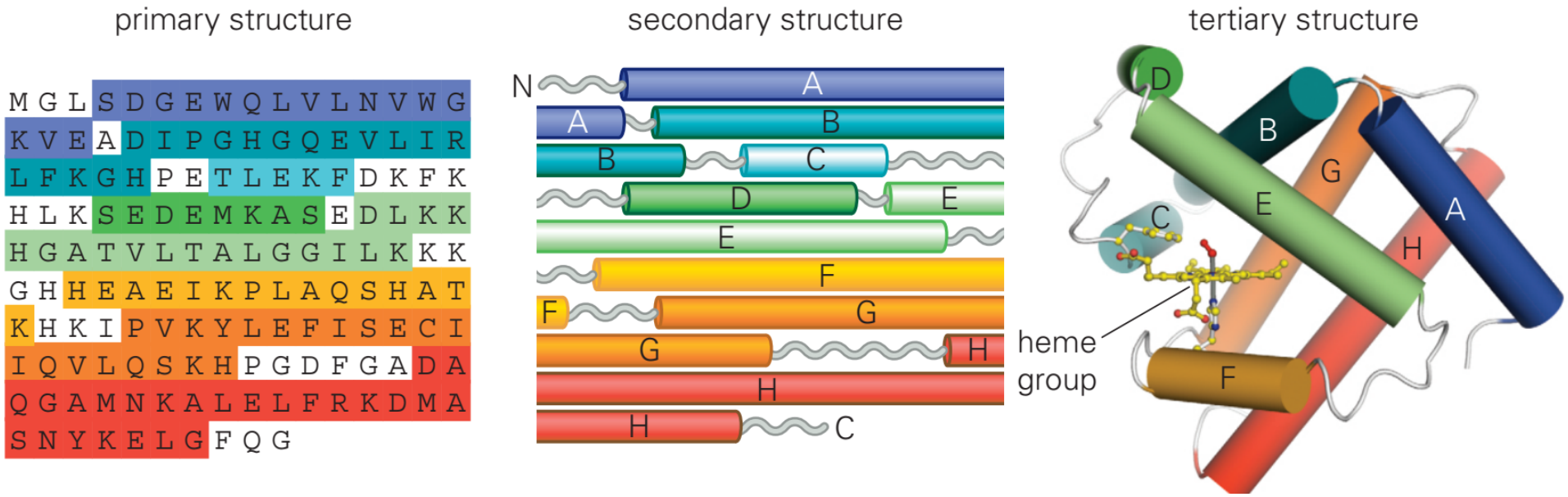
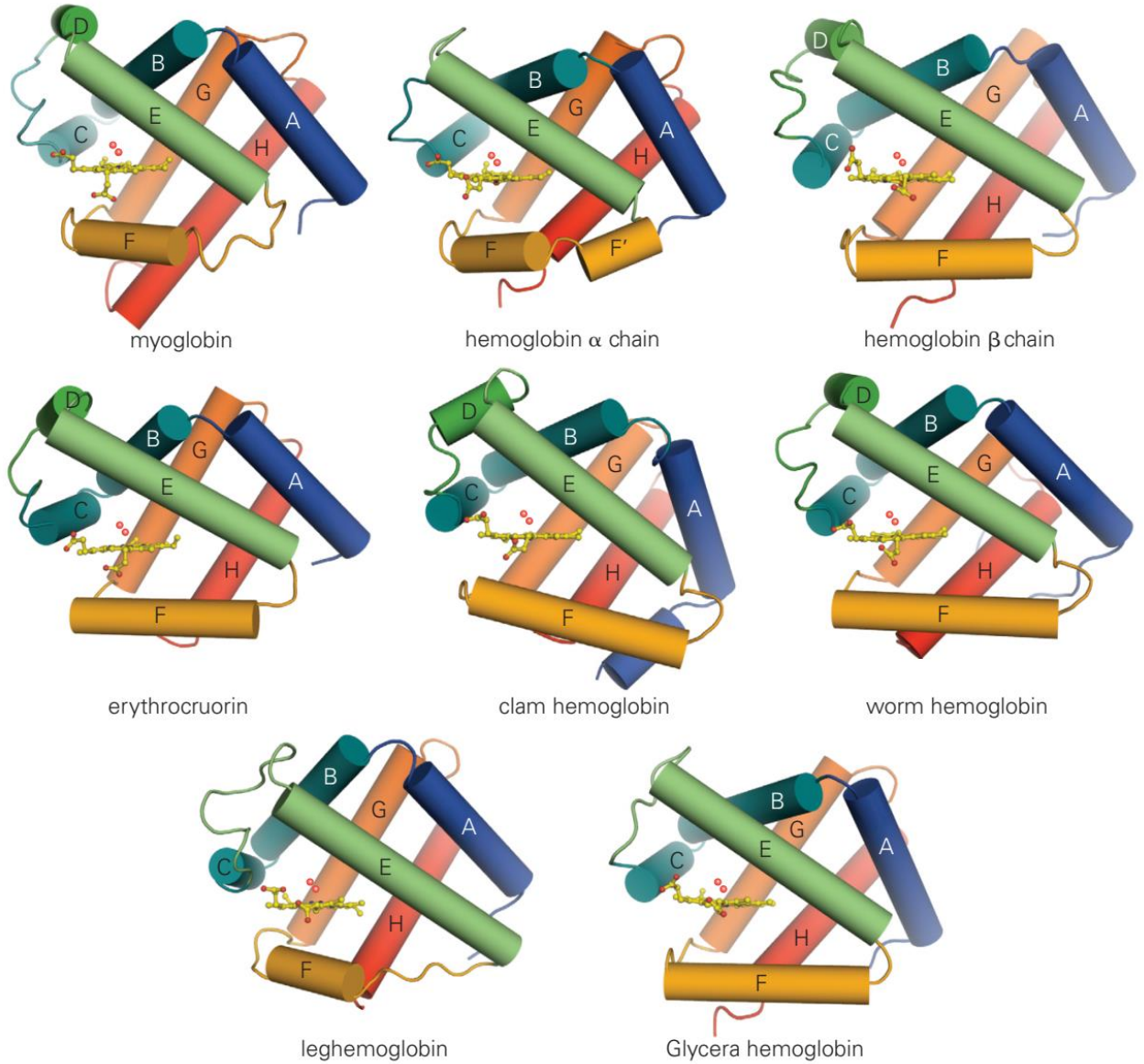
STRUCTURAL BIOINFORMATICS

Basic principles, X-rays, cryo-EM

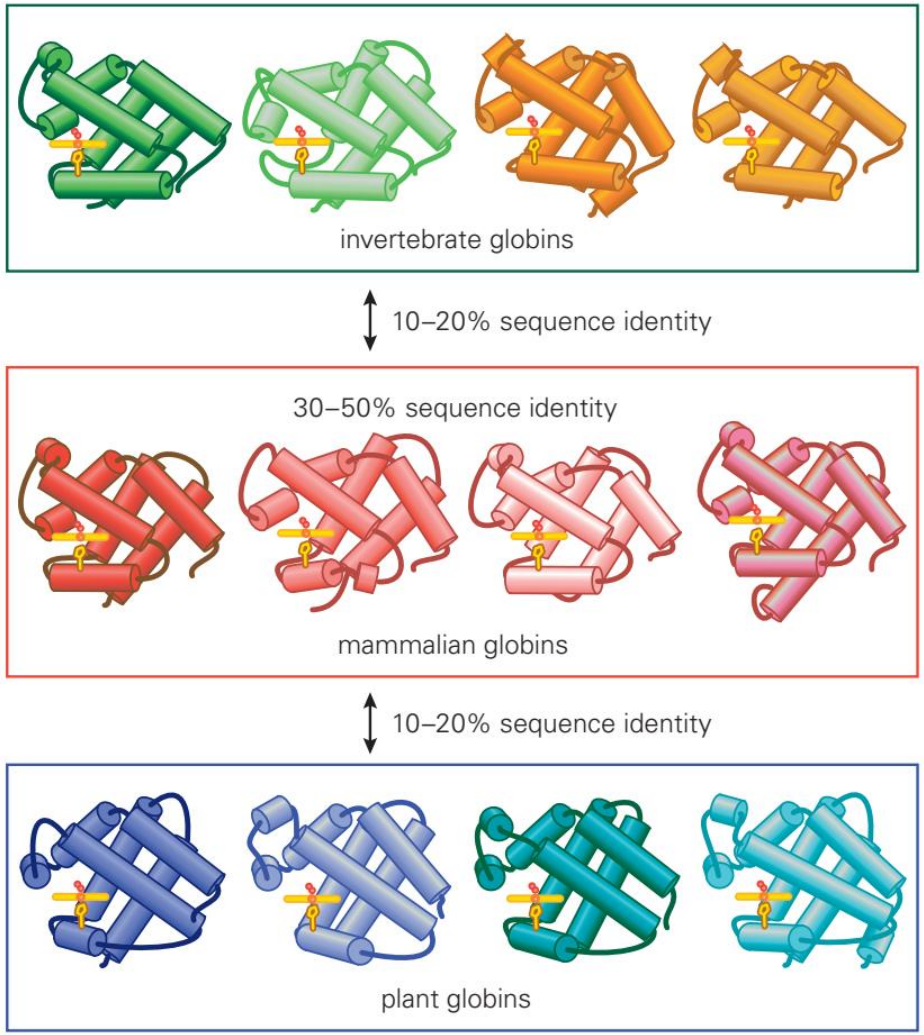
The amino-acid sequence defines the 3D-structure of proteins

