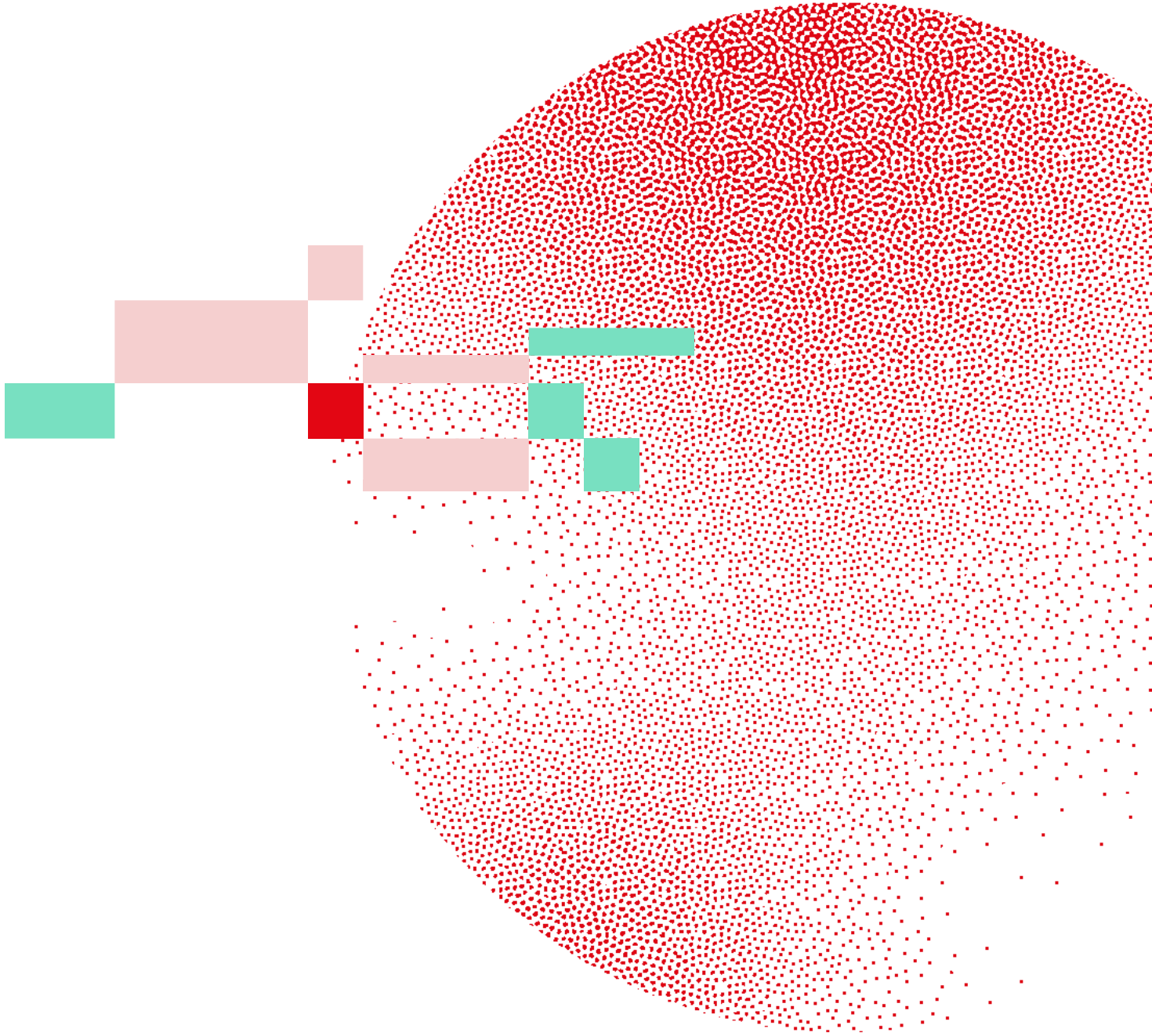


Swiss Institute of Bioinformatics

DAY 1, PART 1

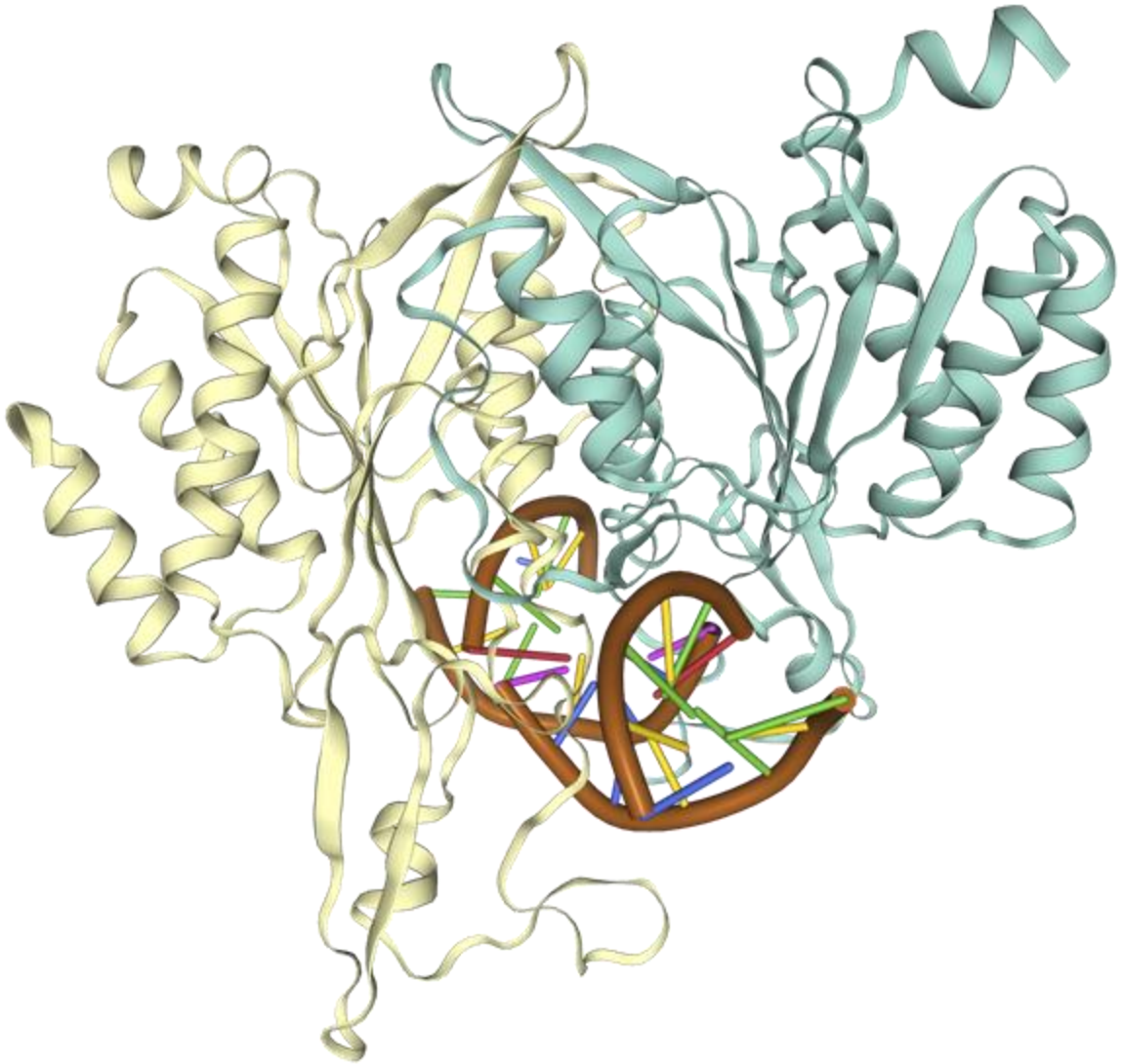
Proteins

Diana Rapota, Rok Breznikar, Janani Durairaj

23-24 June 2026

elixir SWITZERLAND

Proteins drive biomolecular functions

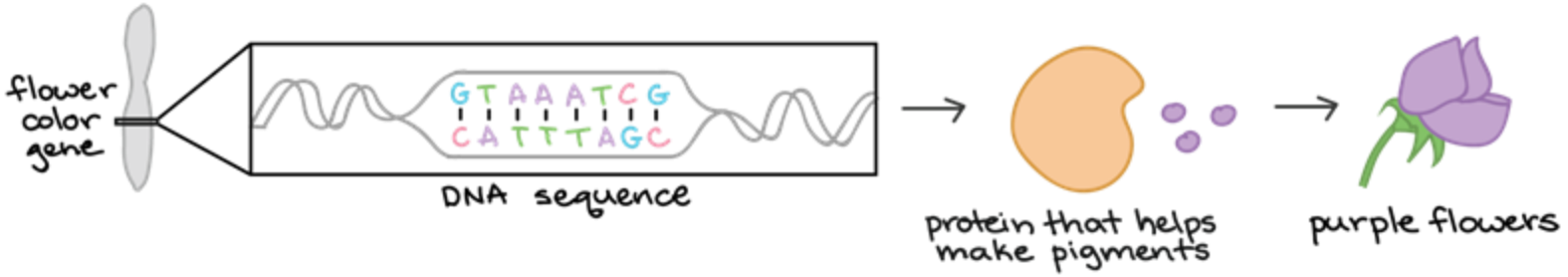


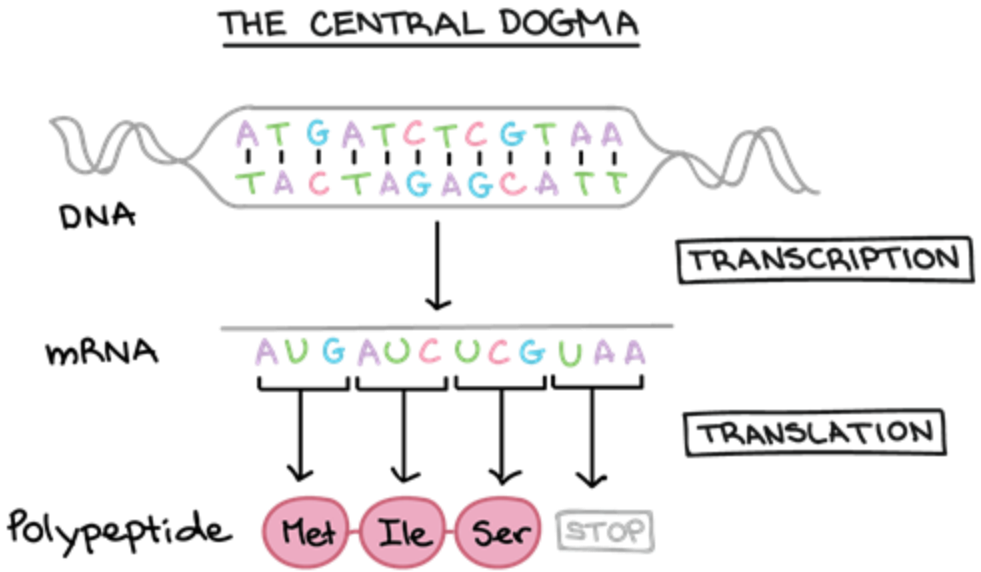
- Cellular processes
- Metabolism
- DNA replication/modification
- Transcription/translation
- Intracellular signalling
- Cell-cell communication
- Protein folding/degradation
- Transport
- Multifunctional proteins
- Defence and immunity
- Miscellaneous functions

Structure of Type-2 restriction enzyme EcoRI, PDB ID 1CL8, source: SWISS-MODEL Repository

Subcategories of the category "Molecular Function" used in the Gene Ontology Project (Ashburner et al. Nature, 2000)

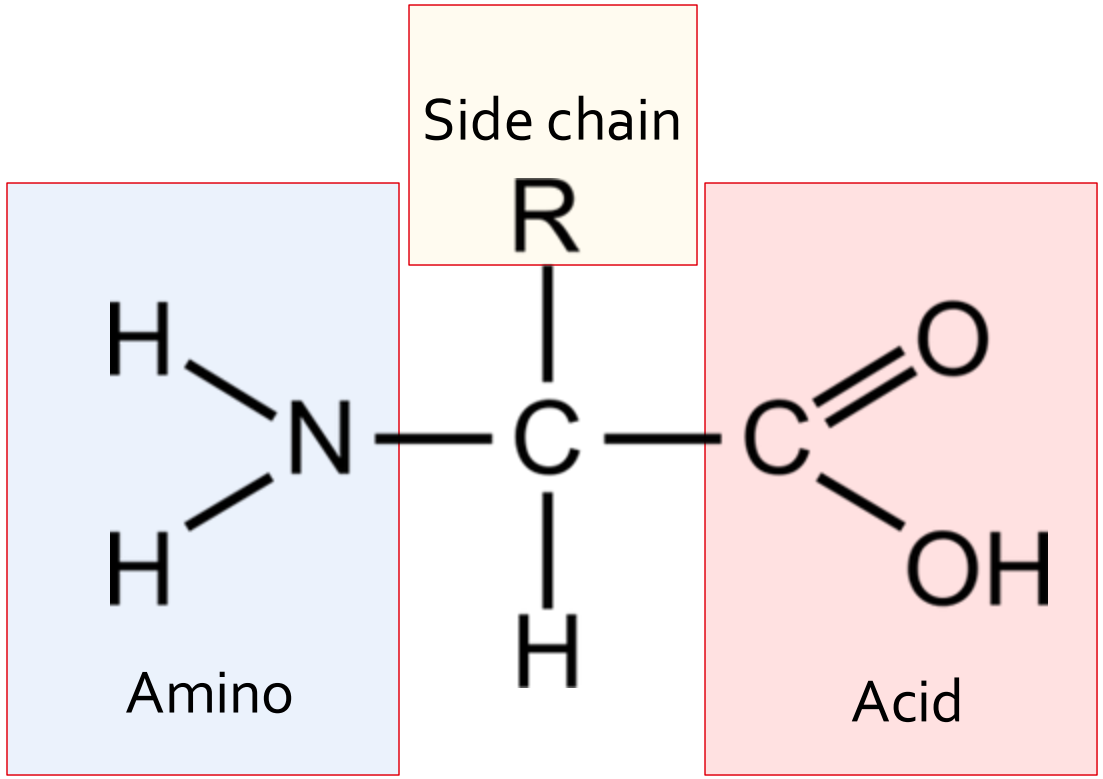
Where do proteins come from: The biological dogma

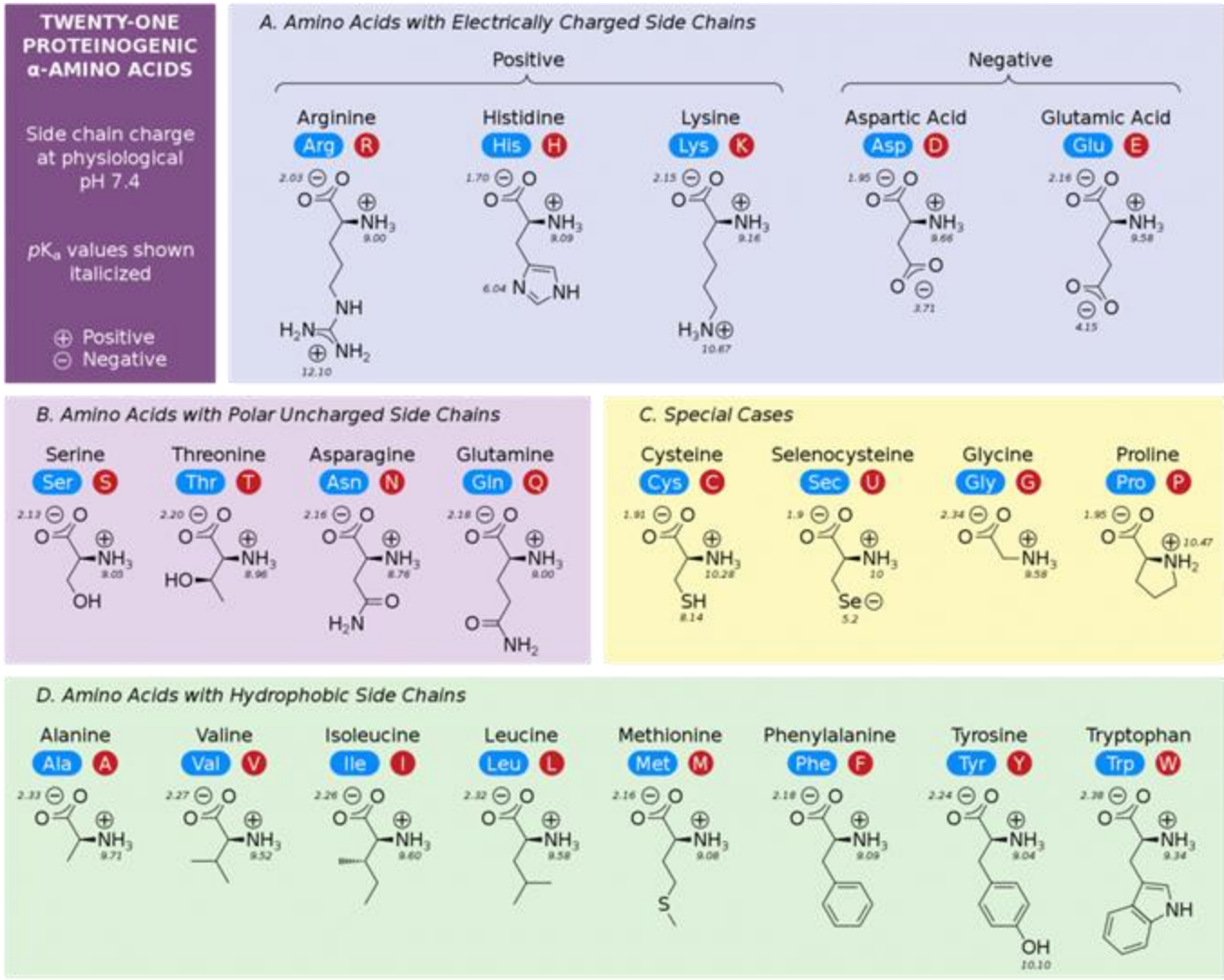




https://khanacademy.org

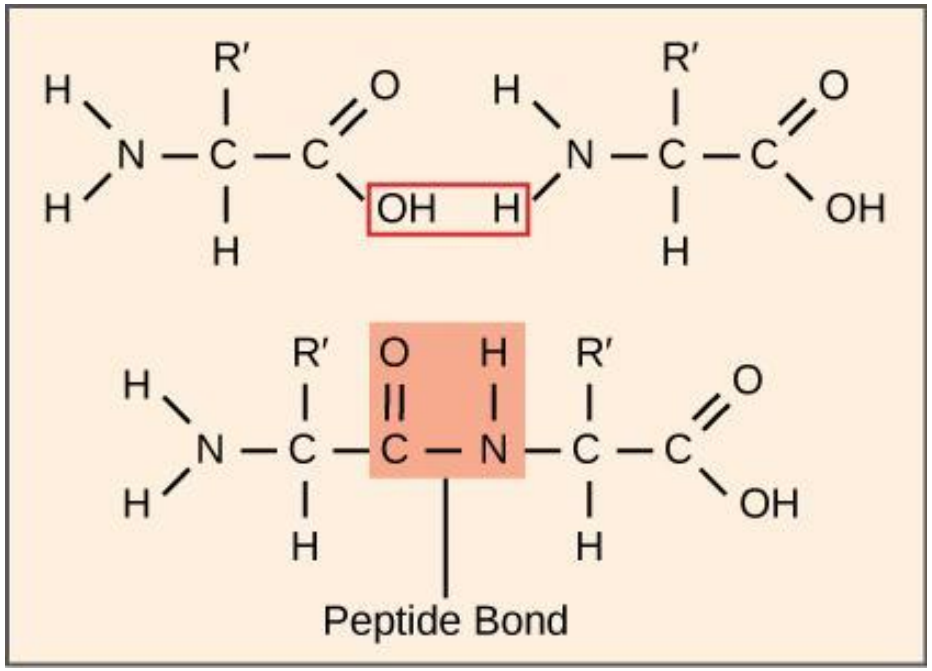
The chemical basis of proteins: Amino acids

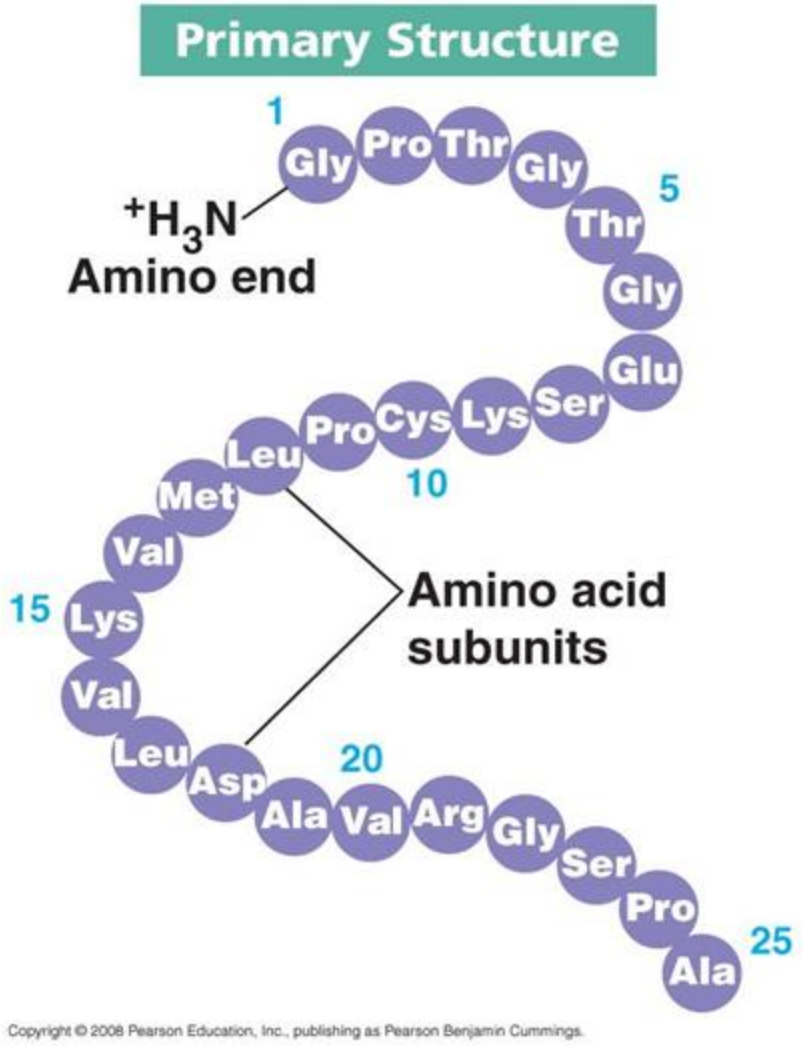




https://chemistrytalk.org/wp-content/uploads/2023/03/ProteinogenicAminoAcids.svg-1024x819-1.png

