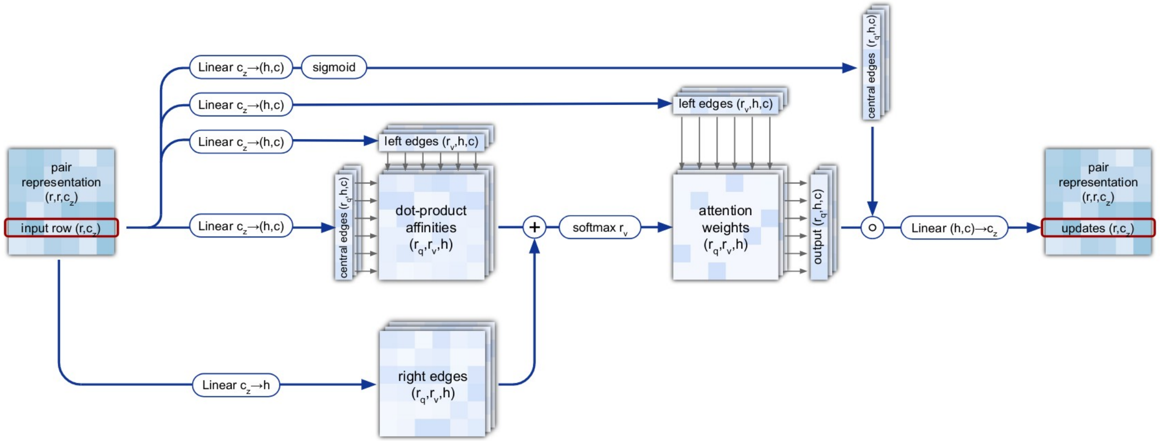
Pair representation update



Similar for incoming edge

Supplementary Figure 6 | Triangular multiplicative update using “outgoing” edges. Dimensions: r: residues, c: channels.