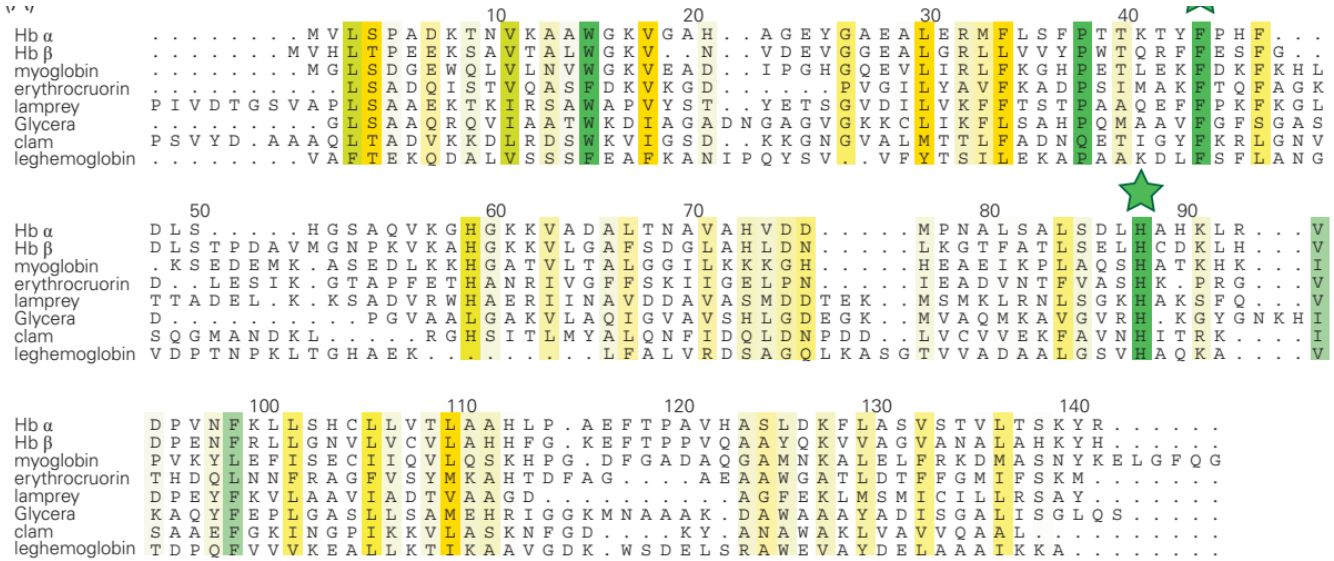


Similar sequences have similar folds

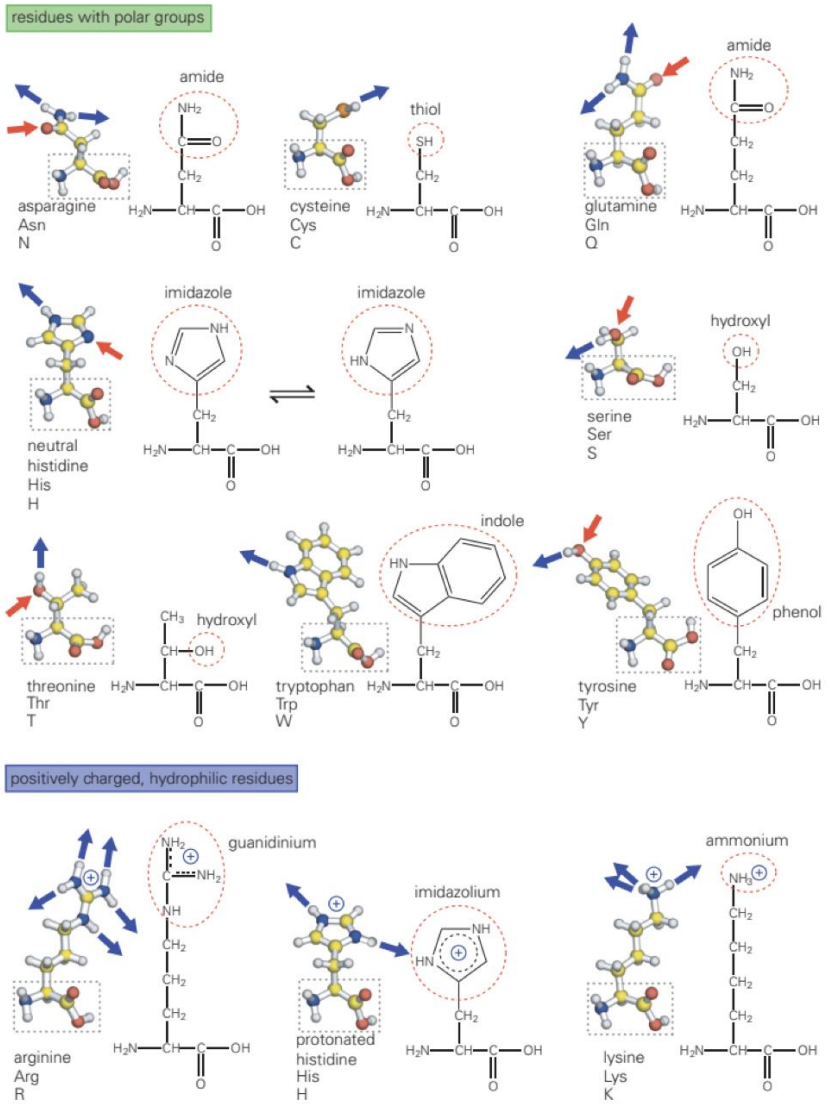


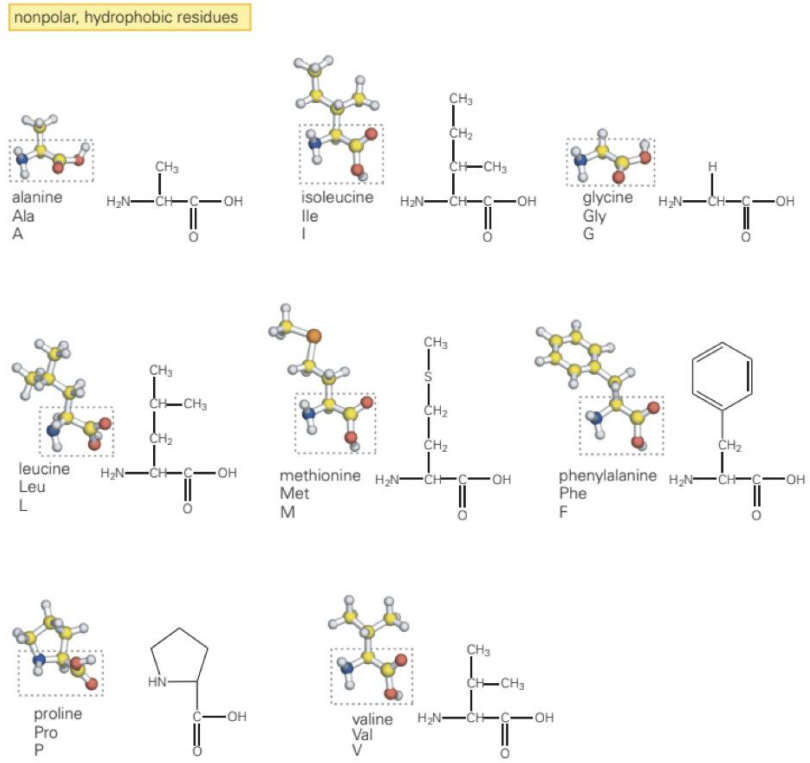
Importance of sequence similarity

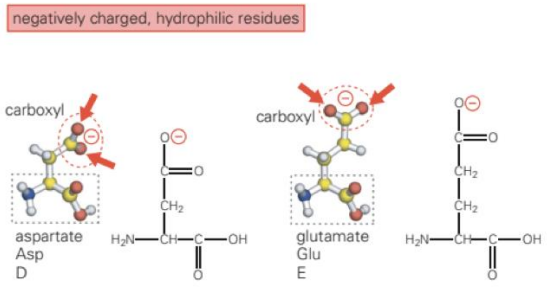




How do we “define” similarity?







Amino acid substitution matrix

- A matrix in which each row and column corresponds to one of the 20 amino acids. Each entry in the matrix is related to the probability that one amino acid is replaced by the other in proteins that are related evolutionarily
- Each entry in the matrix is called a substitution score, Sij, and the value of Sij is related to the frequency with which the ith type of amino acid is replaced by the jth type of amino acid in the alignments of related proteins, relative to the probability that this substitution could occur by random chance, given the abundance of each of the amino acids.

PAM (Dayhoff) matrices

- “Point Accepted Mutation” is an amino acid substitution model derived from empirical observation of mutations among closely related proteins.
- Counts of mutations from one amino acid to the other nineteen amino acids should be proportional to the rates of transition from that one amino acid to the other nineteen

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	Ala																				
R	Arg	30																			
N	Asn	109	17																		
D	Asp	154	0	532																	
C	Cys	33	10	0	0																
Q	Gln	93	120	50	76	0															
E	Glu	266	0	94	831	0	422														
G	Gly	579	10	156	162	10	30	112													
H	His	21	100	226	43	10	243	23	10												
I	Ile	66	20	36	13	17	8	35	0	3											
L	Leu	95	17	37	0	0	75	15	17	40	253										
K	Lys	57	477	322	85	0	147	104	60	23	43	39									
M	Met	29	17	0	0	0	20	7	7	0	57	207	90								
F	Phe	20	7	7	0	0	0	0	17	20	90	167	0	17							
P	Pro	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	Ser	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	Thr	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	Trp	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	Tyr	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	Val	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	

Ala Arg Asn Asp Cys Gln Glu Gly His Ile Leu Lys Met Phe Pro Ser Thr Trp Tyr Val

Figure 80. Numbers of accepted point mutations (X10) accumulated from closely related sequences. Fifteen hundred and seventy-two exchanges are shown. Fractional exchanges result when ancestral sequences are ambiguous.

PAM (Dayhoff) matrices

- This matrix was then transformed to a transition probability matrix
- The unconditional mutation probability, that is $P(i \neq j)$, was set to equal 1 mutation for every 100 sites
- 1 PAM unit is thus defined as the unit of time in which we expect 0.01 mutations to occur per site

ORIGINAL AMINO ACID

REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R Arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N Asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D Asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C Cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q Gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E Glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G Gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H His	1	2	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I Ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L Leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K Lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M Met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F Phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P Pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S Ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T Thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W Trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y Tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V Val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Figure 82. Mutation probability matrix for the evolutionary distance of 1 PAM. An element of this matrix, M_{ij}, gives the probability that the amino acid in column j will be replaced by the amino acid in row i after a given evolutionary interval, in this case 1 accepted point mutation per 100 amino acids. Thus, there is a 0.56% probability that Asp will be replaced by Glu. To simplify the appearance, the elements are shown multiplied by 10,000.

PAM (Dayhoff) matrices

- Different matrices were needed for different expected phylogenetic distances
- Matrices for different PAM can be derived assuming a discrete-time Markov chain
- The PAM250 is used most often.
- Transition probabilities get “flatter” the more time has passed

ORIGINAL AMINO ACID

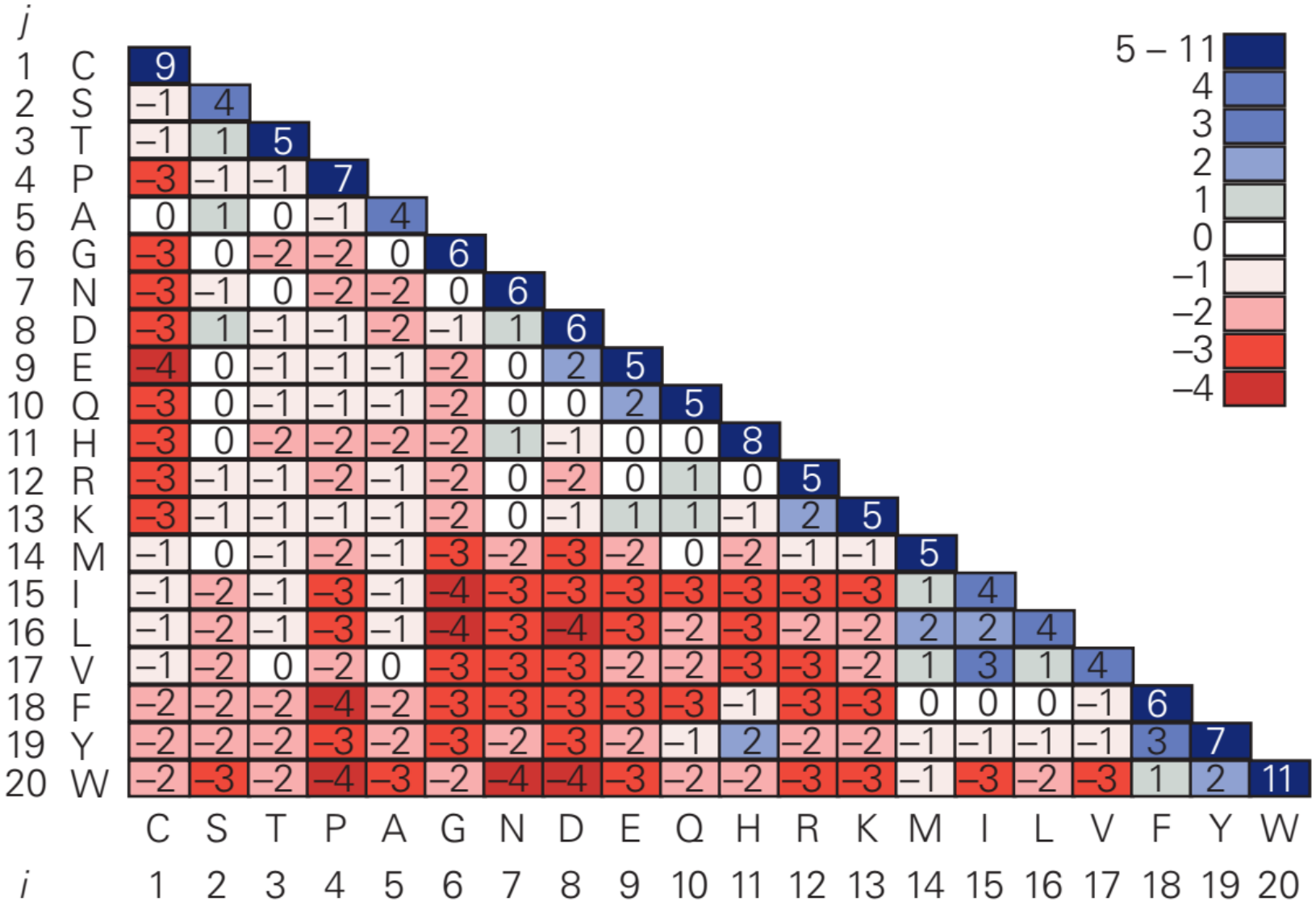
REPLACEMENT AMINO ACID	A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	V Val
A Ala	13	6	9	9	5	8	9	12	5	8	6	7	7	4	11	11	11	2	4	9
R Arg	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N Asn	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D Asp	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C Cys	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q Gln	3	5	5	6	1	10	7	3	7	2	3	5	2	1	4	3	3	1	2	3
E Glu	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G Gly	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H His	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I Ile	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L Leu	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K Lys	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M Met	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F Phe	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P Pro	7	6	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S Ser	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T Thr	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W Trp	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y Tyr	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V Val	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

Figure 83. Mutation probability matrix for the evolutionary distance of 250 PAMs. To simplify the appearance, the elements are shown multiplied by 100. In comparing two sequences of average amino acid frequency at this evolutionary distance, there is a 13% probability that a position containing Ala in the first sequence will contain Ala in the second. There is a 3% chance that it will contain Arg, and so forth. The relationship of two sequences at a distance of 250 PAMs can be demonstrated by statistical methods.

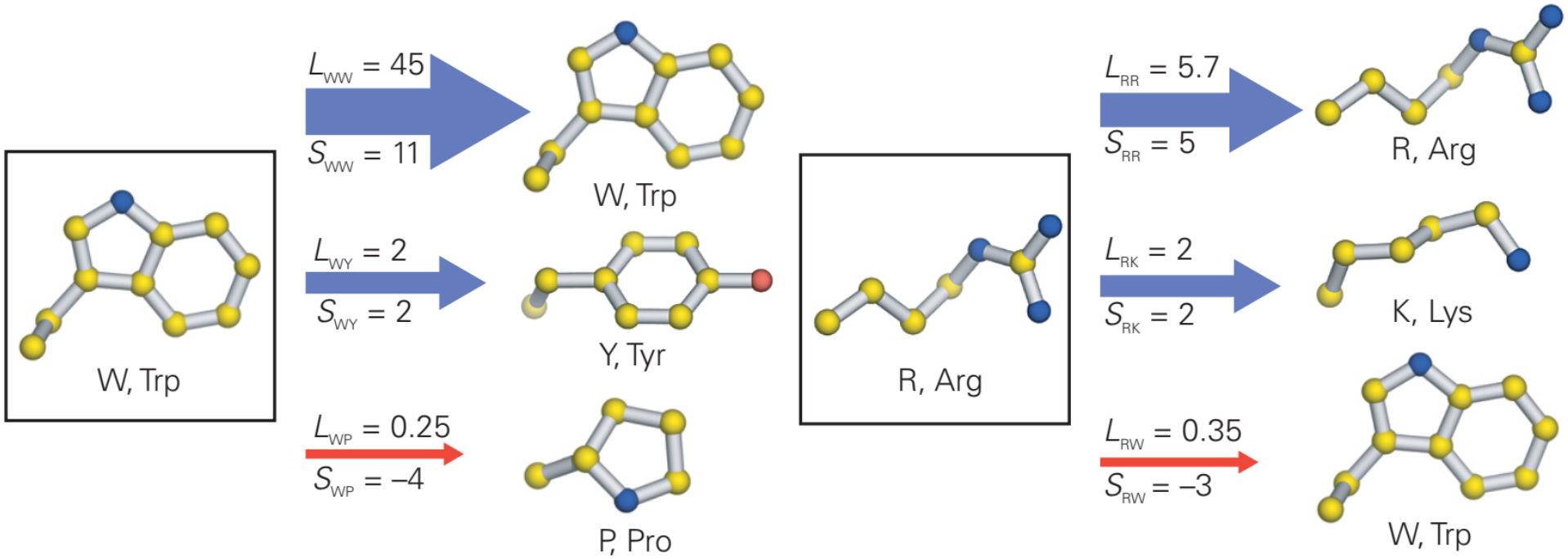
BLOSUM matrices

- BLOcks SUbstitution Matrix
- BLOSUM looks directly at mutations in motifs of related sequences while PAM's extrapolate evolutionary information based on closely related sequences.
- BLOSUM matrices were built based on pre-existing local alignments and optimised iteratively.
- The BLOSUM substitution score for a pairing is defined in terms of the base-2 logarithm of the substitution likelihood
- High BLOSUM numbers denote high expected similarity
  - *BLOSUM 62 is a matrix calculated from comparisons of sequences with a pairwise identity of no more than 62%.*
  - *Remember PAMs do the opposte, the notation denotes time*
- Most programs these days use BLOSUM matrices