Primary structure – sequence of amino acids

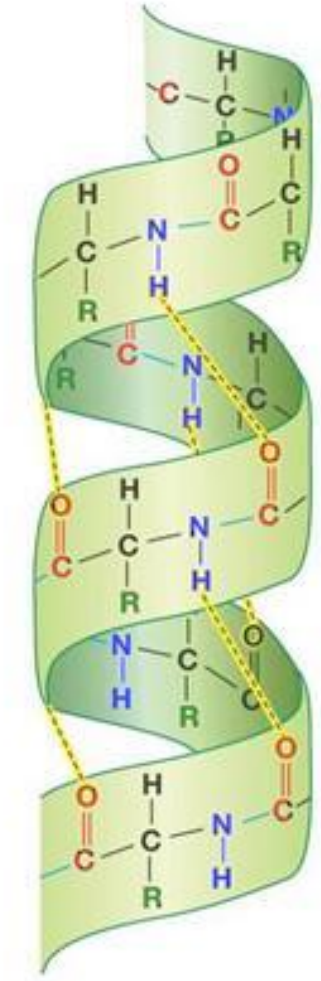




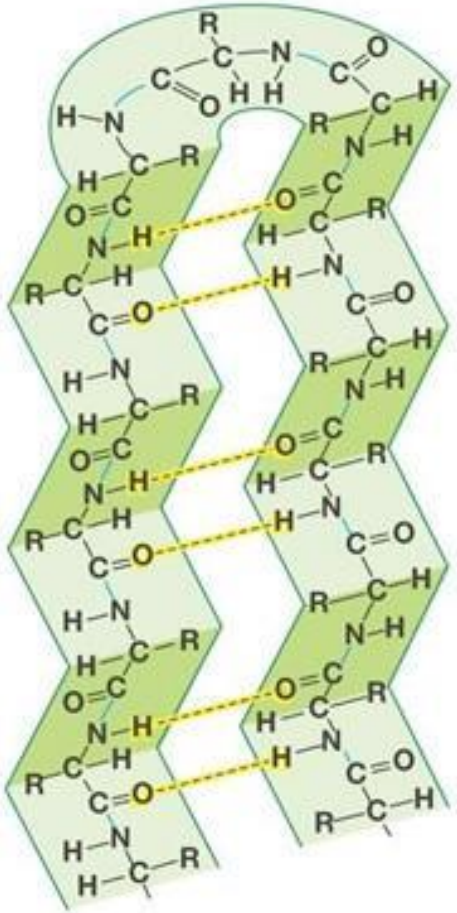
https://biologydictionary.net/peptide-bond/

Secondary structure – 3D local, folded arrangement

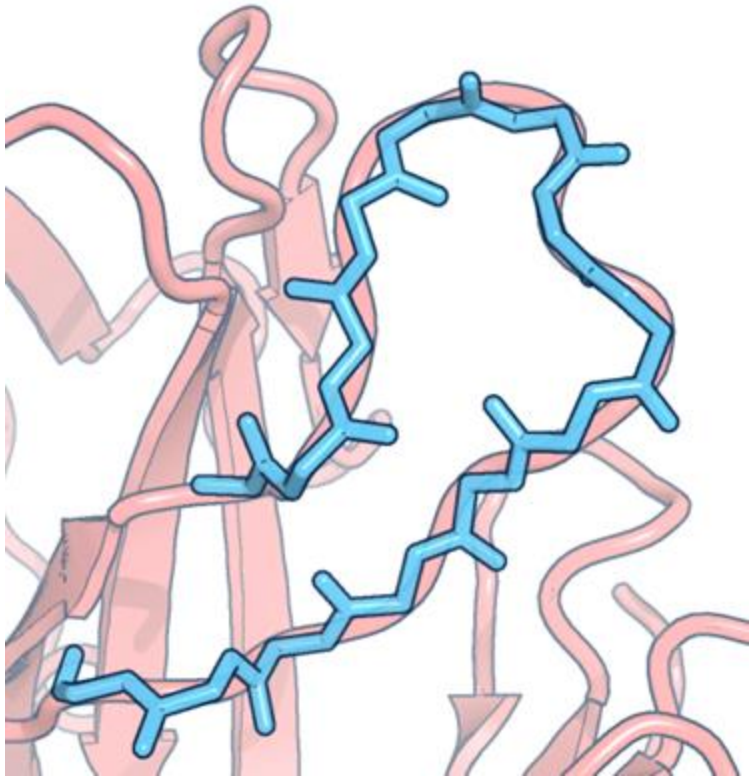
Alpha helix



Beta sheet



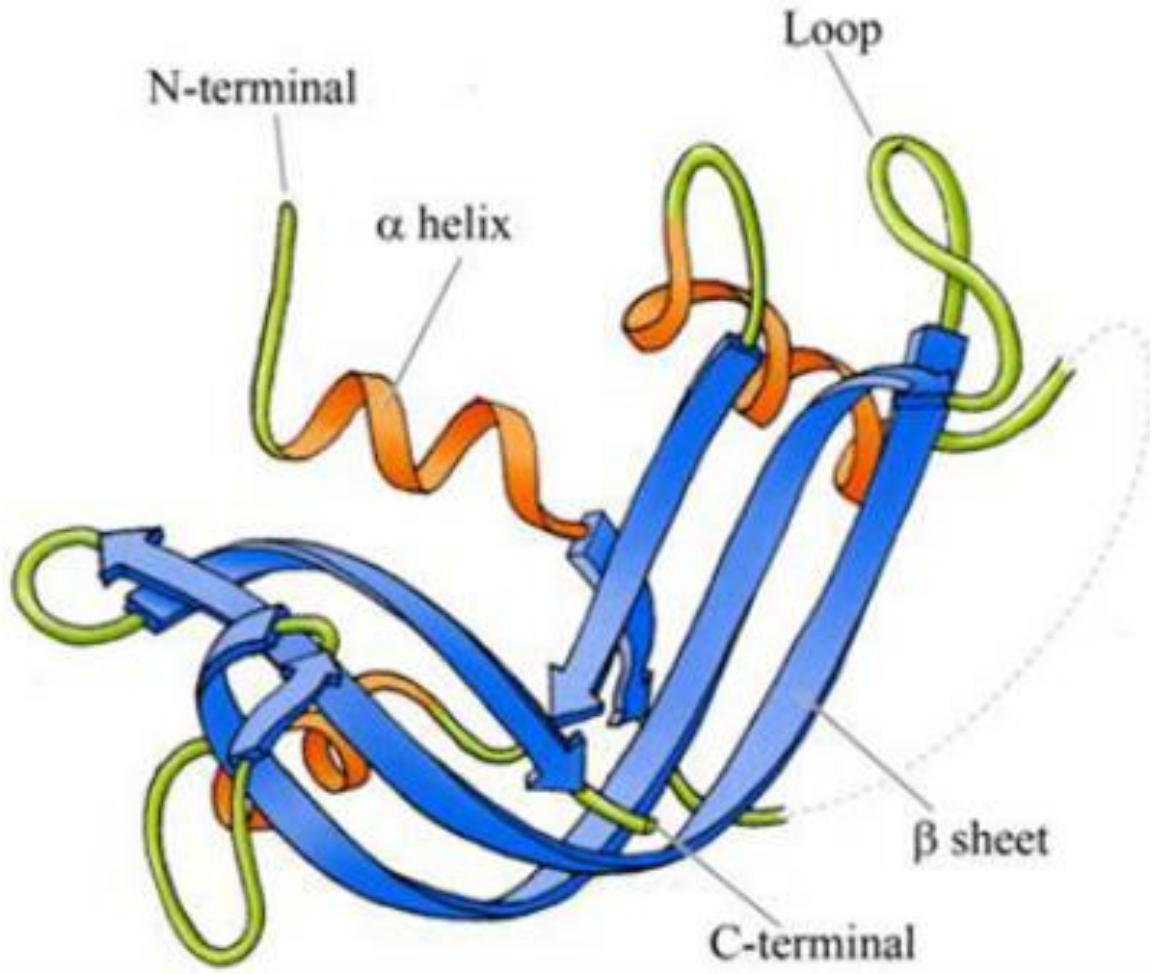
Loop



http://cbm.msoe.edu/teachingResources/proteinStructure/secondary.html

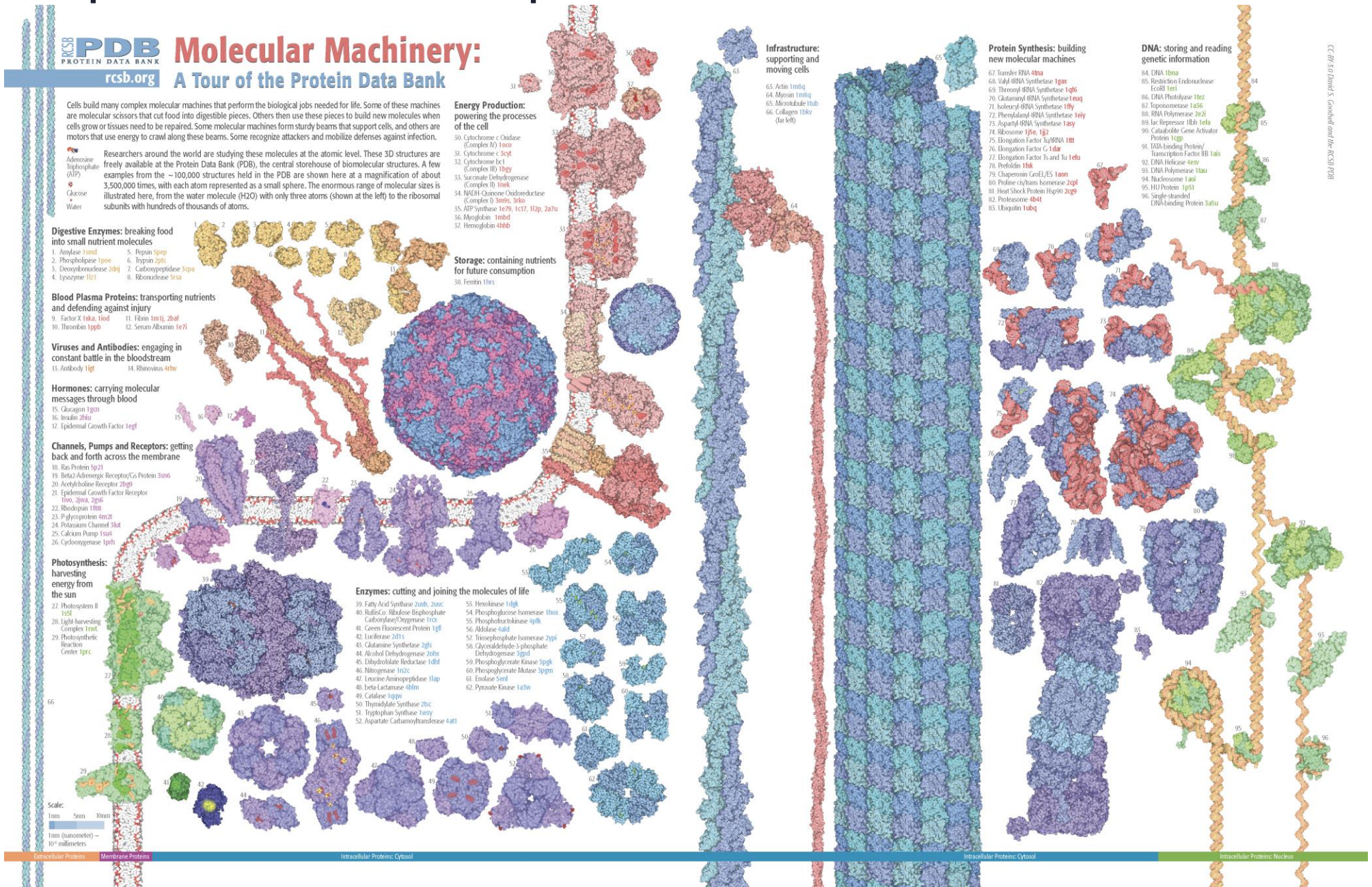
http://opig.stats.ox.ac.uk/webapps/newsabdab/sabpred/fread/

Tertiary structure – 3D shape of one protein chain



Schematic representation of the tertiary structure of a protein with alpha helices, beta sheets and loops,
Figure 8, Lara et al. TripleC, 2009

Multiple different shapes and sizes → function



Sequence determines the structure

Anfinsen's dogma - 3D structure of (most) natural proteins is determined only by its amino acid sequence.

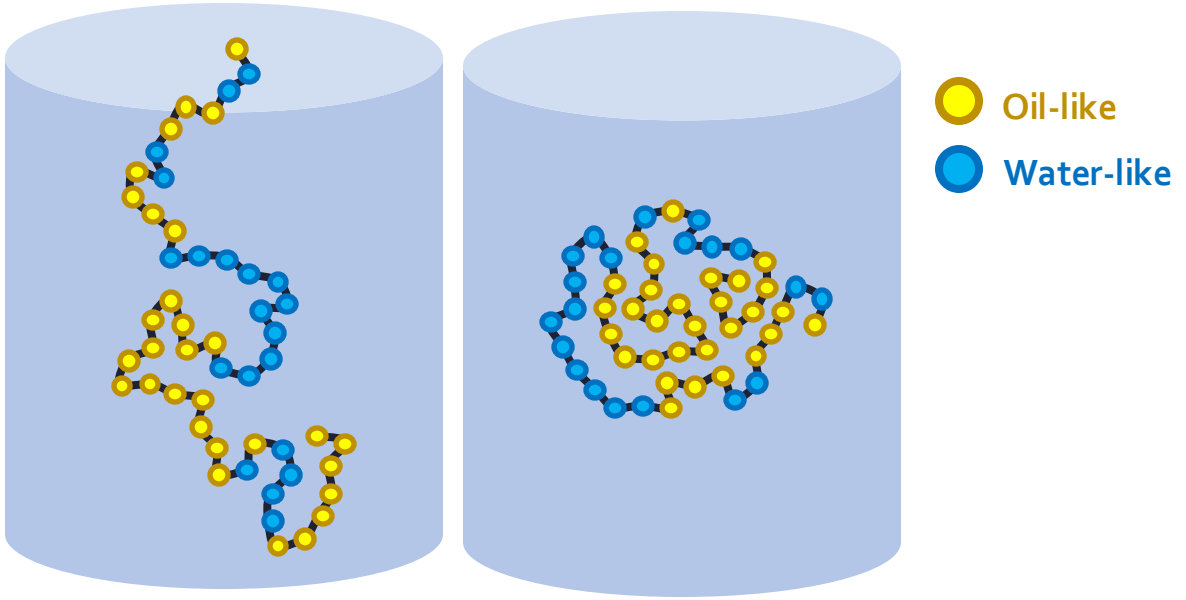
Amino acid sequence (in a given environment)

↓

Interatomic interactions

↓

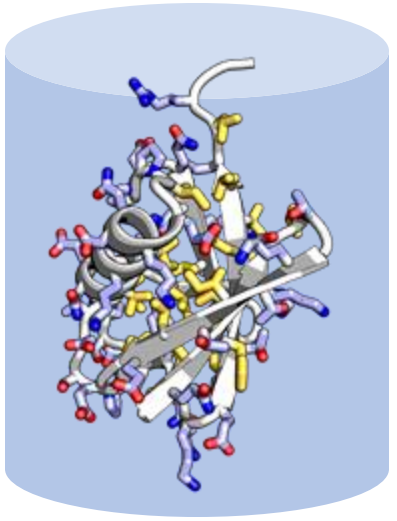
Native conformation (3D structure)



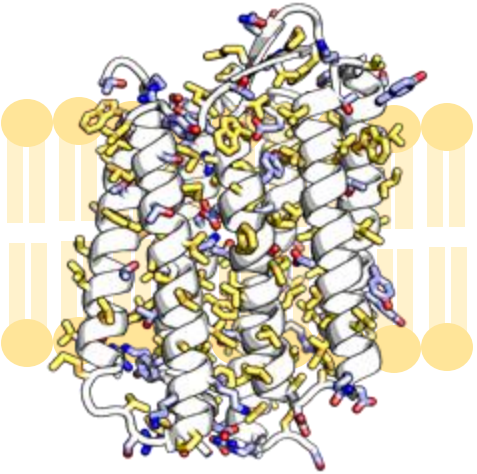
Slide from Sergey Ovchinnikov

Sequence determines the structure

- Hydrophobicity

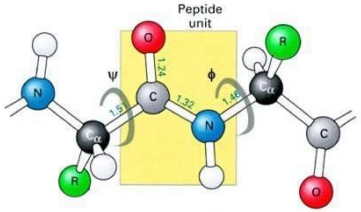


Soluble protein

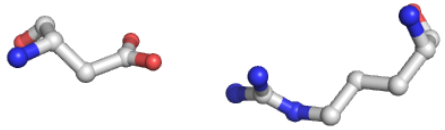


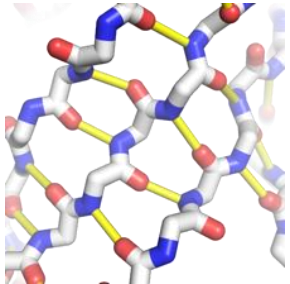
Transmembrane protein

- Sidechain and backbone conformations



- Bonds (hydrogen bonds, ionic bonds, …)

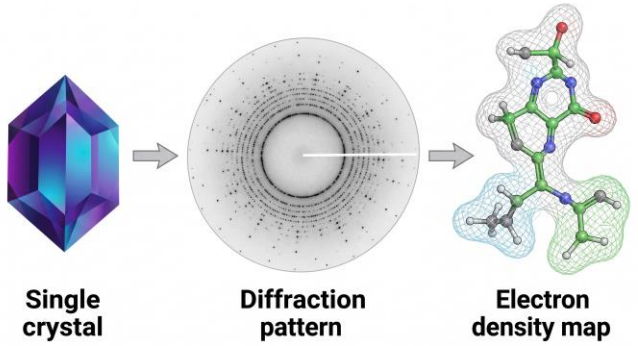




Slide from Sergey Ovchinnikov

Experimental methods to determine structures

X-ray crystallography



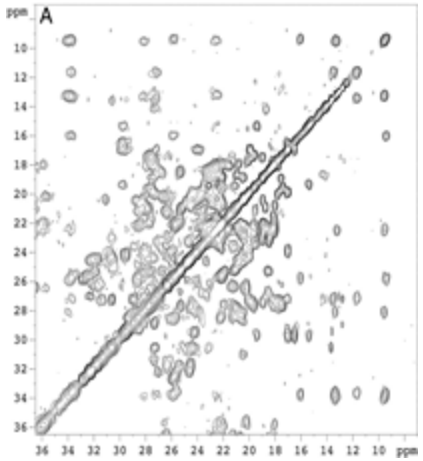
high resolution

not limited by protein size

sample must be crystallizable

protein is in the crystal structure conformation

Nuclear magnetic resonance (NMR) spectroscopy

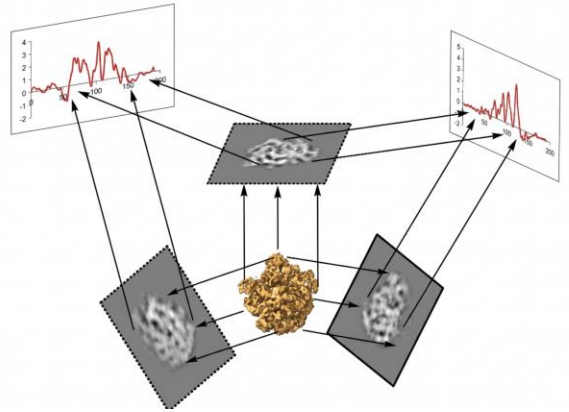


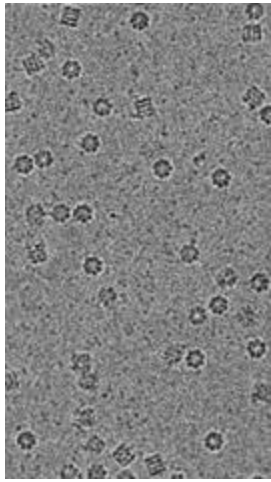
many conformations (measured in solution)

large amount of protein needed

only small proteins

Cryogenic electron microscopy (cryo-EM)





no need to crystallize

small number of samples

small proteins are difficult

flexible proteins are difficult

https://www.creative-biostructure.com/comparison-of-crystallography-nmr-and-em_6.htm, NMR profile of solid apoferritin, Figure 3, modified, Bertini et al., PNAS, 2011, https://casegroup.rutgers.edu/lnotes/cryoEM.pdf, CryoTEM image of GroEL suspended in amorphous ice at 50000× magnification

