Chapter 14 & 15

Protein Structure Prediction

Overview

1. Introduction
2. Introduction to Biological Databases
3. Pairwise Sequence Alignment
4. Database Similarity Searching
5. Multiple Sequence Alignment
6. Profiles and Hidden Markov Models
7. Protein Motifs and Domain Prediction
8. Gene Prediction
9. Promoter and Regulatory Element Prediction
10. Phylogenetics Basics
11. Phylogenetic Tree Construction Methods and Programs
12. Protein Structure Basics
13. Protein Structure Visualization, Comparison and Classification
14. Protein Secondary Structure Prediction
15. Protein Tertiary Structure Prediction
16. RNA Structure Prediction
17. Genome Mapping, Assembly and Comparison
18. Functional Genomics
19. Proteomics

Why is structure prediction useful?

- Protein structure holds the key to understanding protein function
- However, experimental structure determination is slow, labour-intensive and expensive
- Experimental protein structure determination cannot keep up with genome sequencing
- Some proteins are not amenable to any of our experimental structure determination methods

Secondary structure prediction: why?

- Secondary structure predictions can be useful for classification of proteins and for separation of protein domains and functional motifs
- Correctly identifying secondary structure elements (SSE) can help to guide sequence alignment or improve existing sequence alignment of distantly related sequences
- Secondary structure prediction could be seen as a first step towards tertiary structure prediction

Chou-Fasman algorithm (1974)

- The earliest (and simplest) form of protein structure prediction
- Determines the *propensity* ("intrinsic tendency") of each amino acid to be in an α-helix, a β-strand and a turn
- Uses observed frequencies found in protein crystal structures

For example:

- Alanine (A), glutamic acid (E) leucine (L) and methionine (M) are particularly often found in α-helices
- Glycine (G) and proline (P) are rarely found in α-helices

Chou-Fasman algorithm

Propensity for a specific residue **X** to be in a helix is calculated as:

$$P(\mathbf{X}) \text{ within } \alpha\text{-helices} \;/\; P(\mathbf{X}) \text{ in proteins}$$

Number of residues **X** in helices
Number of residues in helices

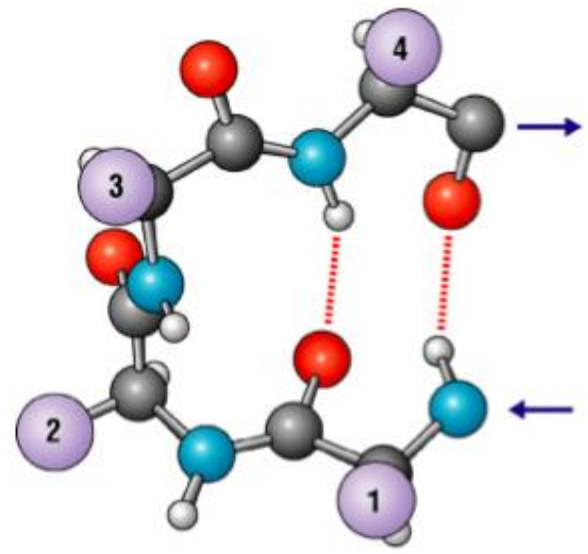
Number of residues **X** in all proteins
Number of residues in all proteins

Chou-Fasman algorithm: propensity values

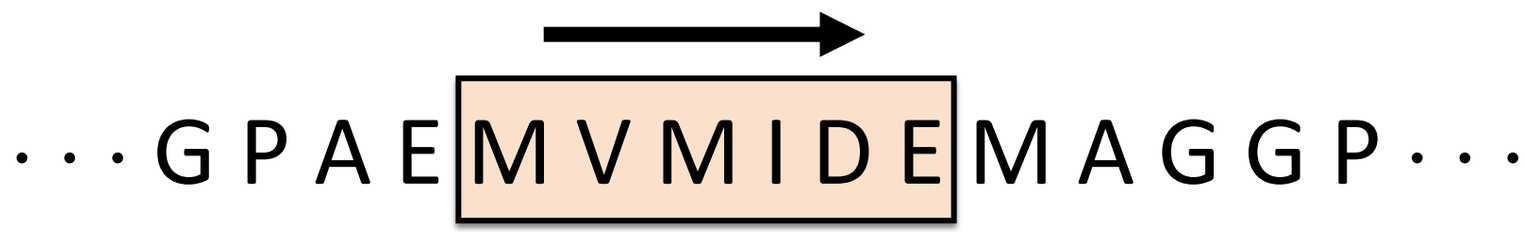
TABLE 14.1. Relative Amino Acid Propensity Values for Secondary Structure Elements Used in the Chou–Fasman Method

Amino Acid	(α-Helix)	P (β-Strand)	P (Turn)
Alanine	1.42	0.83	0.66
Arginine	0.98	0.93	0.95
Asparagine	0.67	0.89	1.56
Aspartic acid	1.01	0.54	1.46
Cysteine	0.70	1.19	1.19
Glutamic acid	1.51	0.37	0.74
Glutamine	1.11	1.11	0.98
Glycine	0.57	0.75	1.56
Histidine	1.00	0.87	0.95
Isoleucine	1.08	1.60	0.47
Leucine	1.21	1.30	0.59
Lysine	1.14	0.74	1.01
Methionine	1.45	1.05	0.60
Phenylalanine	1.13	1.38	0.60
Proline	0.57	0.55	1.52
Serine	0.77	0.75	1.43
Threonine	0.83	1.19	0.96
Tryptophan	0.83	1.19	0.96
Tyrosine	0.69	1.47	1.14
Valine	1.06	1.70	0.50

Turn:

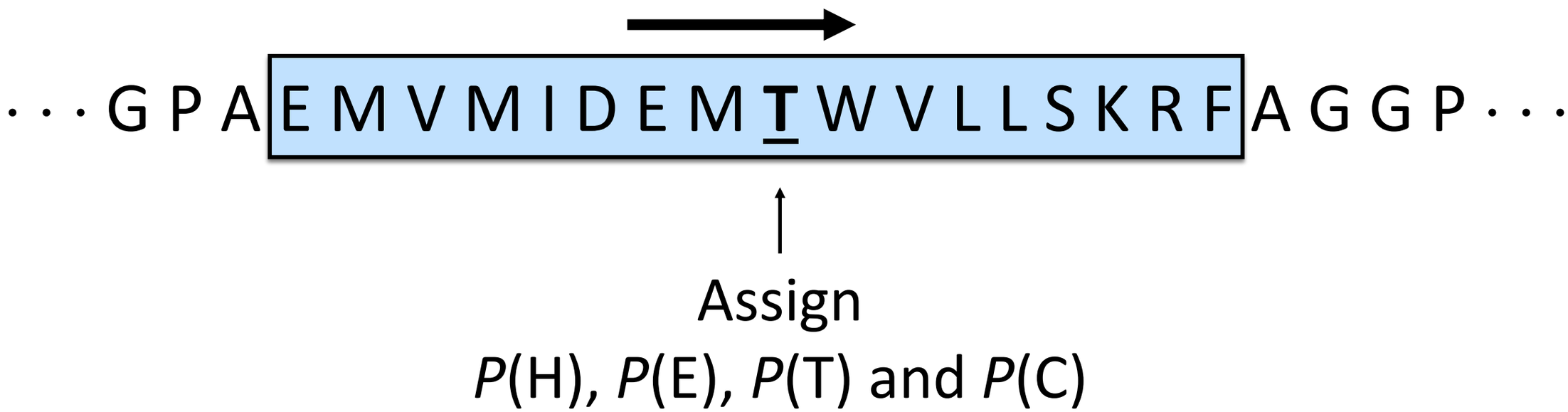


Chou-Fasman algorithm



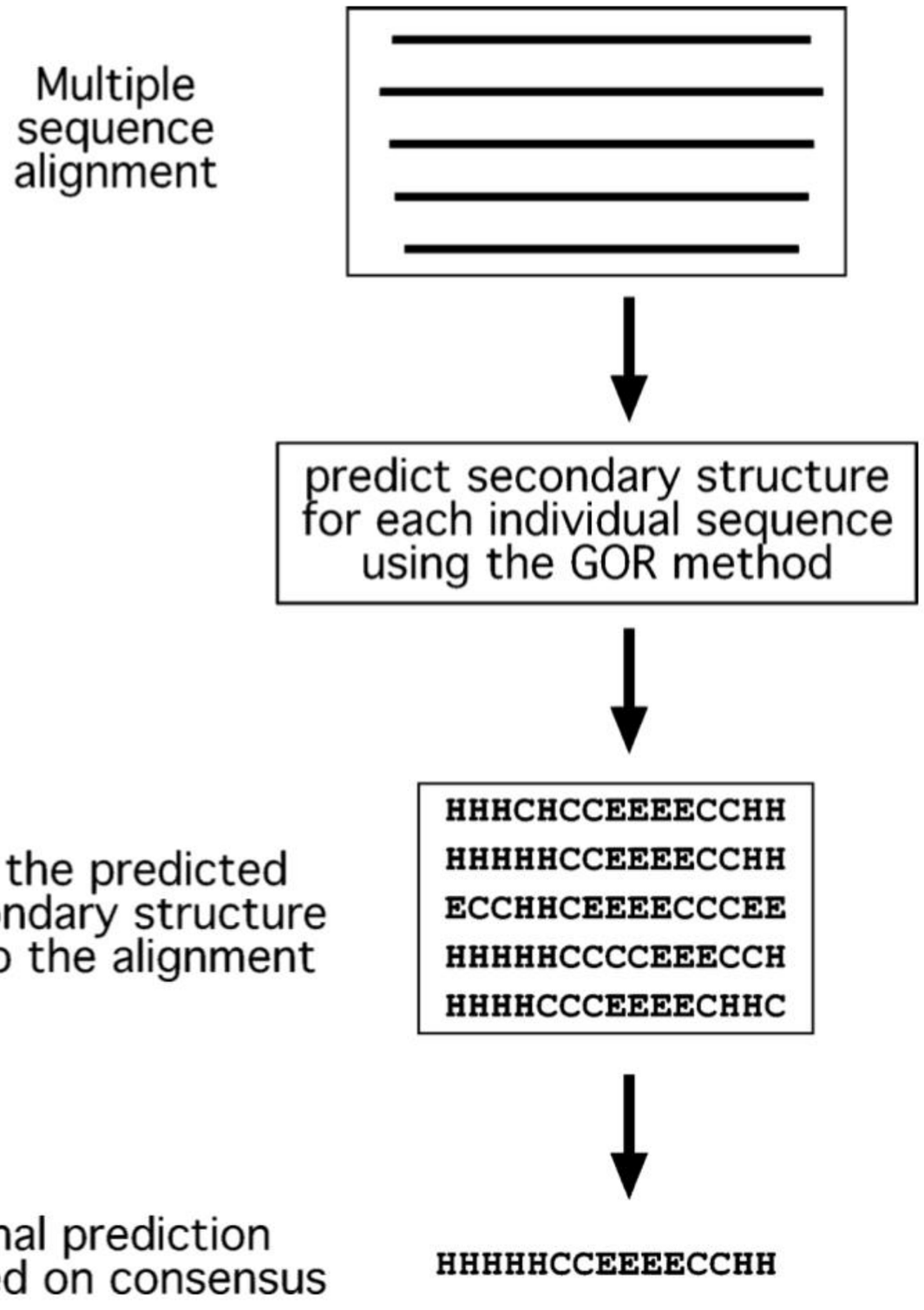
- A window of 6 amino acids with 4 amino acids having *P*(α-helix) > 1.0 is predicted as an α-helix ("nucleation")
- Helical region is extended in both directions until the average *P*(α-helix) score for a window of 4 contiguous amino acids drops below 1.0
- For β-strands the window size for nucleation is 5 residues of which 3 need to favour β-strands
- Overlap: highest score decides, *e.g.* if Σ*P*(α) > Σ*P*(β) the overlap region is declared an α-helix

GOR method (Garnier, Osguthorpe, Robson)



- Examines a 17-residue window and sums up propensity scores for all residues for each of four states: helix (H), strand (E), turn (T) and coil (C)
- The highest score decides the conformational state of the center residue in the window (9^{th} position)