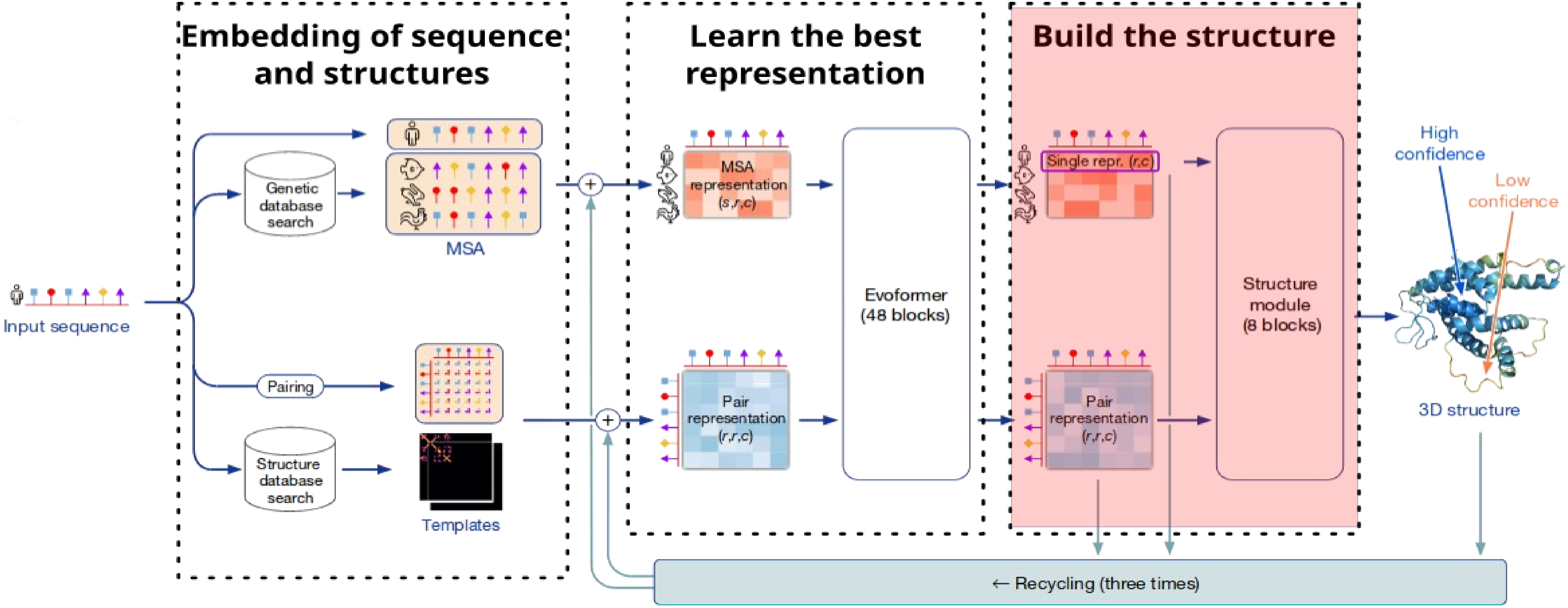




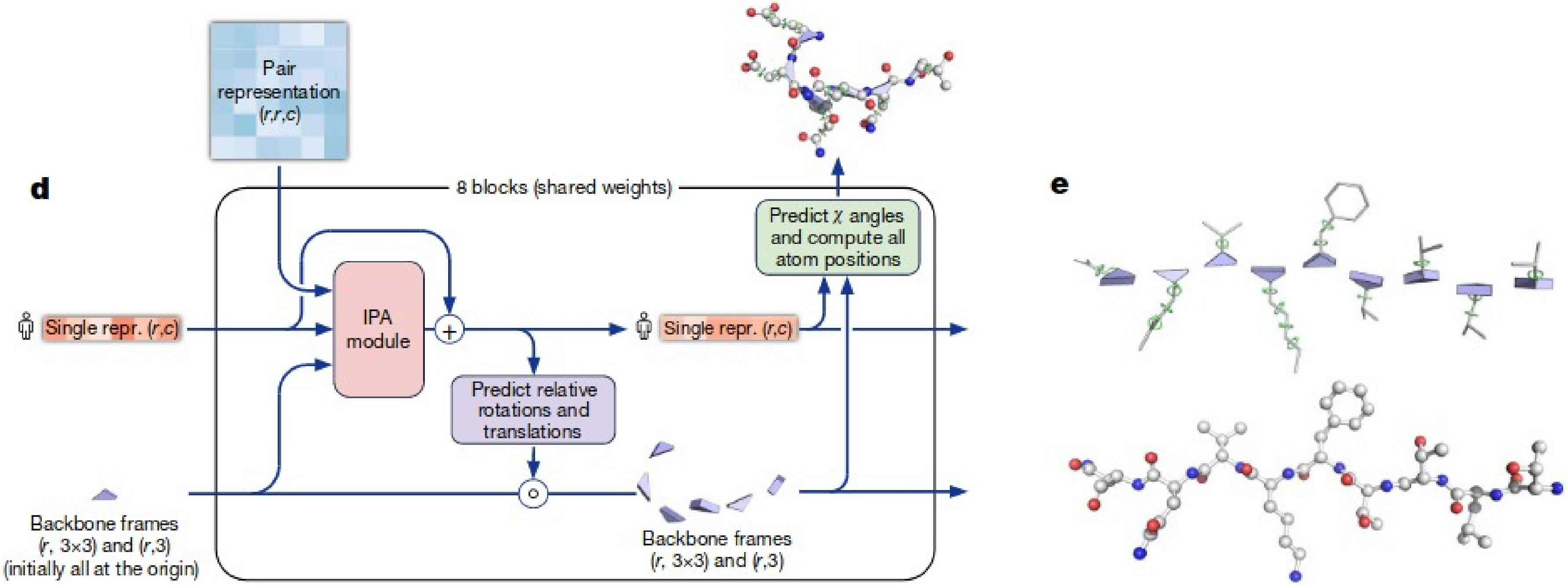
Similar for ending node

Supplementary Figure 7 | Triangular self-attention around starting node. Dimensions: r: residues, c: channels, h: heads