Sequence-structure gap

Billions of sequences

- Modern sequencing methods

~200k structures

- Time and labour consuming

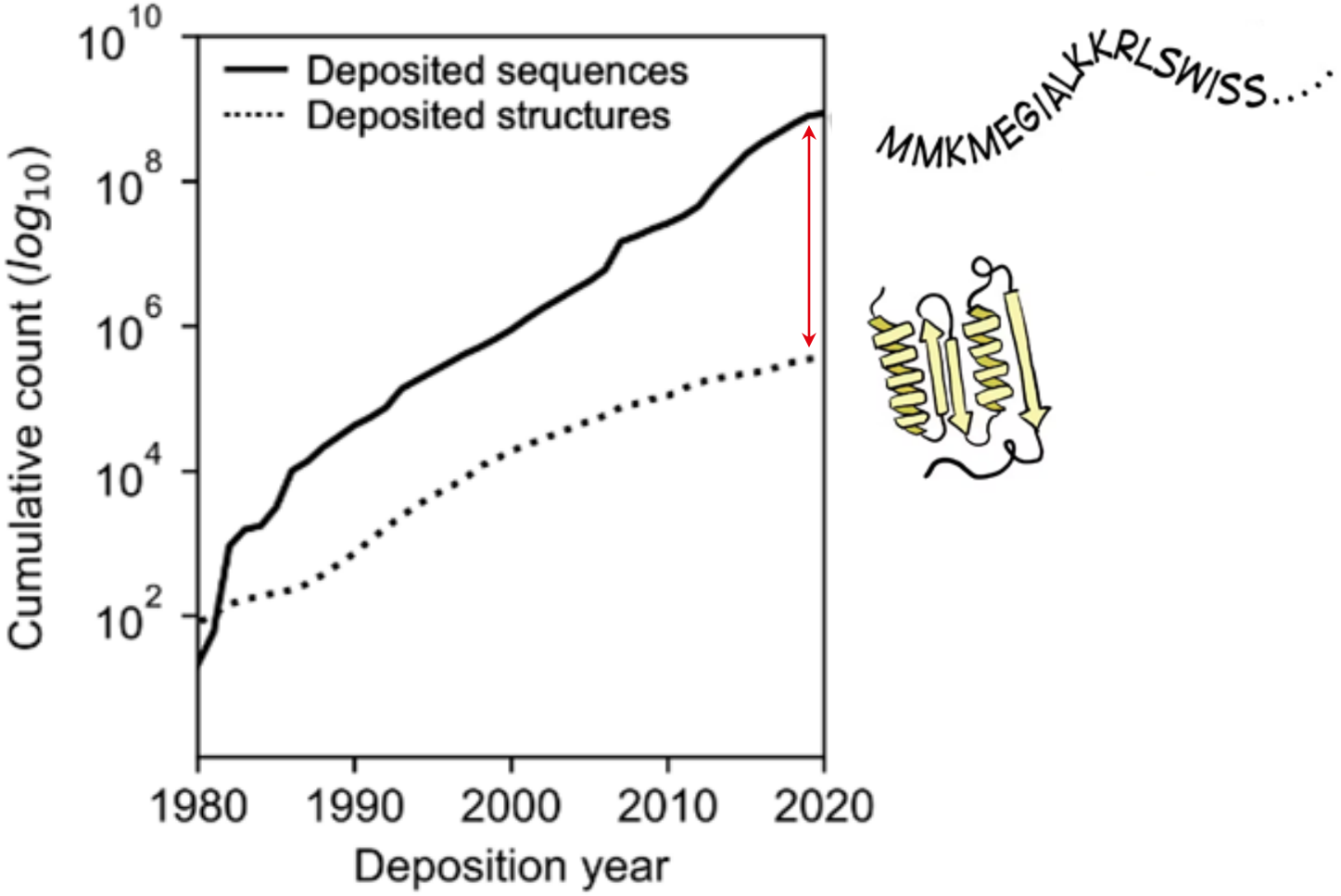
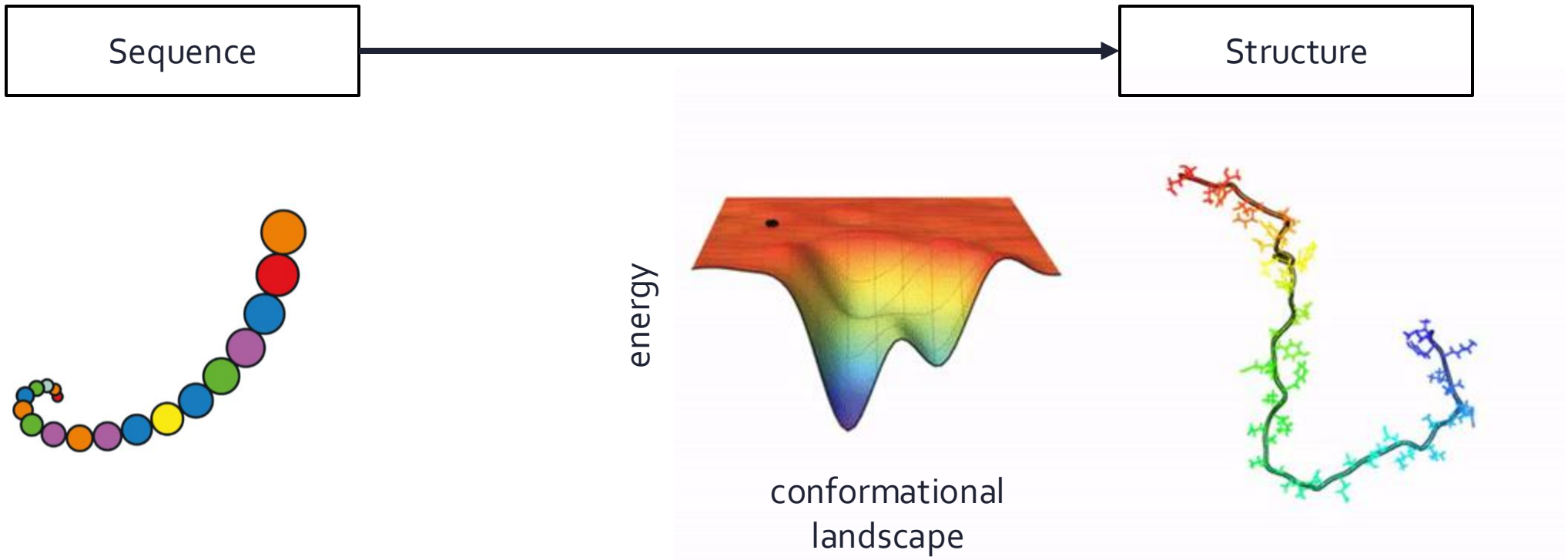


Image adapted from PTNG blog post

The protein folding problem

For a given sequence, find the structure with the lowest free energy



[Video credit: C. Fennell]

Dill, K.A. and MacCallum, J.L., 2012. The protein-folding problem, 50 years on. *science*, *338*(6110), pp.1042-1046.

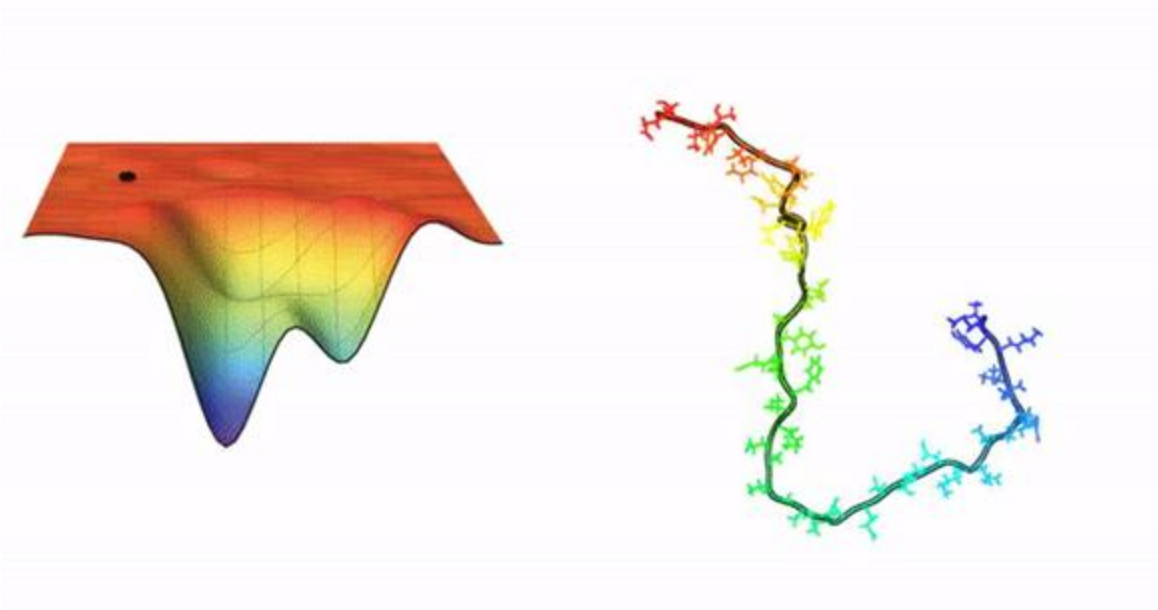
Slide from Sergey Ovchinnikov

The protein folding problem

Levinthal's paradox

- the number of potential configurations a protein could adopt is astronomical
  - searching through all should take longer than the age of the universe

- yet proteins fold rapidly, suggesting efficient folding pathways

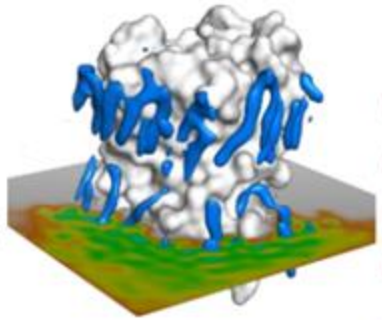


Through the ages

1970s

Template-based Modelling (TBM)

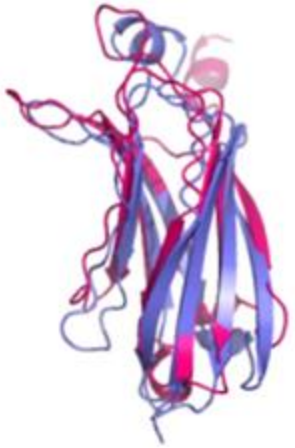
Utilise sequence alignments to "copy" similar residues

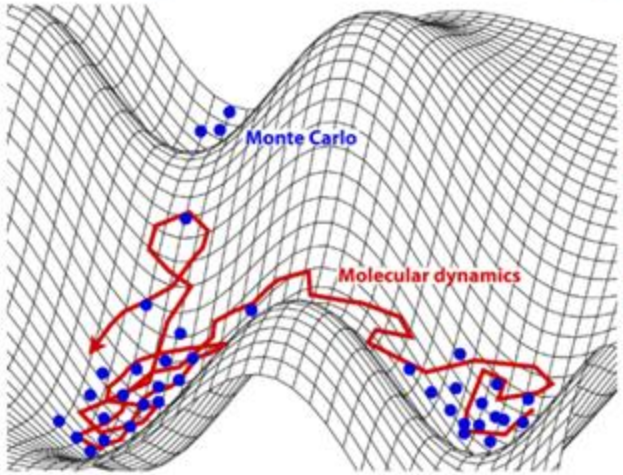


1980s

Molecular Dynamics

AMBER ('81), CHARMM ('83)

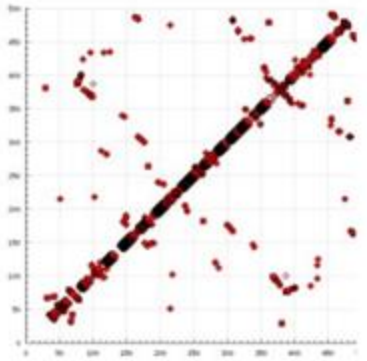




1990s

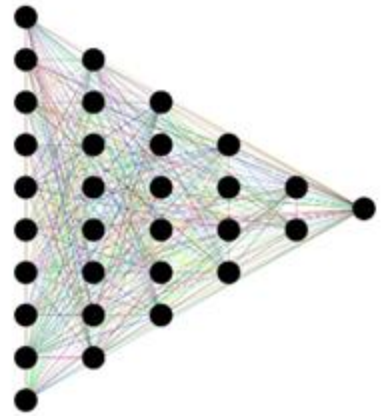
Fragment Assembly

Rosetta ('97), 1st CASP ('94), Threading ('91), BLAST ('90)



2000s

Contact/Distance Map Prediction



2010s

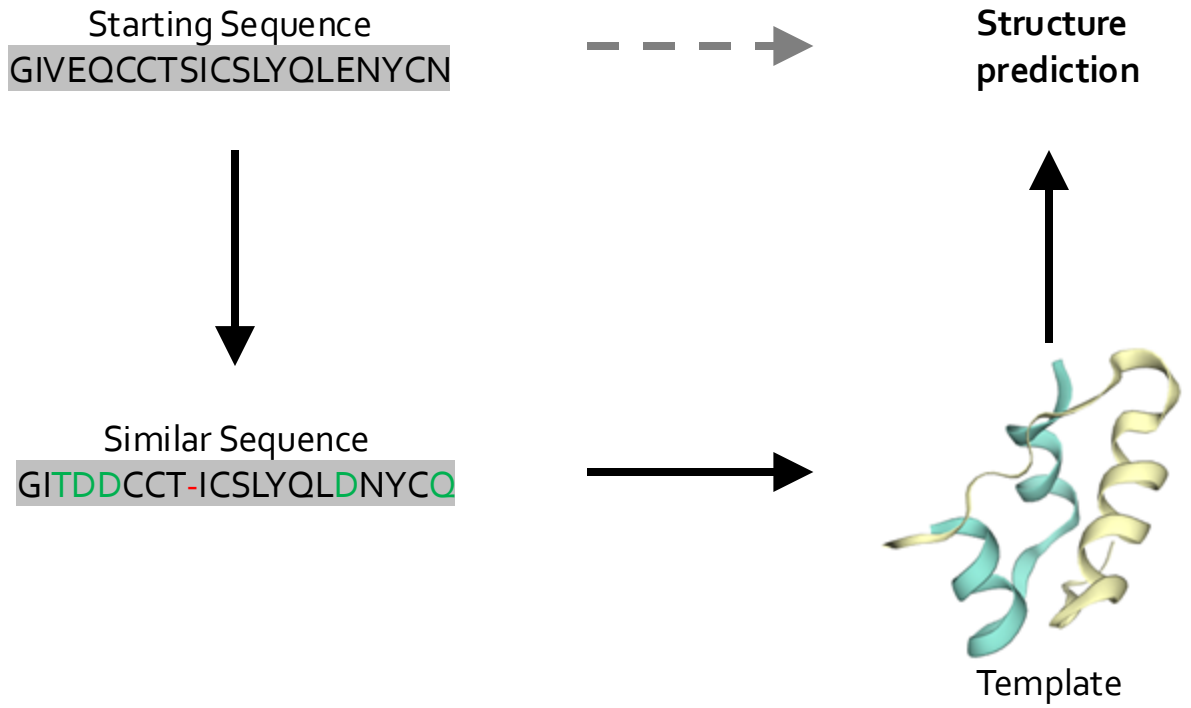
DL, first for maps, then end-to-end

RaptorX ('17), AF ('18), RGN ('19), AF2 ('20)

Slide from Heidelberg University Structural Bioinformatics Course

Homology modelling

- proteins that have **similar sequences** usually have **similar structures**
- protein **structures are more conserved** than their sequence

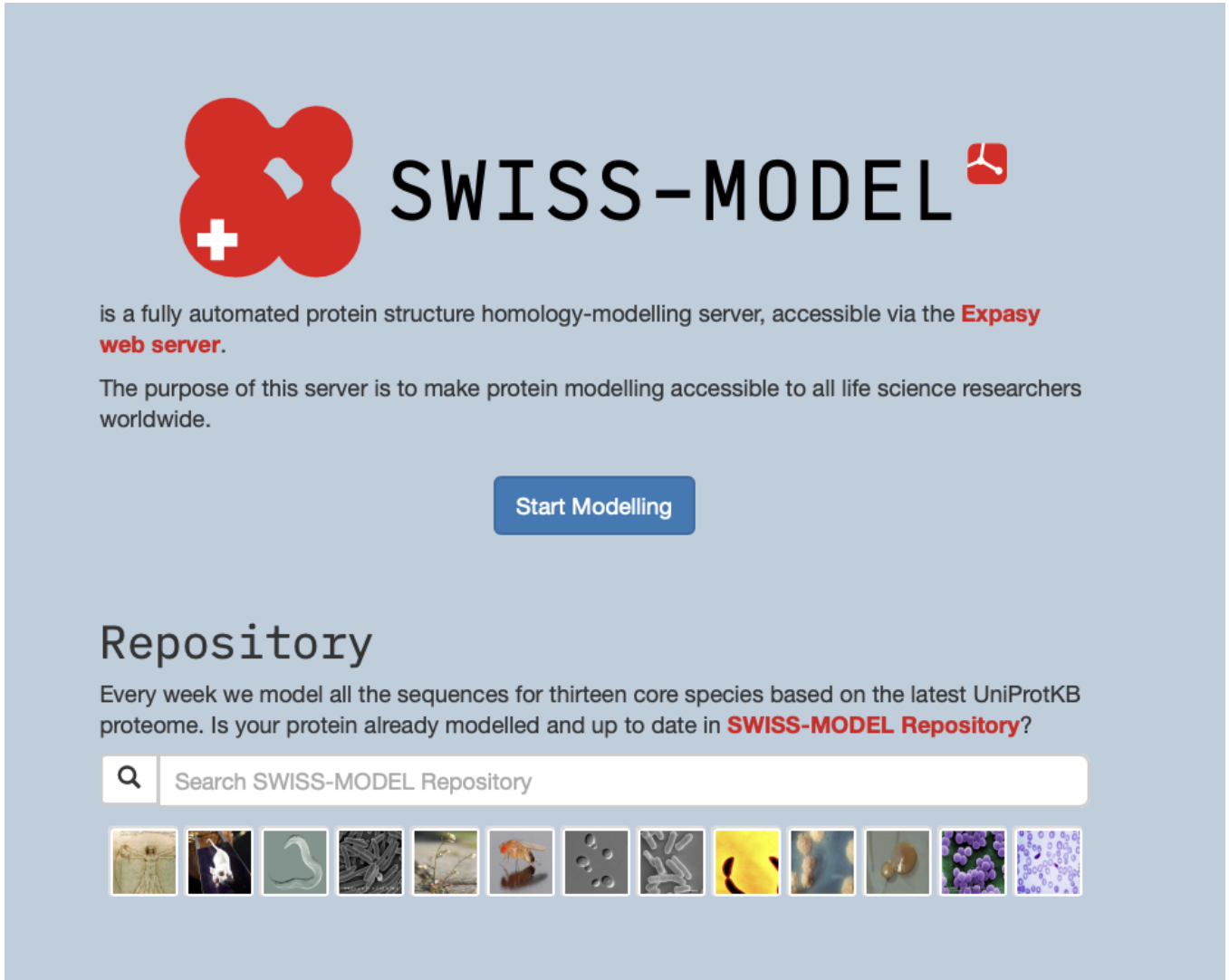


Disadvantages

- Needs available template with high-quality alignment
- Accuracy drops sharply if homologues very remote (e.g. < 30% seq. id.)

Homology modelling

SWISS-MODEL a fully automated protein structure homology-modelling server including pre-computed models in the **SWISS-MODEL Repository**



Overview of the methods covered

(2020)

AlphaFold2

ColabFold

(2021)

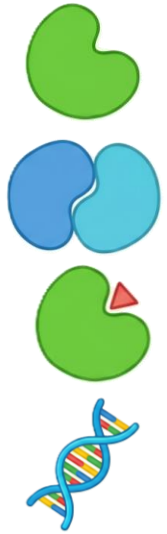
AlphaFold-Multimer

(2024)

AlphaFold3

Boltz

Chai



Swiss Institute of Bioinformatics

DAY 1, PART 2

Protein monomers (AlphaFold2)

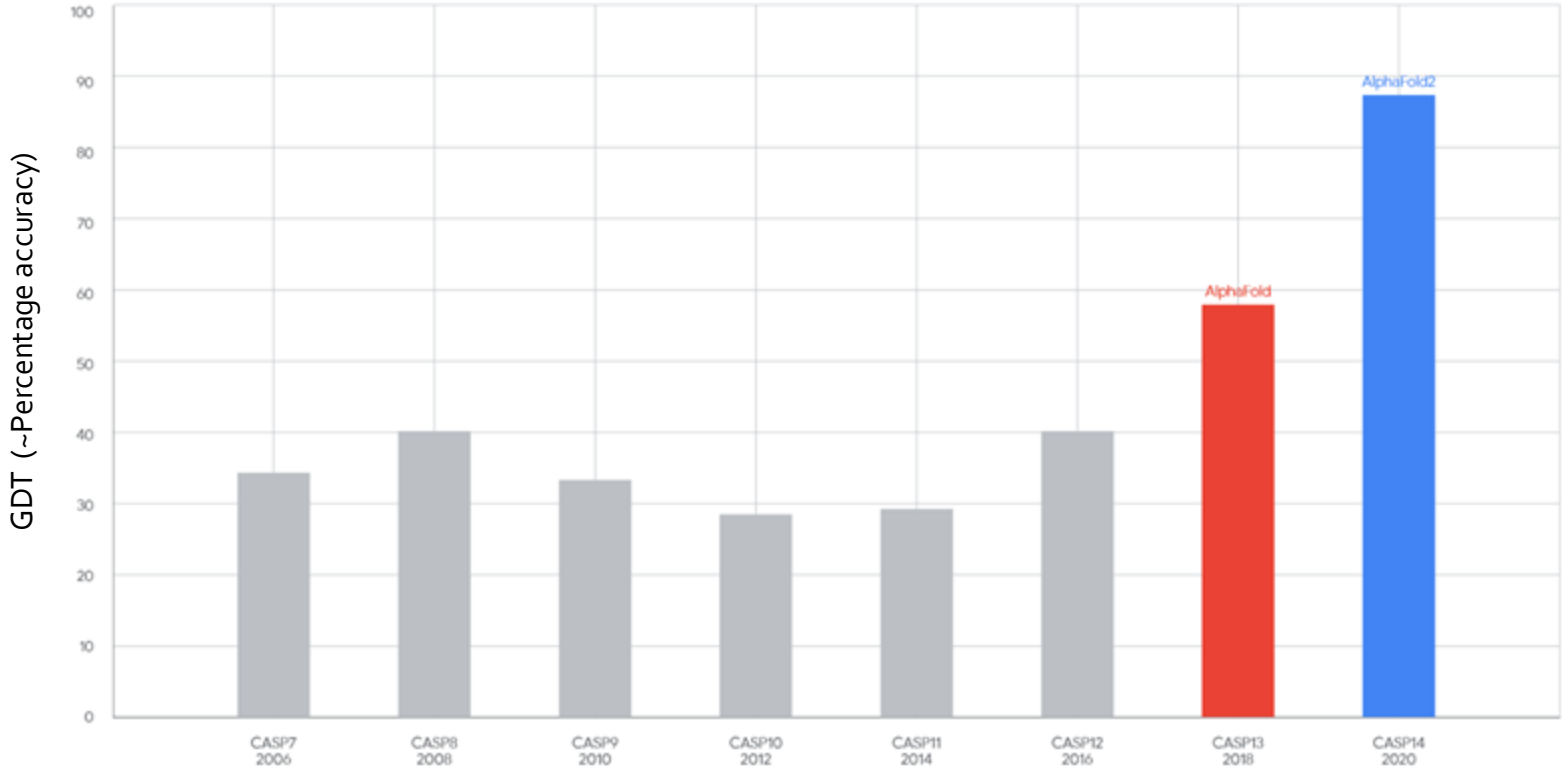
Diana Rapota, Rok Breznikar, Janani Durairaj

23-24 June 2026

elixir SWITZERLAND

Critical Assessment of Structure Prediction (CASP)

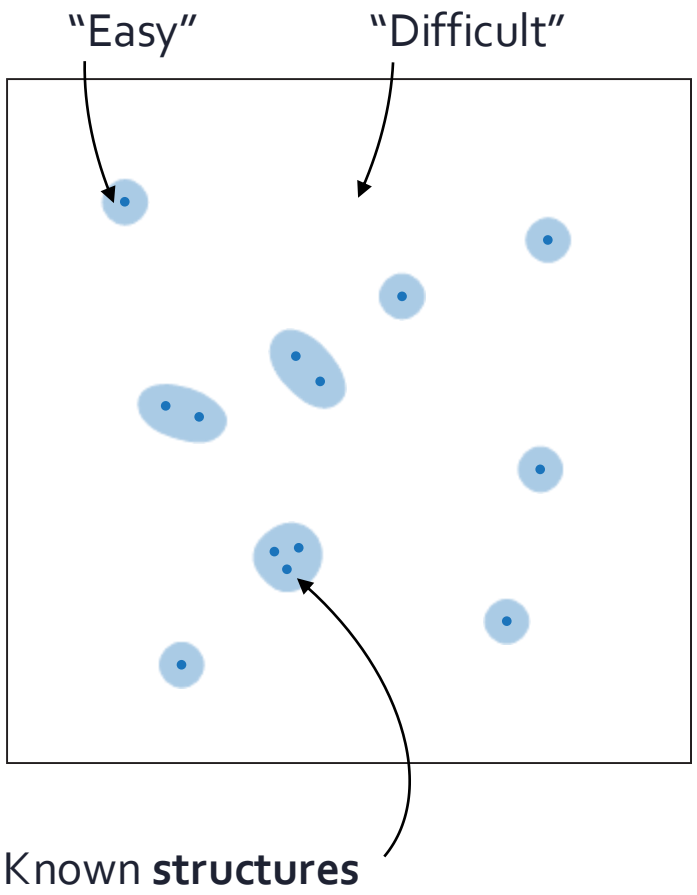
- Test/competition of protein structure predictions
  - On unreleased structures
- Carried out every 2 years