Adding homology information to predictions



- Combine secondary structure predictions of homologous sequences (>35 % identity) in an alignment
- Close homologs are likely to adopt the same structure
- Final secondary structure prediction based on consensus

Improvements in secondary structure prediction

- Using (much) larger structure databases to derive propensities, *e.g.* more recent versions of the PDB
- Neural networks trained on (and using as input) sequence profiles derived from multiple sequence alignments
- However: maximum accuracy remains below ~80%

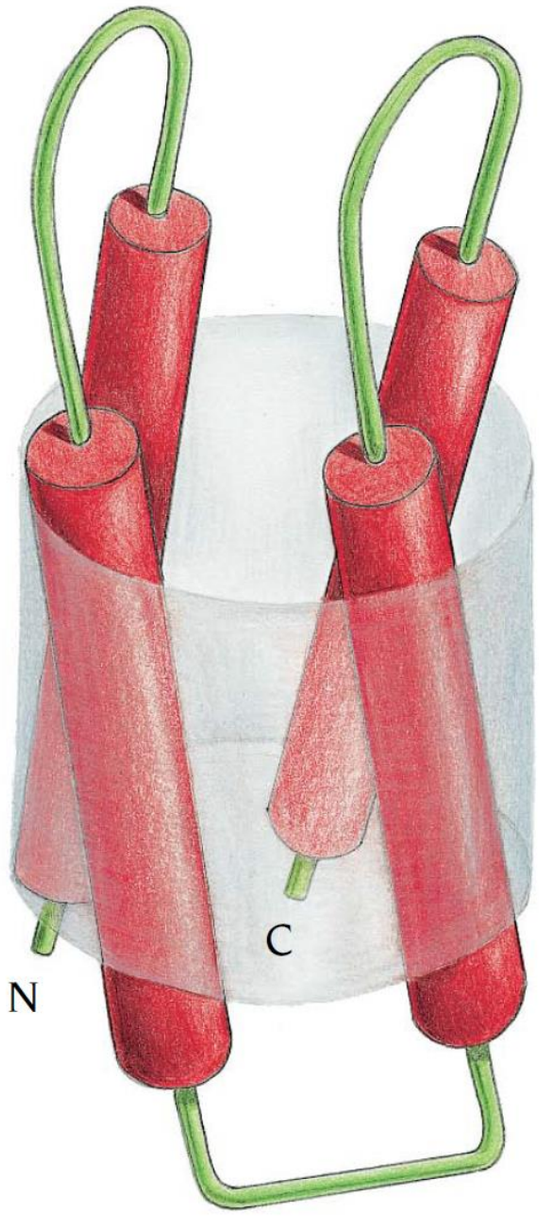
Prediction accuracy remains < 80%

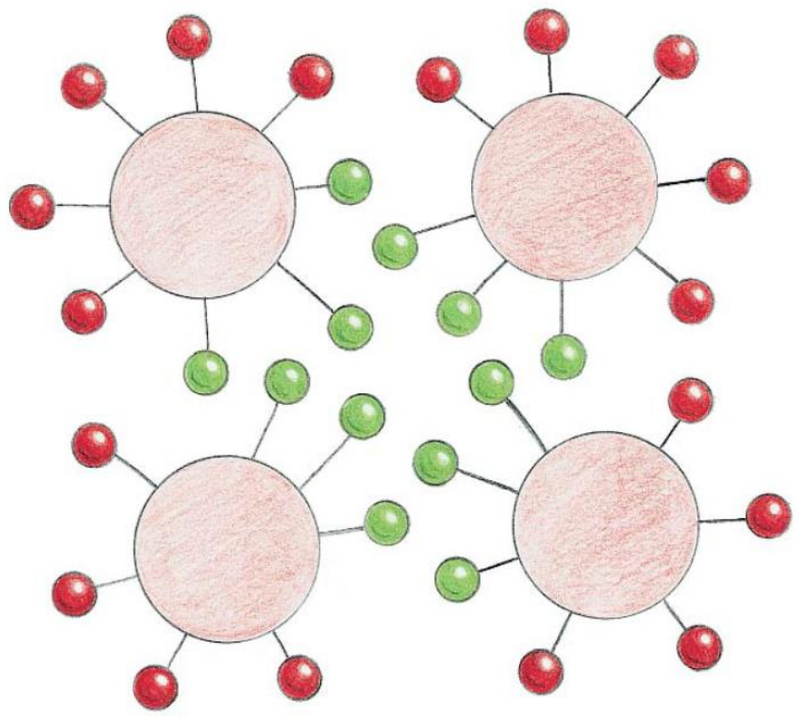
TABLE 14.2. Comparison of Accuracy of Some of the State-of-the-Art Secondary Structure Prediction Tools

Methods	Q_3 (%)
Porter	79.0
SSPro2	78.0
PROF	77.0
PSIPRED	76.6
Pred2ary	75.9
Jpred2	75.2
PHDpsi	75.1
Predator	74.8
HMMSTR	74.3

Note: The Q_3 score is the three-state prediction accuracy for helix, strand, and coil.

Why is secondary structure prediction <u>inherently</u> inaccurate?





Predictions solely rely on <u>local</u> sequence information and do not take into account <u>long-range</u> interactions

Special cases of secondary structure prediction

- Integral membrane proteins
- Coiled coils

Transmembrane proteins

- Around 30% of all cellular proteins are transmembrane proteins
- Transmembrane proteins perform a wide variety of important functions in a cell, *e.g.*
  - signal transduction
  - cross-membrane transport
  - energy conversion
- Transmembrane proteins often serve as drug targets

Categories of integral membrane proteins

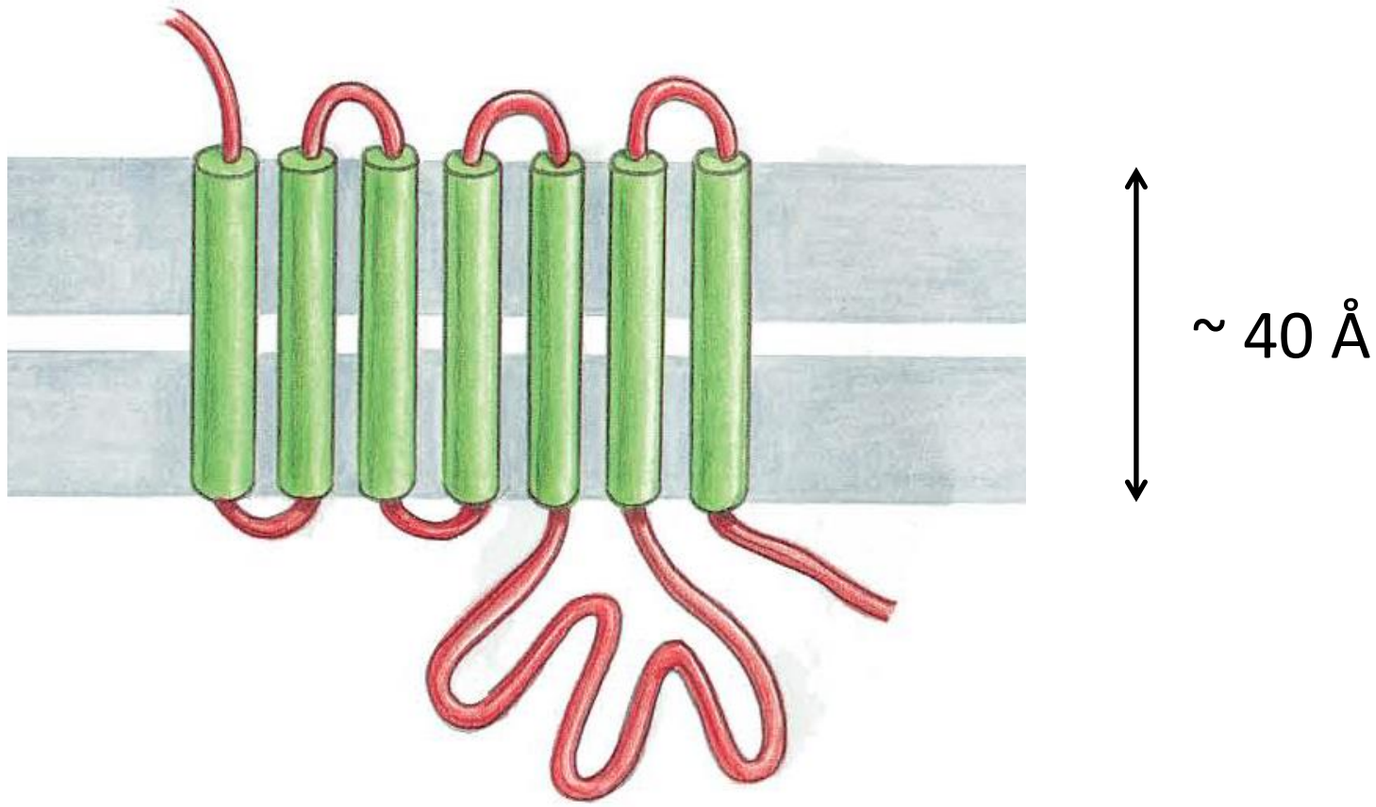
α-helical type

Most transmembrane proteins contain only α-helices

β-barrel type

Some membrane proteins form a cylindrical structure composed of antiparallel β-sheets (*e.g.* in the outer membrane of gram-negative bacteria)

Prediction of α-helical membrane proteins



- Hydrophobic helices are normally separated by hydrophilic (often positively charged) loops
- The α-helices generally run (more or less) perpendicular to the membrane plane
- Average length between 17 and 25 residues

Prediction of α-helical membrane proteins

Positive-inside rule:

- Residues on the cytosolic side are more positively charged than those at the luminal/periplasmic side
- This allows the prediction of the <u>orientation</u> of the secondary structure elements

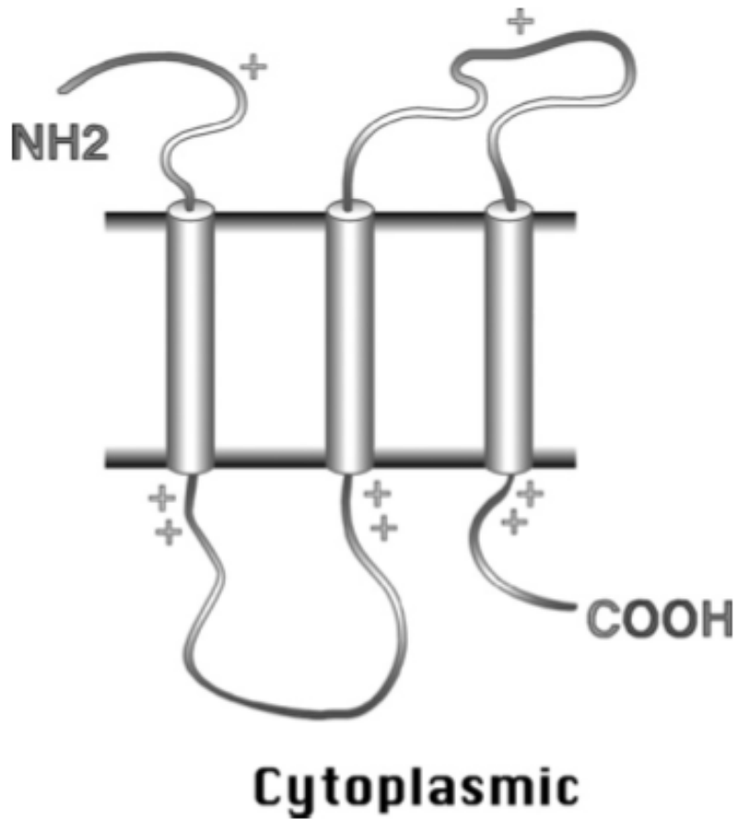
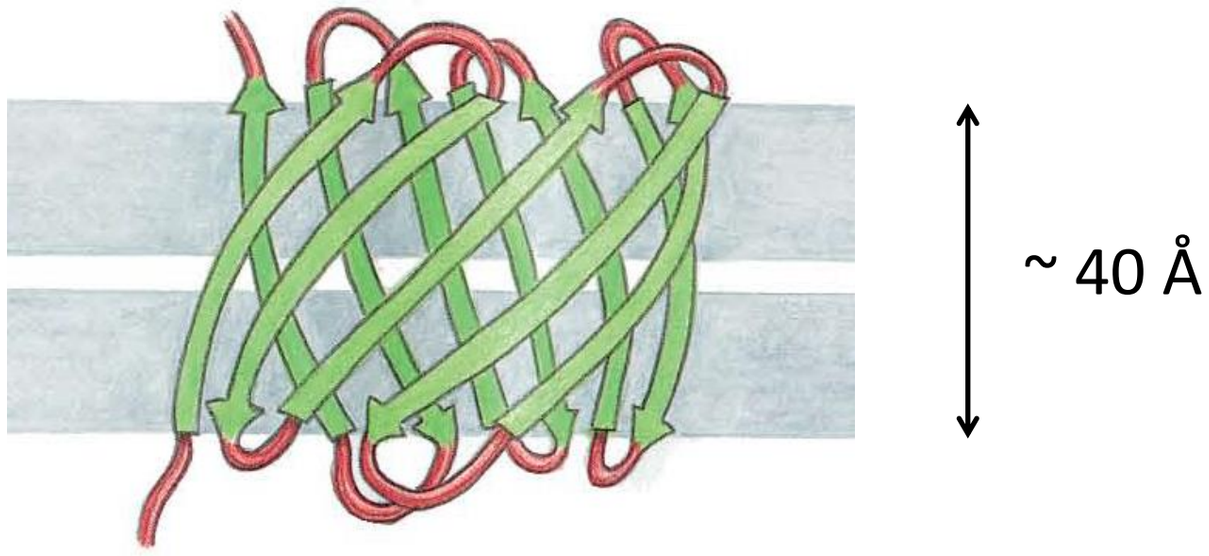