The Blosum62 matrix

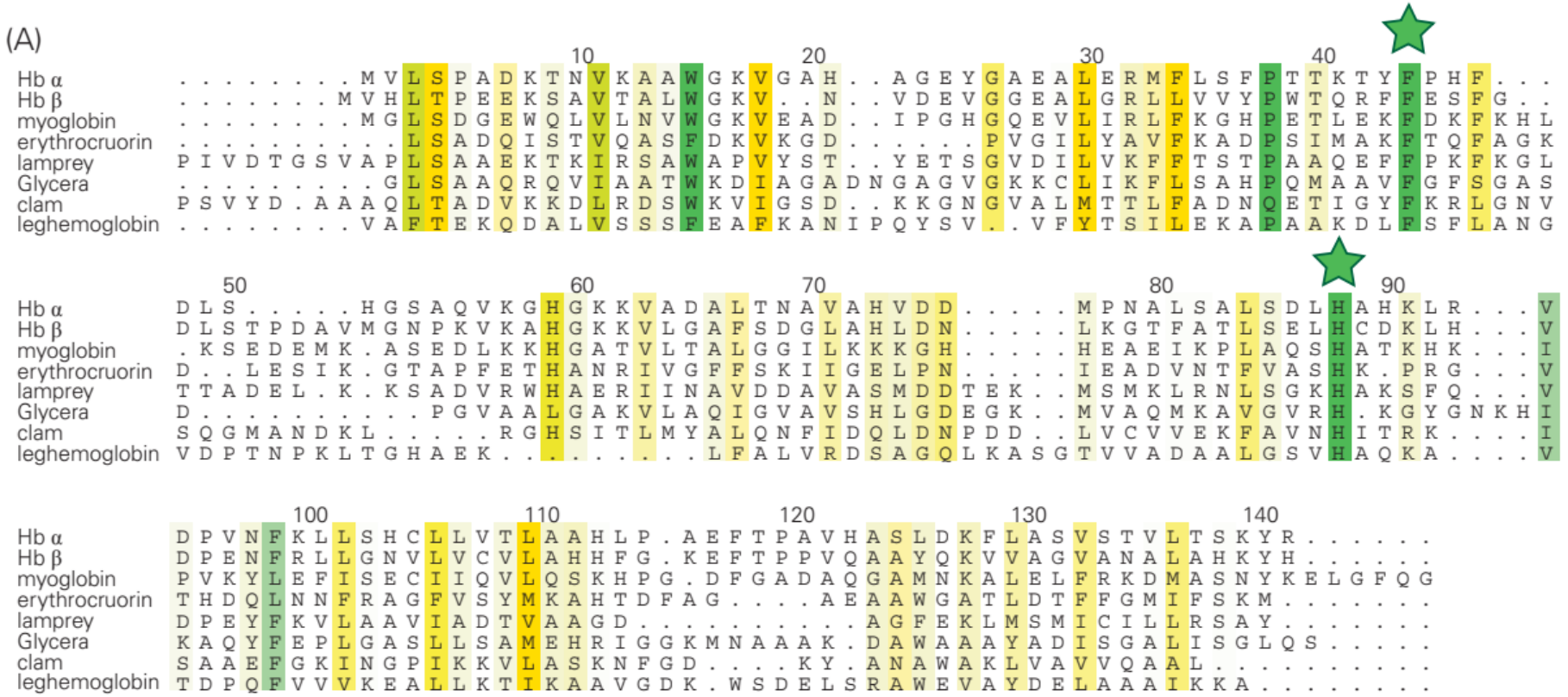


Substitution scores reflect the chemical properties of the amino acids

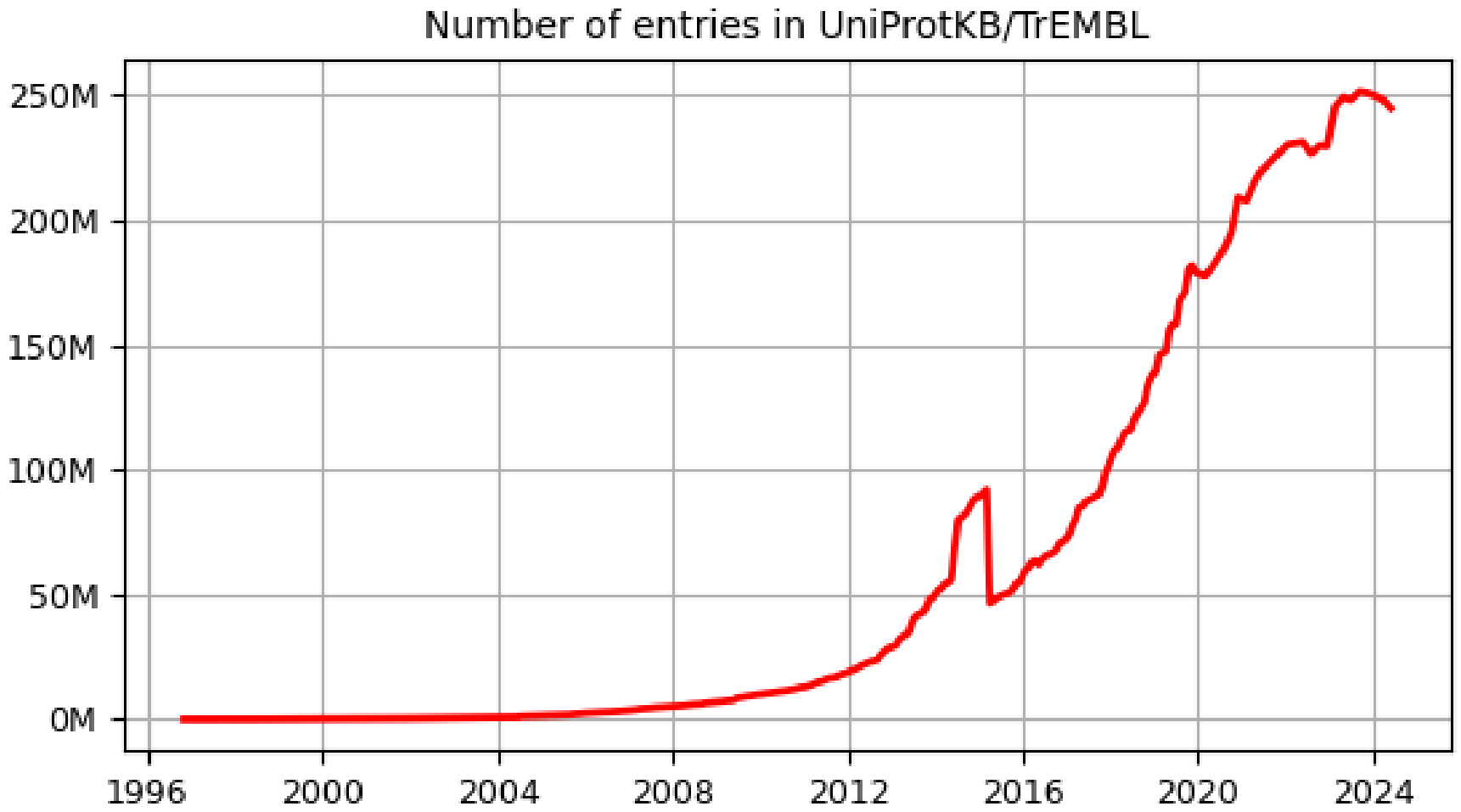


Multiple Sequence Alignments

- One of the most important tools to understand structure function and structure
- Important to align many (all) avilable sequences to understand conservation.
- CLUSTAL, MUSCLE, T-Cofee are useful alignment tools

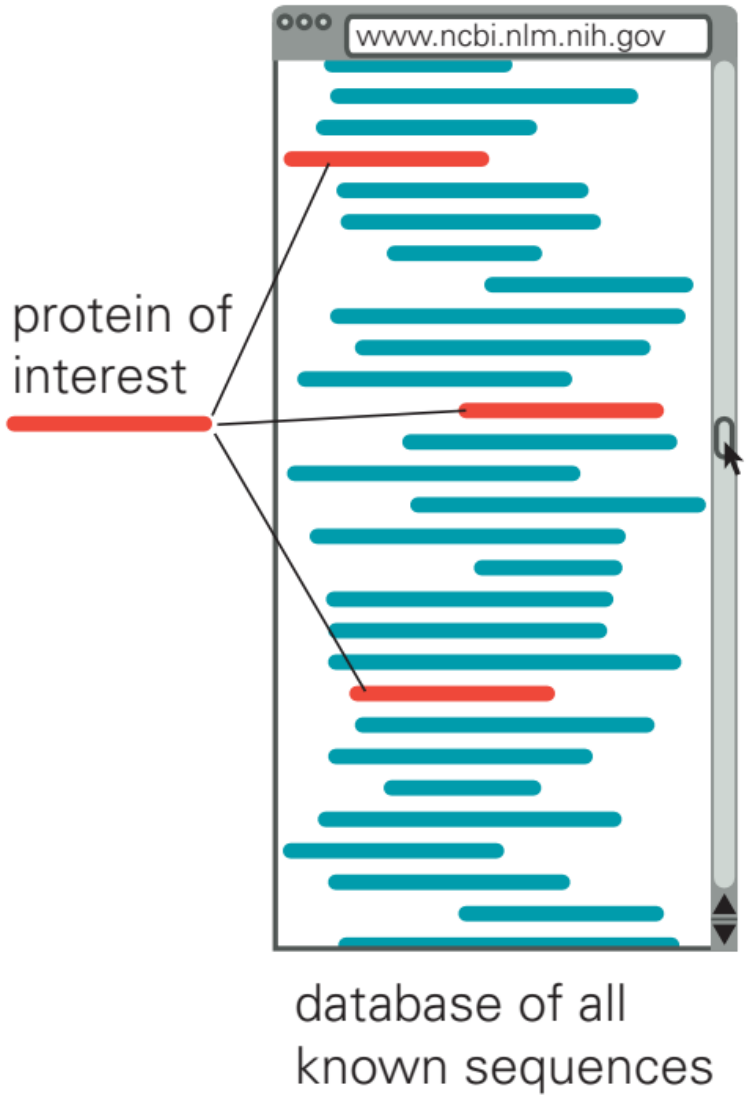


Number of known protein sequences



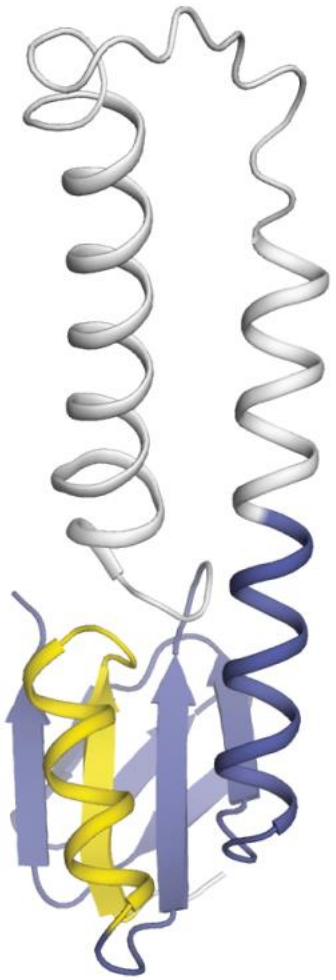
Structure variation is related to sequence variation

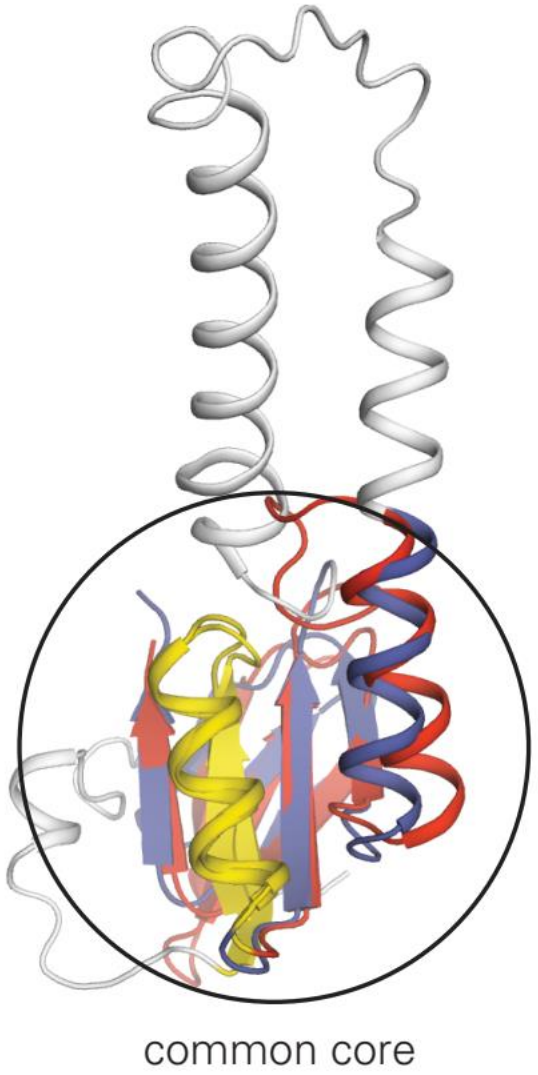
- The protein of interest (is aligned against all known proteins.
- The probability of each alignment being due to related proteins is evaluated using substitution matrices.
- Such searcher are typically performed in “fragments”
- Score are typically returned as the probability for an alignment course to not be due to random chance
  - *E-scores*
- Popular software
  - *Basic Local Alignment Search Tool (BLASTP)*
  - *Hidden Markov Model (HMMER)*