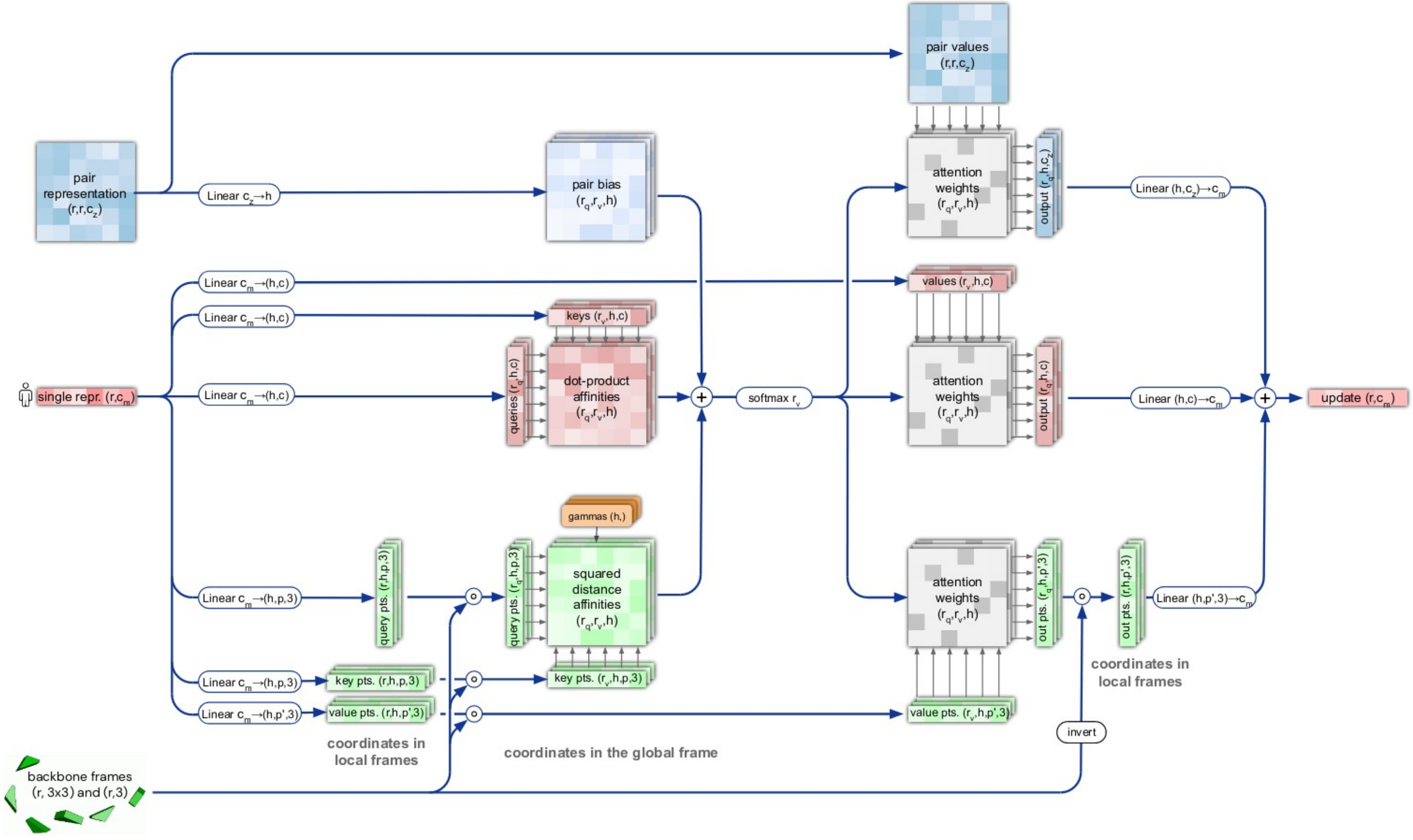
Alphafold model


Embedding of sequence and structures
Learn the best representation
Build the structure
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