Figure 14.3: Schematic of the positive-inside rule for the orientation of membrane helices. The cylinders represent the transmembrane α–helices. There are relatively more positive charges near the helical anchor on the cytoplasmic side than on the periplasmic side.

DeepTMHMM

- https://services.healthtech.dtu.dk/services/DeepTMHMM-1.0
- Program based on an HMM algorithm
- Trained based on a set of well-characterized helical membrane proteins
- Probability of an α-helical domain, orientation, number of transmembrane helices and boundaries of the helices are predicted
- Can be used to distinguish between globular proteins and membrane proteins

Prediction of β-barrel membrane proteins



The β-strands forming a transmembrane pore are *amphipathic*:

- Each β-strand contains 10 to 22 residues
- Every second residue is hydrophobic and faces the lipid bilayer
- Residues facing the inside of the β-barrel are hydrophilic

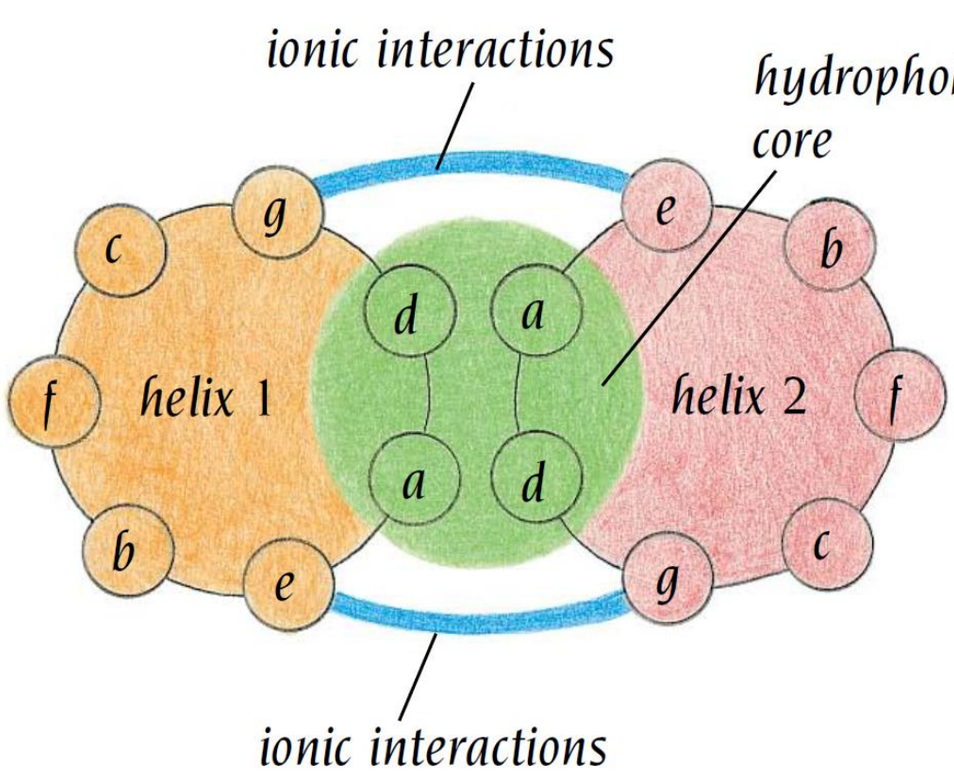
TBBpred

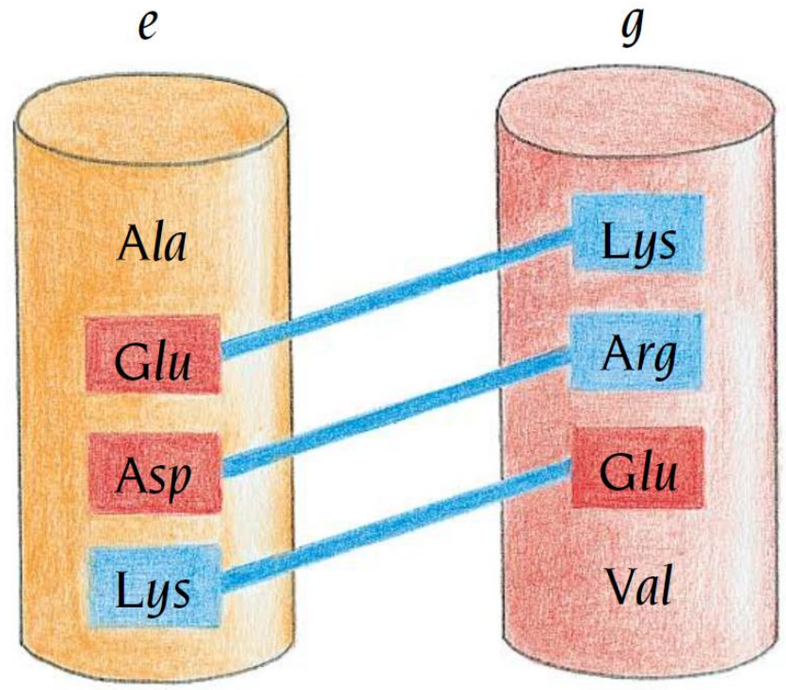
- https://webs.iiitd.edu.in/raghava/tbbpred/index.html
- Uses a neural network to predict transmembrane β-barrel regions
- Network is trained with a limited number of transmembrane β-barrel protein structures

Coiled coils

- Superhelical structures involving two to more interacting α-helices
- Individual α-helices twist and wind around each other to form a coiled bundle
- Often involved in protein dimerisation
- Often involved in transcription regulation (dimeric transcription factors) or in the maintenance of cytoskeletal structures

Coiled coil prediction





- Coiled coils have repeats of seven residues (*heptads*) enabling side-chain interactions at the interface
- The 1st and 4th residues of every heptad are <u>hydrophobic</u> facing the helical interface
- Sequence periodicity is the basis for prediction

DeepCoil

- https://toolkit.tuebingen.mpg.de/tools/deepcoil
- Based on a neural network

2ZIP

- https://www.lirmm.fr/2zip
- Specifically predicts *leucine zippers*:

L-X(6)-L-X(6)-L-X(6)-L

<u>Tertiary</u> structure prediction

The three computational approaches to protein three-dimensional structure modelling and prediction:

- **Homology modelling**: structures are predicted based on experimentally determined structures of related sequences
- **Threading**: identifies proteins that are structurally similar, independent of sequence similarities
- ***Ab-initio***: based on fundamental principles governing protein folding without the use of structural templates

Homology modelling

- Also known as *comparative* modelling
- The principle: if two proteins share high sequence similarity (> 30%), they are likely to have virtually identical three-dimensional structures
- Modifies an existing model (the *template*) based on sequence alignment

Homology modelling

Steps required:

1. Template selection: identification of homologous sequences with corresponding structures in the PDB
2. Alignment of target and template sequences
3. Main chain atom framework construction for target
4. Addition and optimisation of side chain atoms and loops
5. Refinement and optimisation of entire model according to energy criteria
6. Evaluation of overall quality of model obtained

Homology modeling

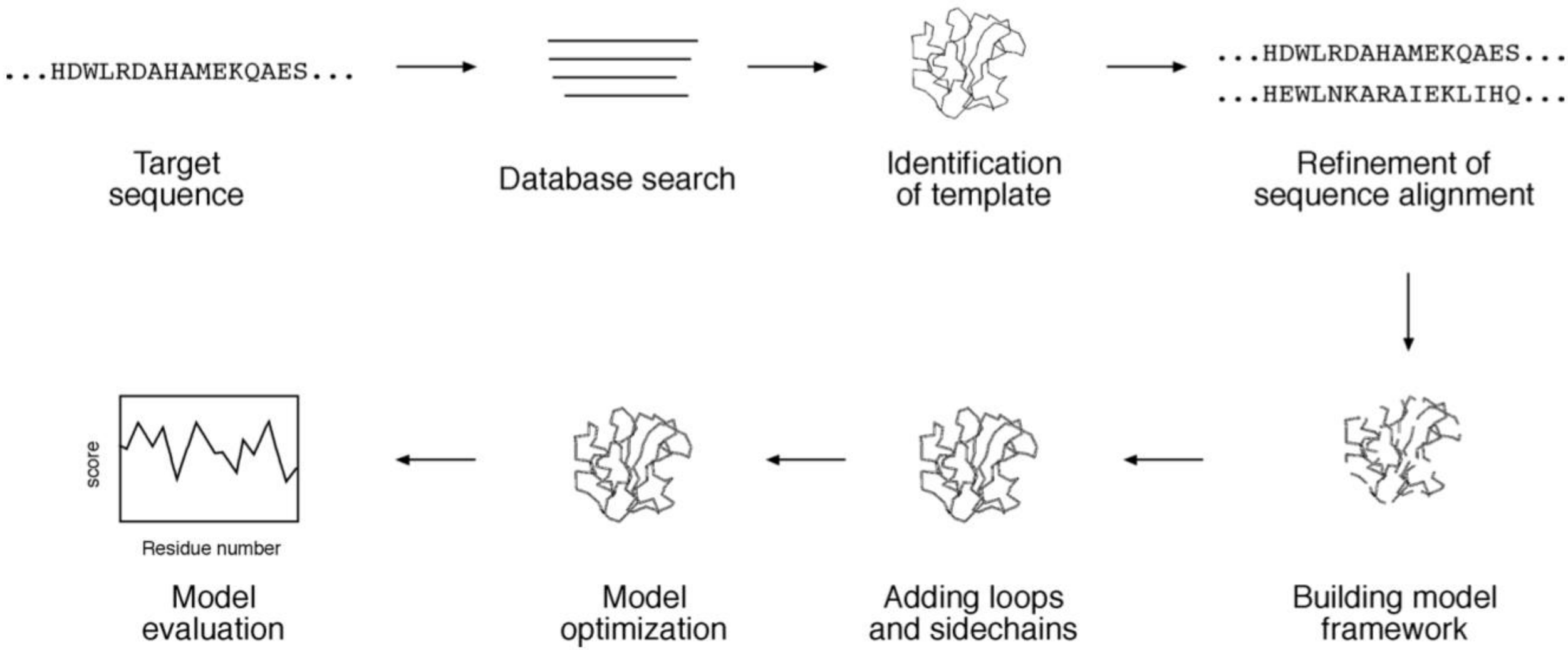


Figure 15.1: Flowchart showing steps involved in homology modeling.