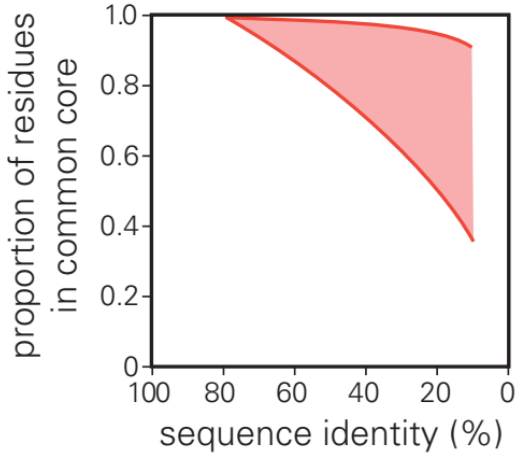


Common structural core

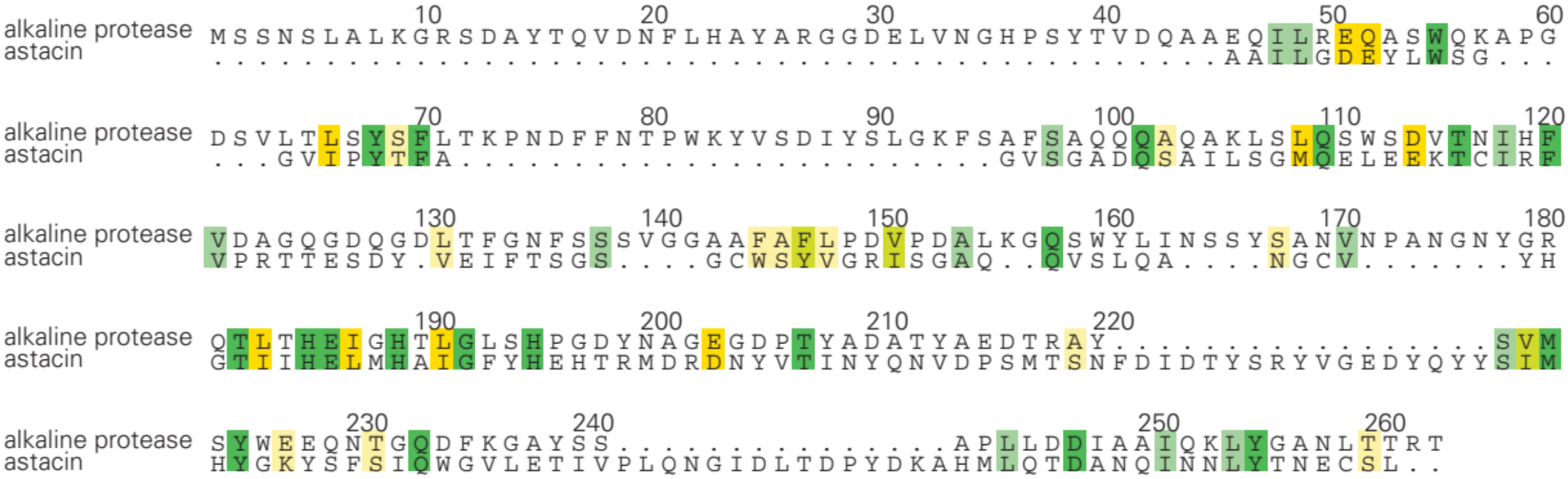
- What can we say about proteins that are not so similar in their sequences?
- We can partition the structures into multiple regions, with and without similarities.
- Proteins can share a common structural core



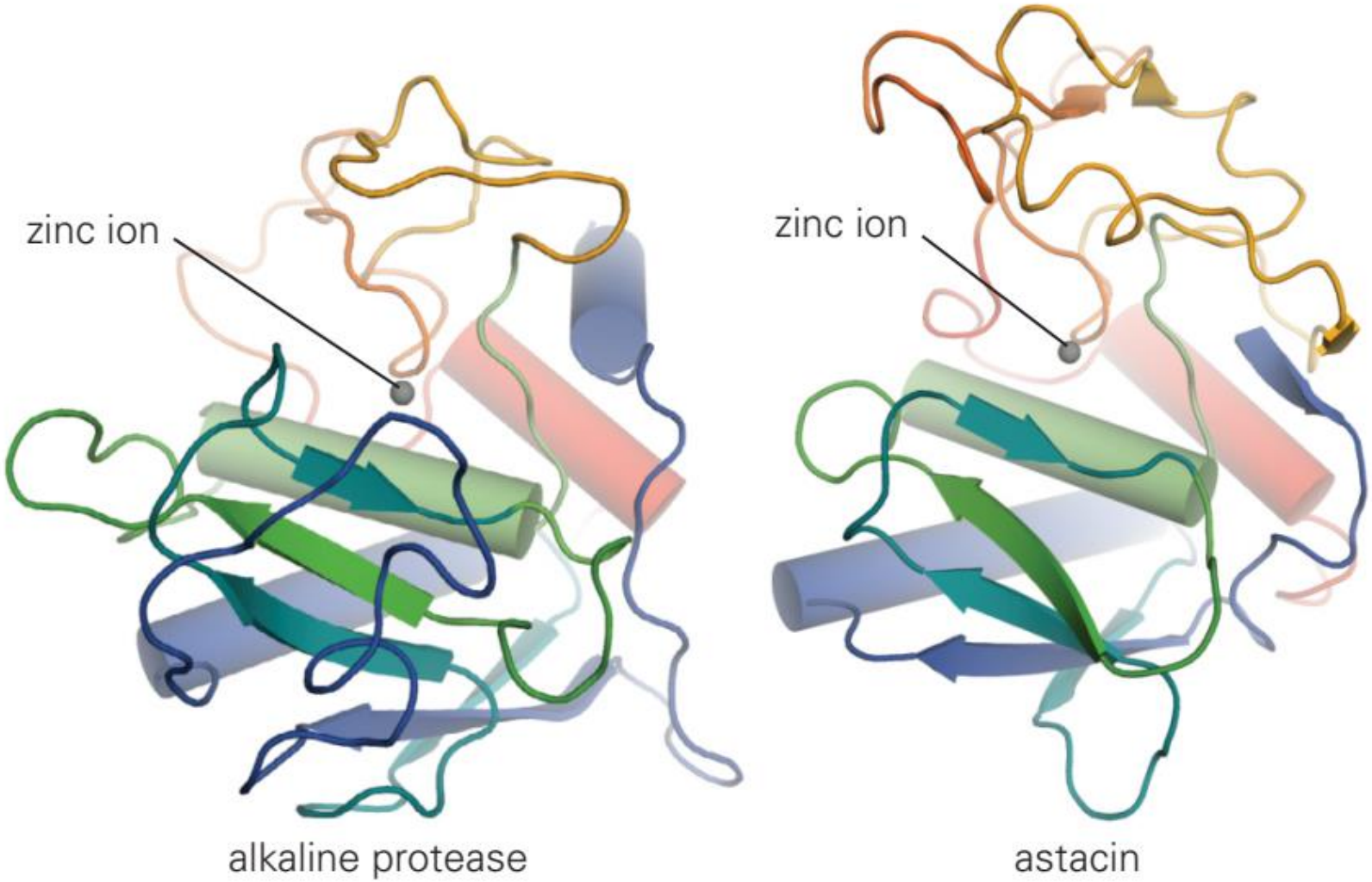




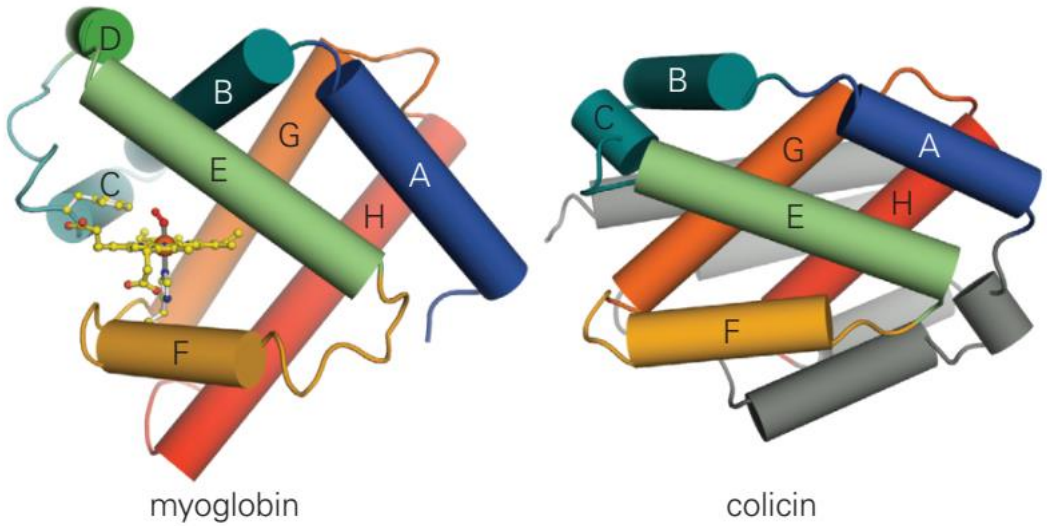
Conservation along the whole sequence can be missleading