Structure block

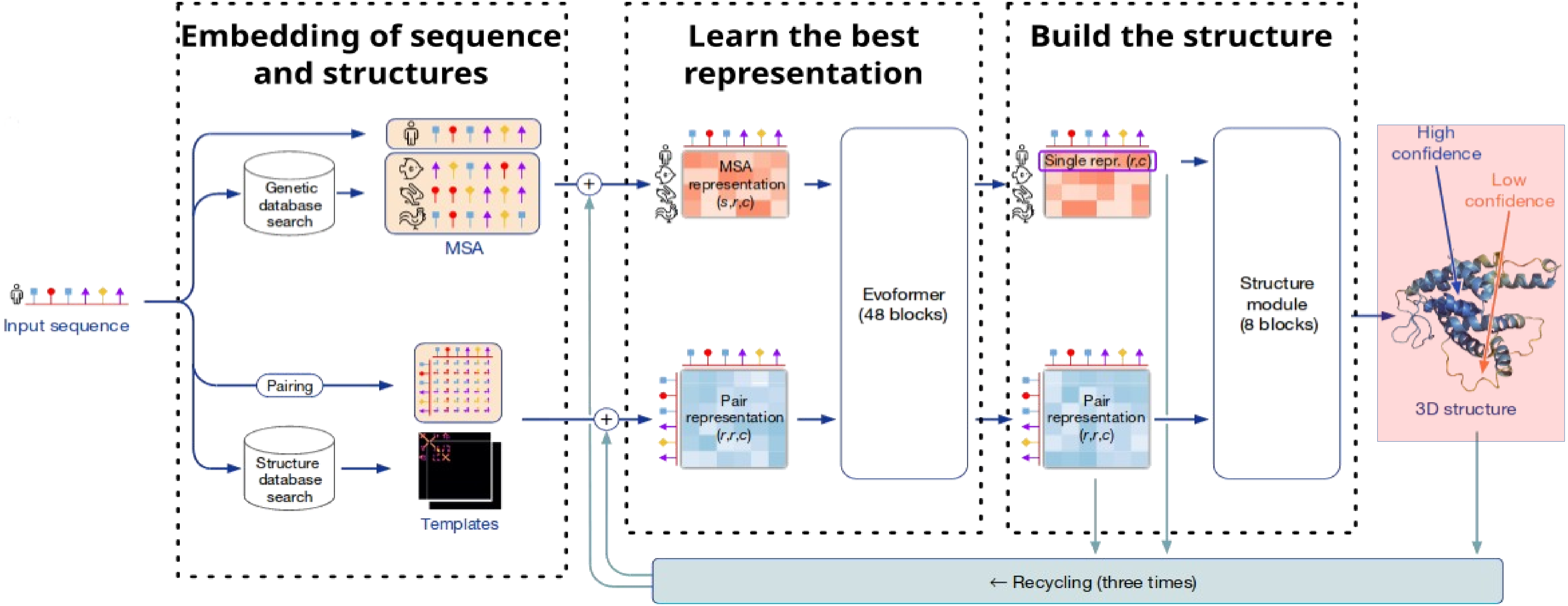

d
Pair representation (r,r,c)
8 blocks (shared weights)
Single repr. (r,c)
IPA module
+
Predict relative rotations and translations
∘
Single repr. (r,c)
Predict χ angles and compute all atom positions
Backbone frames (r, 3×3) and (r,3) (initially all at the origin)
Backbone frames (r, 3×3) and (r,3)
e