Template selection

- *E.g.* BLAST is used to search the sequences of all protein structures in the PDB
- The database protein should have <u>at least</u> 30% sequence identity with the query sequence to be selected as template, ideally more
- If multiple database structures with comparable sequence similarity to the protein of interest are found:
  - ➢ Choose the one with the highest resolution
  - ➢ Choose the one with most appropriate cofactors, *etc.*

Sequence alignment

- Sequences of the template and target proteins need to be (re)aligned to obtain the best possible alignment
- This is often the most critical step, directly affecting the quality of the final model
- Best possible (exhaustive) alignment algorithms should be used
- Visual inspection and manual refinement

Model building

- For aligned residues that are the same, the coordinates of the template protein are simply copied
- Non-identical residues: only the backbone atoms are taken
- For non-identical residues, side chains are initially obtained from a rotamer library and later refined using a force field
- Insertions in the alignment are treated as loops and modelled either *ab initio* or by searching the PDB for a loop of the right length that "fits", *i.e.* can bridge the gap

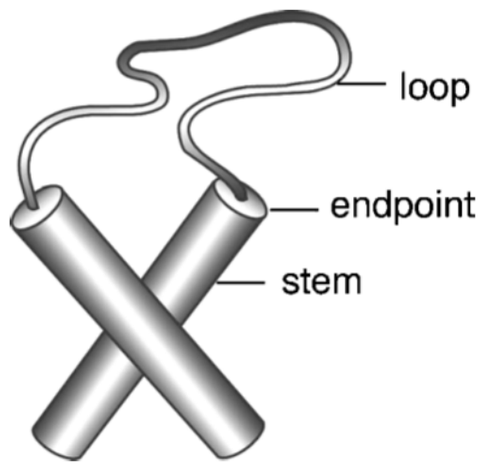


Figure 15.2: Schematic of loop modeling by fitting a loop structure onto the endpoints of existing stem structures represented by cylinders.

Completely automated modelling servers/programs

Swiss-Model (server): https://swissmodel.expasy.org

Modeller (for download): https://salilab.org/modeller

Homology model databases

Protein models for entire sequence databases have been performed automatically:

ModBase: https://modbase.compbio.ucsf.edu

- Database of protein models generated by Modeller program

3Dcrunch: Swiss-Model Repository

- Models derived from Swiss-Prot database derived using the Swiss-Model program

Model evaluation

- The features of the model should be consistent with basic "physicochemical rules": φ–ψ angles (Ramachandran plot!), reasonable bond lengths, no close contacts ("collisions"), *etc*.

- Comparison of such features with statistical profiles derived from experimentally determined structures can be done by a variety of programs and servers:

MolProbity: http://molprobity.biochem.duke.edu

WHAT IF: https://swift.cmbi.umcn.nl/servers/html/index.html

PROCHECK: https://www.ebi.ac.uk/thornton-srv/software/PROCHECK

ANOLEA: http://melolab.org/anolea

VERIFY3D: https://www.doe-mbi.ucla.edu/verify3d

Threading

- Protein structures tend to be more conserved than protein sequences; many proteins share a similar fold, even in the absence of sequence similarity
- The idea behind threading: structural fold prediction by fitting a sequence into a structural database and selecting the best-fitting fold, *e.g.* based on energy calculations
- Approach can identify structurally similar proteins even without detectable sequence similarity

Threading

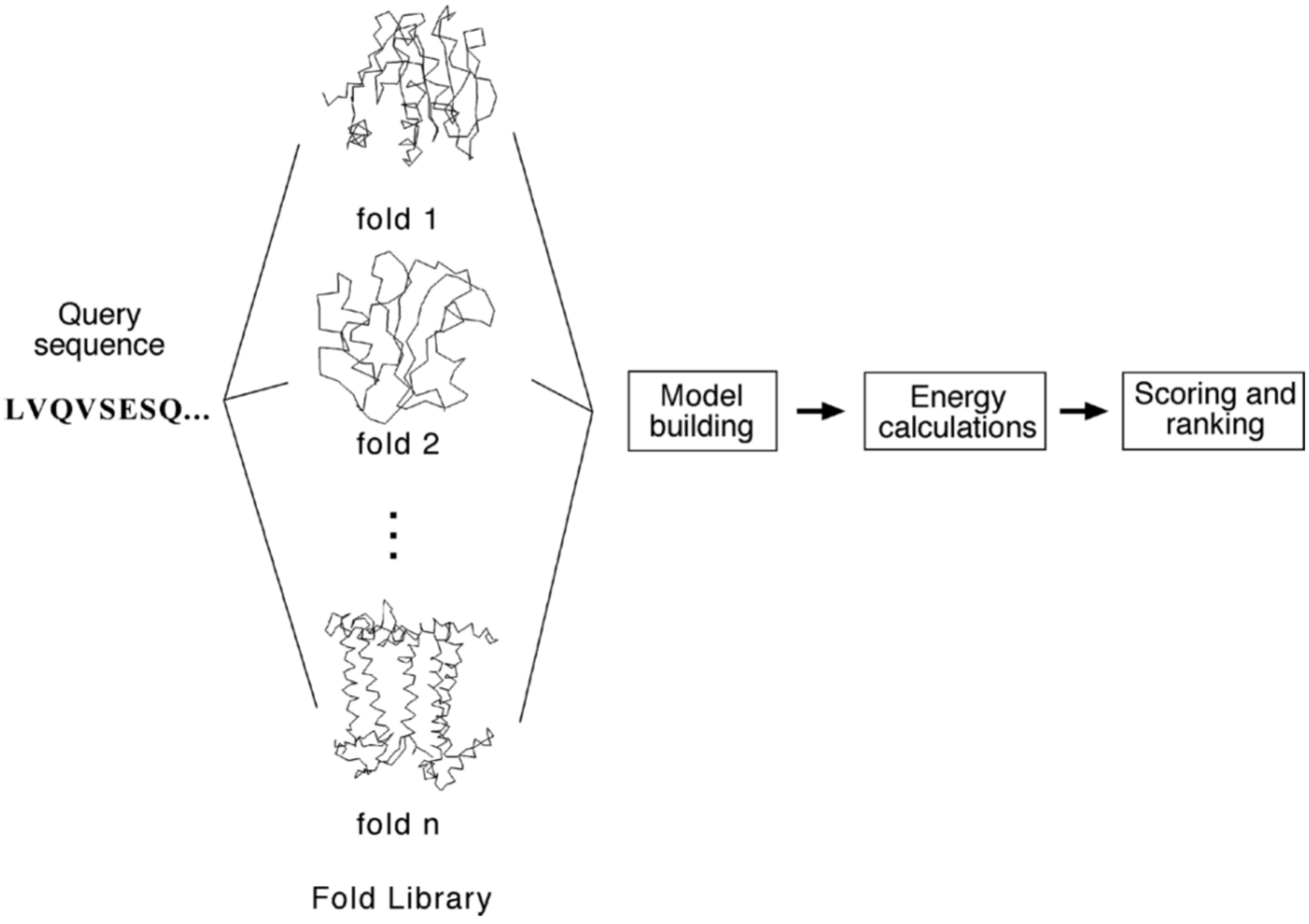


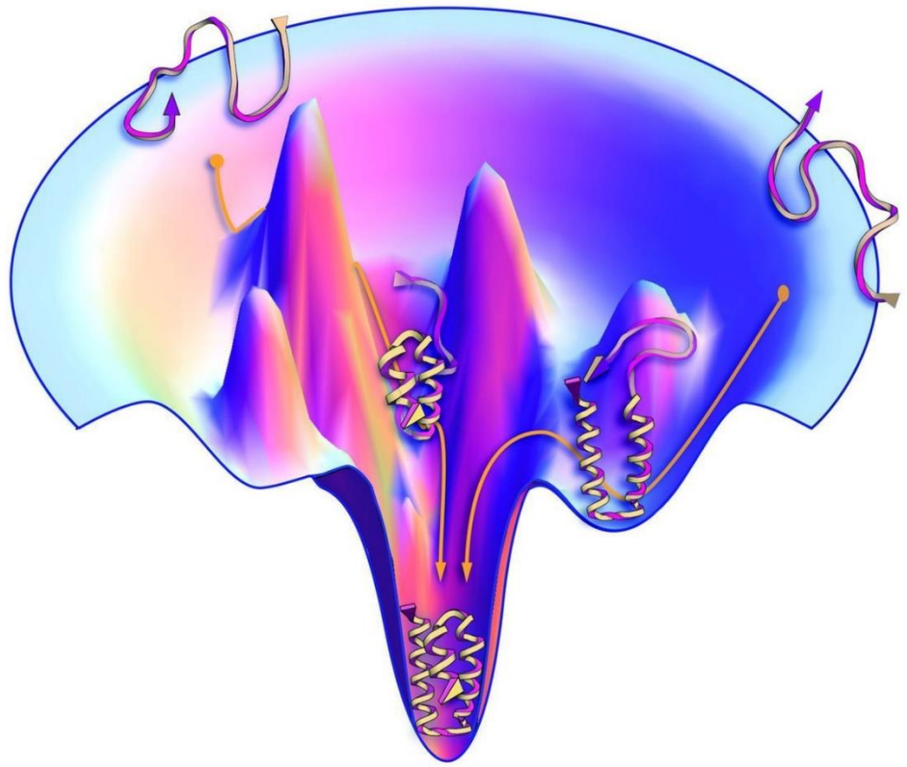
Figure 15.4: Outline of the threading method using the pairwise energy approach to predict protein structural folds from sequence. By fitting a structural fold library and assessing the energy terms of the resulting raw models, the best-fit structural fold can be selected.

Ab initio protein structure prediction

- Searching all possible conformations to find the most stable (*i.e.* energetically favourable) state is computationally impossible
- Estimate: the world's fastest supercomputer would require ~ 10^{20} years to sample all possible conformation for even a small (~40-residue) protein

Protein folding

- Proteins <u>do not sample all possible conformations</u> when they fold; rather, they appear to slide down an “energy funnel” towards a minimum:



- Physicochemical laws governing this process are not well understood and current algorithms are not able to accurately simulate the protein folding process

The Nobel Prize in Chemistry 2024

Ill. Niklas Elmehed © Nobel Prize Outreach

David Baker

Prize share: 1/2

Ill. Niklas Elmehed © Nobel Prize Outreach

Demis Hassabis

Prize share: 1/4

Ill. Niklas Elmehed © Nobel Prize Outreach

John M. Jumper

Prize share: 1/4

The Nobel Prize in Chemistry 2024 was divided, one half awarded to David Baker "for computational protein design", the other half jointly to Demis Hassabis and John M. Jumper "for protein structure prediction"

AI-based structure prediction

Can be accessed *via* various web servers, *e.g.*

AlphaFold:

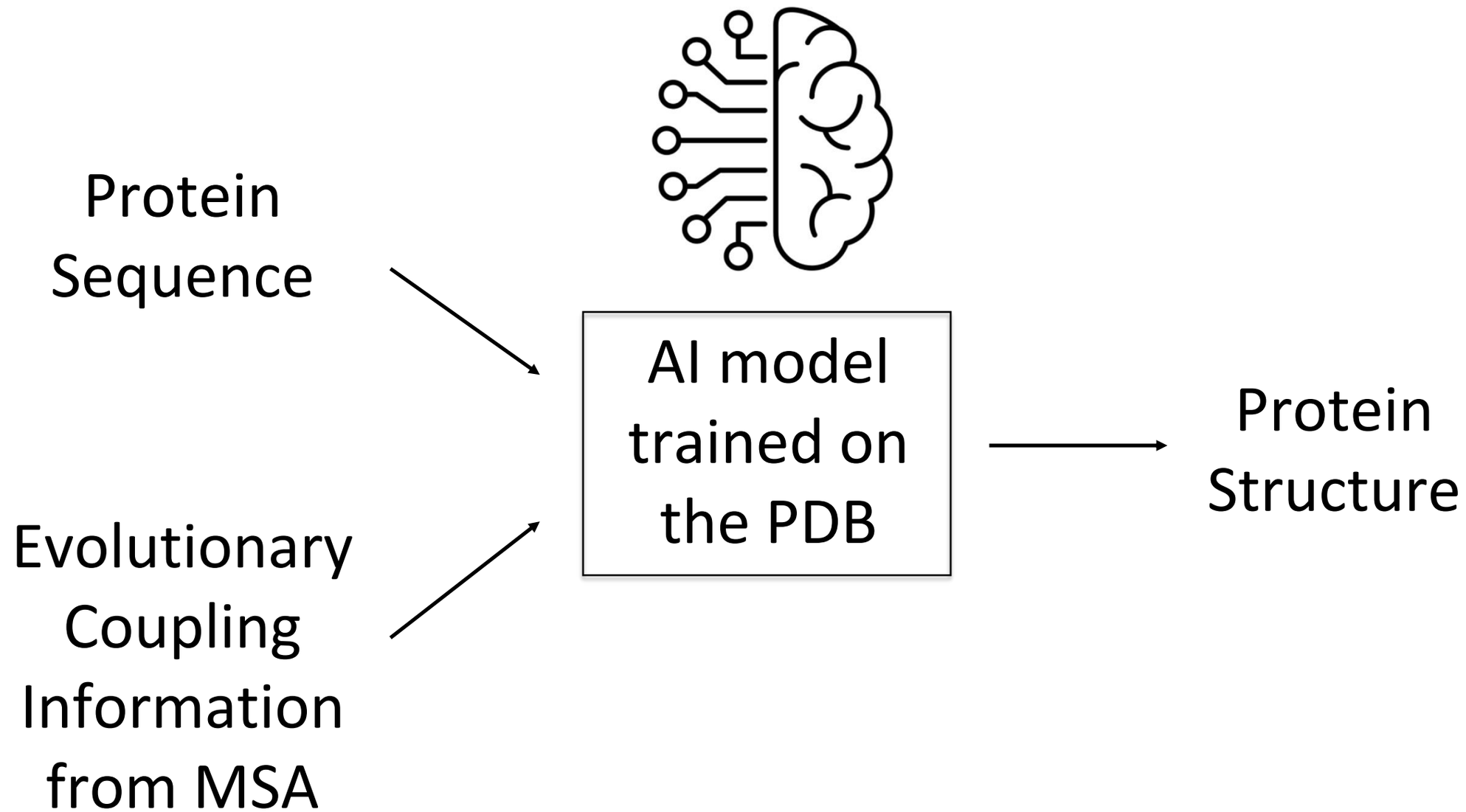
https://deepmind.google/technologies/alphafold

RoseTTAFold:

https://neurosnap.ai/service/RoseTTAFold%20All-Atom

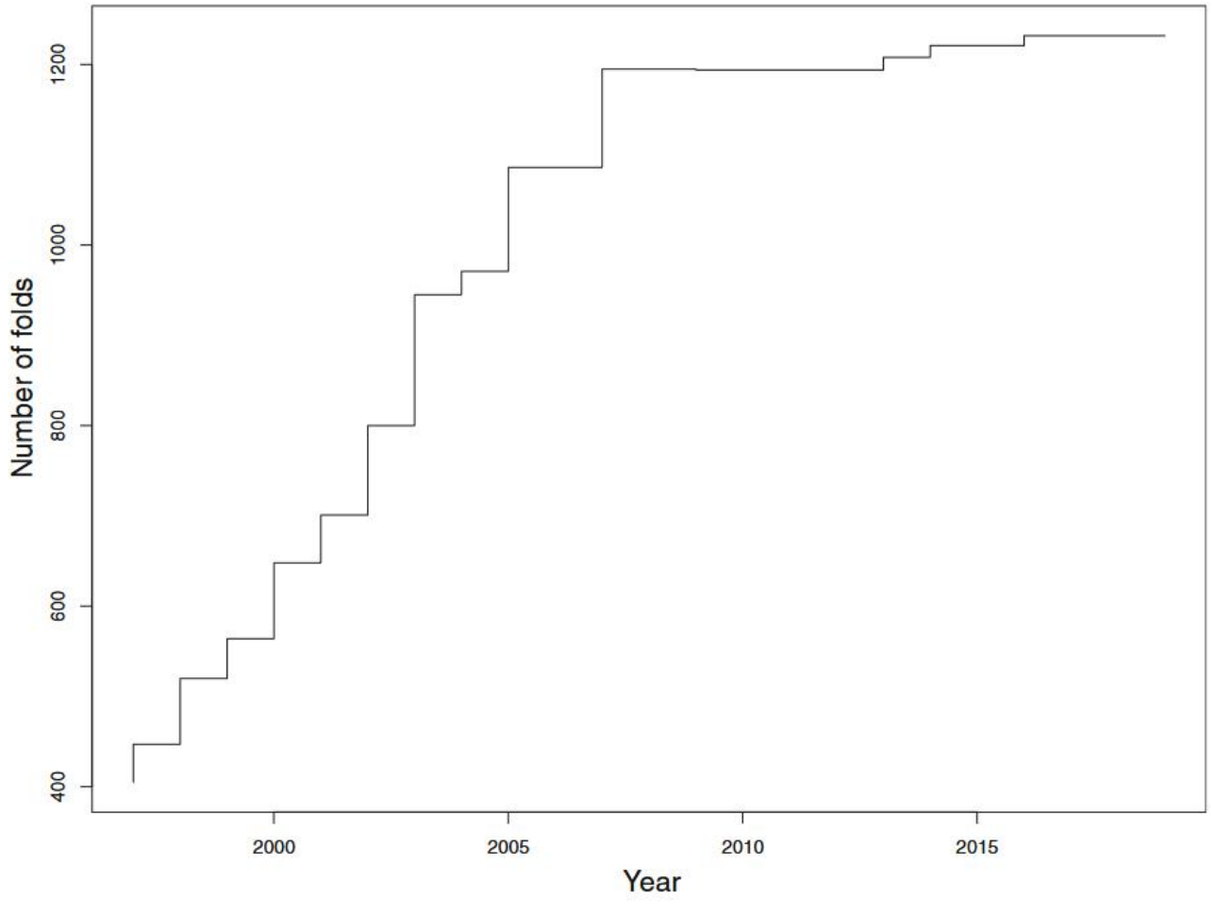
Database of pre-calculated AlphaFold structures:
https://alphafold.ebi.ac.uk

What information does AI-based prediction use?



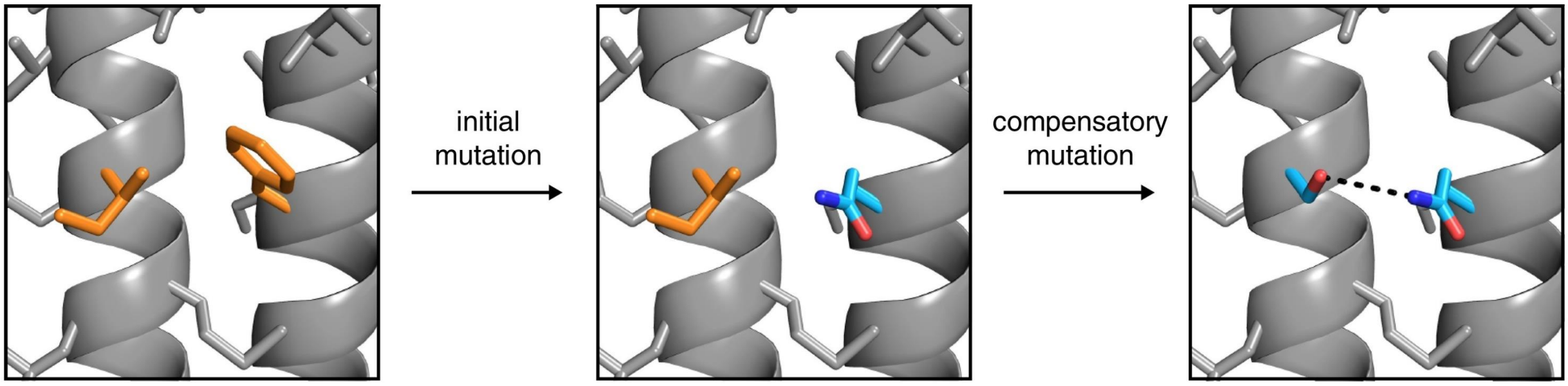
The PDB appears to cover the natural protein universe

- The PDB currently has > 200 000 entries and the number is steadily increasing
- However: there are still only ~1200 truly different protein folds, and this number appears to be flatlining:



Evolutionary coupling

In evolution, one mutation often compensates for another:

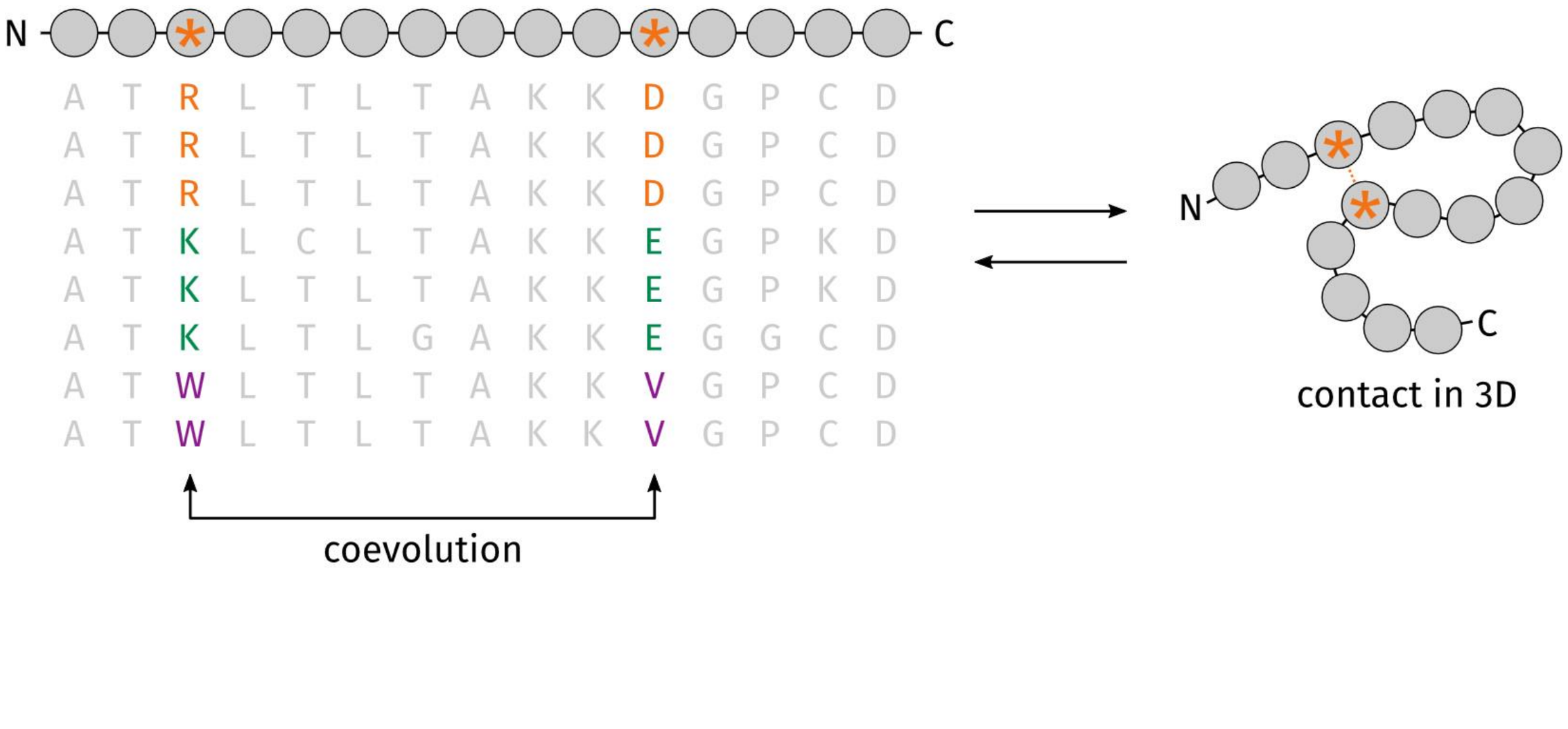


Hydrophobic interaction

Hydrogen bond

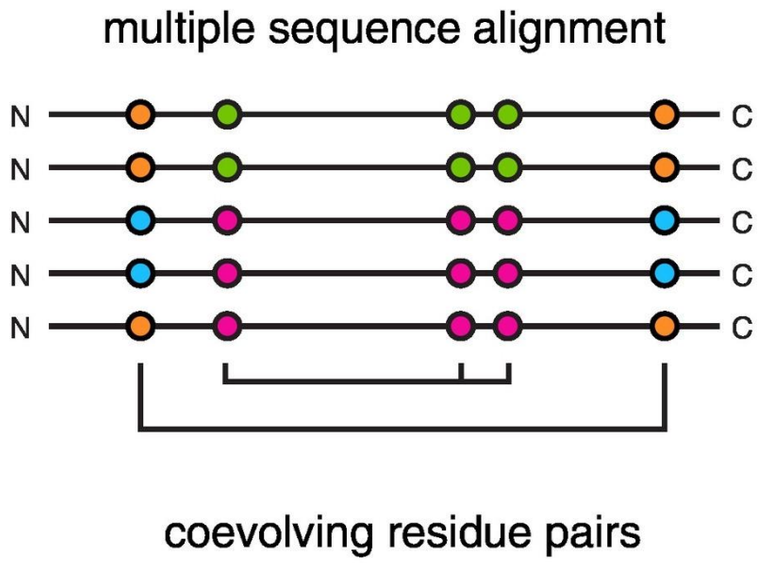
Evolutionary coupling

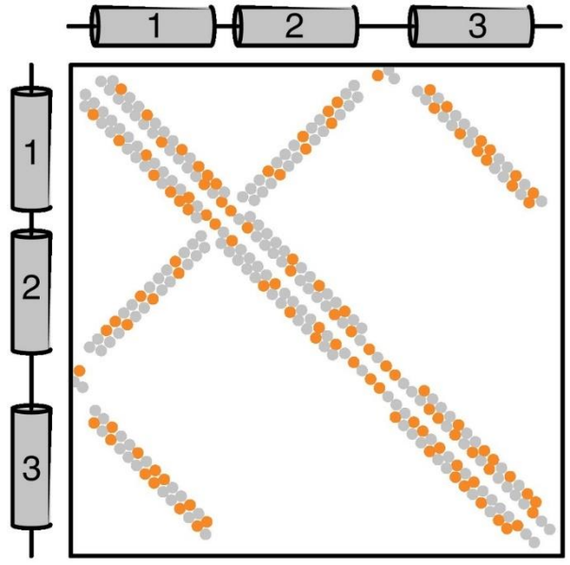
Residues that are in close contact in a structure, will show *evolutionary coupling* (*i.e.* correlation) in a multiple sequence alignment (MSA):

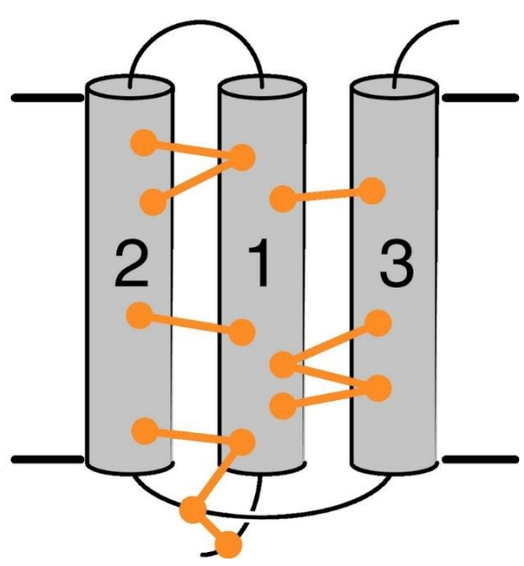


Evolutionary coupling

From evolutionary coupling information, a *contact matrix* can be constructed, which helps to guide the reconstruction of the 3D structure:





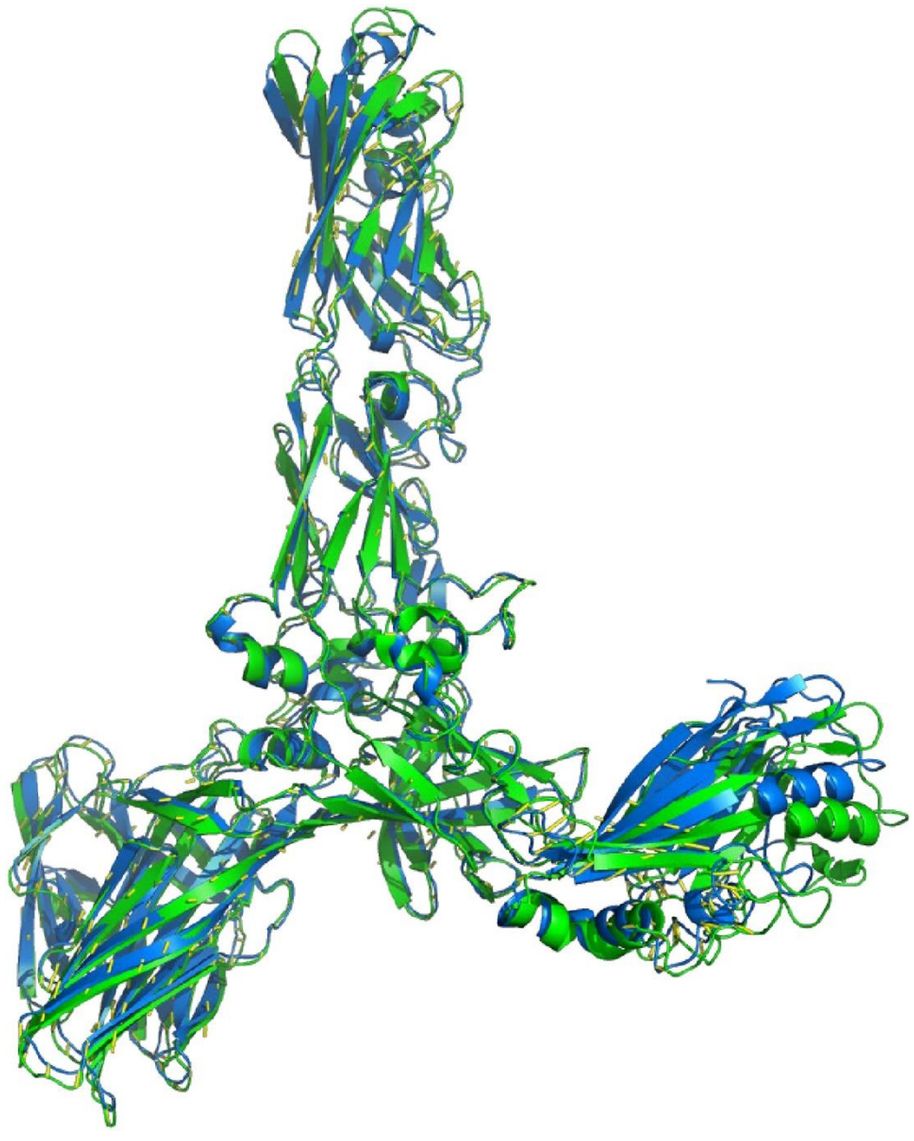


CASP (Critical Assessment of Structure Prediction)

- Biannual contest initiated 1994
- CASP contestants are given protein sequences whose structures have been solved by X-ray crystallography, EM or NMR, but have not yet been published
- Each contestant predicts the structures and submits the results to the CASP organisers
- Results of the predictions are compared with the newly determined structures using structure alignment programs

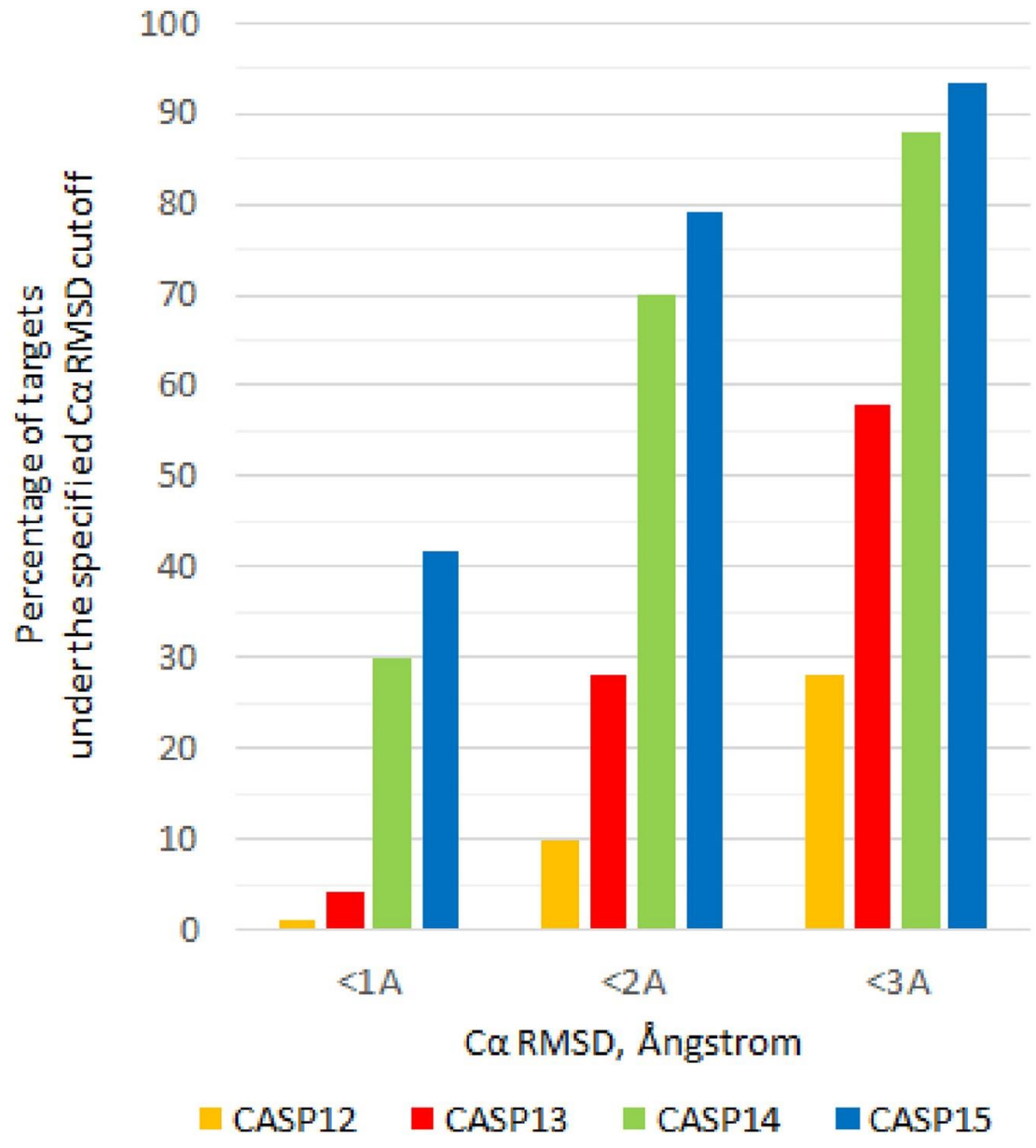
Protein Structure Prediction Center:
https://predictioncenter.org/

Example of a CASP15 target protein



Superposition of a large protein target, T1154, a 1040 residue archaeal S-layer protein (green) and the closest calculated structure (blue).