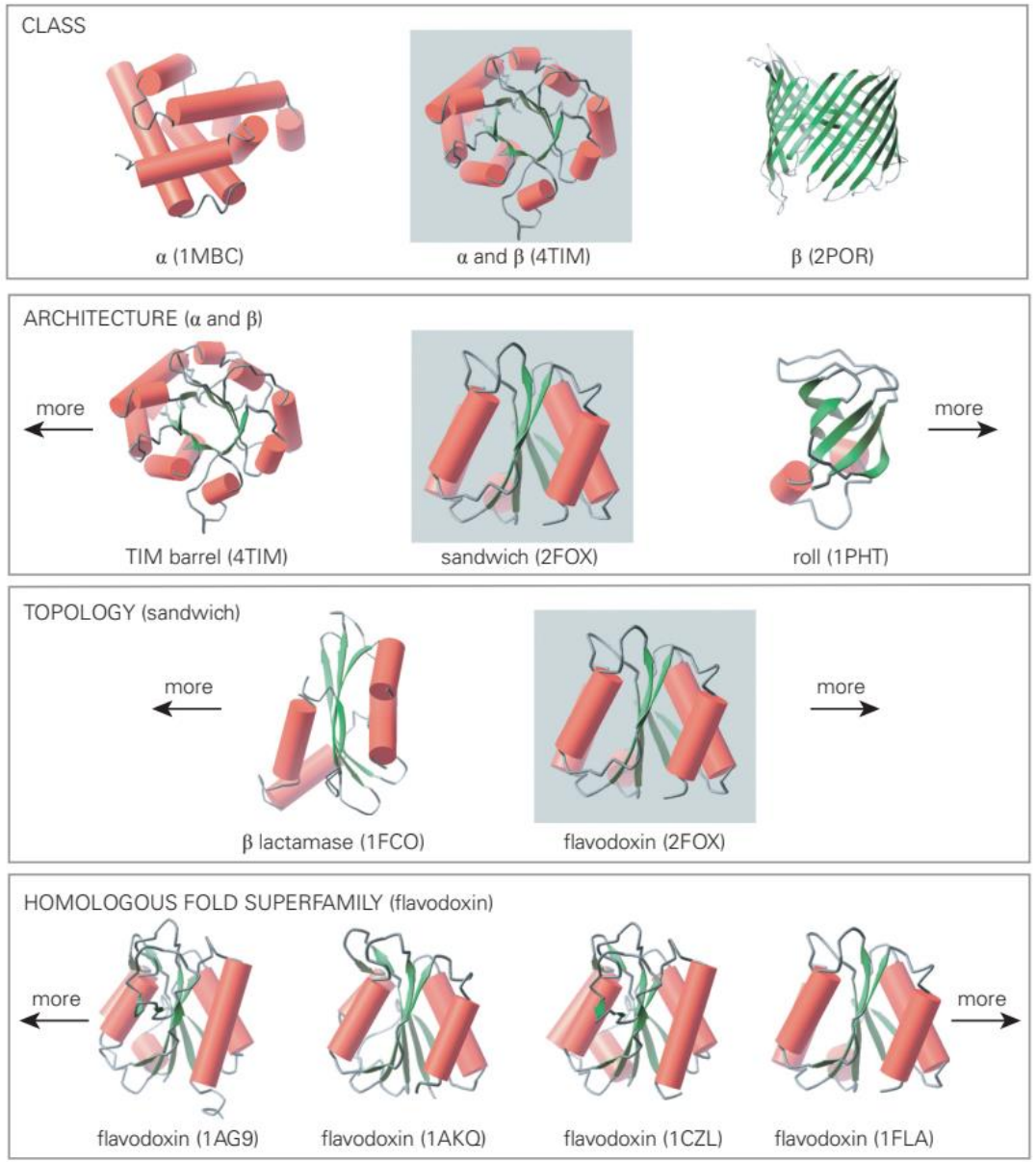


Conservation within the conserved core is important for e.g. catalysis

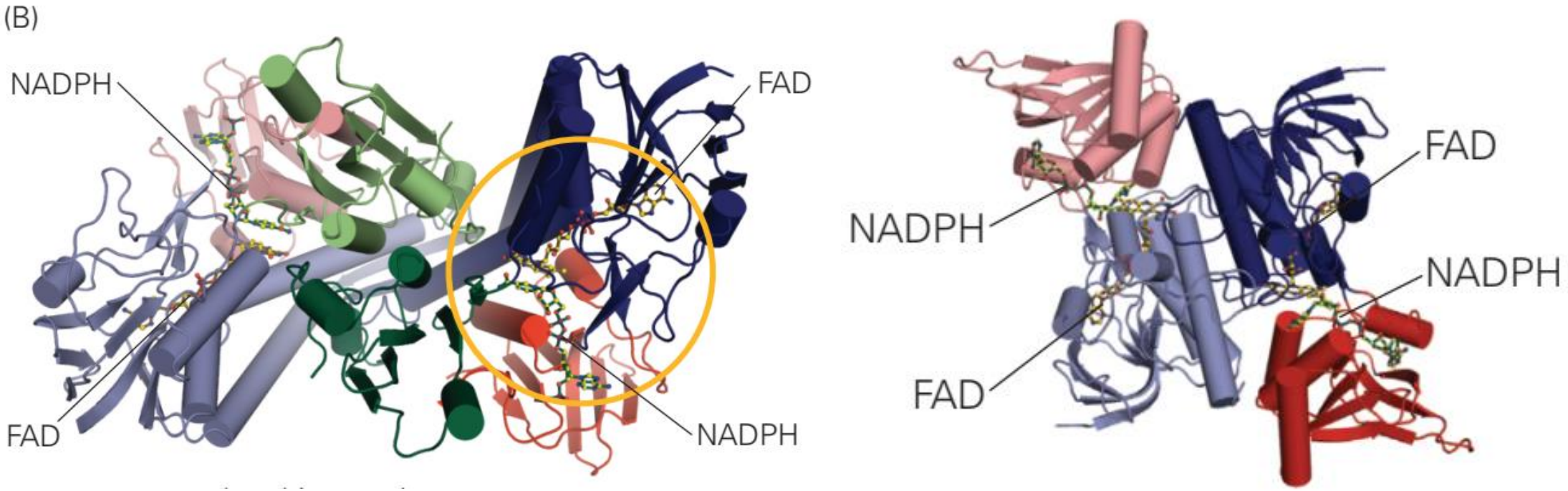


Conserved cores are often reused for different functions

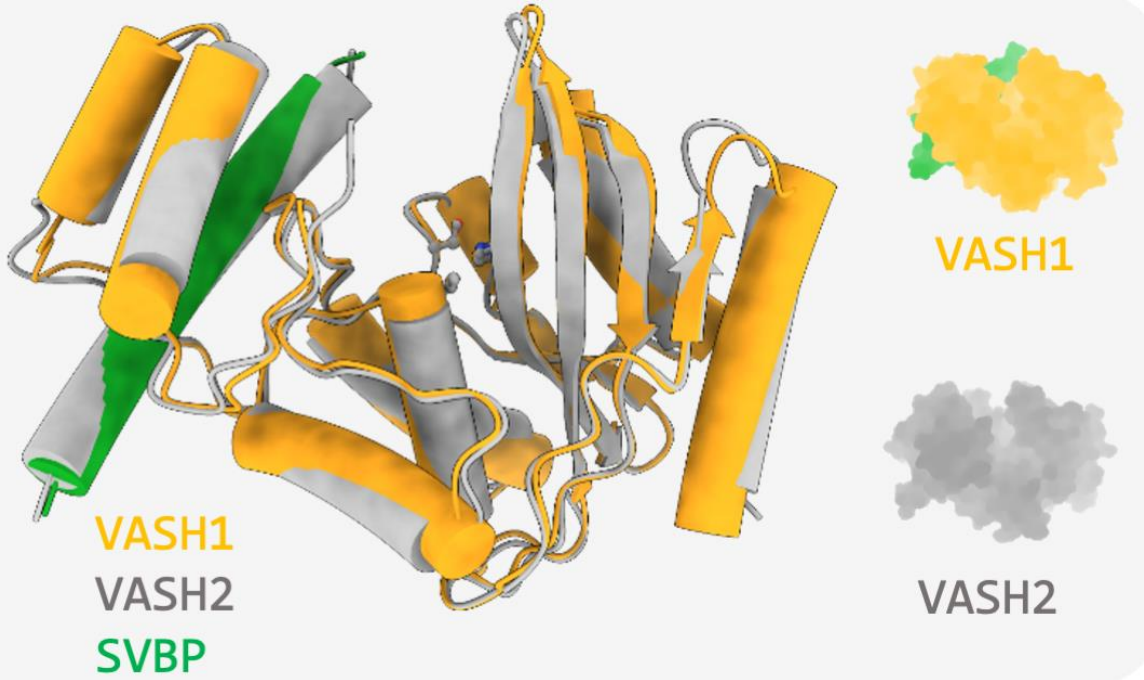


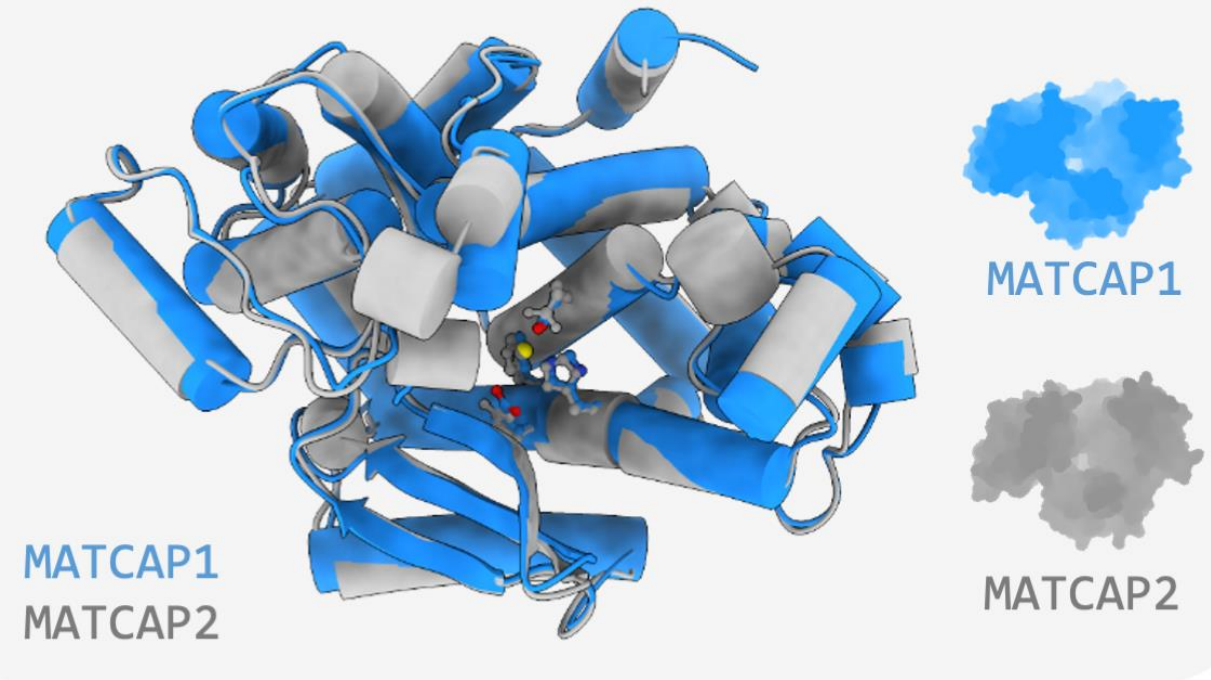


The same function can be served by different folds: convergent evolution



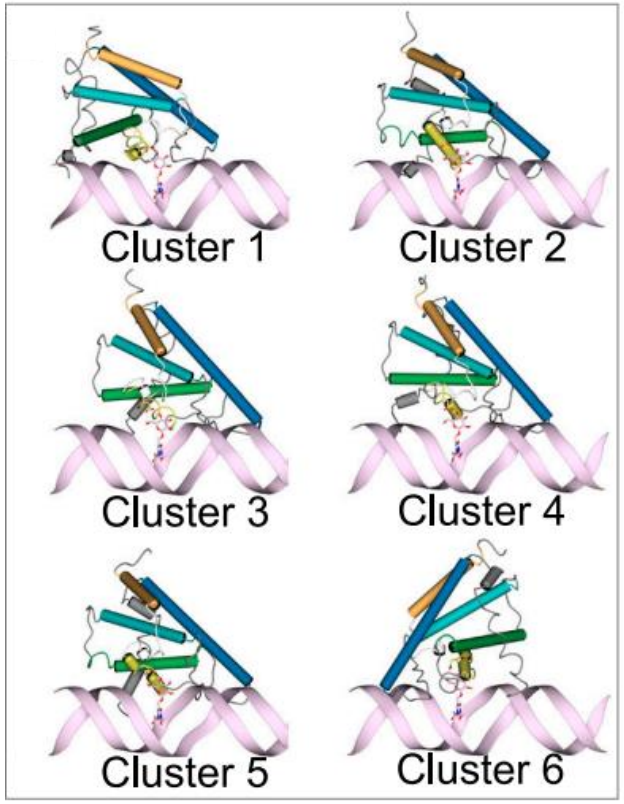
The same function can be served by different folds: convergent evolution

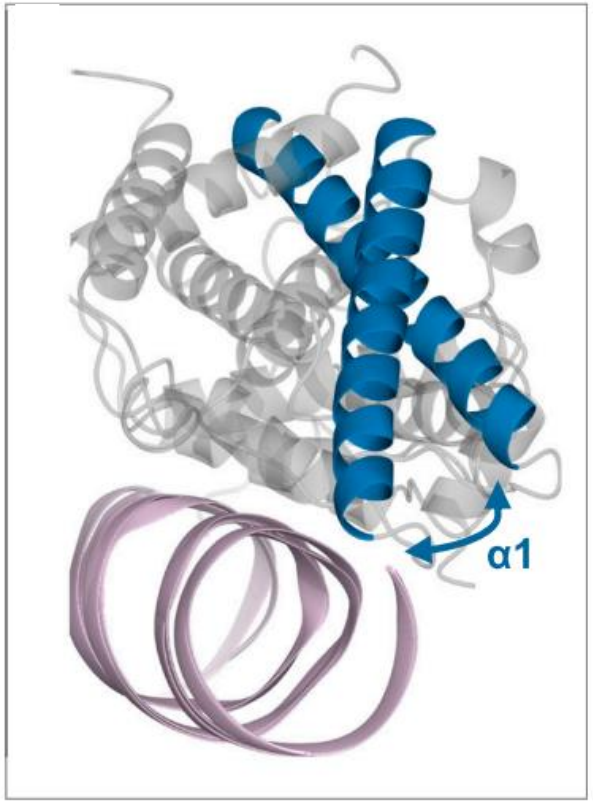




Aligning (superposing) structures

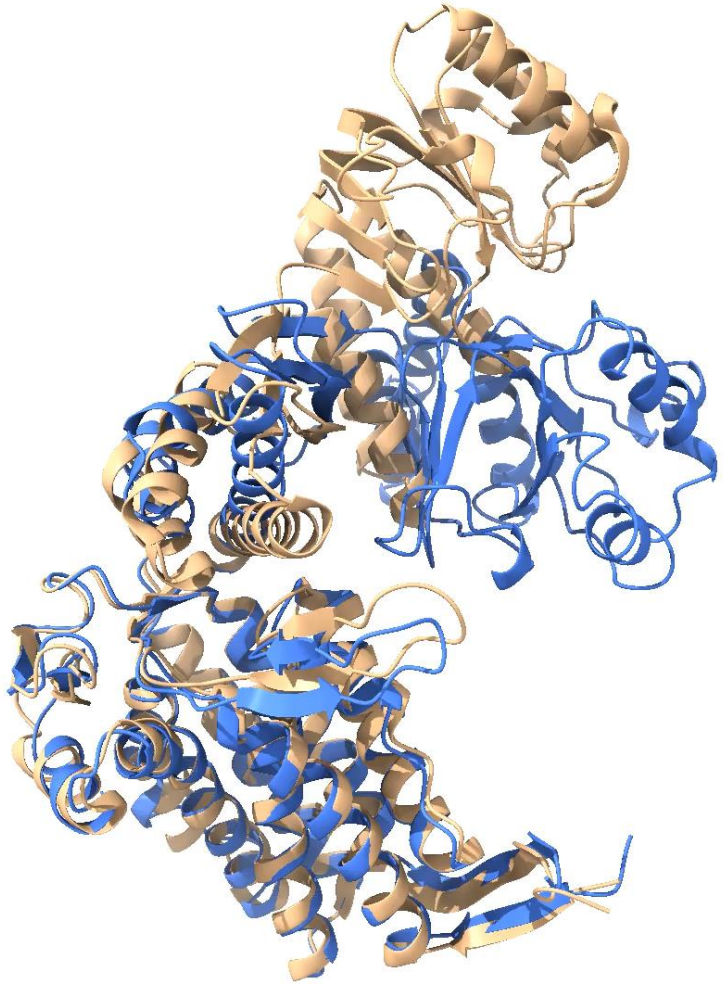
- Structure alignemnt is not identical to sequence alignment.
- Which one is preferred depends to the case study.
- Both can be informative.
- Automated structure alignments (e.g. Chimera-X “matchmaker” or Coot “SSM superpose”) will align automatically two structures.
- You can show superposed structures seperately or overlapped.





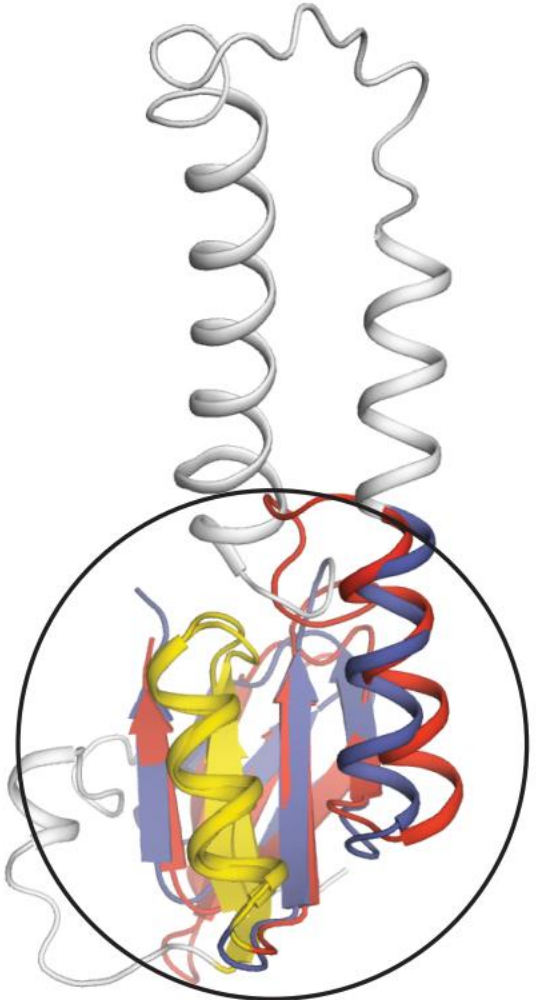
Aligning (superposing) multidomain proteins

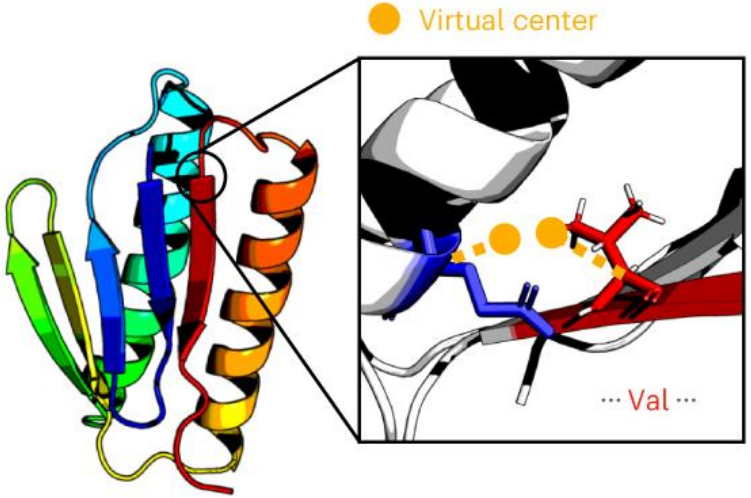
- Proteins are often put together by stiching together different domains.
- Domain movement needs to be taken into account during a structural alignment.
- Tools like RAPIDO, THESEUS, or [*your brain*] can be used to decide how to superpose.
- Superposition-independent structure comparisons are sometimes a good choice.