Invariant Point Attention



Institut du Développement et des Ressources en Informatique Scientifique

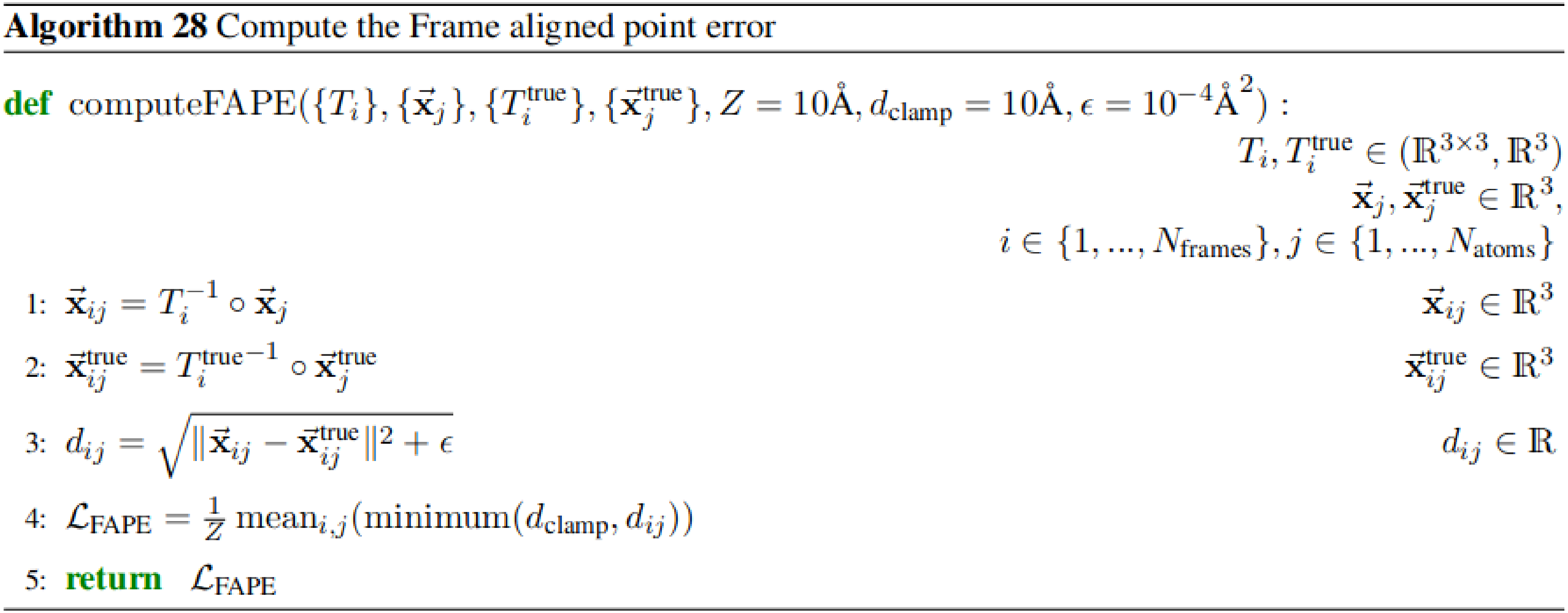
Alphafold model



Embedding of sequence and structures
Learn the best representation
Build the structure
Input sequence
Genetic database search
MSA
Pairing
Structure database search
Templates
MSA representation (s,r,c)
Pair representation (r,r,c)
Evoformer (48 blocks)
Single repr. (r,c)
Pair representation (r,r,c)
Structure module (8 blocks)
High confidence
Low confidence
3D structure
← Recycling (three times)

The loss function

- Introduced FAPE (Frame Aligned Point Error)