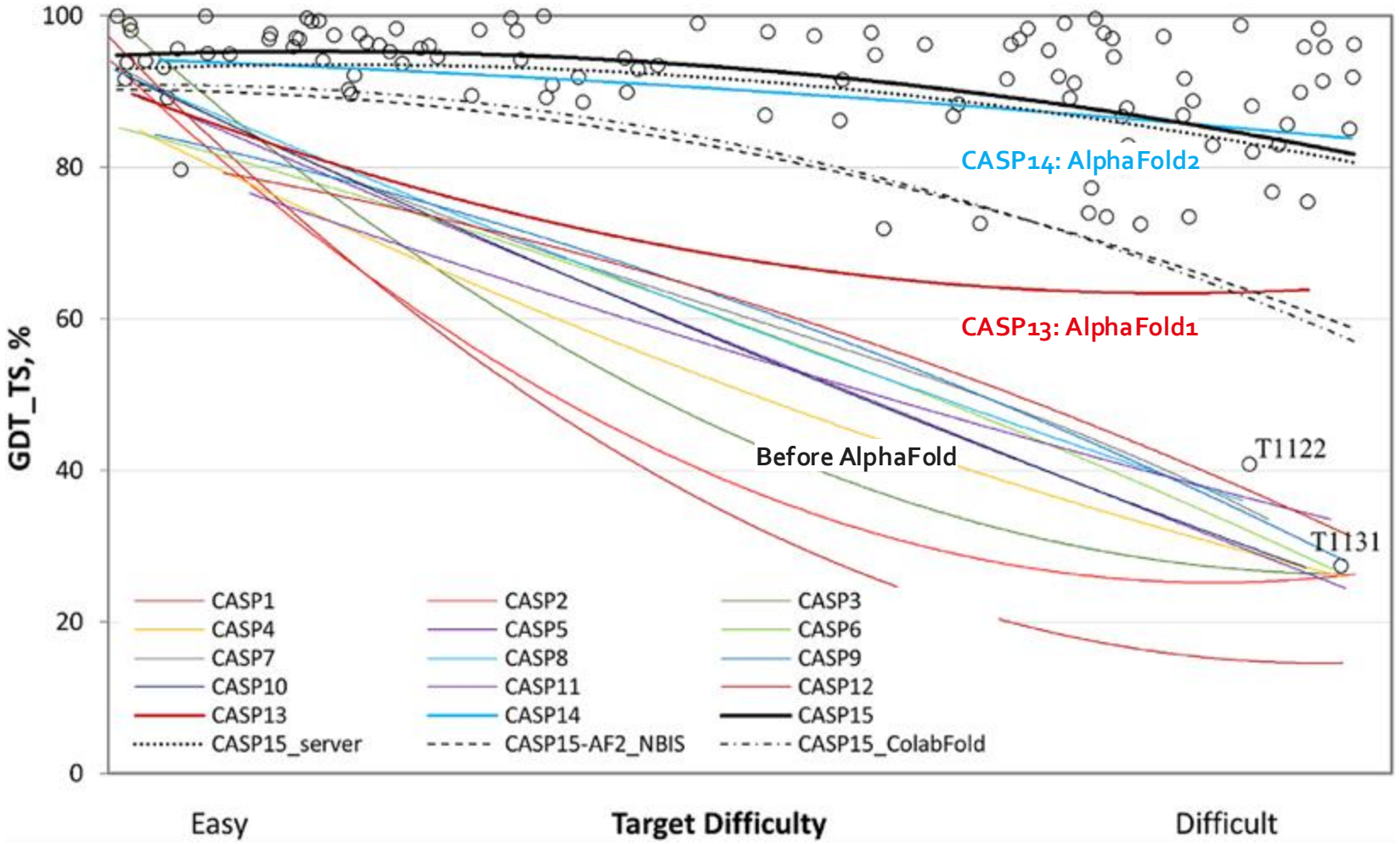


Overall success at protein structure prediction in CASPs over the years.

Critical Assessment of Structure Prediction (CASP)

- Similarity of structures to already known structures





AlphaFold2

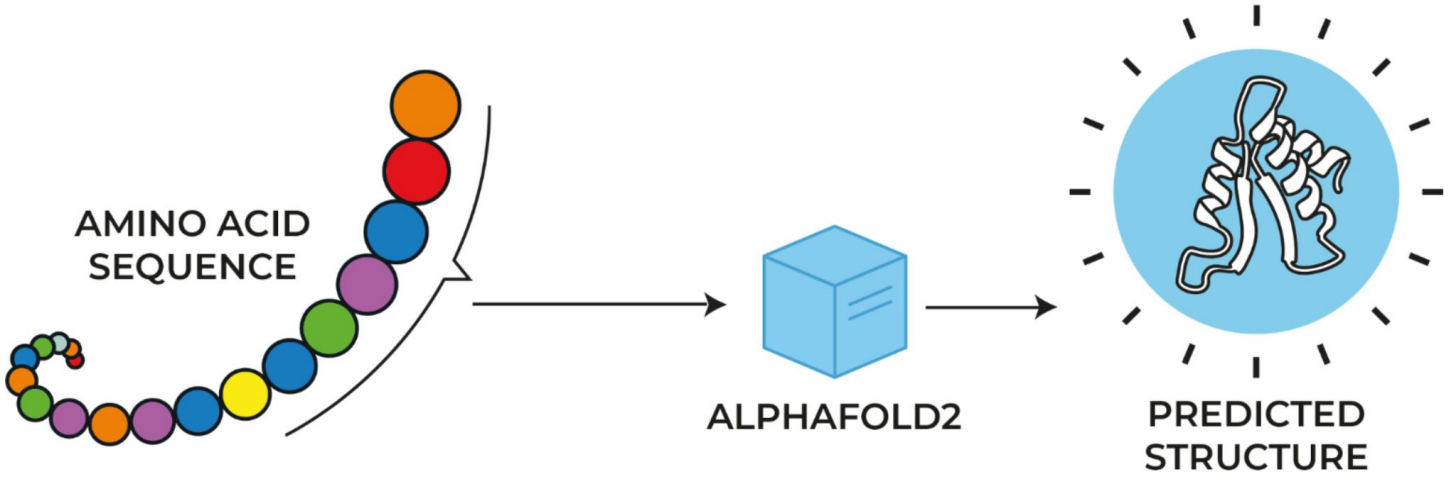
Developed by Google DeepMind

Input:

- sequence

Output:

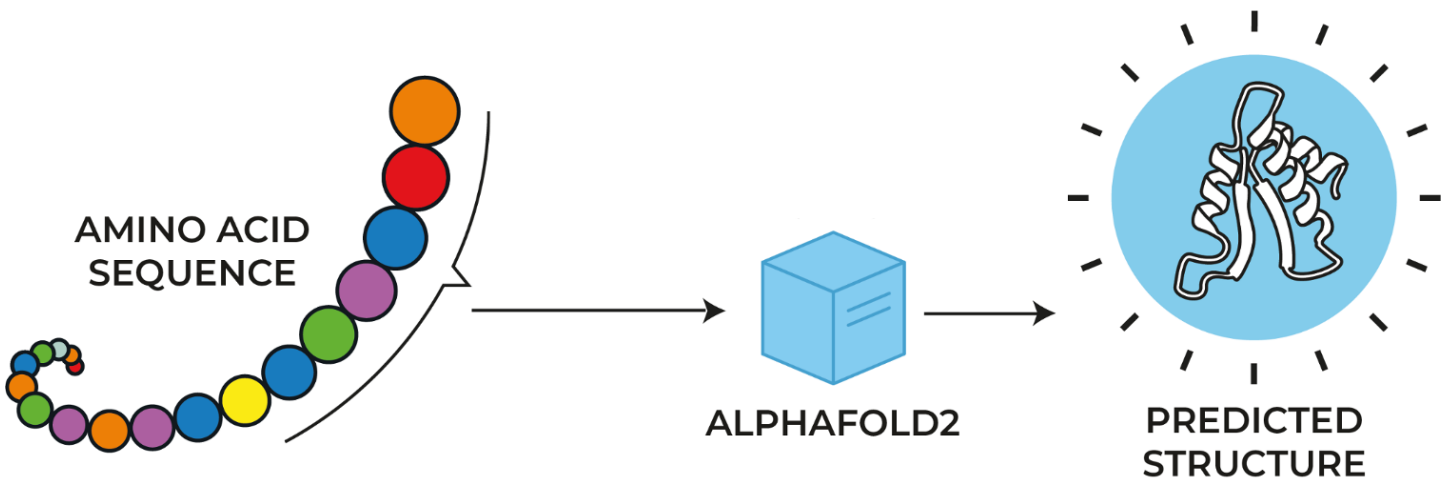
- structure (static, not multiple conformations)
- confidence metrics (pLDDT, PAE, …)

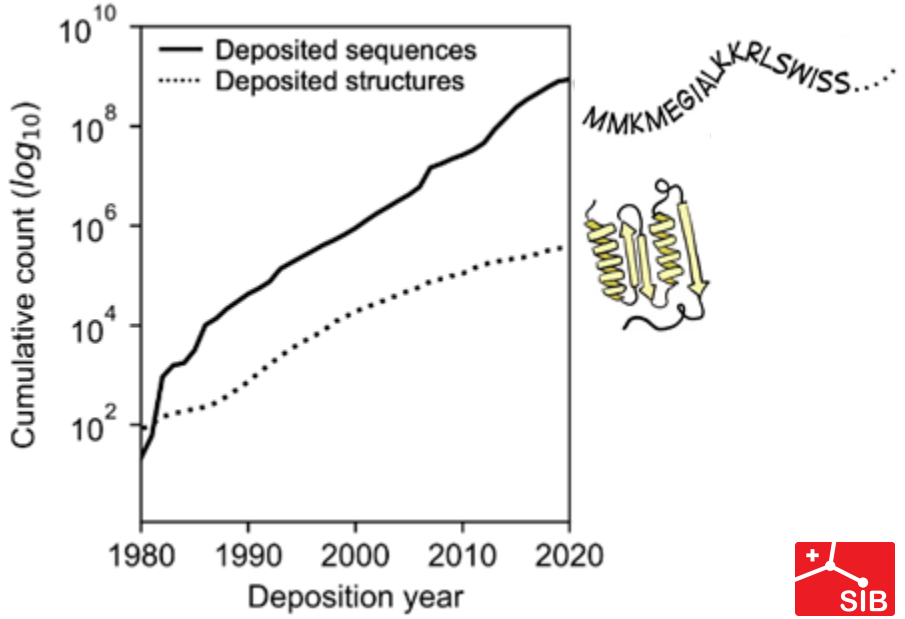


Images from the Nobel Prize explainer

AlphaFold2

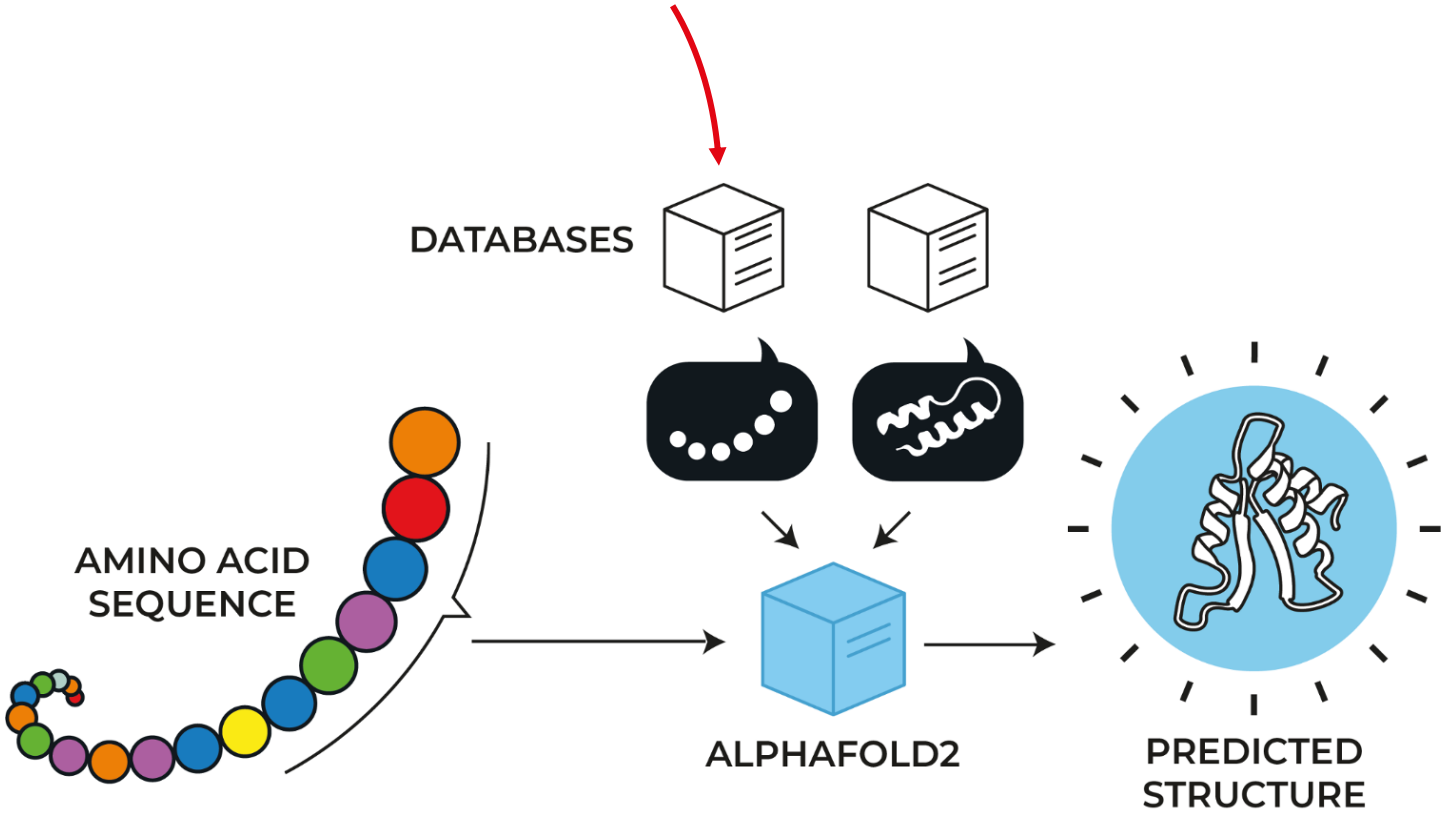
- Deep Leaning method
- Only ~200k examples of protein structures to train on

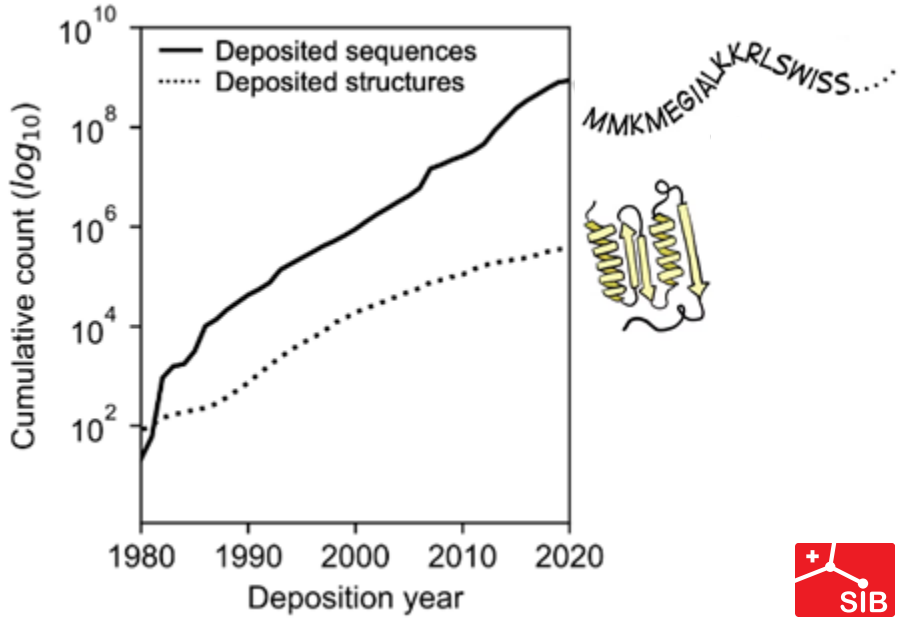




AlphaFold2

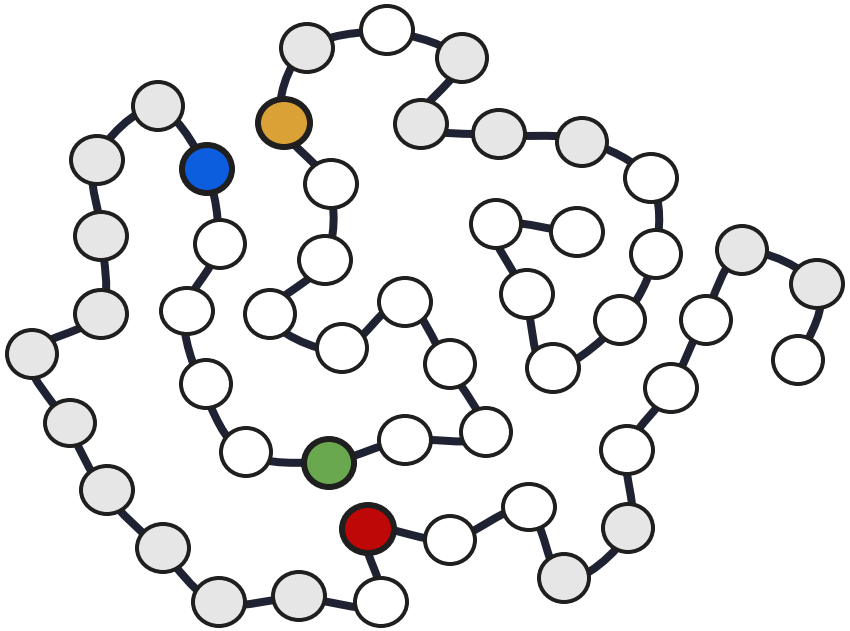
- Deep Leaning method
- Only ~200k examples of protein structures to train on
- Gets help from sequence-only data





Coevolution / Multiple Sequence Alignment (MSA)

- Proteins have long-range **dependencies** between residues
- **Not obvious** when looking at a single sequence



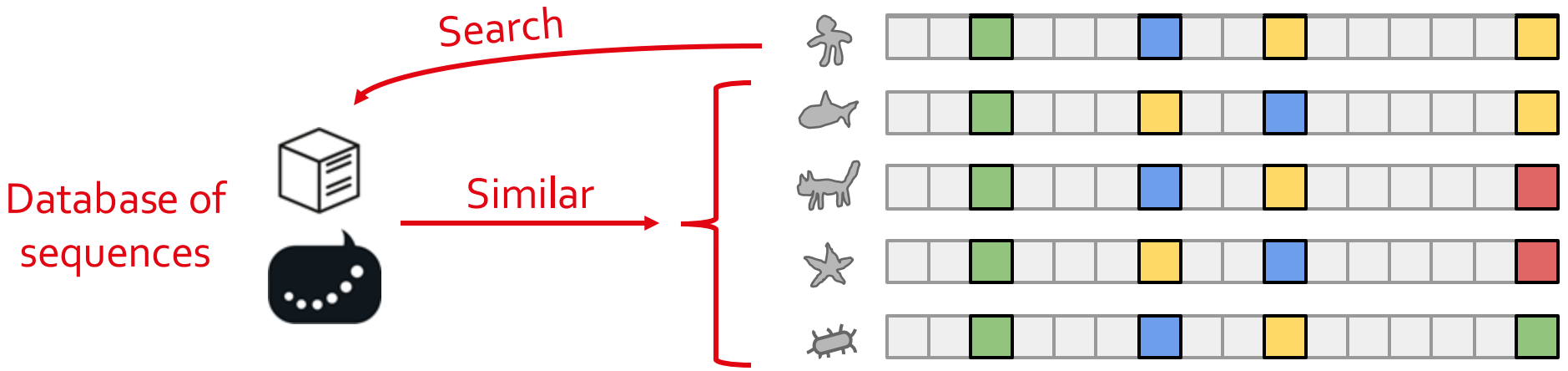
Slide from Sergey Ovchinnikov

Coevolution / Multiple Sequence Alignment (MSA)

- Proteins have long-range **dependencies** between residues
- **Not obvious** when looking at a single sequence

Coevolution / Multiple Sequence Alignment (MSA)

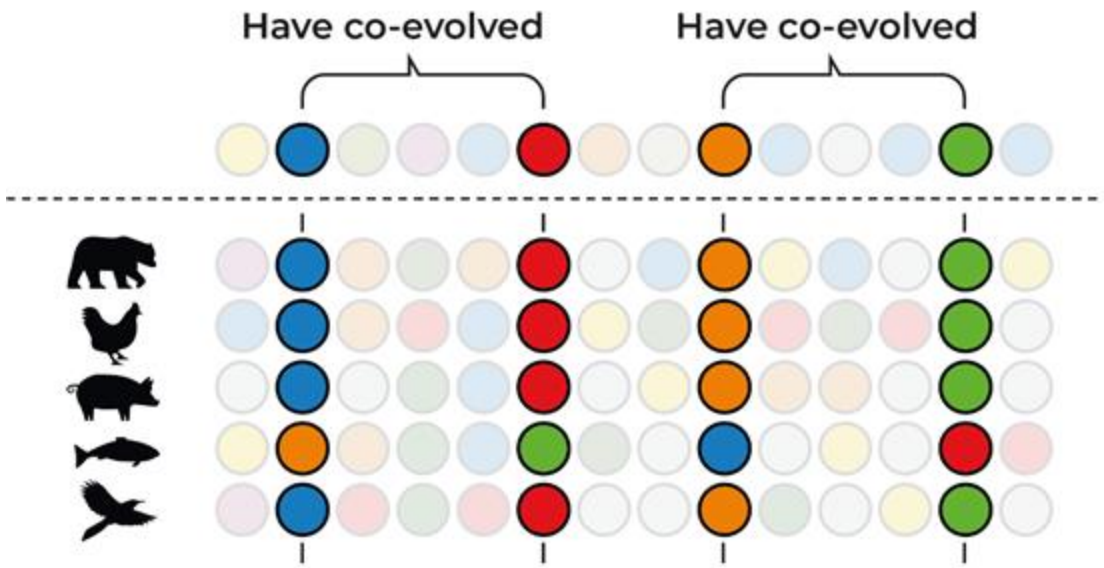
- Proteins have long-range **dependencies** between residues
- **Not obvious** when looking at a single sequence
- We can learn this by looking at **multiple sequences**

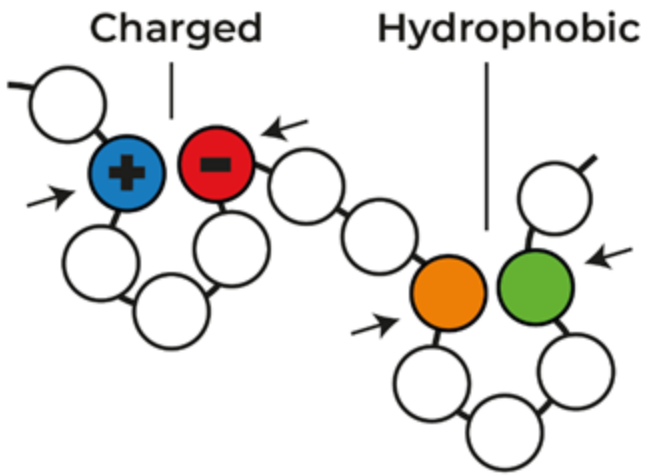


Coevolution / Multiple Sequence Alignment (MSA)

If two amino acids in a protein are in **close contact**, mutations in one of them will probably be followed by **mutations of the other** (to preserve the structure).