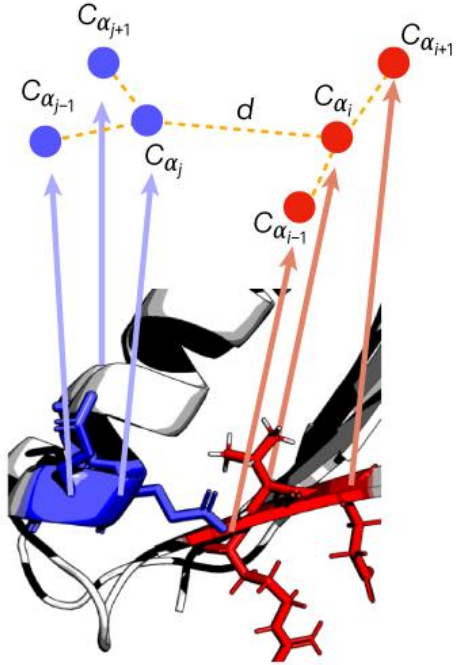


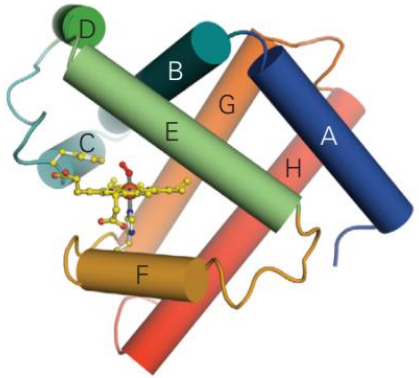
Searching for structure similarity

- Looking for similar folds, regardless of sequence, can detect low similarity or cases that only a very small conserved core is conserved.
- DALI, CE, TM-align: iterative or stochastic optimisation of superpositions.
- CLE, 3D-BLAST, Protein Blocks: describe C-α conformations in a “structural alphabet” and perform sequence alignment.
- FoldSeek: describe tertiary structure as a structural alphabet (1000x faster).

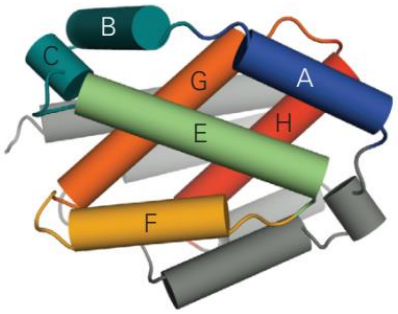








myoglobin



colicin

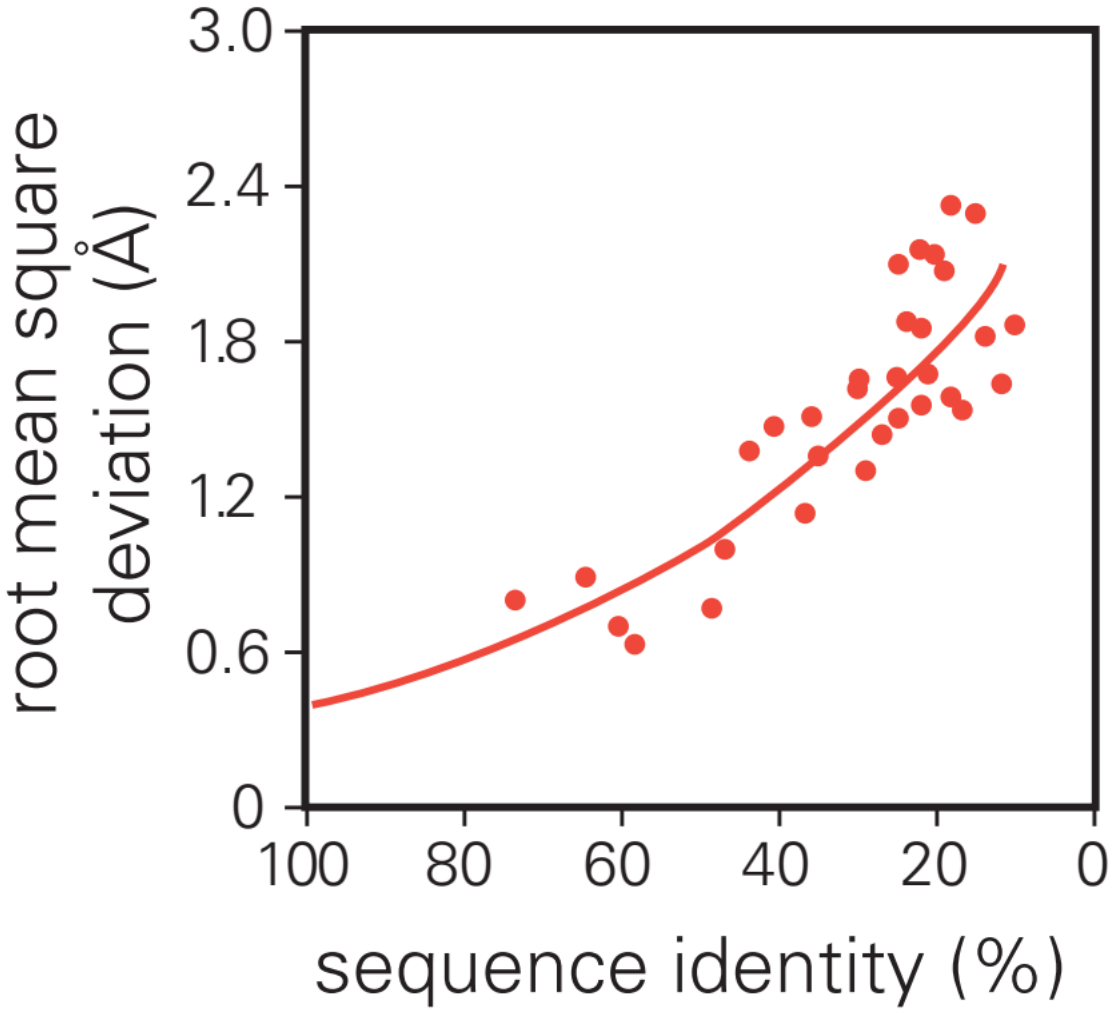
Sequence and structure divergence

- Structure similarity can exist even with non-detectable sequence similarity.
- Structure similarity can exist even without homology
  - *Homology: common evolutionary origin.*
  - *Similarity: based on amino-acid substitution matrix*
  - *Identity: what is says.*
- Defining sequence similarity becomes complex in multi-domain proteins.
- Structure superpositions can help understand conservation.

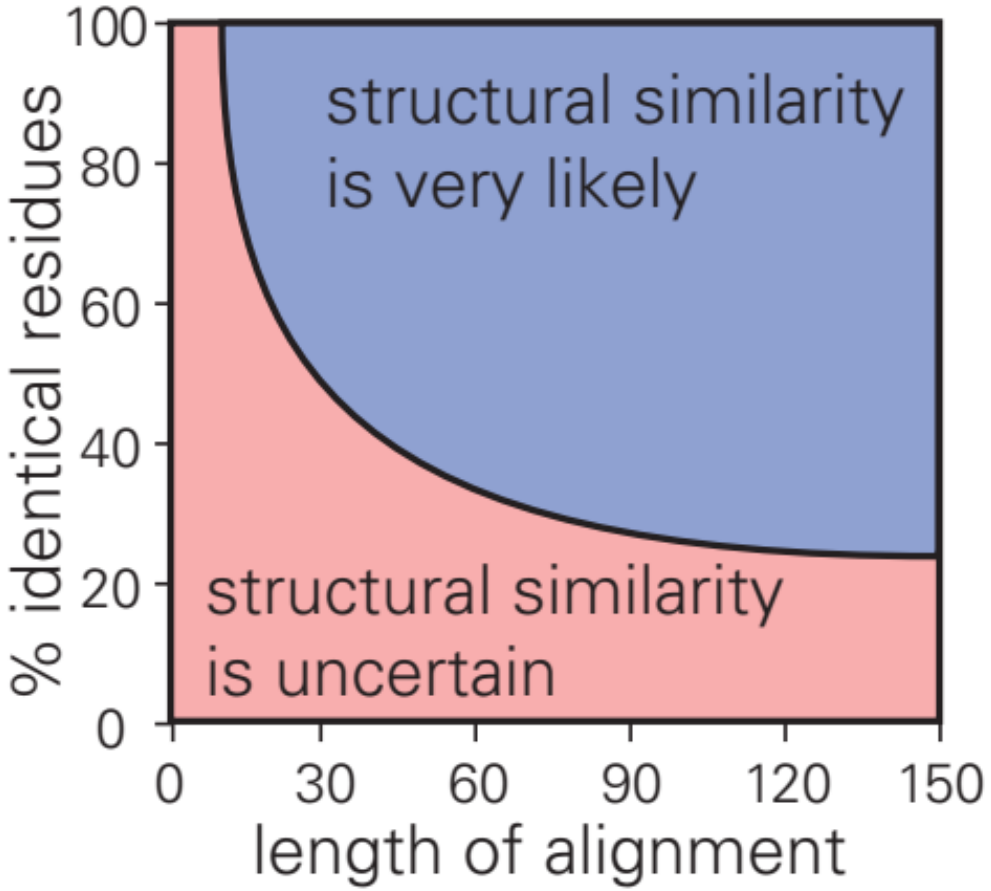
- At what degree does sequence similarity guarantee structure similarity?

Sequence and structure divergence

- A measure of the degree of structural overlap between two proteins is the root mean square (rms) deviation in the positions of Cα atoms in the common core between the two structures.
- The rms deviation should be calculated using residues within the common core, to be meaningful.
- Proteins that are related by ~50% or more sequence identity have common cores that are structurally very similar, with rms deviations tha are less than ~1 Å.
- The rms deviation in backbone positions rises steadily as the level of sequence identity drops, and the deviations can exceed 2Å when the two proteins being compared share less than 20% sequence identity.

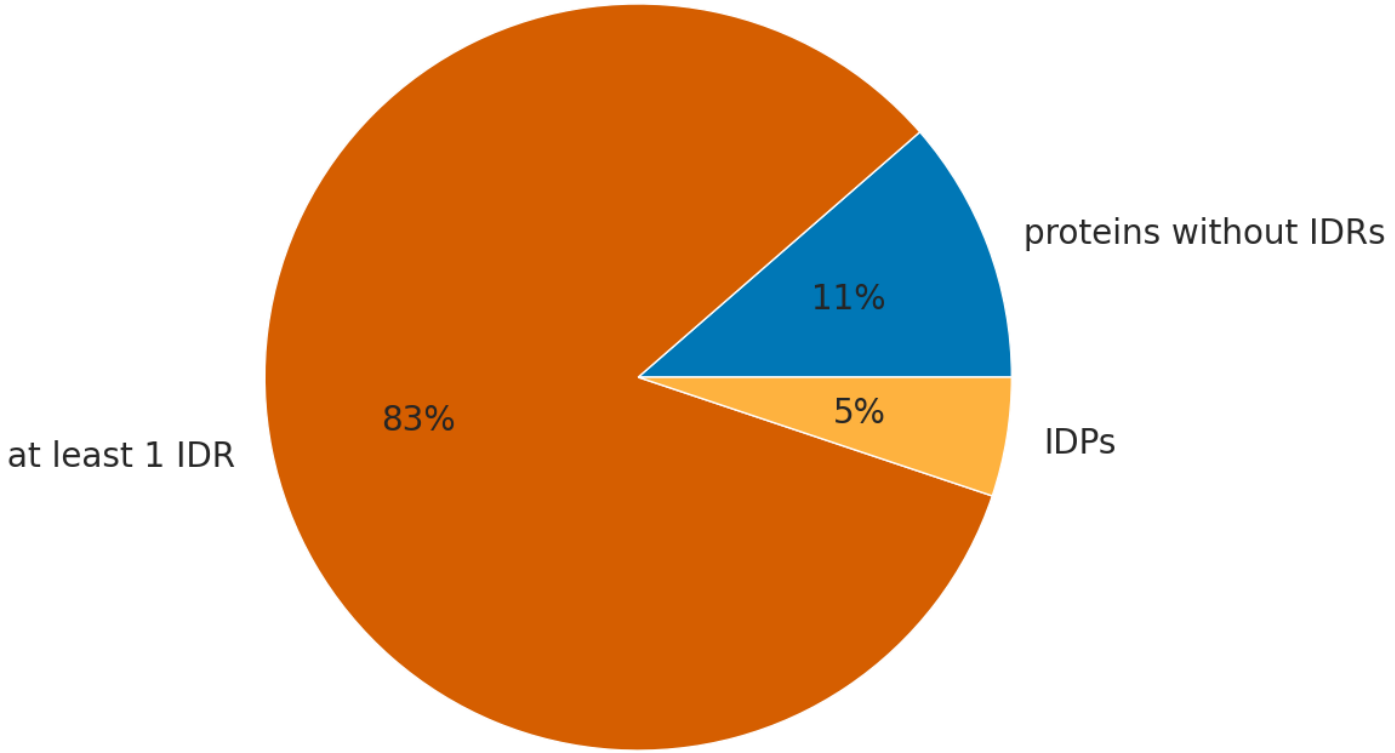


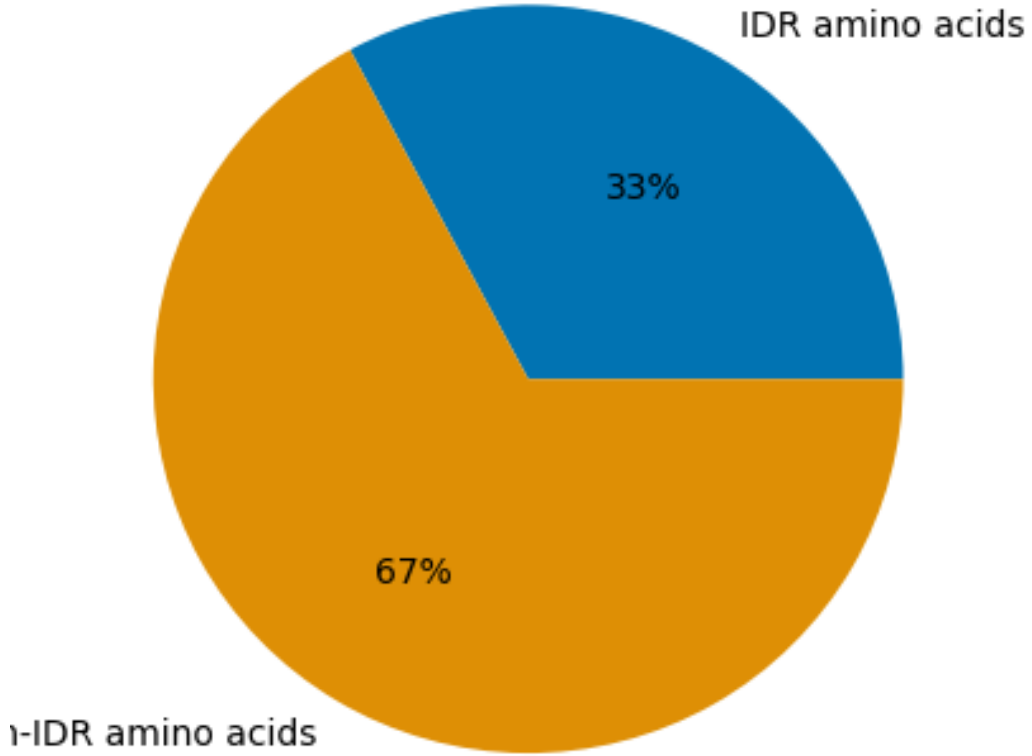
How much can we tell about structure from sequence?