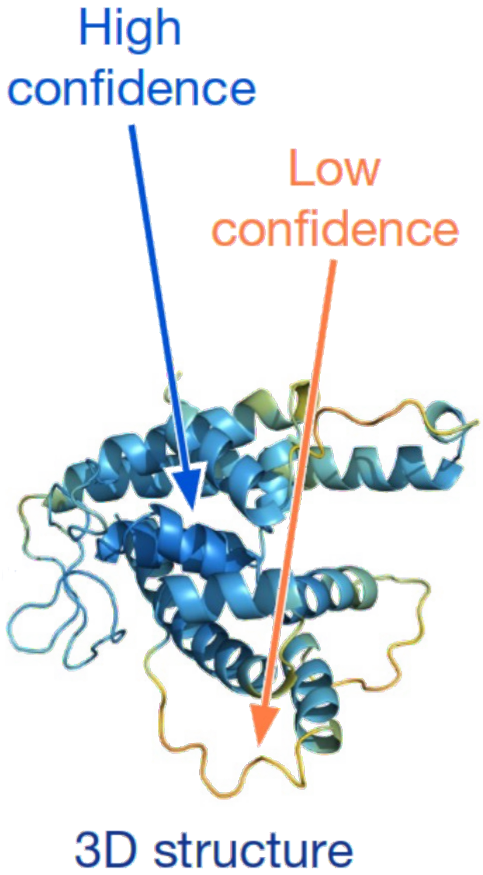


Algorithm 28 Compute the Frame aligned point error

```
def computeFAPE($\{T_i\}, \{\vec{\mathbf{x}}_j\}, \{T_i^{\text{true}}\}, \{\vec{\mathbf{x}}_j^{\text{true}}\}, Z = 10\text{Å}, d_{\text{clamp}} = 10\text{Å}, \epsilon = 10^{-4}\text{Å}^2$) :
                                                            $T_i, T_i^{\text{true}} \in (\mathbb{R}^{3\times3}, \mathbb{R}^3)$
                                                            $\vec{\mathbf{x}}_j, \vec{\mathbf{x}}_j^{\text{true}} \in \mathbb{R}^3,$
                                                            $i \in \{1, \ldots, N_{\text{frames}}\}, j \in \{1, \ldots, N_{\text{atoms}}\}$
    $\vec{\mathbf{x}}_{ij} = T_i^{-1} \circ \vec{\mathbf{x}}_j$                              $\vec{\mathbf{x}}_{ij} \in \mathbb{R}^3$
    $\vec{\mathbf{x}}_{ij}^{\text{true}} = T_i^{\text{true}-1} \circ \vec{\mathbf{x}}_j^{\text{true}}$                              $\vec{\mathbf{x}}_{ij}^{\text{true}} \in \mathbb{R}^3$
    $d_{ij} = \sqrt{\|\vec{\mathbf{x}}_{ij} - \vec{\mathbf{x}}_{ij}^{\text{true}}\|^2 + \epsilon}$                              $d_{ij} \in \mathbb{R}$
    $\mathcal{L}_{\text{FAPE}} = \frac{1}{Z}\, \text{mean}_{i,j}(\text{minimum}(d_{\text{clamp}}, d_{ij}))$
    return $\mathcal{L}_{\text{FAPE}}$
```