RMSD values for CASP 12, 13, 14 and 15



AlphaFold produces per-residue confidence scores

The pLDDT scores range from 1 to 100 (higher is better) and are stored in the B-factor column of the PDB file:

structure annotation

amino acid field

cofactor field

```
HEADER    LYASE (CARBON-CARBON)                   03-JUL-95   1DNP
TITLE     STRUCTURE OF DEOXYRIBODIPYRIMIDINE PHOTOLYASE
... ...
SOURCE   2 ORGANISM_SCIENTIFIC: ESCHERICHIA COLI
KEYWDS    DNA REPAIR, ELECTRON TRANSFER, EXCITATION ENERGY TRANSFER,
KEYWDS   2 LYASE, CARBON-CARBON
... ...
ATOM     21  ND1 HIS A   3      55.365  27.866  62.971  1.00 11.07           N
ATOM     22  CD2 HIS A   3      57.200  28.354  61.894  1.00 13.12           C
ATOM     23  CE1 HIS A   3      56.124  26.783  62.981  1.00 13.03           C
ATOM     24  NE2 HIS A   3      57.243  27.052  62.334  1.00  8.19           N
ATOM     25  N   LEU A   4      55.580  32.694  59.656  1.00 12.61           N
ATOM     26  CA  LEU A   4      54.799  33.803  59.113  1.00 11.56           C
ATOM     27  C   LEU A   4      53.552  33.269  58.374  1.00  7.76           C
ATOM     28  O   LEU A   4      53.650  32.363  57.532  1.00  6.99           O
ATOM     29  CB  LEU A   4      55.656  34.683  58.174  1.00  9.03           C
ATOM     30  CG  LEU A   4      54.946  35.887  57.518  1.00  2.00           C
ATOM     31  CD1 LEU A   4      54.623  36.920  58.550  1.00  6.21           C
... ...
HETATM 7641 AN7  FAD B 472      27.855  78.556  29.073  1.00  4.55           N
HETATM 7642 AC5  FAD B 472      28.524  78.026  27.955  1.00  2.00           C
HETATM 7643 AC6  FAD B 472      29.848  77.609  27.724  1.00  3.40           C
HETATM 7644 AN6  FAD B 472      30.787  77.757  28.664  1.00  6.22           N
```

atom number

atom name

residue name

polypeptide chain identifier

residue number

x, y, z coordinates

occupancy

temperature factor

atom type

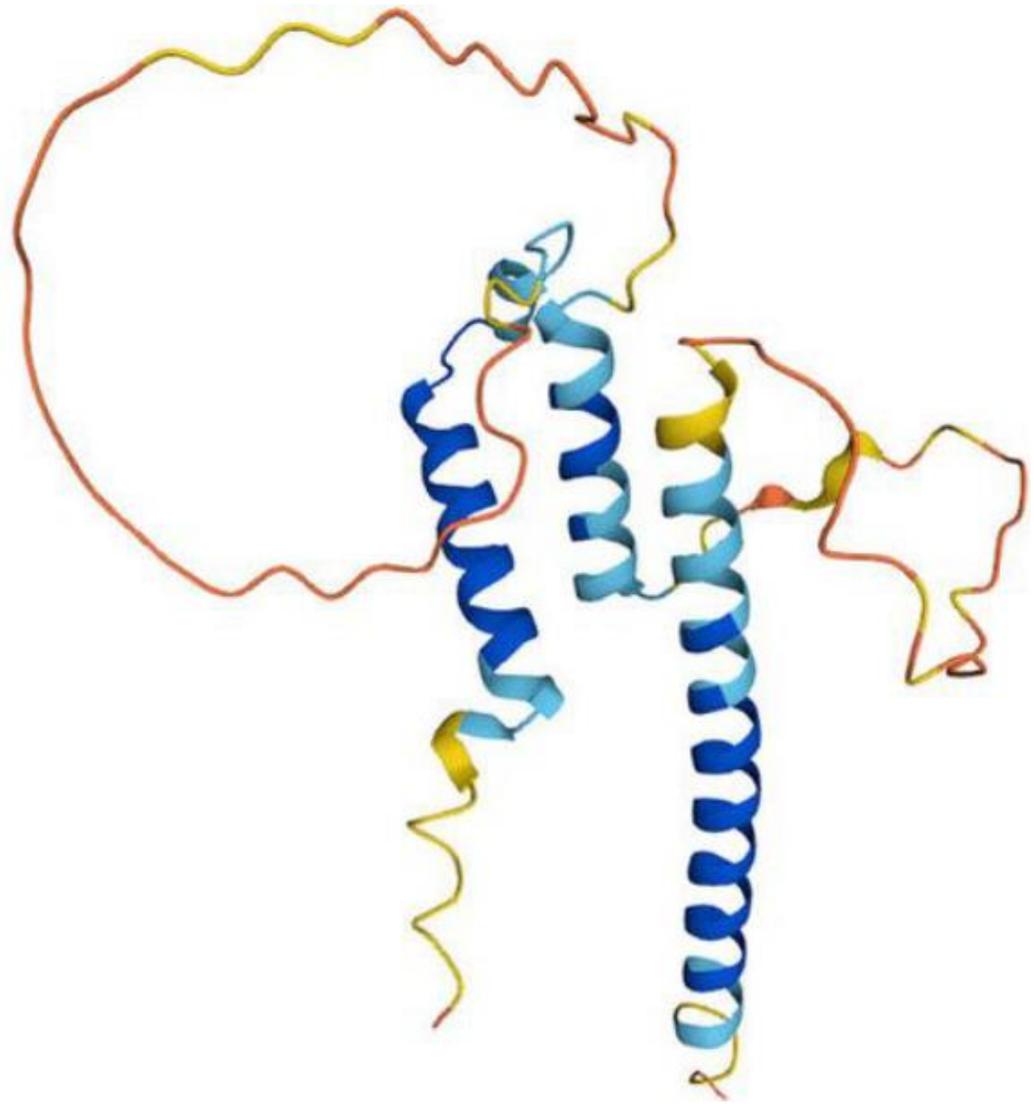
AlphaFold produces per-residue confidence scores

The pLDDT (predicted Local Distance Difference Test) scores range from 1 to 100 (higher is better)

Model Confidence:

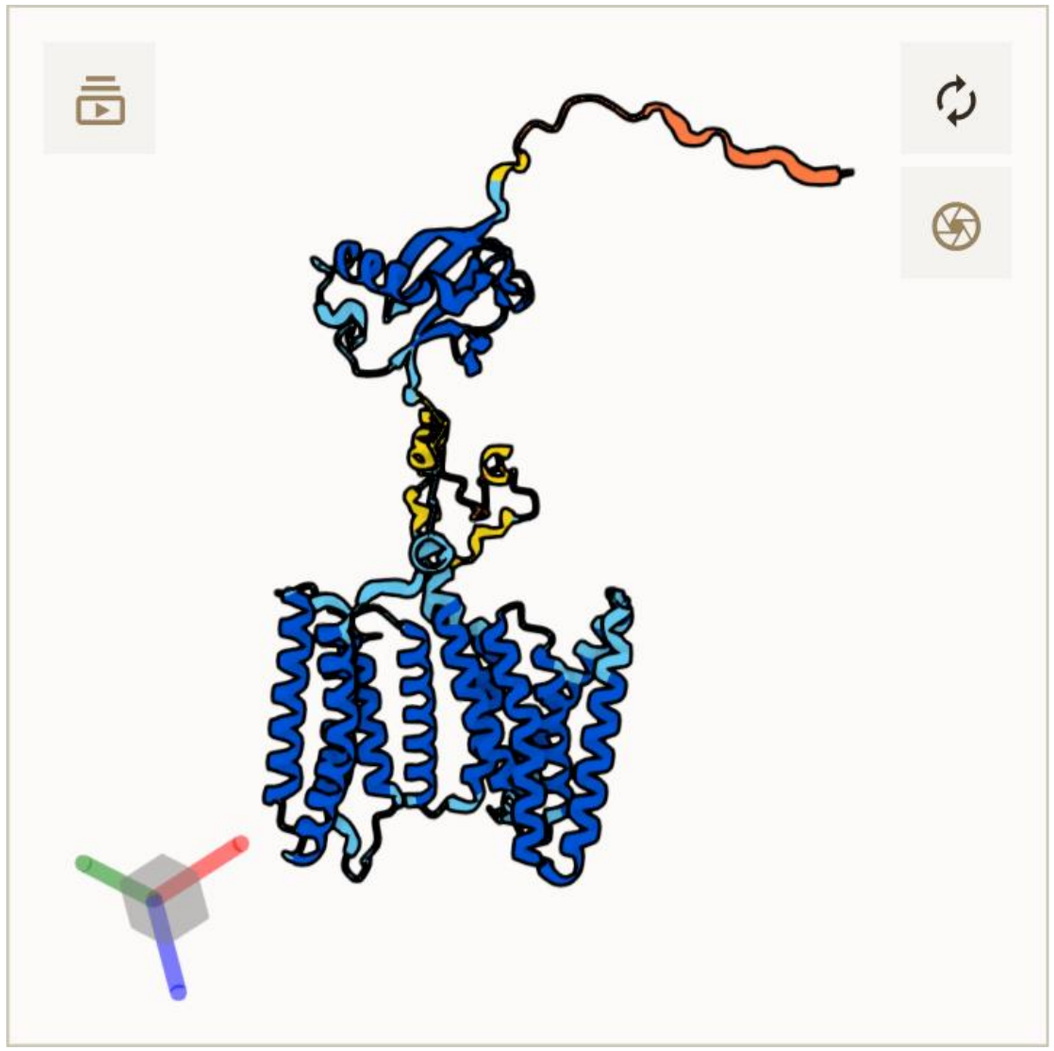
- Very high (pLDDT > 90)
- Confident (90 > pLDDT > 70)
- Low (70 > pLDDT > 50)
- Very low (pLDDT < 50)

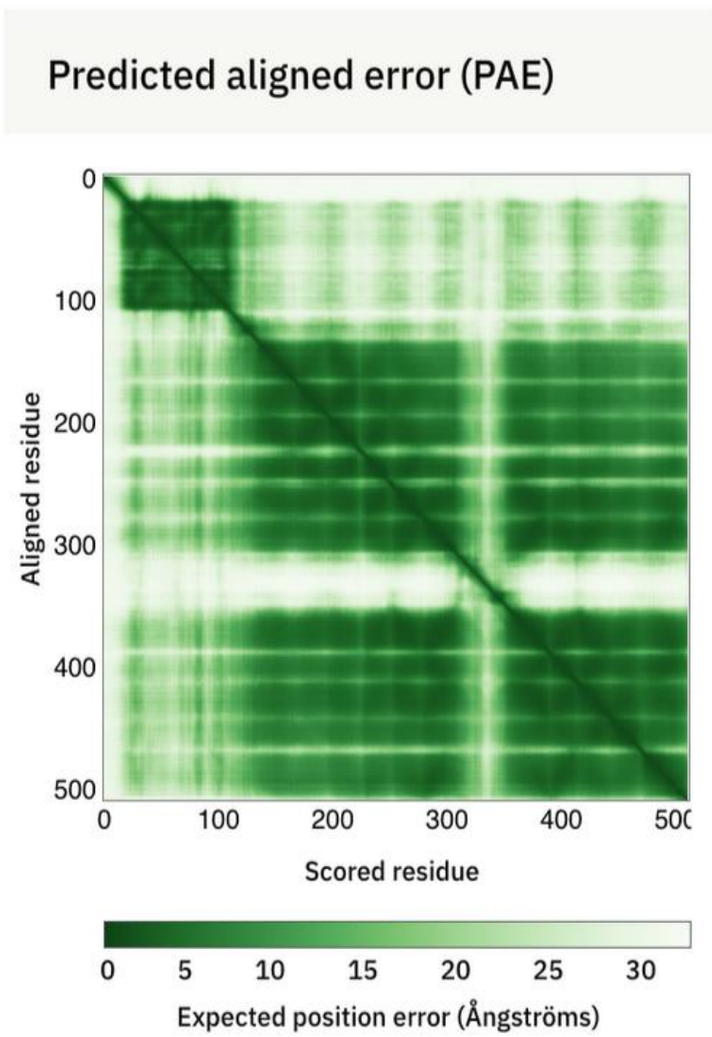
AlphaFold produces a per-residue confidence score (pLDDT) between 0 and 100. Some regions with low pLDDT may be unstructured in isolation.