- Sequence comparisons become more informative when the lengths of the segments being compared are ~50 residues or greater.
- Similarity is virtually assured when the sequence identity between longer segments is greater than ~25%.
- Proteins that are related by less than ~25% sequence identity can have significantly different three-dimensional structures.

Intrinsincly disordered proteins & residues



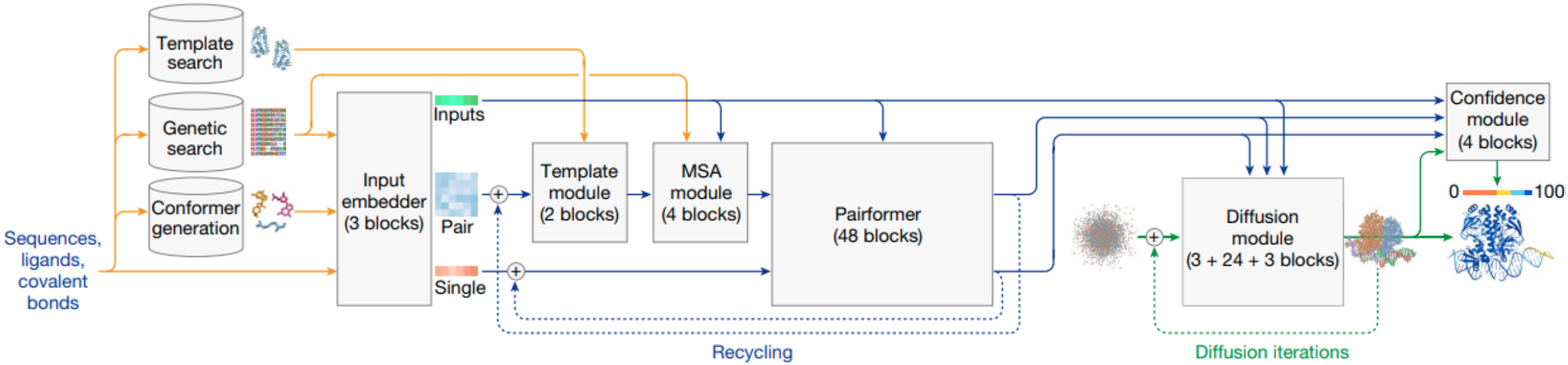


AI redefines the limits on the information that can be extracted from sequences.

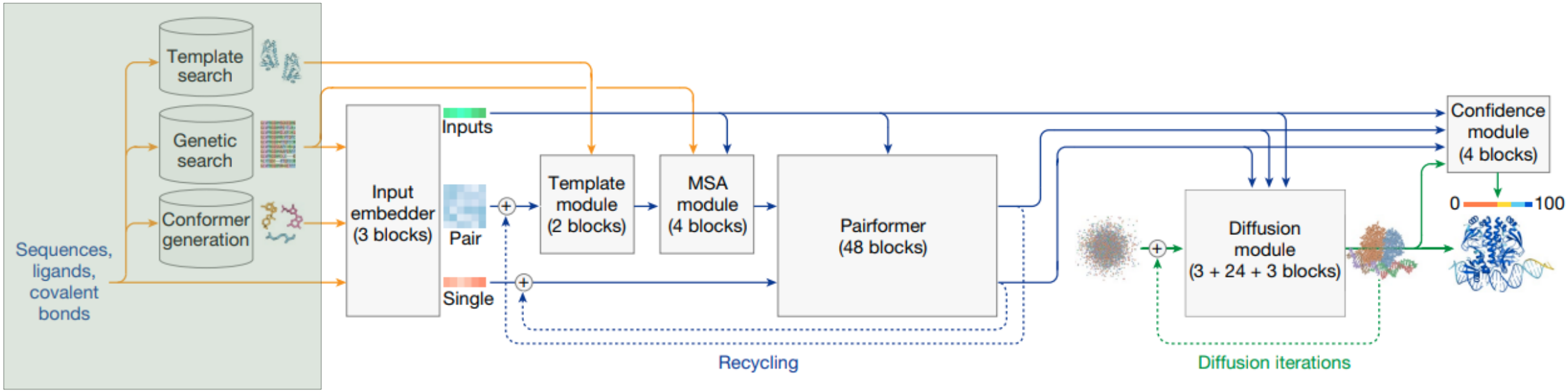
- AlphaFold3 and RosettaAllAtoms do an outstandign job modeling structures from sequence, even with shockingly little similarity.
- All these program learned from the public, free PDB data that costed billions to create and costs millions of public money to extend and maintain.
- You must however, be very critical on what you do with predicted models.

- AlphaFold3 provides excellent validation criteria.

AlphaFold Architecture

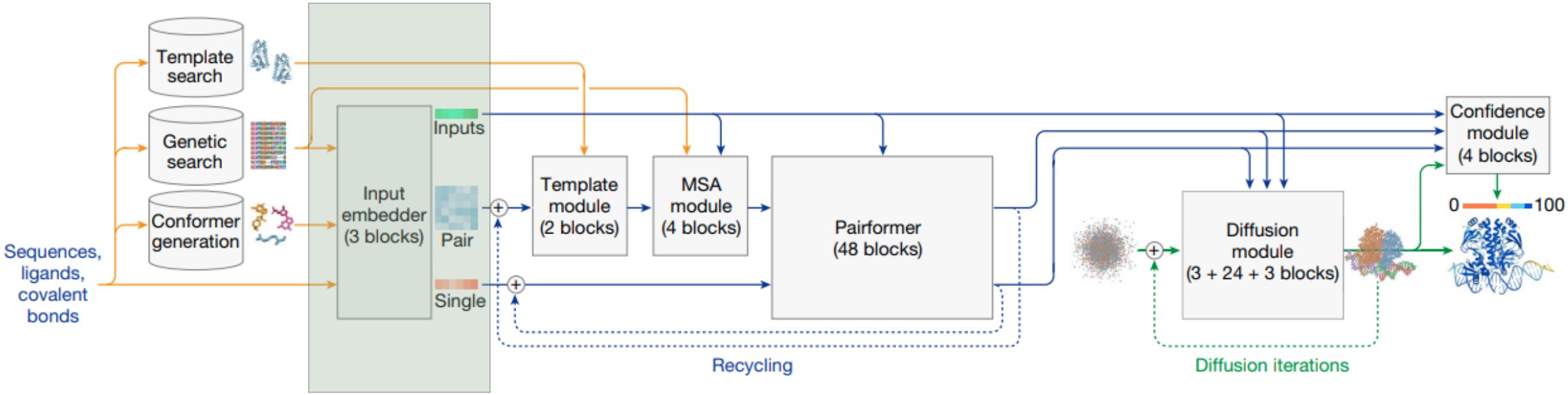


AlphaFold Architecture: input

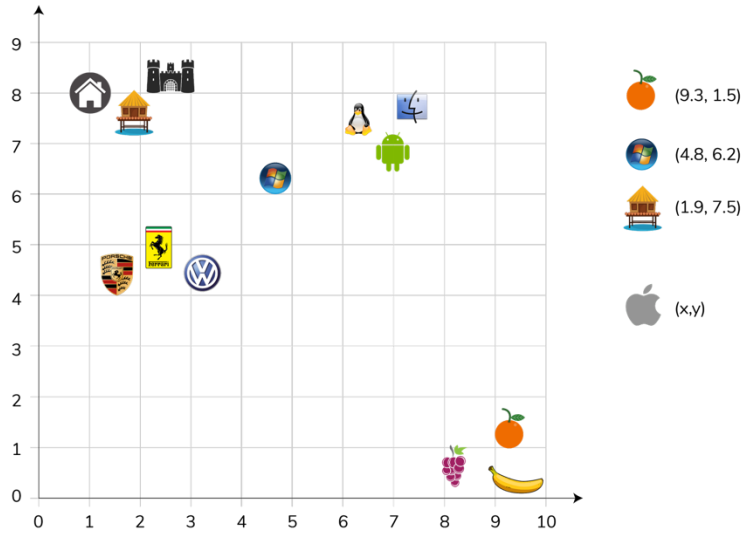


- **Intrinsic Features**
  - Sequence, type, chain, etc.
- **Conformer Search**
  - Representative 3D internal coordinates
- **Genetic Search (MSA)**
  - Evolutionary context
- **Template Search**
  - Known structural matches

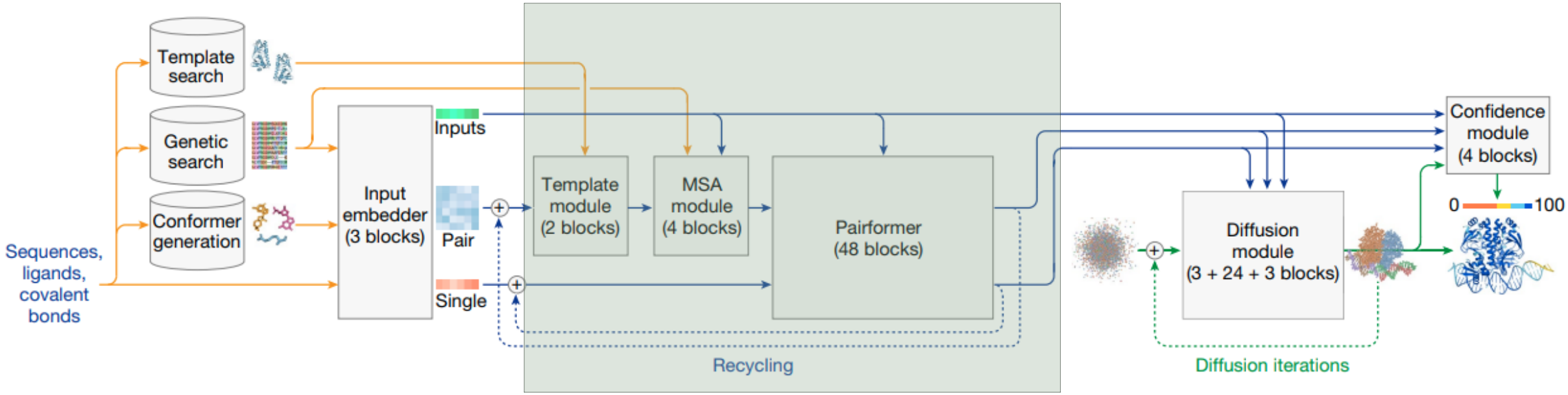
AlphaFold Architecture: embedding



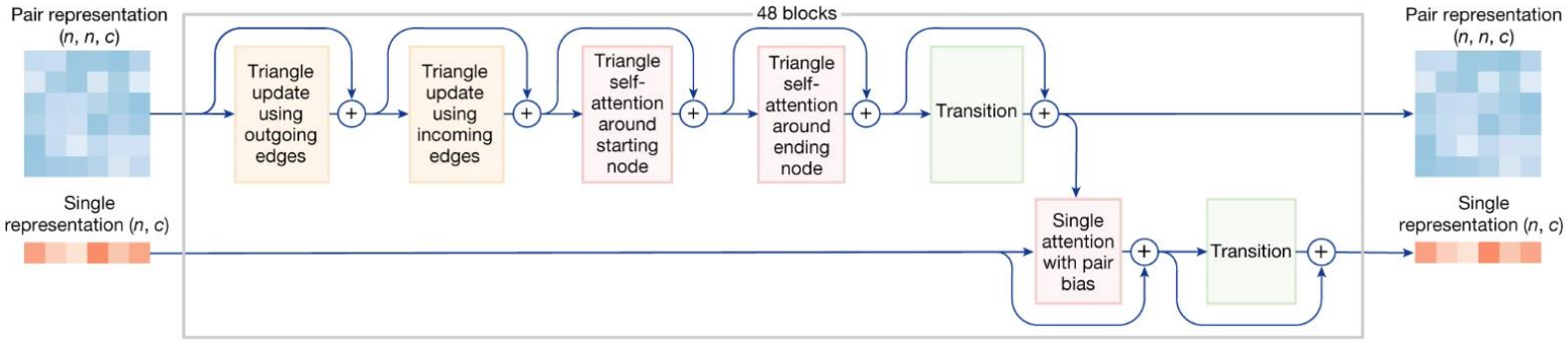
- We need to remember what is our sequence
  - Combining sequence and evolutionary data
- Encodes relation between tokens within the complex
- Represents retrieved annotation for all tokens within the complex