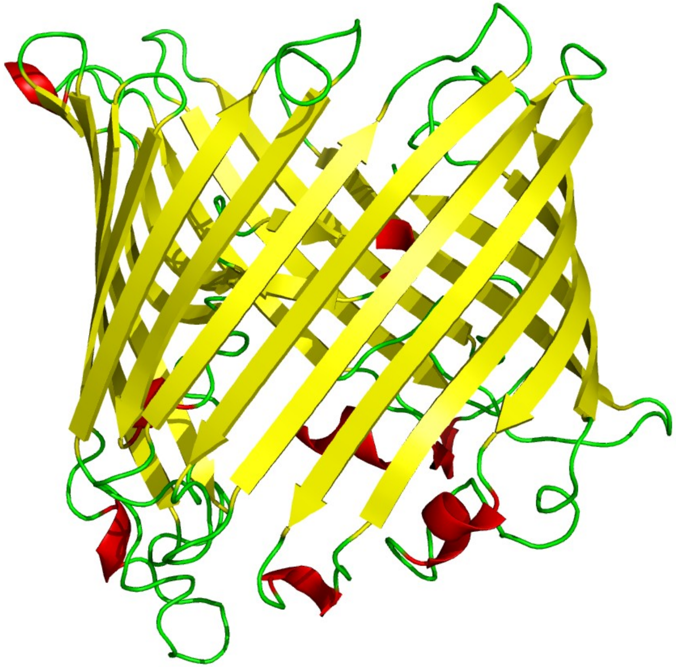
Inference: the structures

- At the end:
  - 3D structures
  - “confidence” score for each residue
- Refinement with parametrized physics software possible (OpenMM with Amber Force Field)

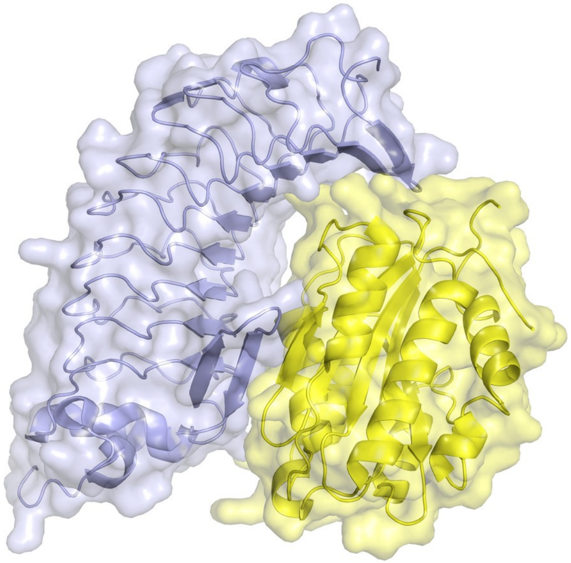


Features

- Monomer



- Multimer



Other software

- RoseTTA
- Openfold
- Colabfold (uses alphafold model but MSA is done with Mmseqs)
- OmegaFold (pytorch port of Alphafold)
- ESMFold (based on OpenFold but no MSA needed)
- ...

Pictures Attribution

(1) Thomas Shafee, CC BY 4.0 via Wikimedia Commons (secondary and tertiary structures of proteins)

(2) Muskid, CC BY-SA 3.0 via Wikimedia Commons (beta turn secondary structure)

(3) TungstenEinsteinium, CC BY-SA 4.0 via Wikimedia Commons (Table of amino acids)

(4) Simoncaulton, CC BY-SA 4.0 via Wikimedia Commons (Hetero dimer cloating factor)

(5) EMBL-EBI, CC BY 4.0 via Wikimedia Commons (tertiary structure FAM151A)

(6) Humphrey, W., Dalke, A. and Schulten, K., "VMD - Visual Molecular Dynamics", J. Molec. Graphics, 1996, vol. 14, pp. 33-38. (Homotrimer)

(7) VMD was developed by the Theoretical and Computational Biophysics Group in the Beckman Institute for Advanced Science and Technology at the University of Illinois at Urbana-Champaign.

(8) Jeff Dahl, CC BY-SA 3.0 via Wikimedia Commons (X-ray diffraction pattern)

(9) Loteralle, CC BY-SA 3.0 via Wikimedia Commons (NMR spectrum of calmodulin)

(10) Simon, Kailene & Pollock, Naomi & Lee, Sarah. (2018). Membrane protein nanoparticles: The shape of things to come. Biochemical Society Transactions. 46. BST20180139. 10.1042/BST20180139. (Cryoelectron microscopy of AcrB-SMALP)

(11) https://www.deepmind.com/blog/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology (GDT CASP)

(12) https://bitesizebio.com/38005/homology-modeling-proteins/ (Thomas Warwick, homology of LytR)

(13) Opabinia regalis - Self-created from PDB ID 1A0S using PyMol, CC BY-SA 3.0 (Sucrose Porin)