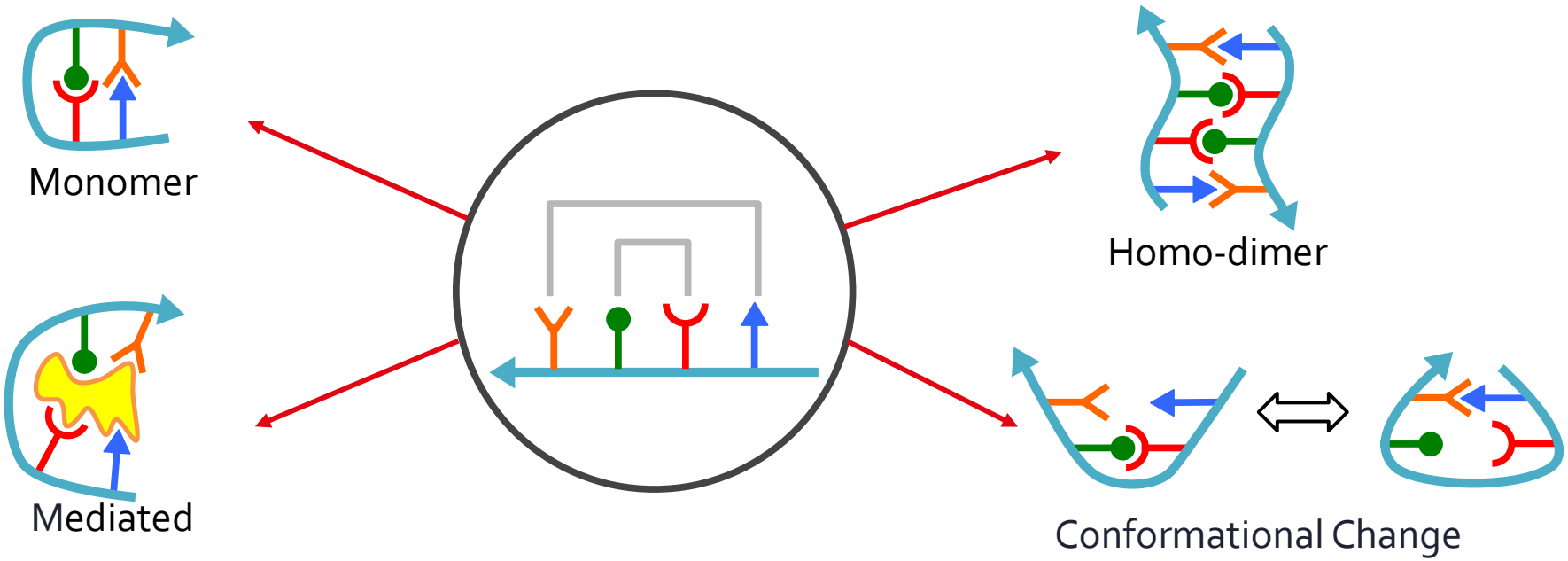
The opposite is also true: if two regions of a protein are changing and **evolving independently** from each other, it is likely that they are **not in direct contact**.





Images from the Nobel Prize explainer

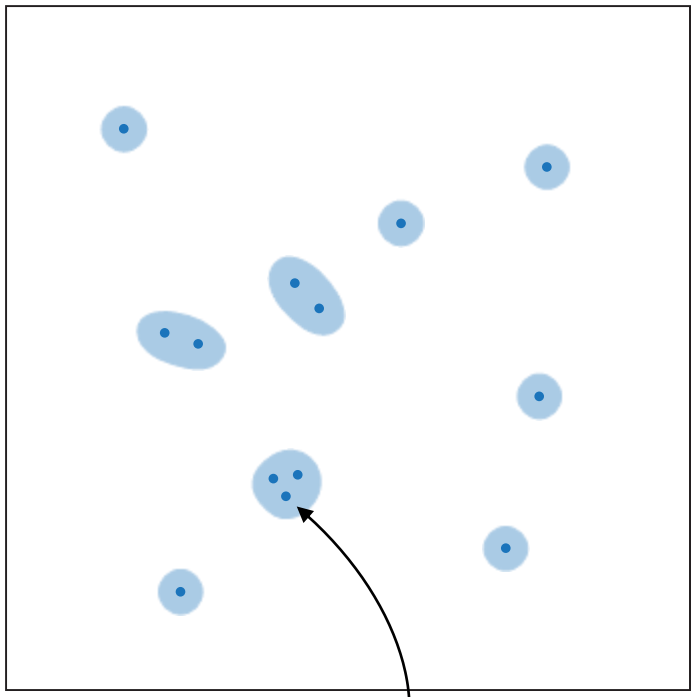
Coevolution / Multiple Sequence Alignment (MSA)



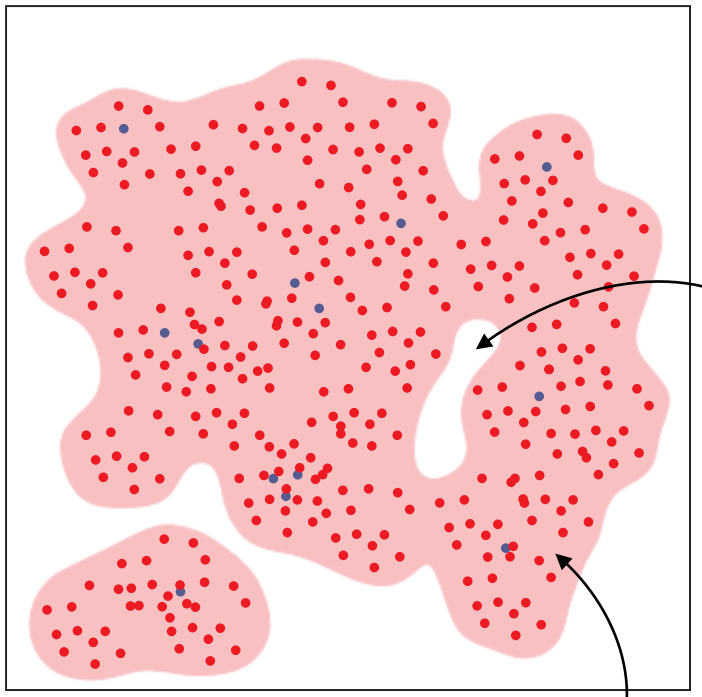
Slide from Sergey Ovchinnikov

Coevolution extends the applicability domain

- Where AlphaFold2 works well:



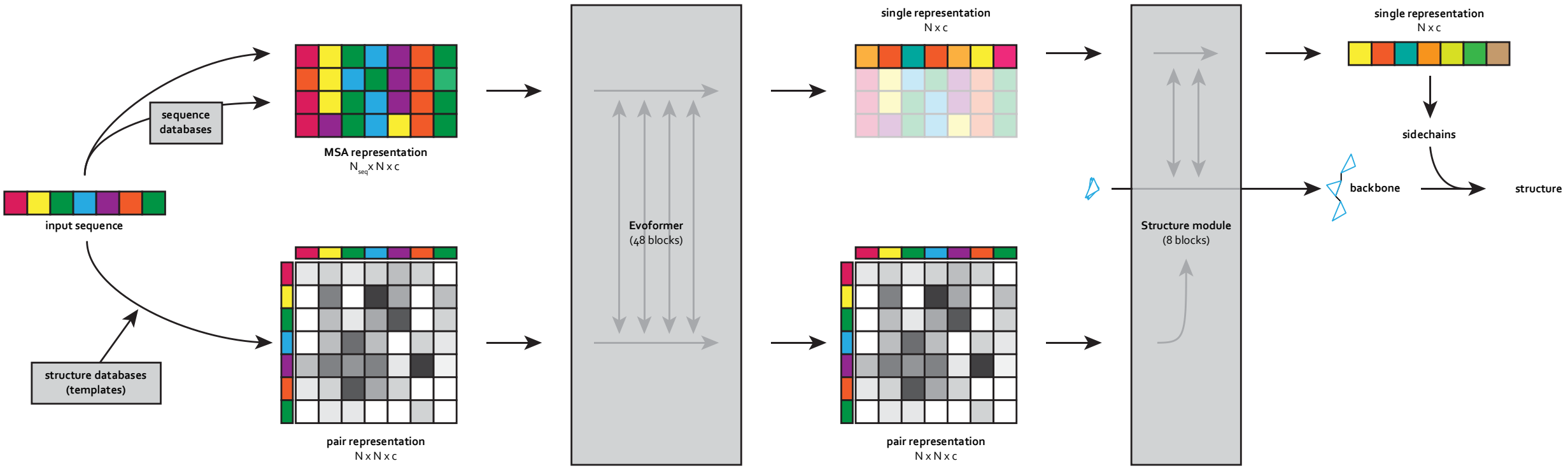
Known **structures**



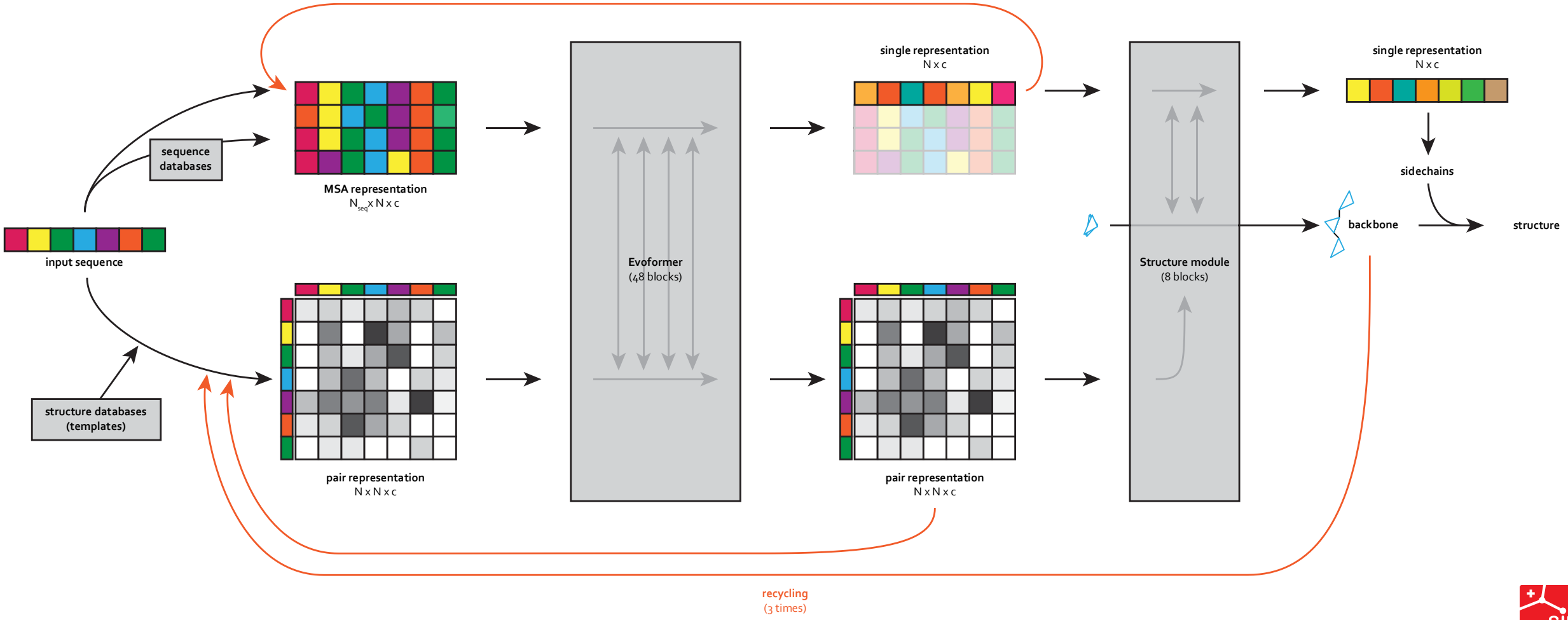
Orphan proteins
Designed sequences
Fast-evolving families
(not many close relatives
→ shallow MSA)

Known **sequences**
(and many close relatives → deep MSA)

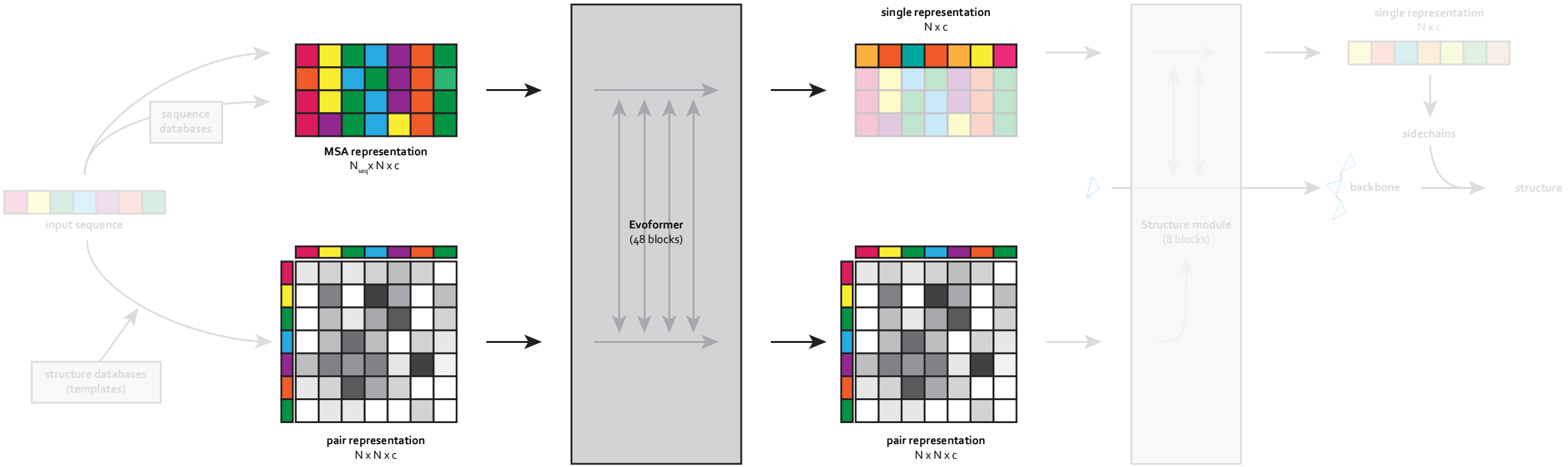
Architecture



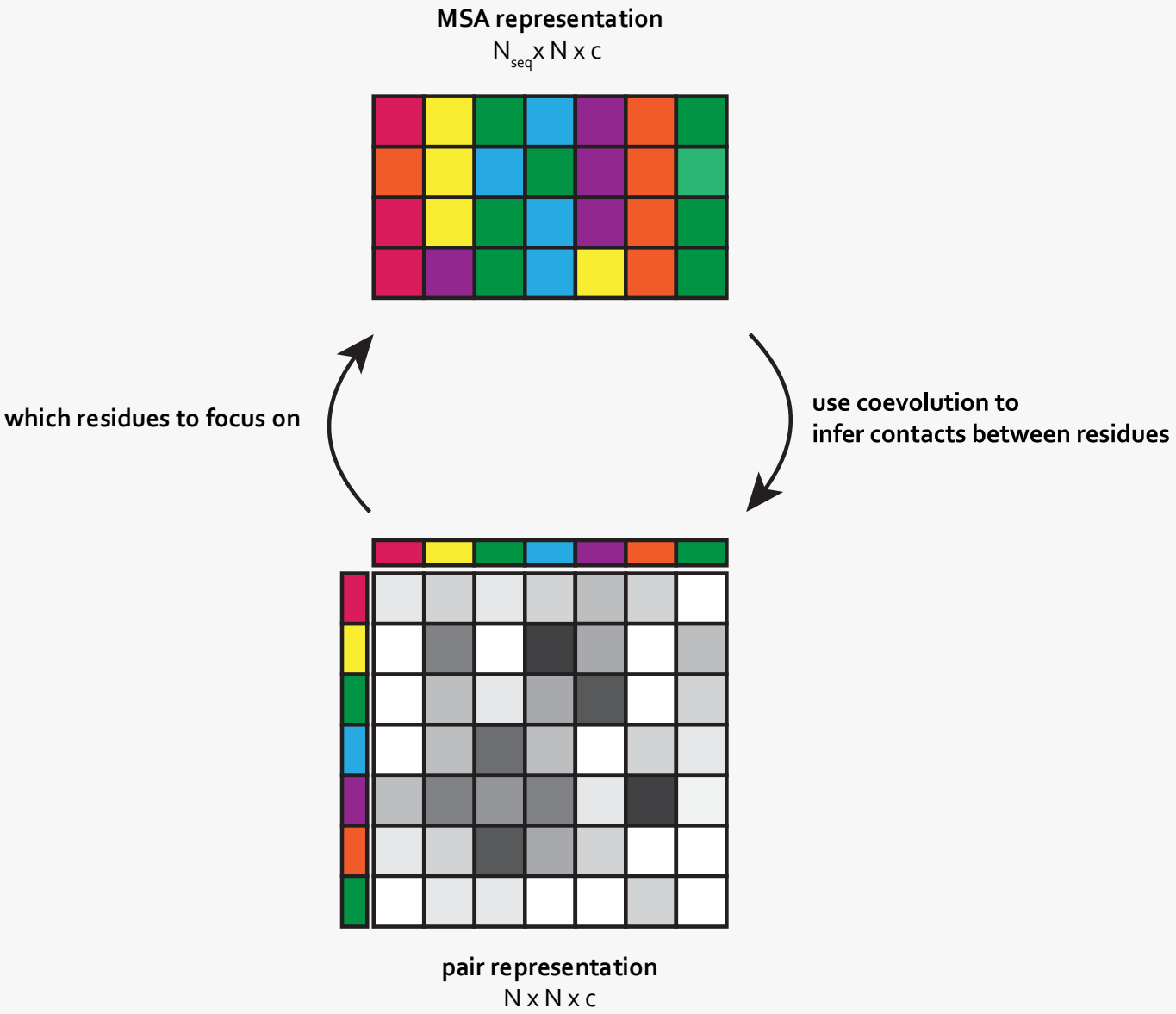
Architecture

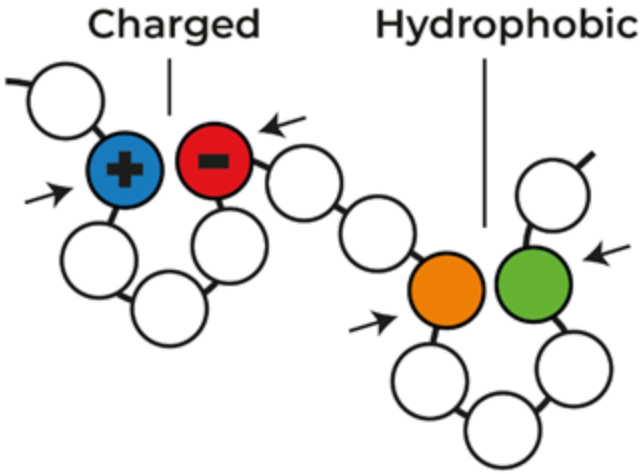


Evoformer

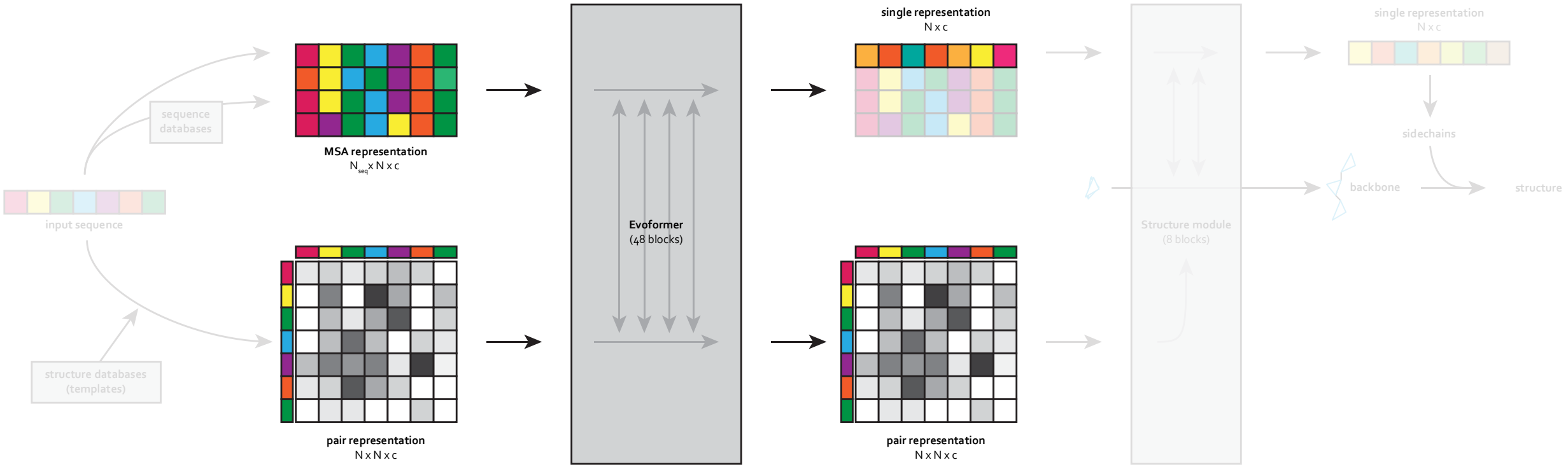

sequence databases
input sequence
structure databases (templates)
MSA representation
N_{seq} x N x c
pair representation
N x N x c
Evoformer
(48 blocks)
single representation
N x c
pair representation
N x N x c
Structure module
(8 blocks)
single representation
N x c
sidechains
backbone
structure