AlphaFold also produces a 2D confidence plot

- The PAE (Predicted Alignment Error) plot shows confidence in inter-residue distances
- Well-predicted domains show up as dark blocks:





AI-based prediction: current shortcomings

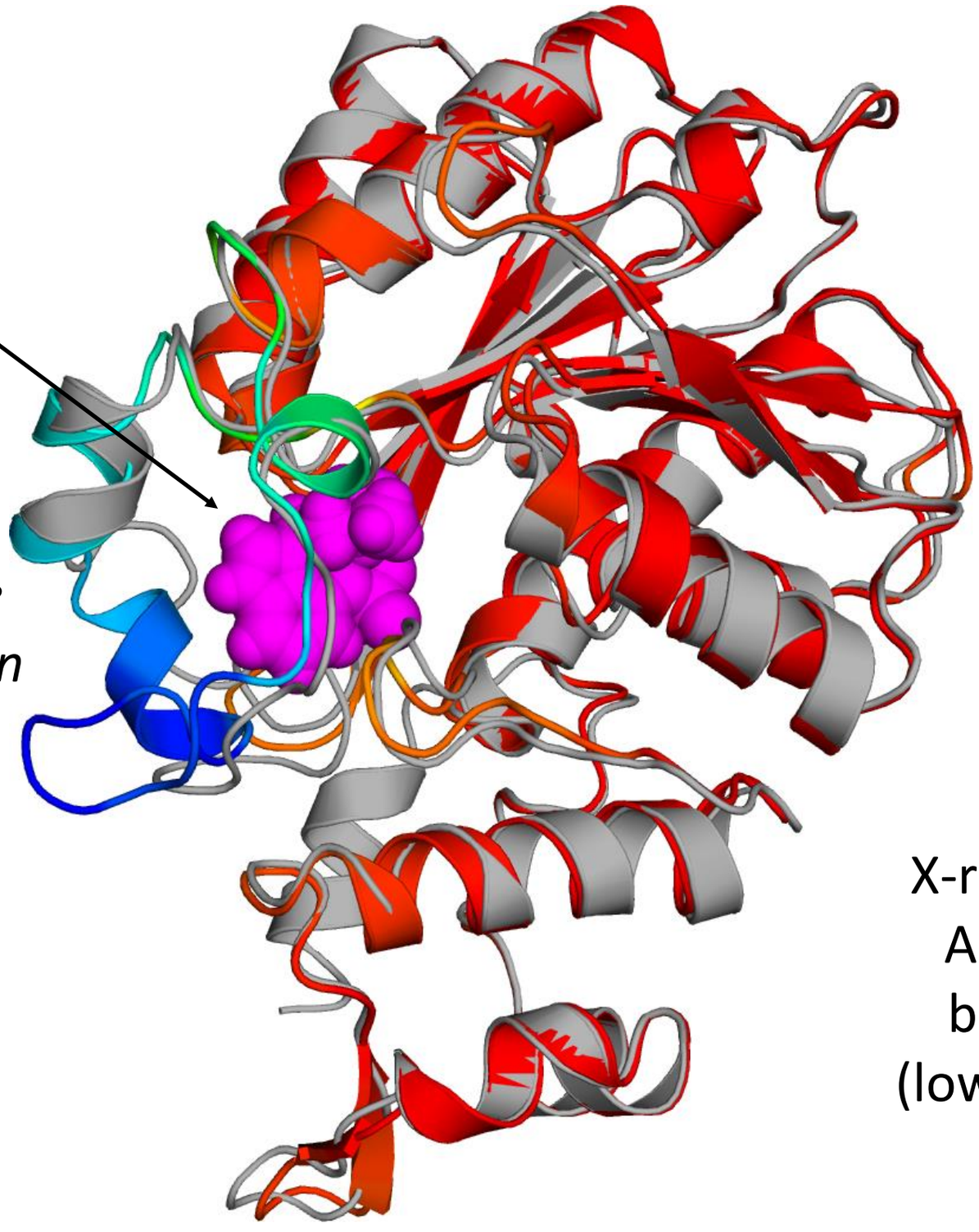
AI predictor is trained on the PDB, it does not know the laws of physics!

- Bias towards well-structured, crystallisable proteins
- Does not deal well with unusual structures (the "final 5-10%"), particularly if there are only few homologs (MSA is *very* important!)
- Does not deal with (most) small-molecule ligands
- Will not correctly predict the effect of mutations, deletions, *etc*.
- Does not consider structural changes, domain movement, *etc*.

AlphaFold prediction of PsiM (RMSD: 1.8 Å)

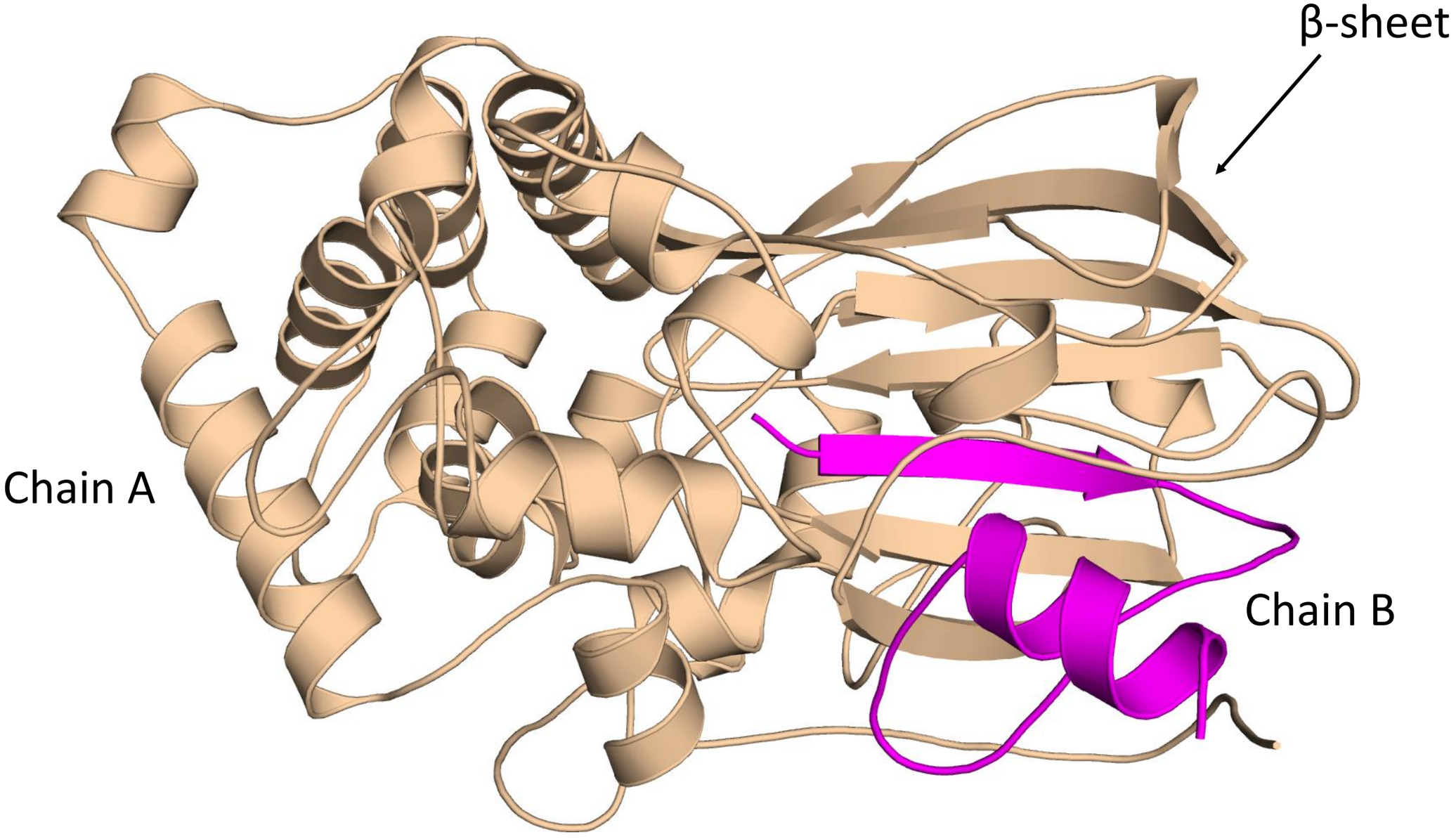
Purple: psilocybin in the X-ray structure

Substrate recognition loop

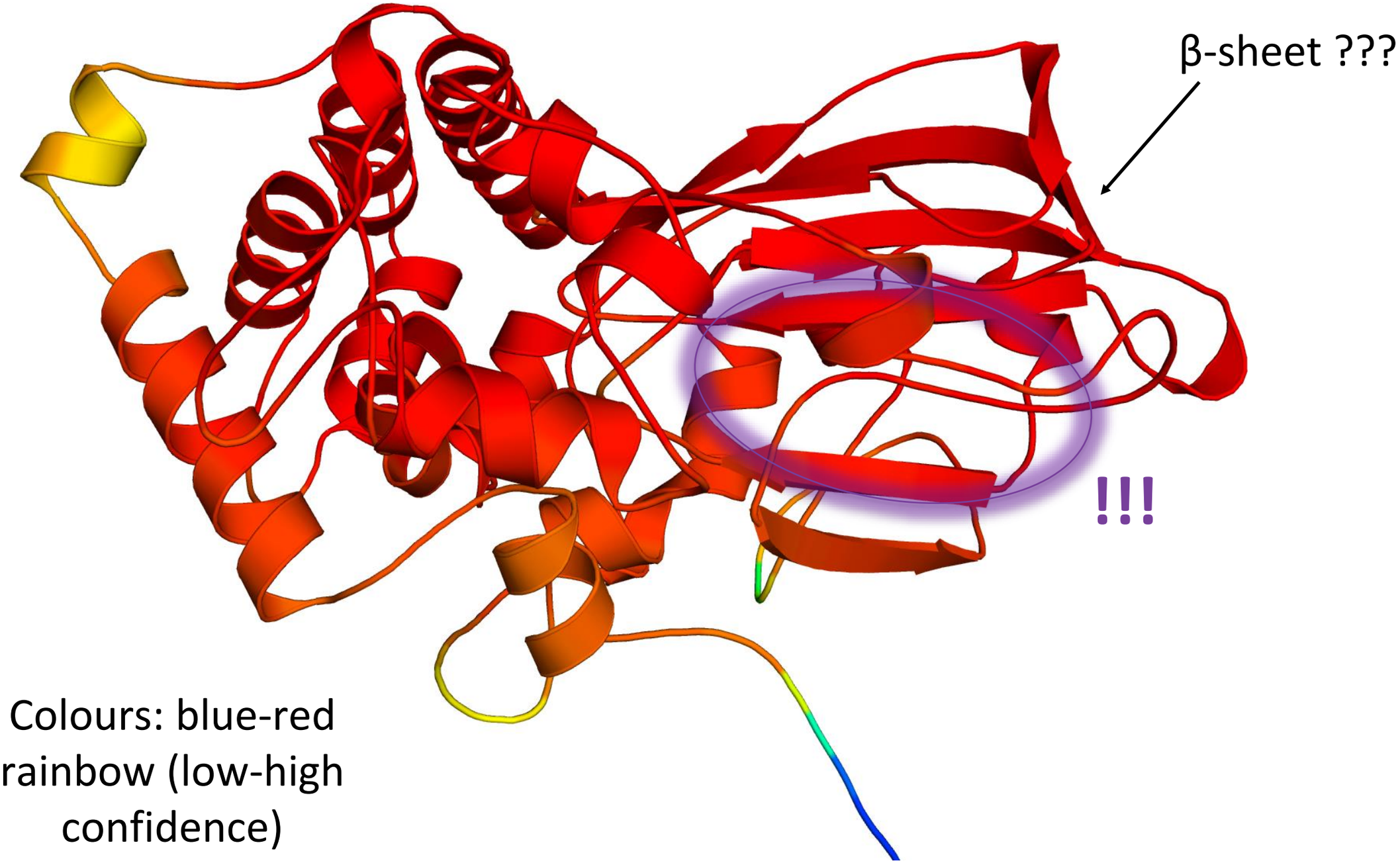


X-ray structure: grey
AlphaFold model: blue-red rainbow (low-high confidence)

AlphaFold prediction of PsiD



AlphaFold prediction of PsiD - chain A alone



Colours: blue-red rainbow (low-high confidence)