Evoformer

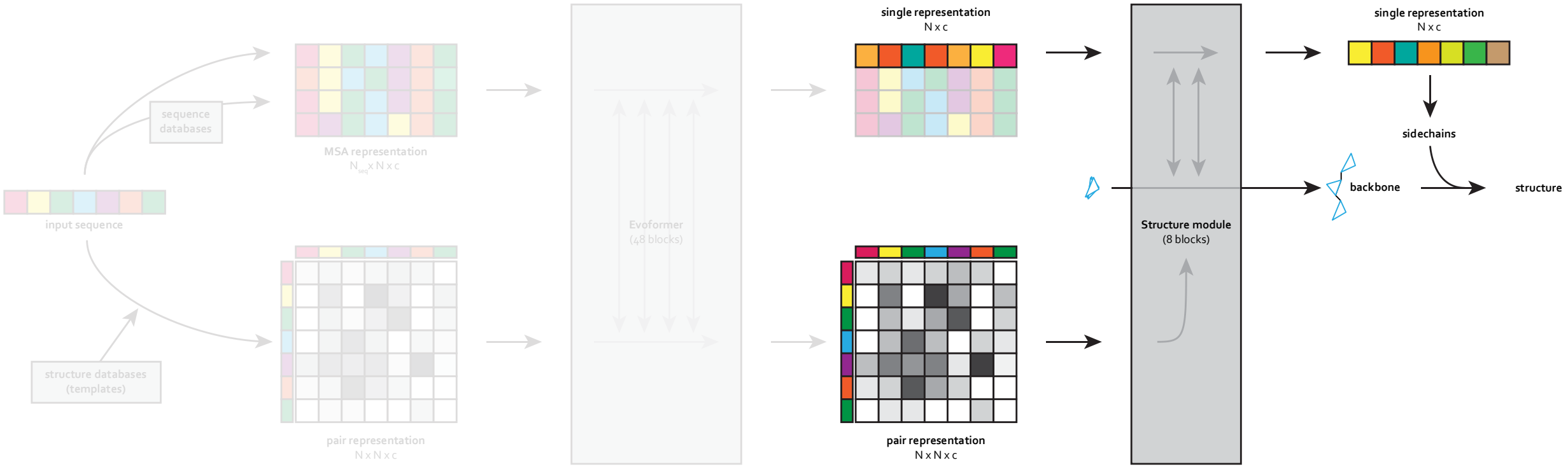

MSA representation
N_{seq} x N x c
which residues to focus on
use coevolution to
infer contacts between residues
pair representation
N x N x c


Charged
Hydrophobic

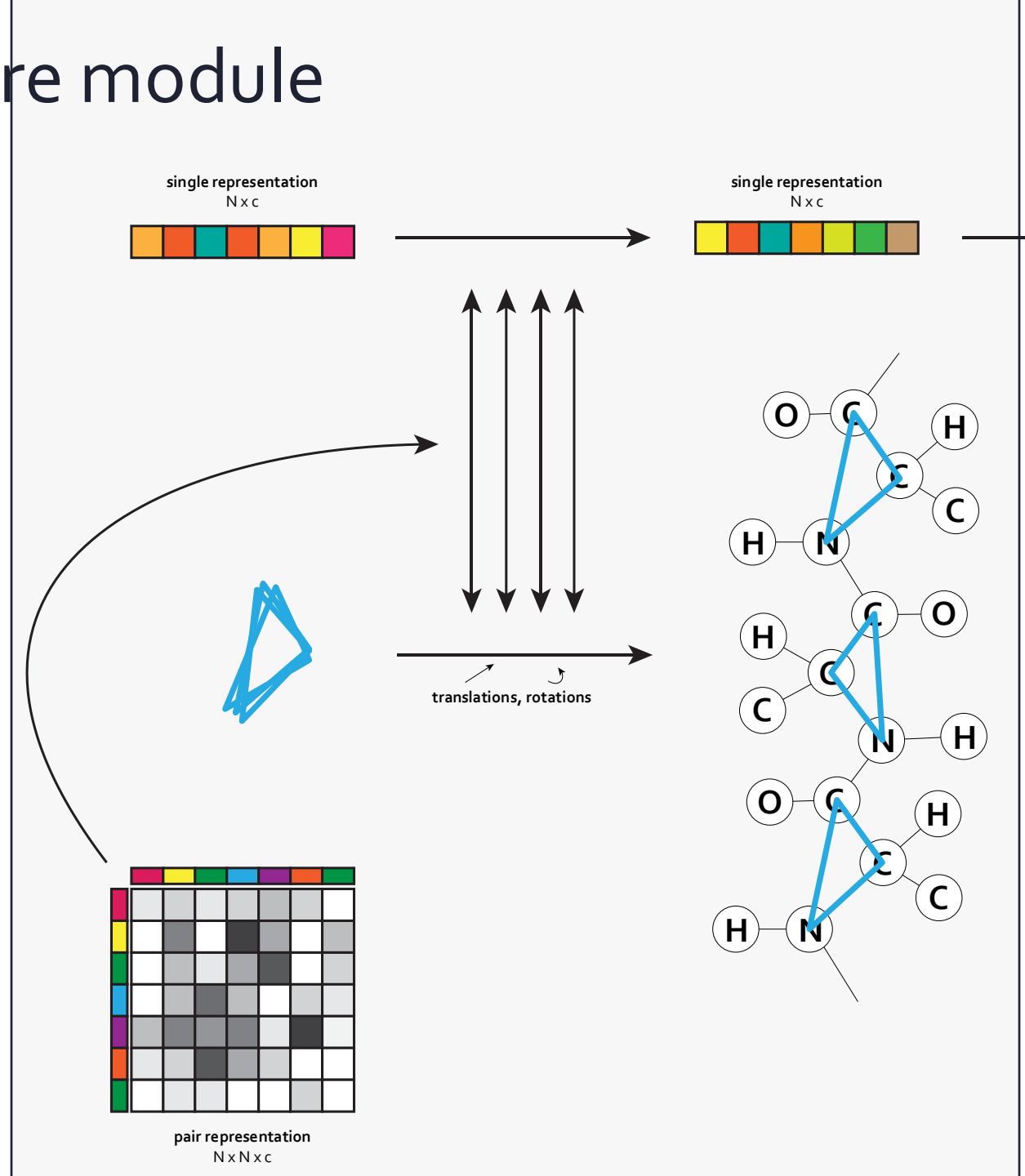
Evoformer


sequence databases
input sequence
structure databases (templates)
MSA representation
N_{seq} x N x c
pair representation
N x N x c
Evoformer
(48 blocks)
single representation
N x c
pair representation
N x N x c
Structure module
(8 blocks)
single representation
N x c
sidechains
backbone
structure

Structure module



Structure module



5 separate AlphaFold2 models

- AlphaFold2 outputs 5 structures
- Each is a result of a different model (same architecture, different training)

Model 1	Model 2	Model 3	Model 4	Model 5
Uses templates		No templates		
More MSA seq	Less MSA seq	More MSA seq		Less MSA seq

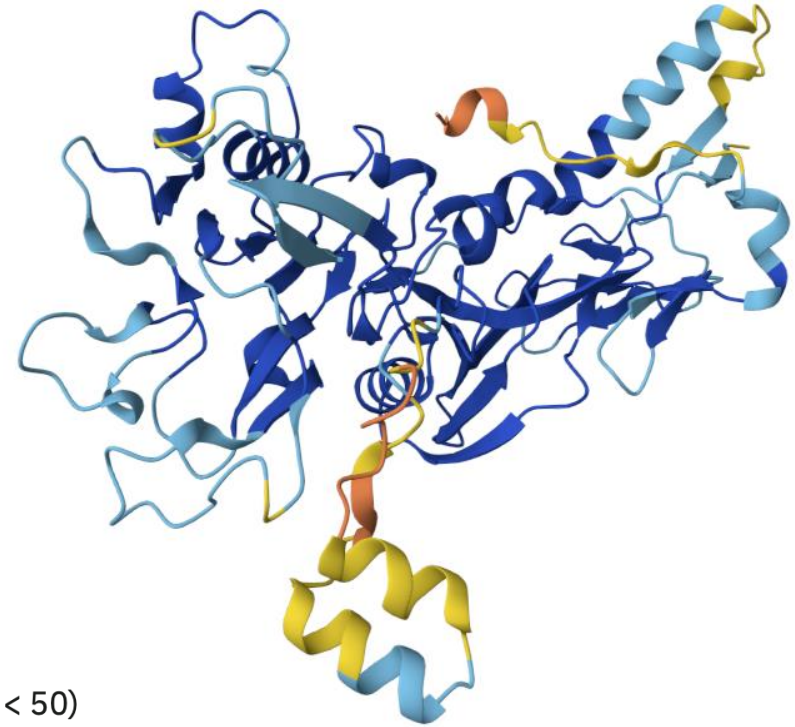
Confidence metrics

- AlphaFold2 was trained to also output confidence metrics alongside predicted structures
- Predicted metrics that tell how close to the real answer it is,
  in other words, how confident the model is

- pLDDT
  - Local confidence
- PAE
  - Relative position of two residues

pLDDT – local confidence

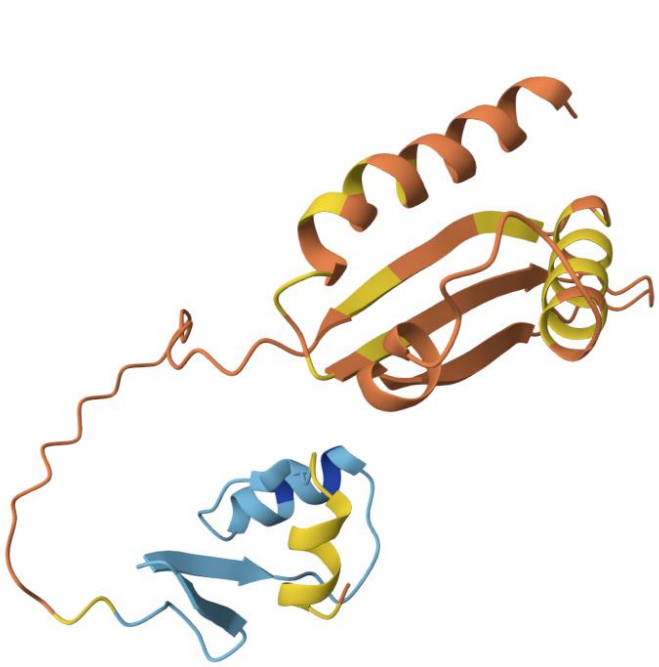
- Predicted local distance difference test (pLDDT)
  - Measures correctness of distances from an atom (Cα) to its close neighbours (also Cα)
- Per-residue measure of local confidence
- Range: 0-100
  higher score = higher confidence
- Low scores:
  - AF2 doesn't have enough information
  - Disorder/flexibility

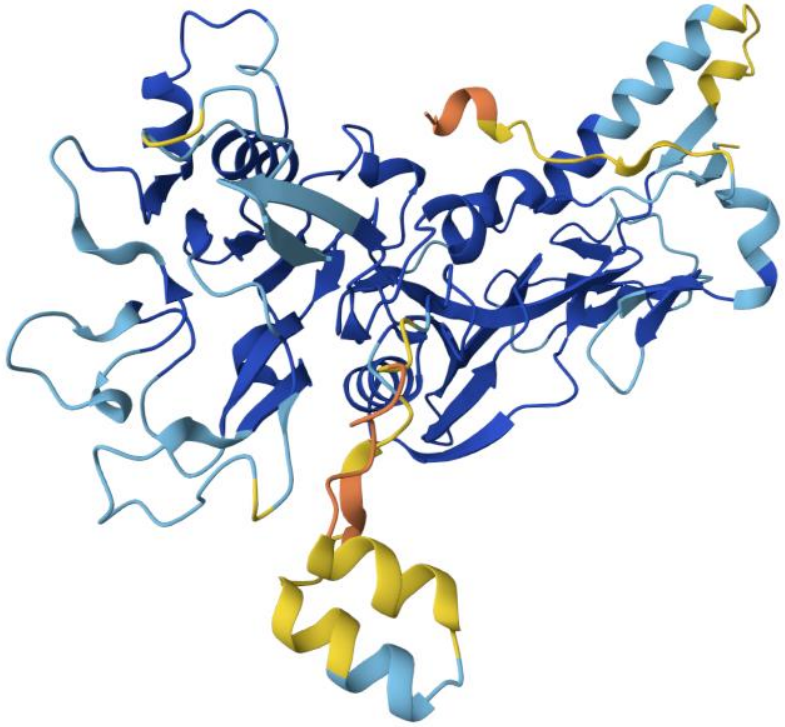


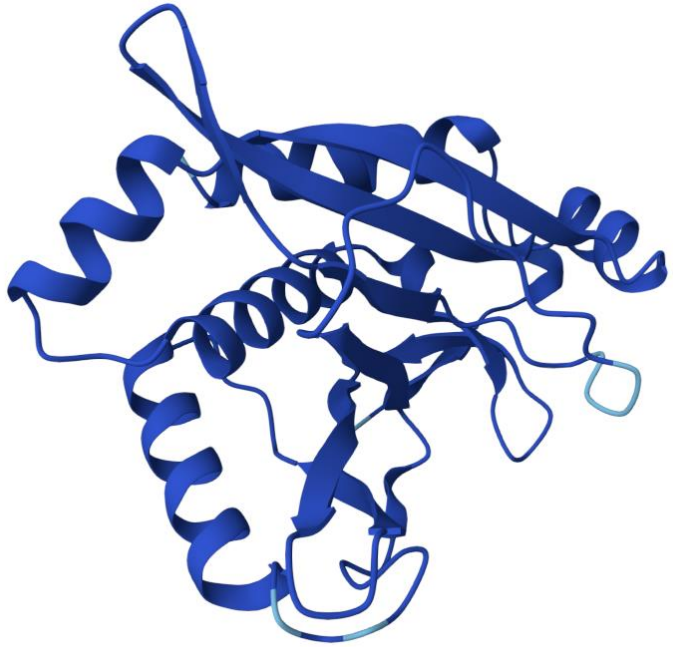
Very high (pLDDT > 90) | Confident (90 > pLDDT > 70) | Low (70 > pLDDT > 50) | Very low (pLDDT < 50)

AlphaFold ID: AF-Q3S2X4-F1

pLDDT – local confidence







Very high (pLDDT > 90)

Confident (90 > pLDDT > 70)

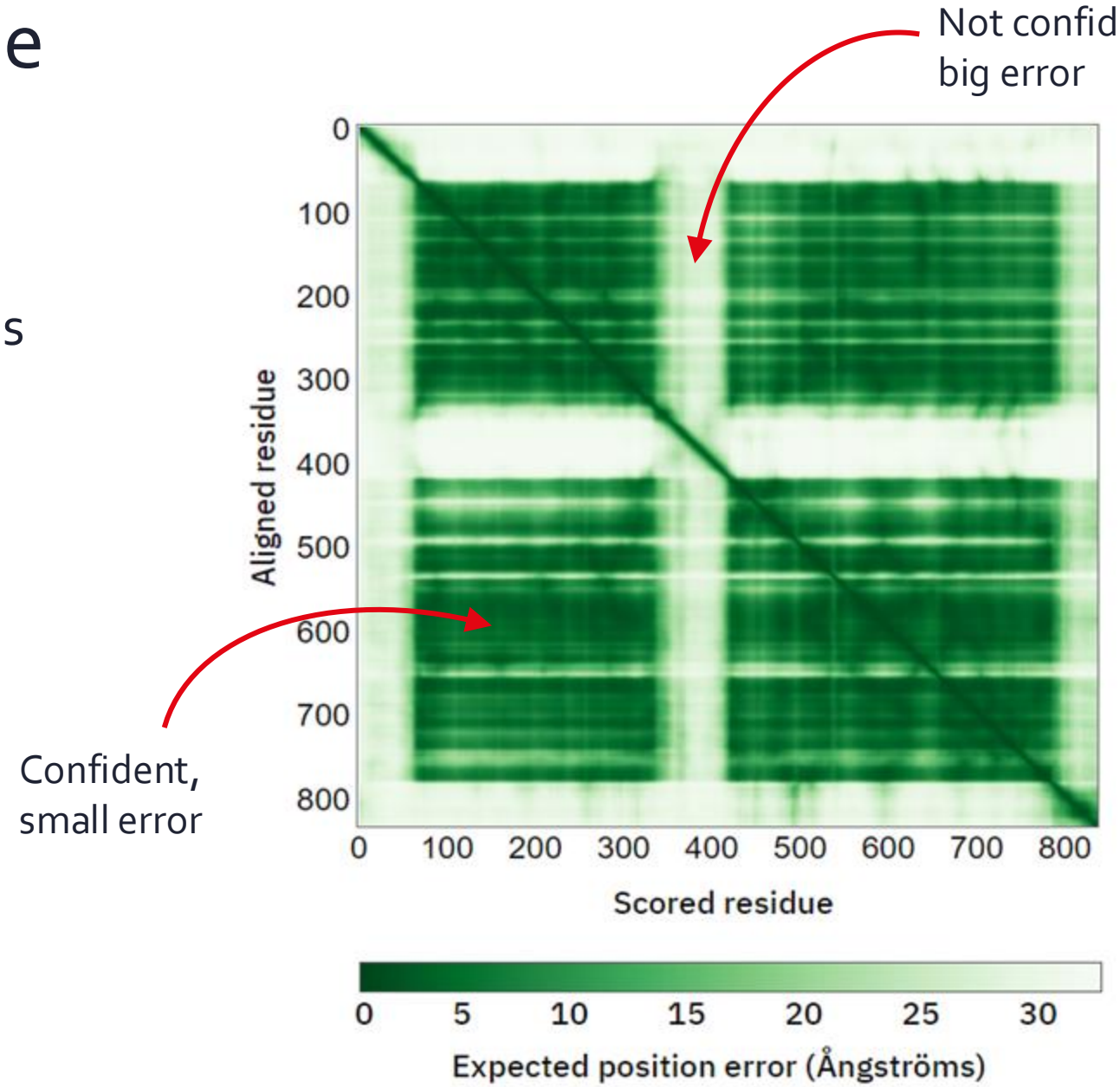
Low (70 > pLDDT > 50)

Very low (pLDDT < 50)

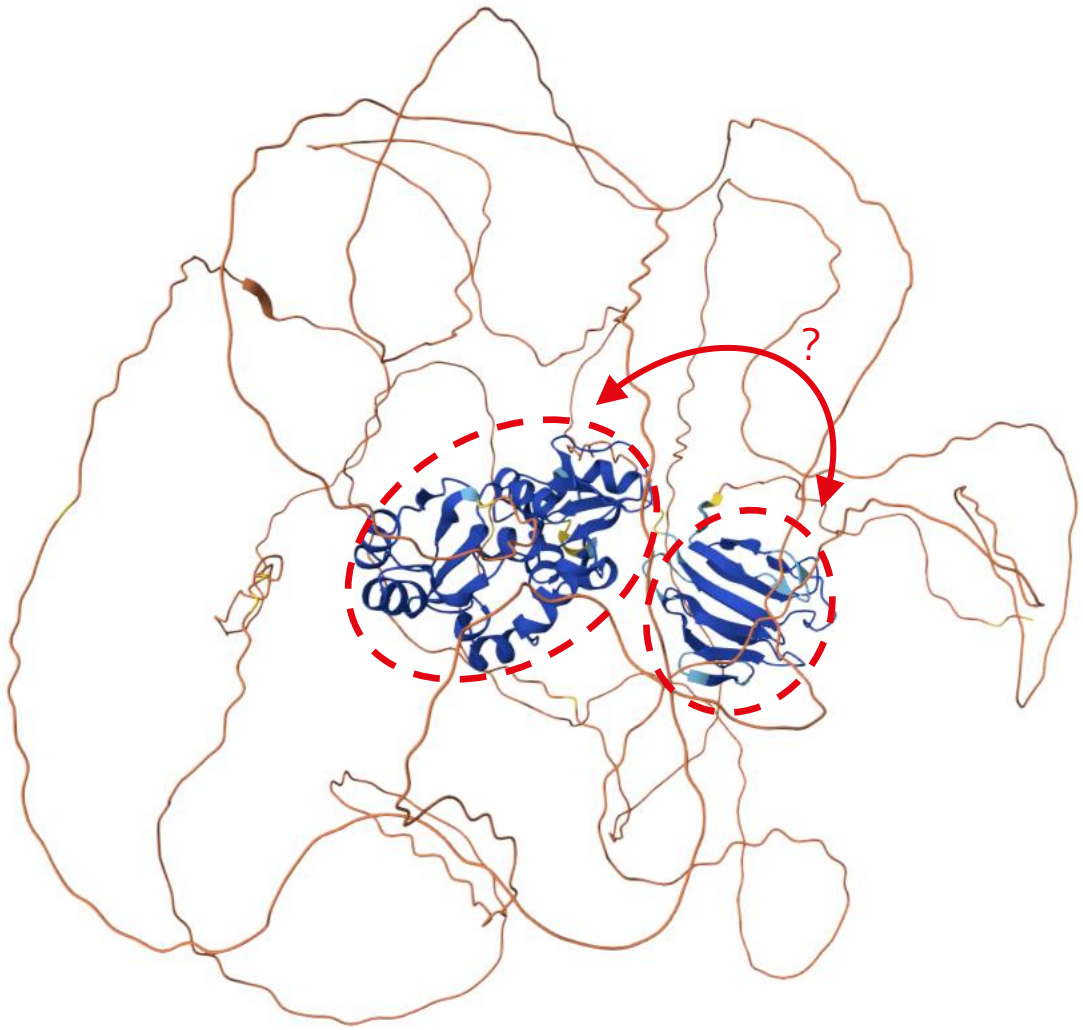
AlphaFold ID: AF-Q8IZS7-F1, AF-Q3S2X4-F1, AF-P36883-F1

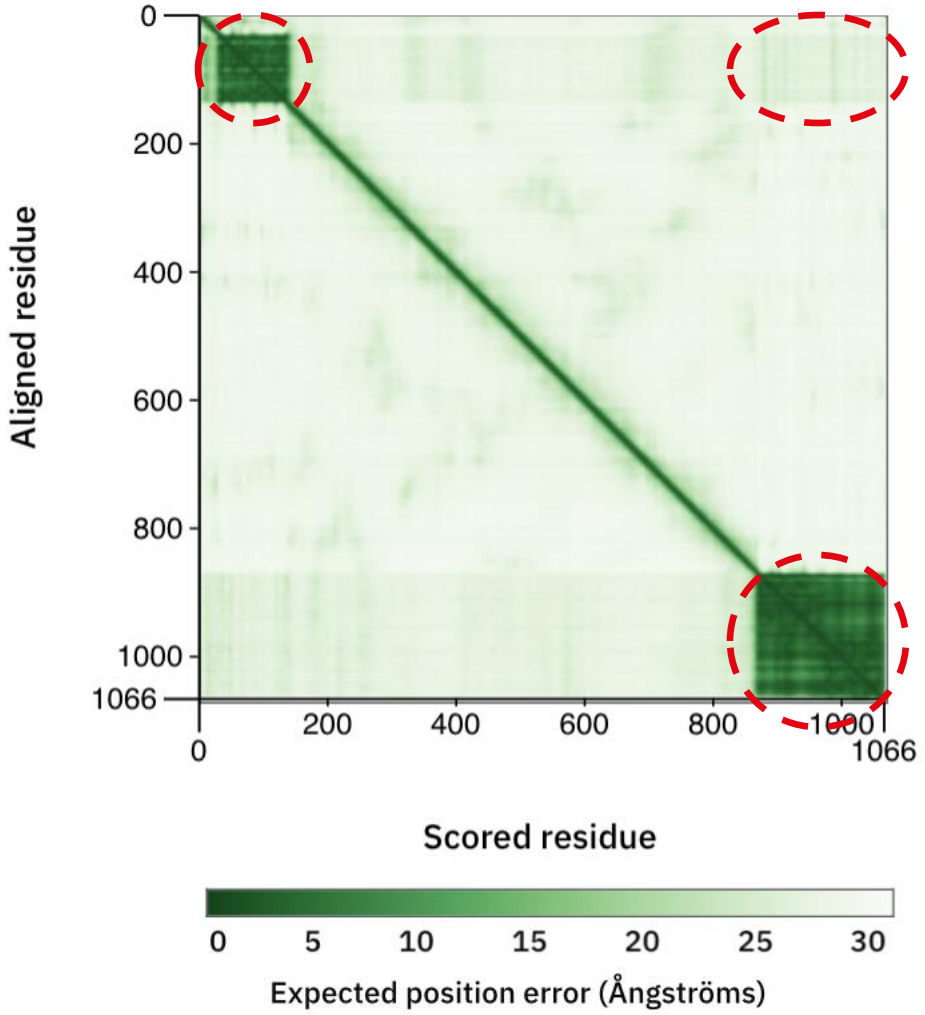
PAE – global confidence

- Predicted aligned error (PAE)
- Relative position of two residues
- N x N matrix



PAE – global confidence

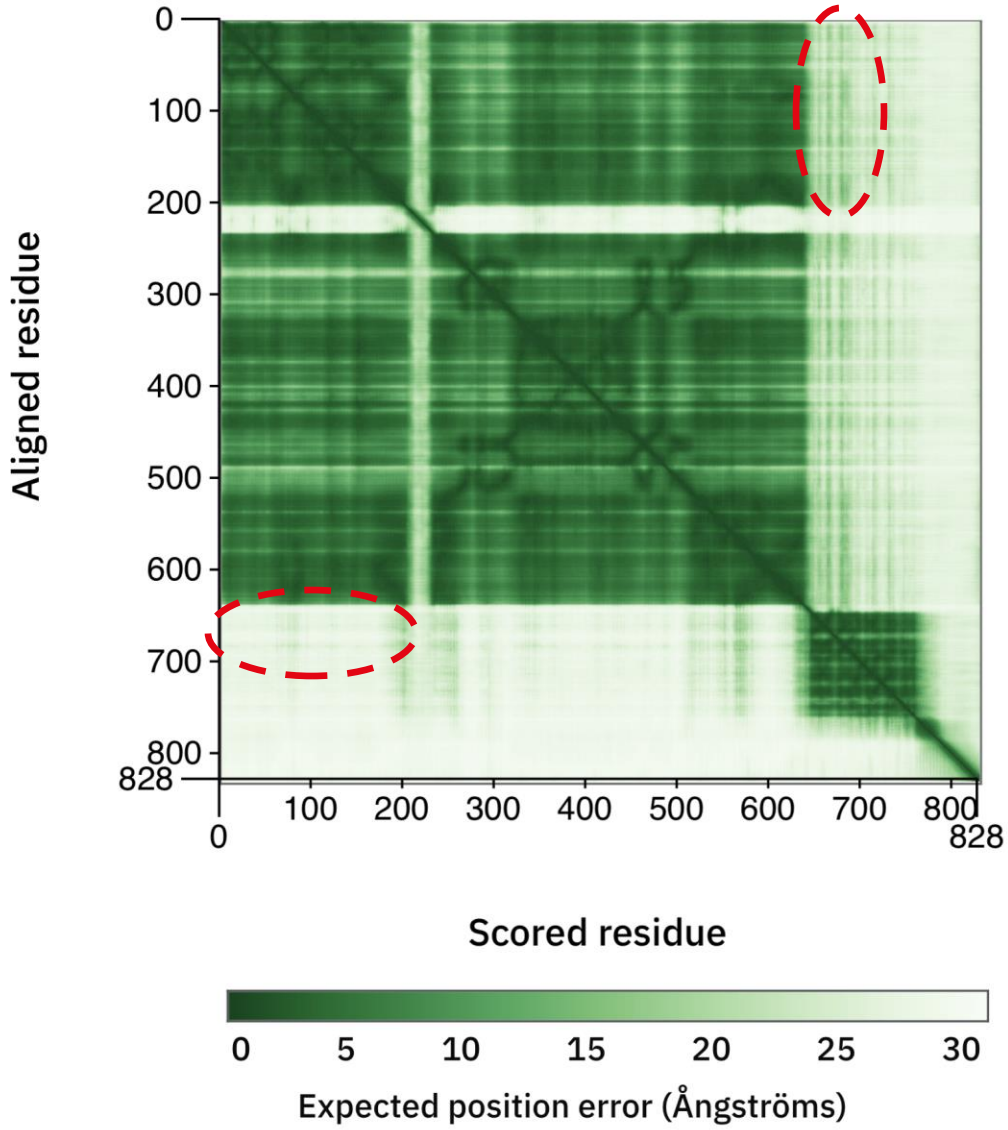




AlphaFold ID: AF-Q14676-F1

PAE – global confidence

- Not symmetrical

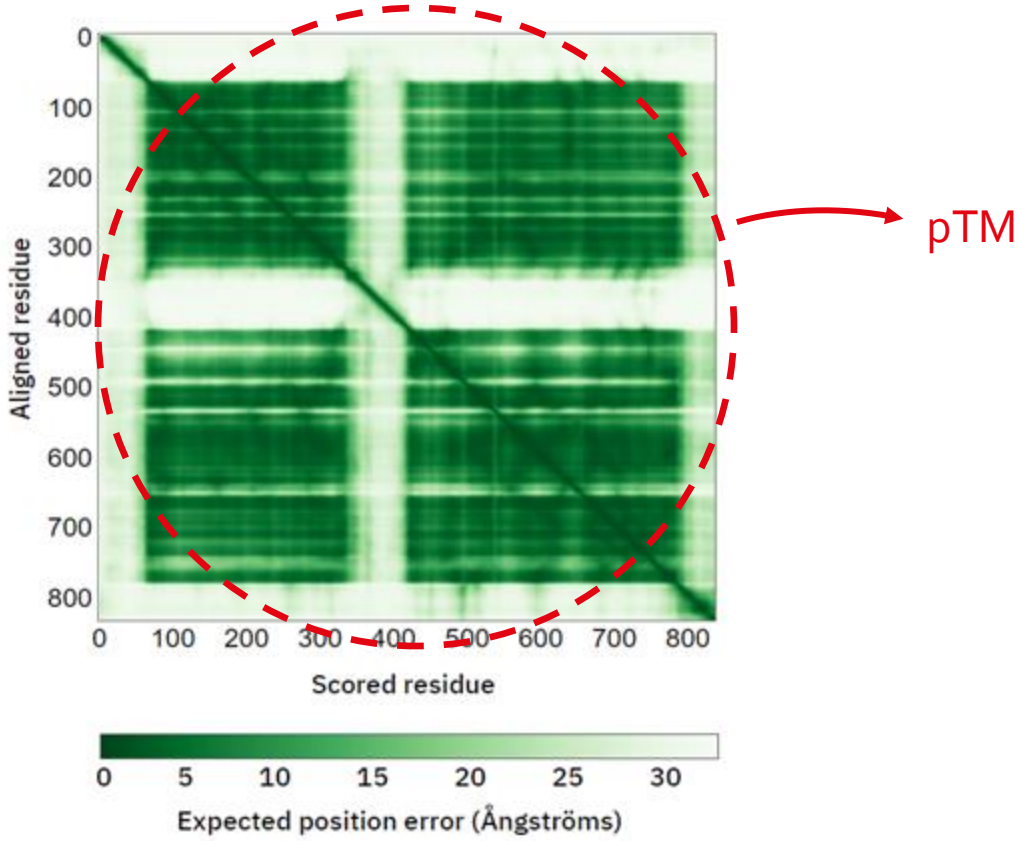


AlphaFold ID: AF-Q8T2T7-F1

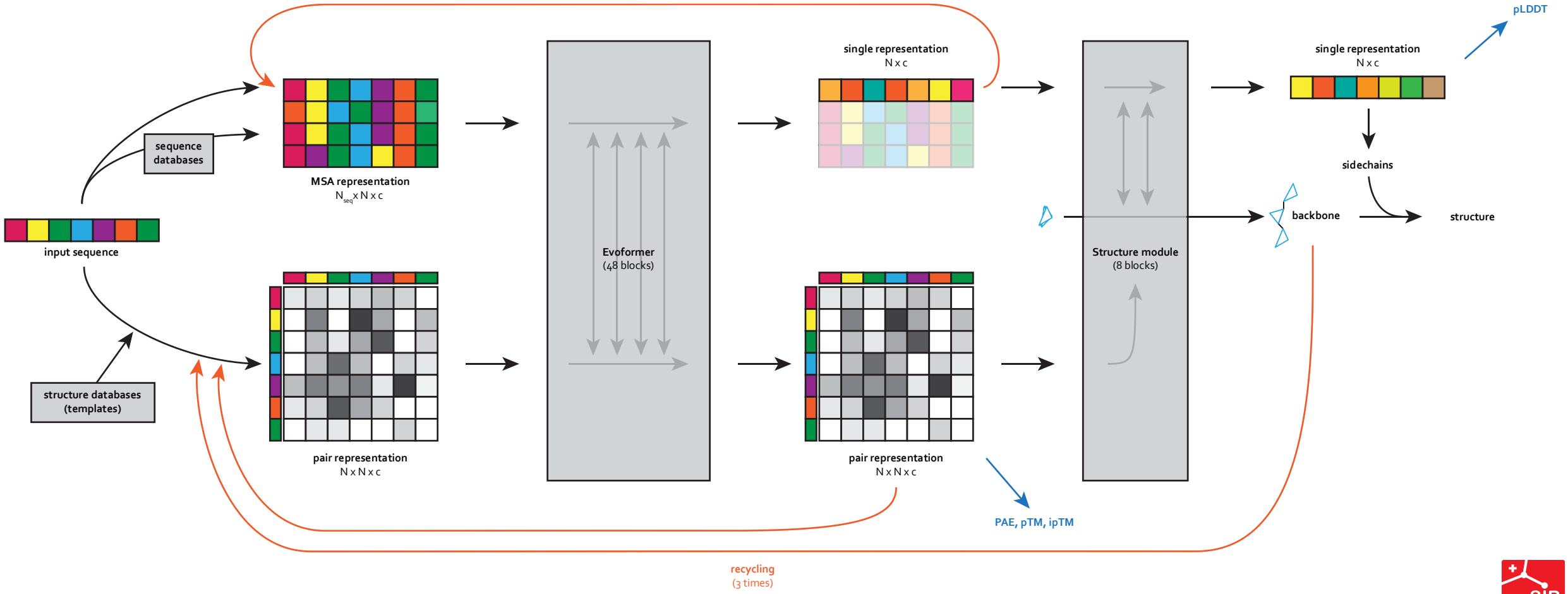
pTM – predicted template modelling score

- One number for the whole structure
- Accuracy of the global structure of the protein, relatively insensitive to localized inaccuracies
- Can be thought of as being calculated from PAE

- Range: 0-1
  higher is better



Confidence metrics in the architecture



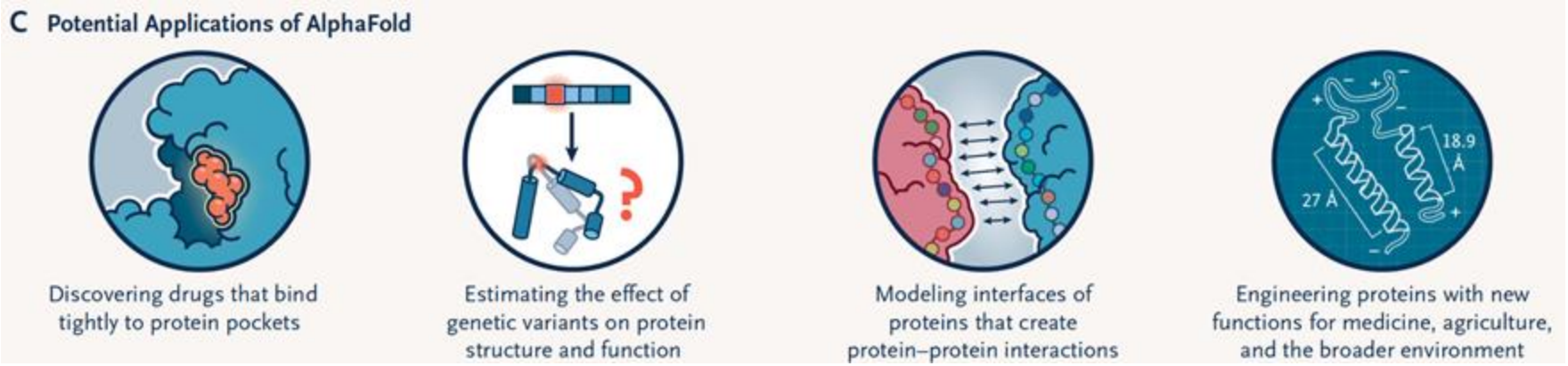
Outputs

- AlphaFold2 outputs 5 structures (one for each of the 5 models)

- Ranks them according to the mean pLDDT

Limitations of AlphaFold2

- Not sensitive to point mutations
- Struggles with a weak MSA
  - Highly variable sequences (e.g. antibodies, viruses)
  - "Orphan" proteins (those with few close relatives / similar sequence)
- Only single protein chains; no protein-protein complexes, ligands, nucleic acids

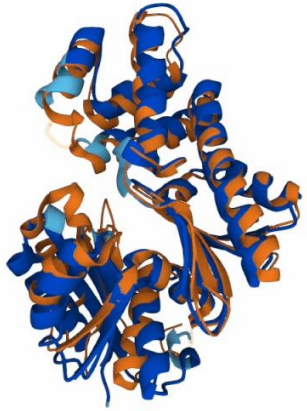


A Holy Grail — The Prediction of Protein Structure (Russ B. Altman, New England Journal of Medicine, 2023)

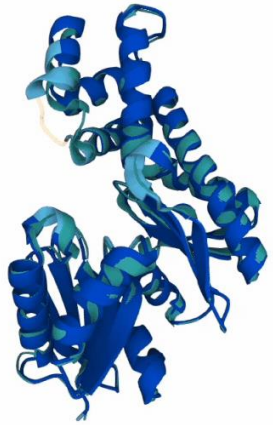
Limitations of AlphaFold2

Conformational diversity

- AlphaFold2 is trained on static structures
- Tricks for diversity (templates; smaller MSA)



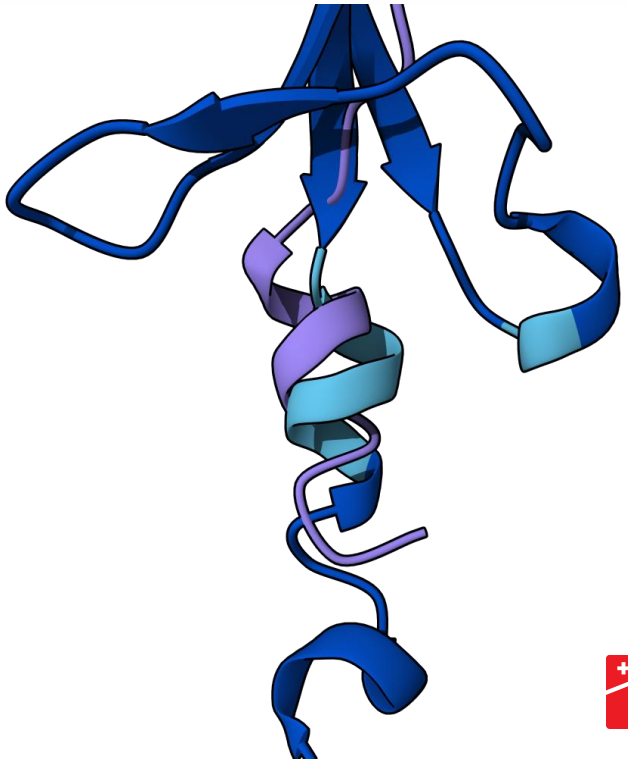
Sugar-bound
RMSD 3.02 Å



Sugar-free
RMSD 0.67 Å

Memorization artefacts

- eukaryotic translation initiation factor 4E-binding protein 2 (4E-BP2, UniProt ID: Q13542)
  - Lacks a structure in unbound state
  - Adopts a helical structure in bound state (PDB ID: 3AM7)
  - AF2 was trained on the bound state, so it predicts that



Swiss Institute of Bioinformatics

DAY 1, PART 3

Protein-protein complexes (AlphaFold-Multimer)

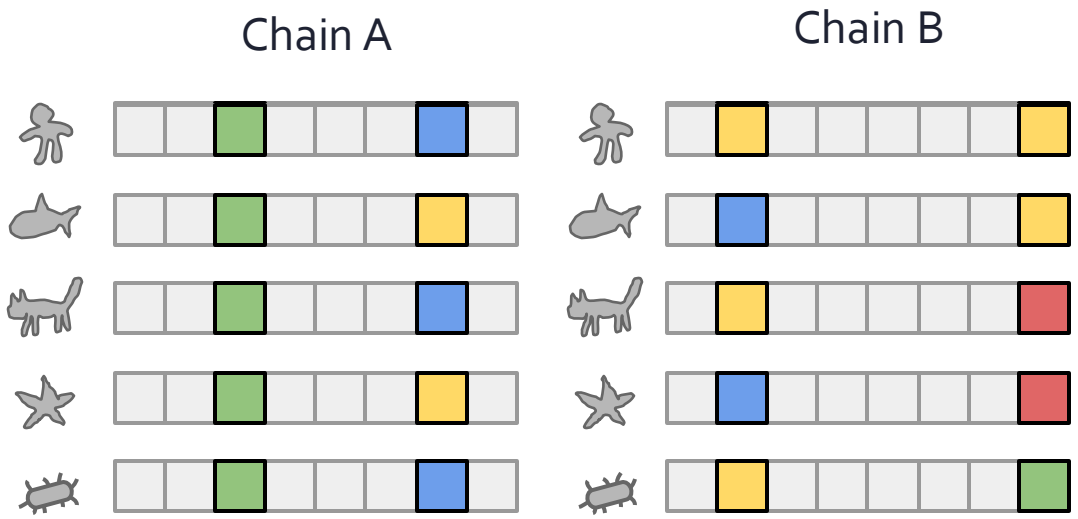
Diana Rapota, Rok Breznikar, Janani Durairaj

23-24 June 2026

elixir SWITZERLAND

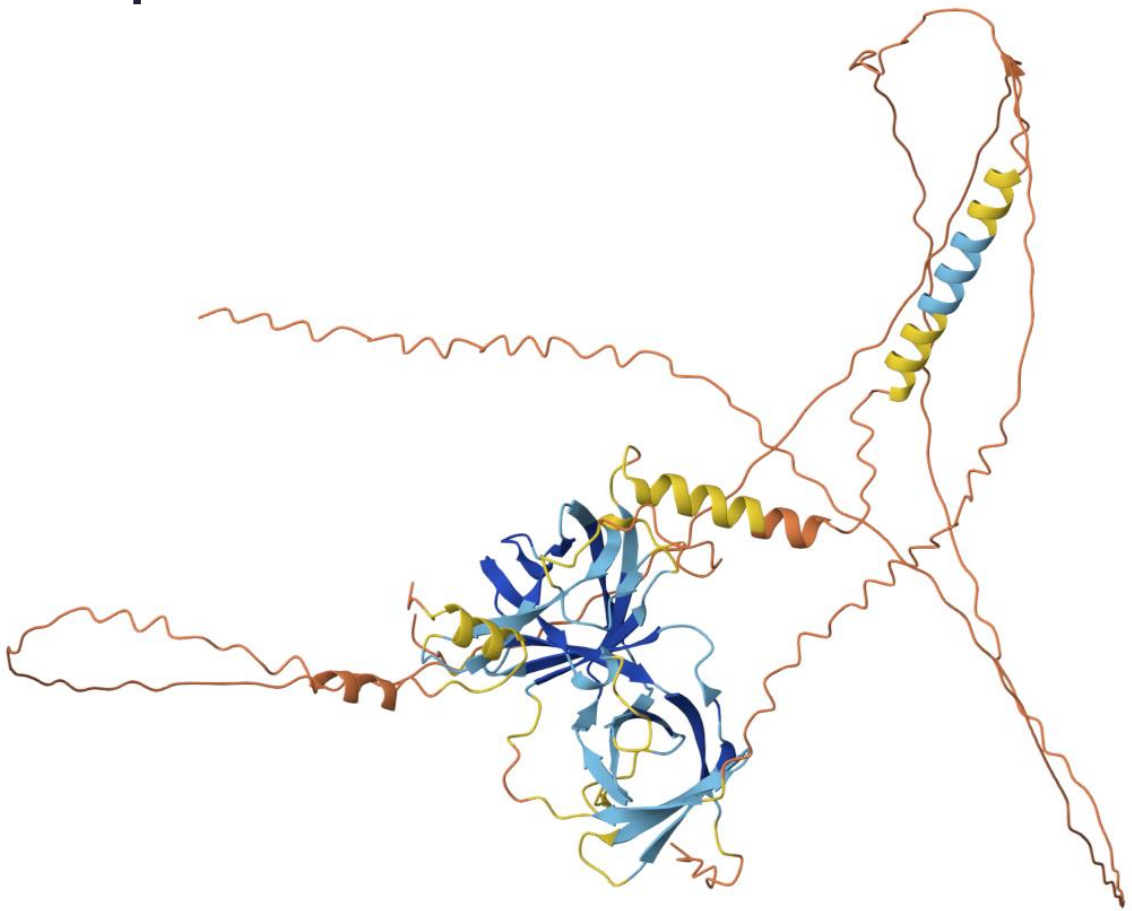
AlphaFold-Multimer

- Same architecture as AlphaFold2
- Same databases
- Retrained, different training data cutoff (more structures)
- Builds **paired MSA**



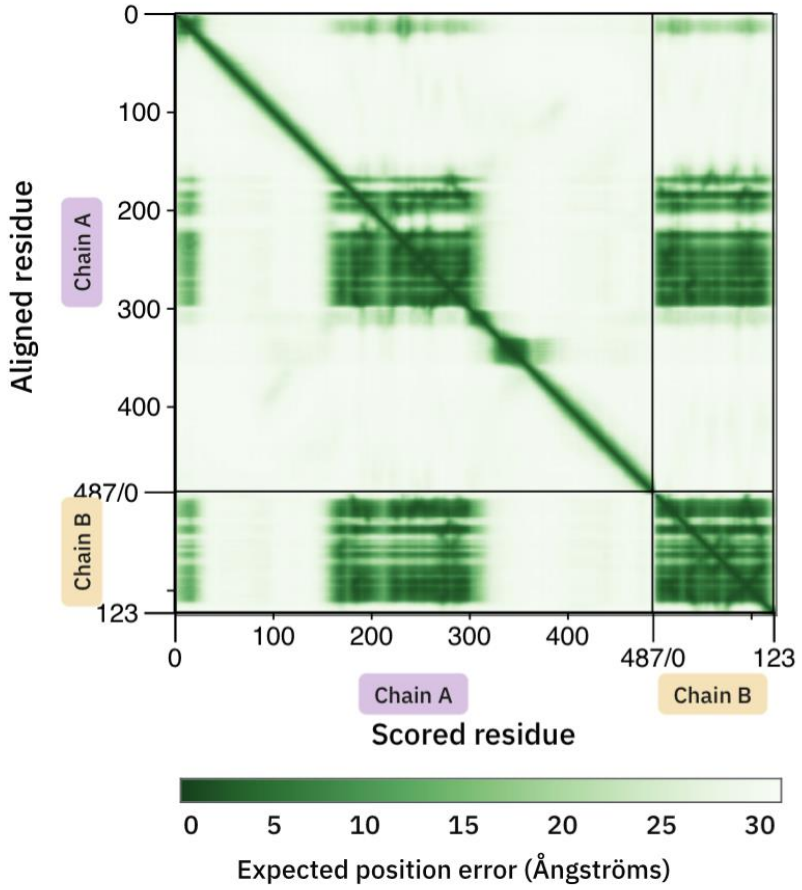
Confidence metrics

pLDDT (local confidence)



Very high (pLDDT > 90) | Confident (90 > pLDDT > 70) | Low (70 > pLDDT > 50) | Very low (pLDDT < 50)

PAE (global confidence)



AlphaFold ID: AF-0000000203959827

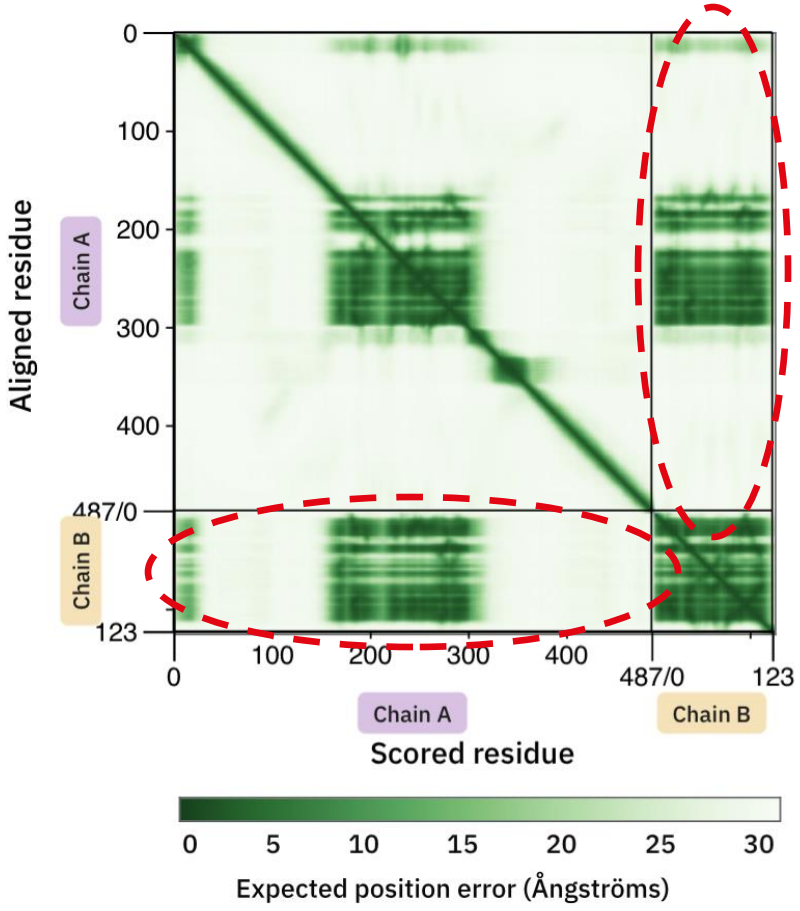
Confidence metrics

pTM – predicted template modelling score

- derived from PAE
- if one of the interacting partners is larger than the other, it dominates the pTM
- pTM above **0.5** suggests the complex fold may be broadly correct

ipTM – interface predicted template modelling score

- only for the interface
- measures the accuracy of the predicted relative positions of the subunits within the complex
- ipTM **> 0.8**: confident; **0.6-0.8**: gray zone; **< 0.6**: possibly failed
- regions with low pLDDT and disordered regions may negatively affect ipTM



Confidence metrics

ipSAE – interaction prediction Score from Aligned Errors

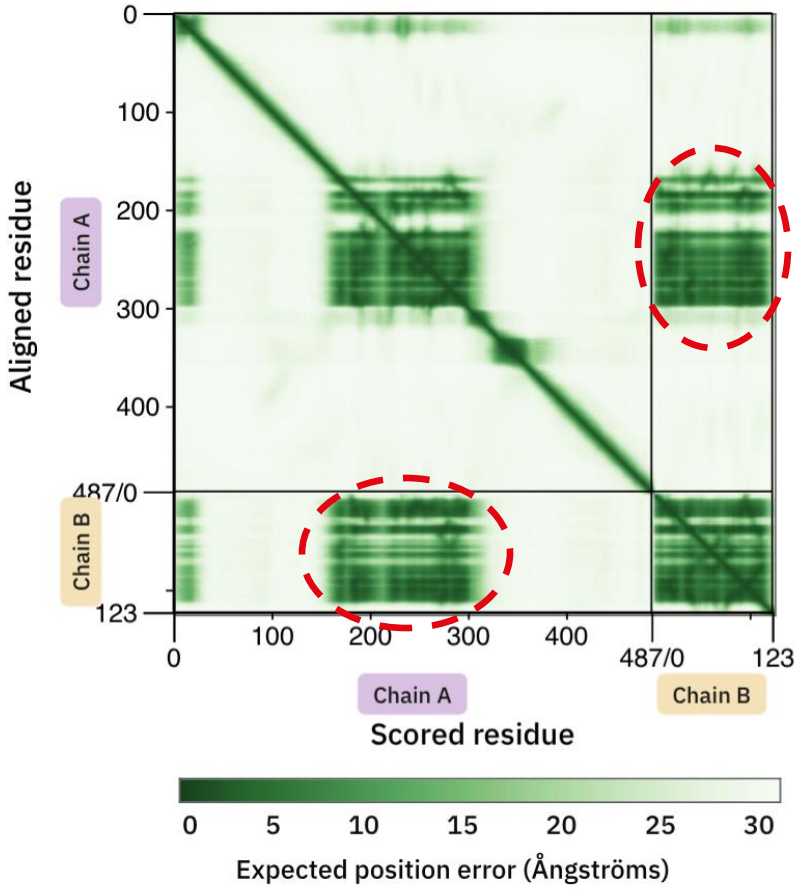
- improved ipTM
- not calculated by AlphaFold, calculated separately

ipSAE ≥ 0.8: The predicted homodimer interface is modelled with high accuracy

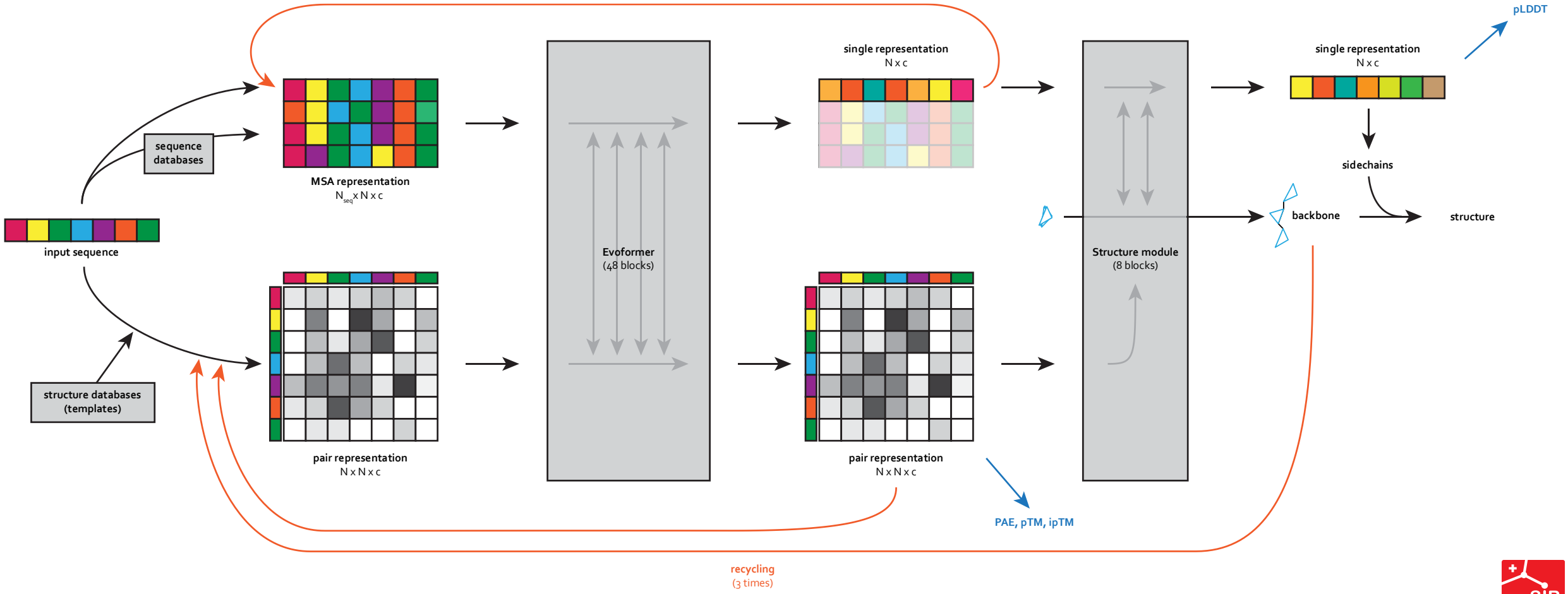
0.7- 0.8: The predicted interface is likely correct with a well-arranged interaction

0.6 - 0.7: The interaction may be correct but should be interpreted cautiously

< 0.6: The predicted interaction is unlikely to represent a reliable complex



Confidence metrics in the architecture



Ranking of structures

- For complexes, the structures are ranked mostly by ipTM

$$\text{model confidence} = 0.8 \cdot \text{ipTM} + 0.2 \cdot \text{pTM}$$

Swiss Institute of Bioinformatics

DAY 1, PART 4

Accessing AlphaFold2

Diana Rapota, Rok Breznikar, Janani Durairaj

23-24 June 2026

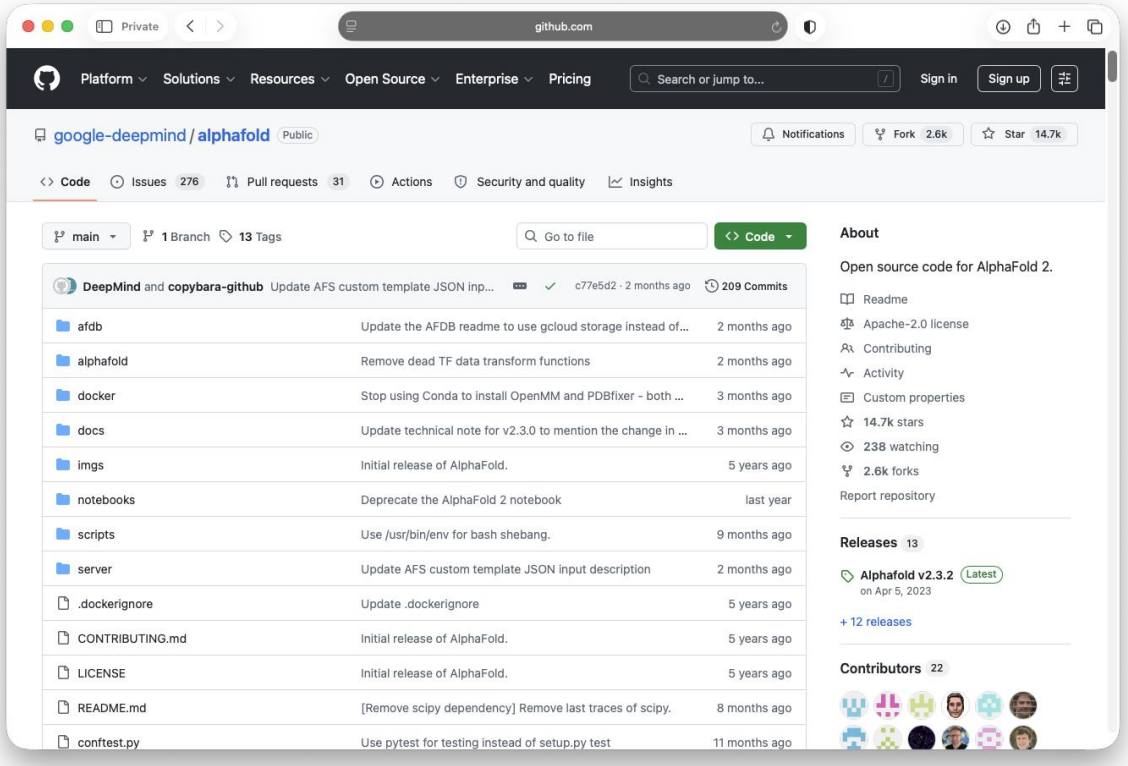
elixir SWITZERLAND

3 ways to access AlphaFold2 predictions

1. Installing and running AlphaFold2
2. Running online implementations of AlphaFold2 (ColabFold)
3. Searching a database of structures that have already been predicted using AlphaFold2 (AFDB)

AlphaFold2 open-source code

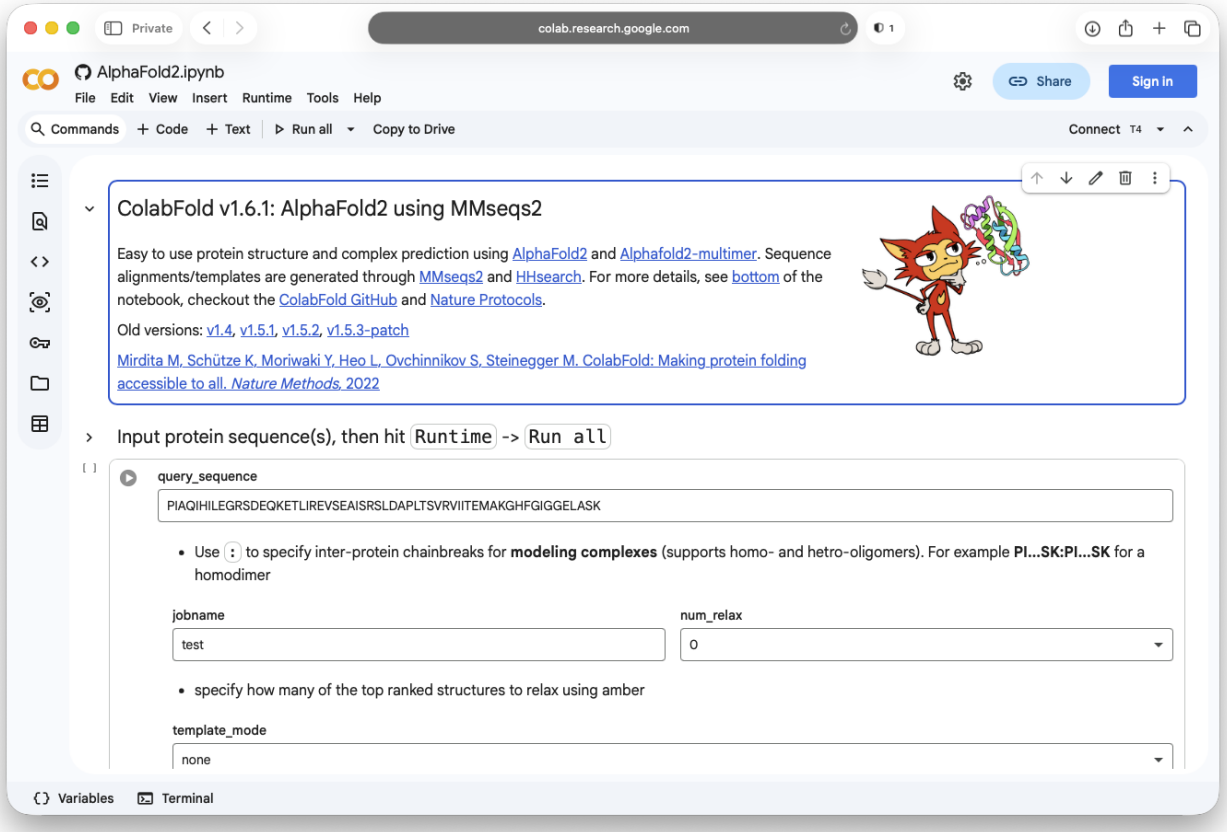
- Can be found on GitHub



- Requirements:
  - Linux
  - 3 TB of disk space for genetic databases
  - Modern GPU
    - GPU RAM determines the maximum size of a protein
    - 40 GB: ~5000 residues

ColabFold

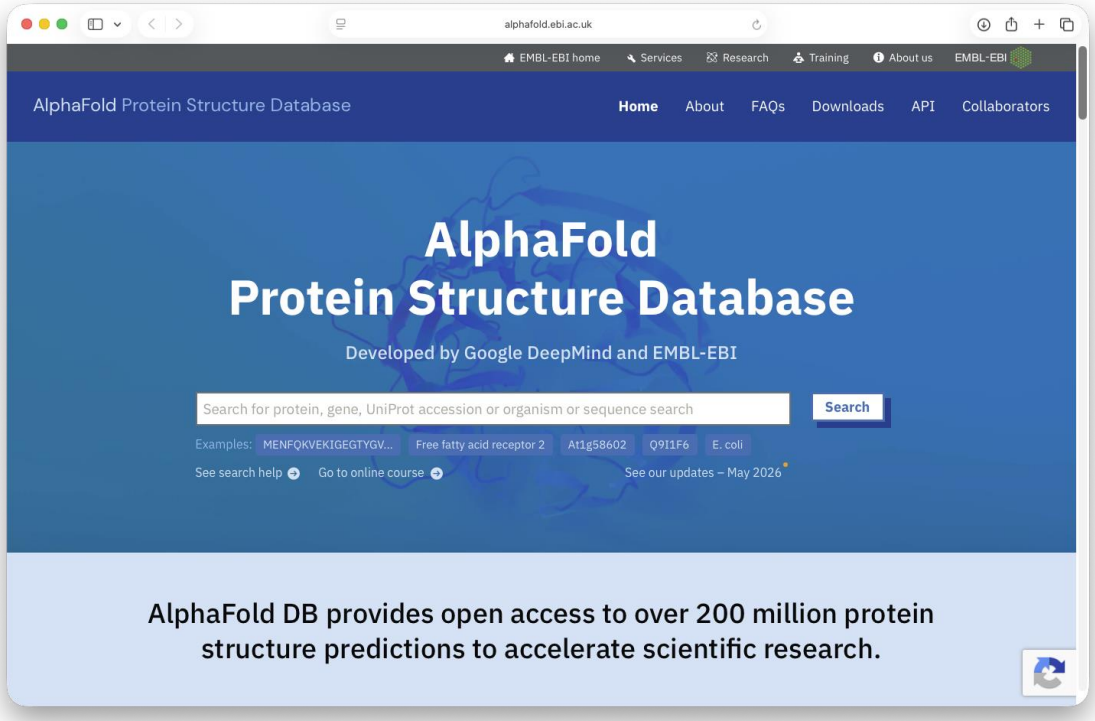
- No need to install the software
- No need to have powerful computing resources

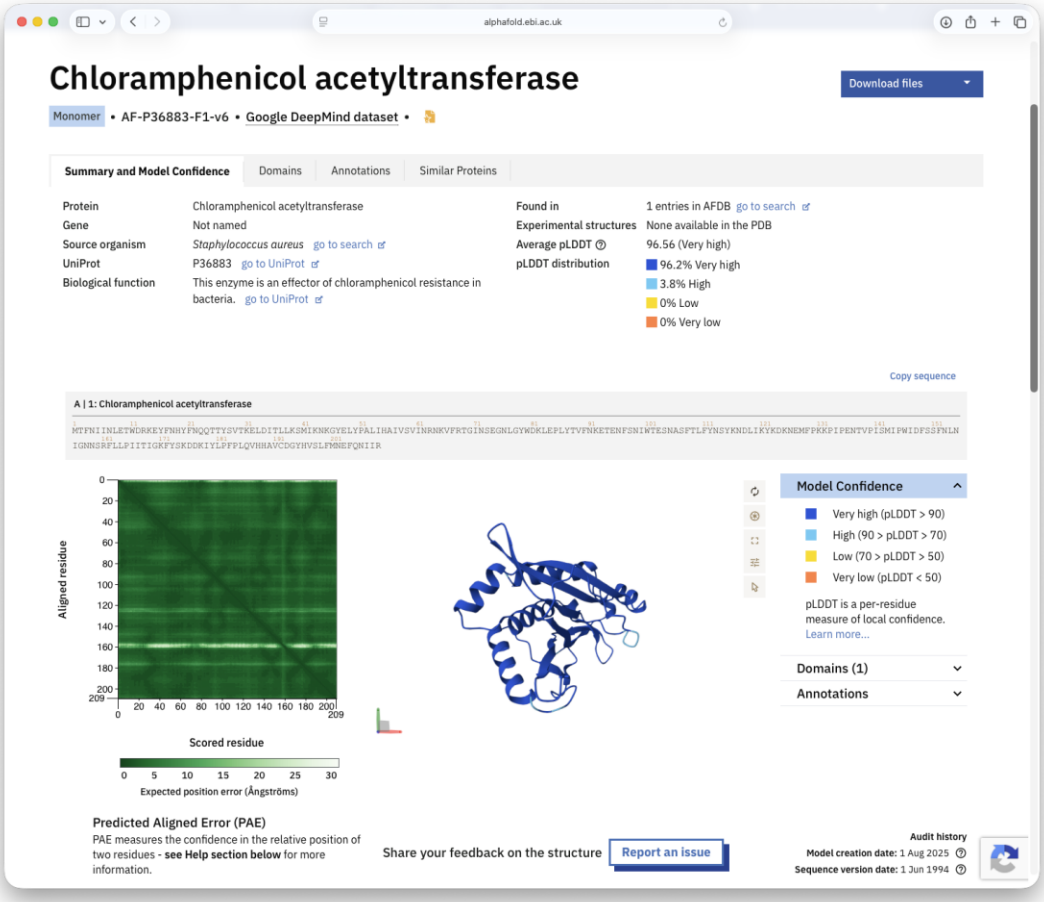


- Limits on the size of the protein (based on the GPU allocated to you)
  - 2500 residues for monomers, 4000 residues for complexes

AlphaFold Protein Structure Database (AFDB)

Over 200 million predicted protein structures





AlphaFold Protein Structure Database (AFDB)

- Over 200 million predicted protein structures (monomers)
- Millions of protein complexes (since 2026)[1]

- Up to a certain size of the protein
  - 2700 residues (Swiss-Prot) or 1280 residues (UniProt)
- Not so many viral proteins (they evolve fast → hard to make MSA; they are polyproteins; viral complexes sometimes too big to model)
  - Different database for viral proteins: Viral AlphaFold Database (VAD)[2]

- Only 1 conformation per protein
- No control over the prediction
- Only model 2 of AlphaFold2

[1] https://research.nvidia.com/labs/dbr/assets/data/manuscripts/afdb.html
[2] https://www.science.org/doi/10.1126/sciadv.adz8560