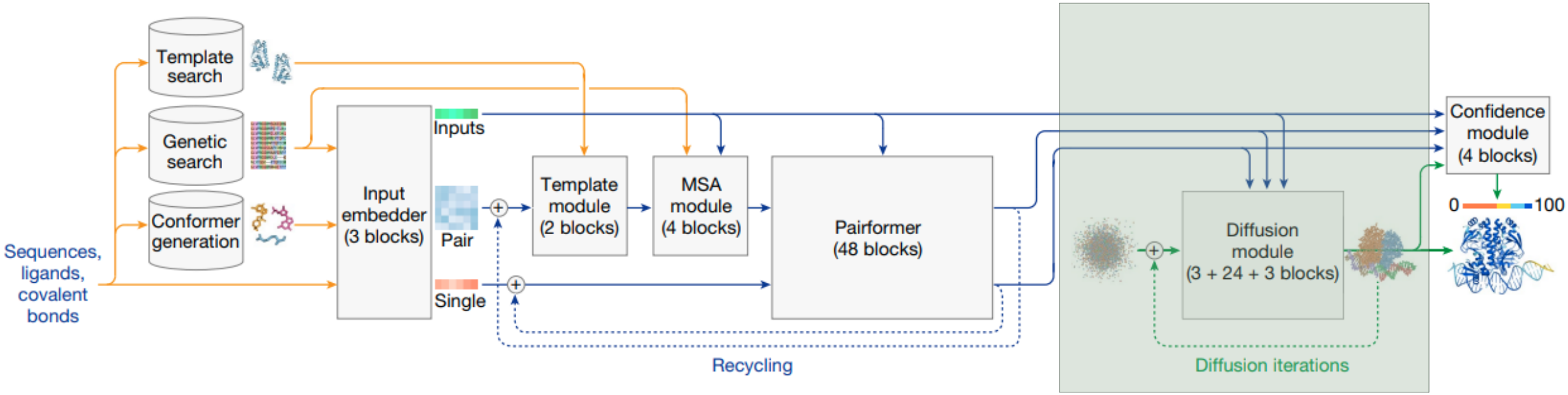
AlphaFold Architecture: transformer



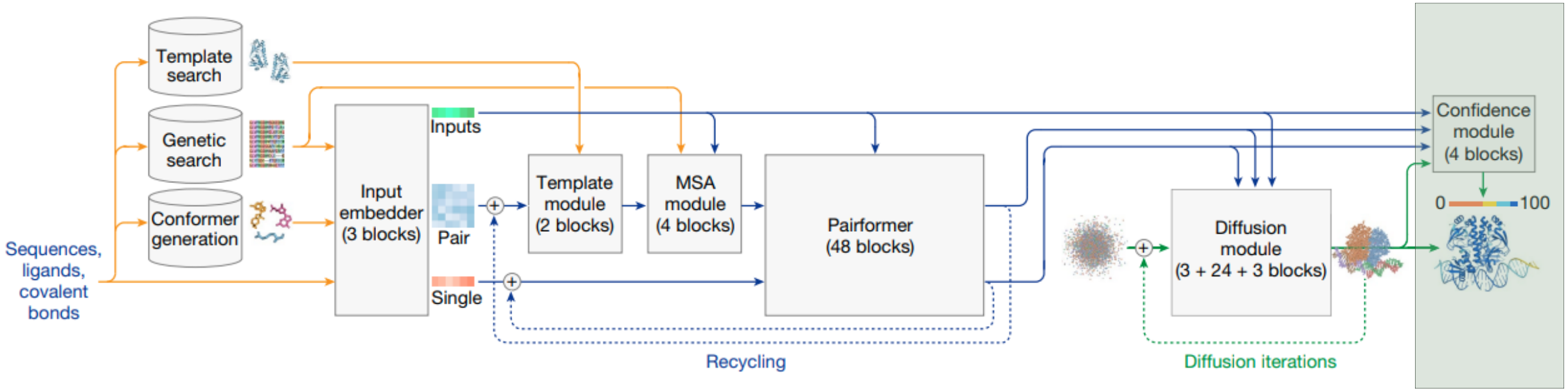
- **Bringing context to the input to focus attension**
  - Template structure of similar sequences
  - Attention to each row of the alignment
  - Pair representation to bring attention to tokens keeping triangular relationships;



AlphaFold Architecture: diffusion



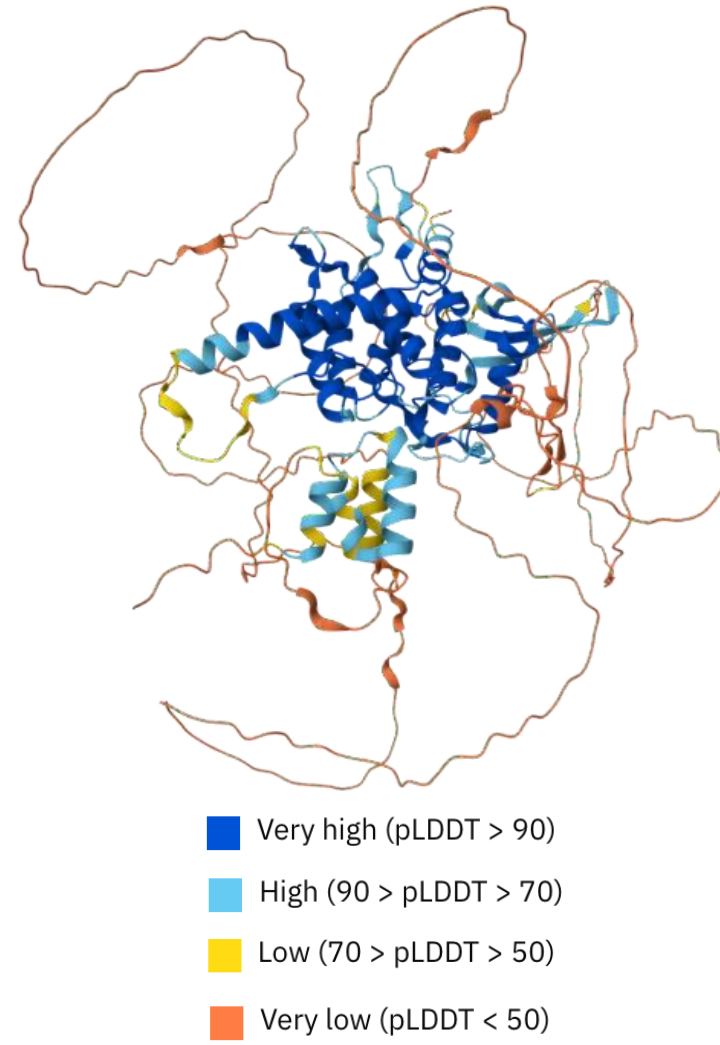
AlphaFold Architecture: output



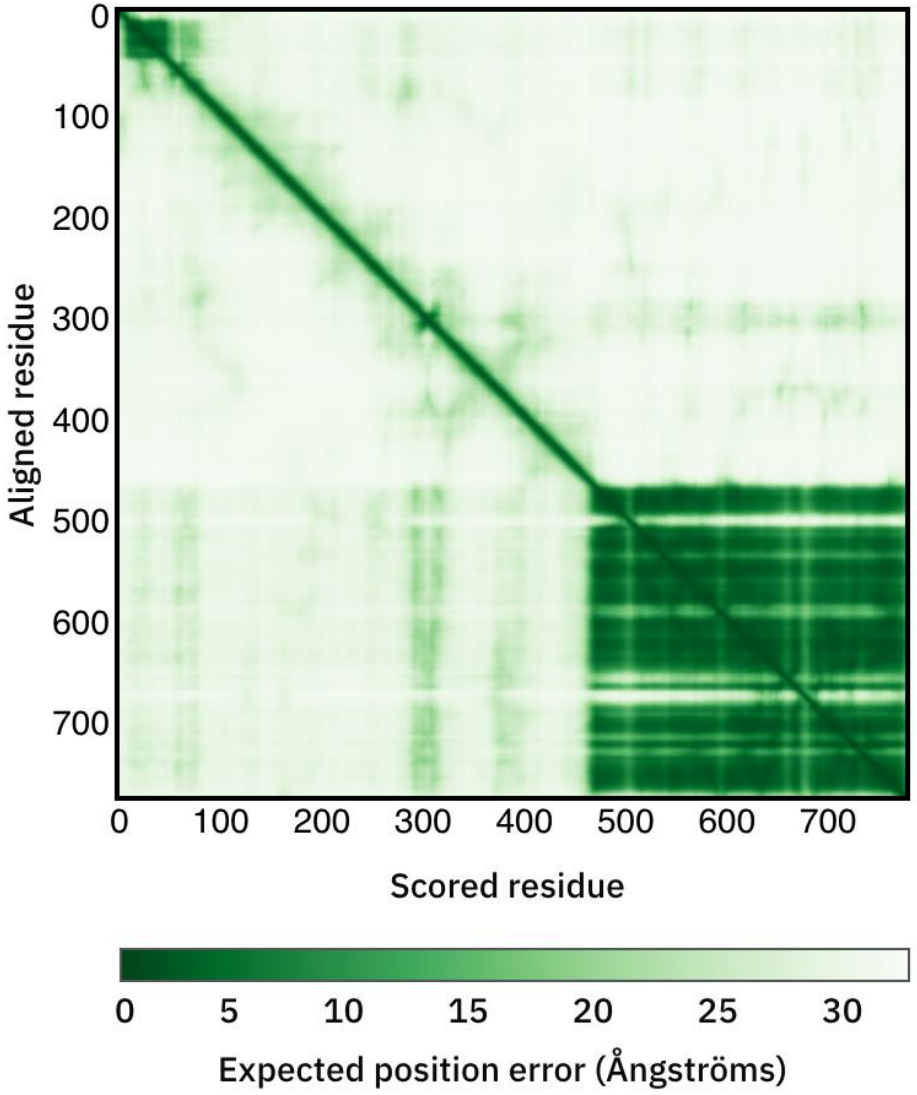
- Confidence metrics!
  - PLDTT
  - PAE

pLDDT:predicted local distance difference test

- AlphaFold produces a per-residue estimate of its confidence on a scale from 0 - 100. This confidence measure is called pLDDT
- AlphaFolδ models are colored by pLDDT
  - Regions with pLDDT > 90 are expected to be modelled to high accuracy. These should be suitable for any application that benefits from high accuracy (e.g. characterising binding sites).
  - Regions with pLDDT between 70 and 90 are expected to be modelled well (a generally good backbone prediction).
  - Regions with pLDDT between 50 and 70 are low confidence and should be treated with caution.
  - The 3D coordinates of regions with pLDDT < 50 often have a ribbon-like appearance and should not be interpreted.

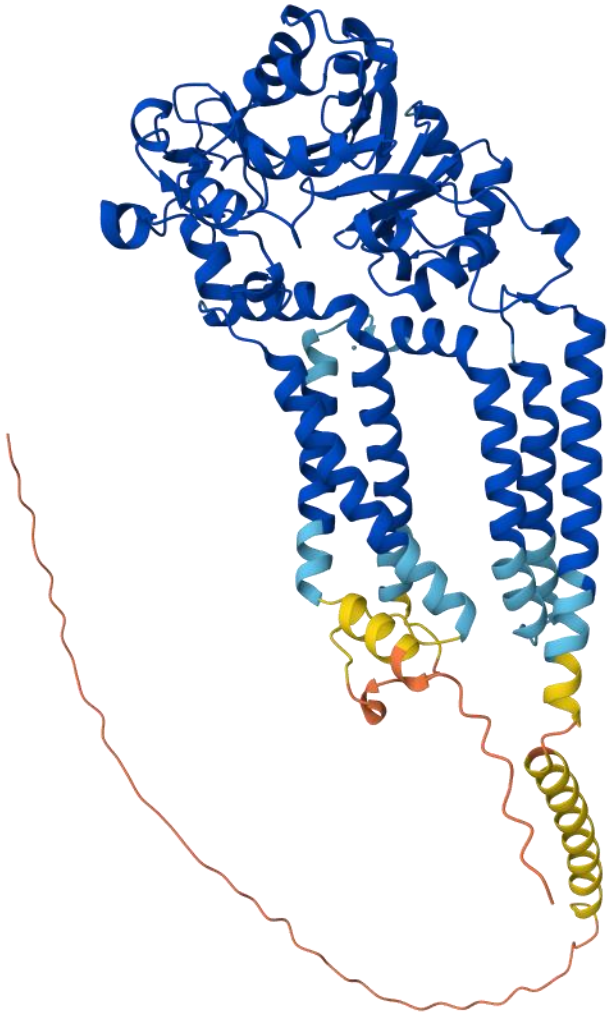


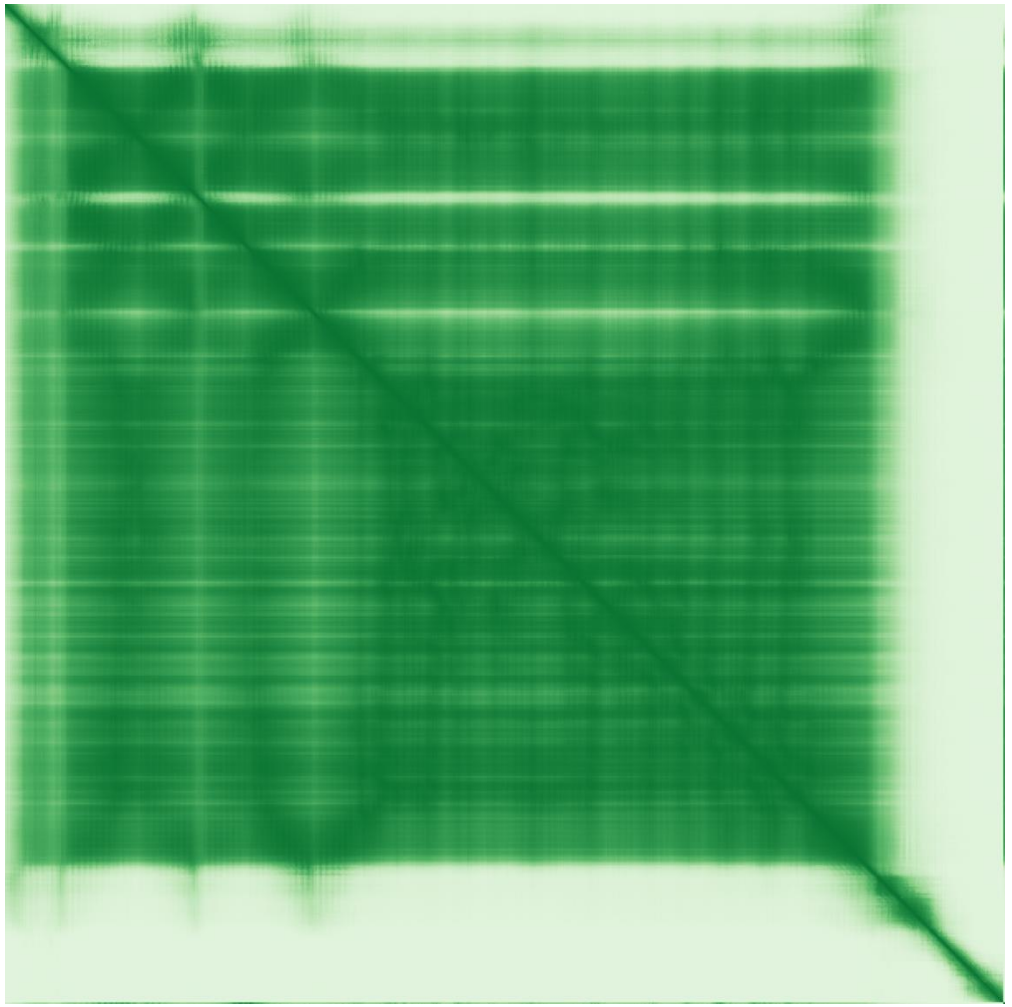
PAE: predicted alignment error



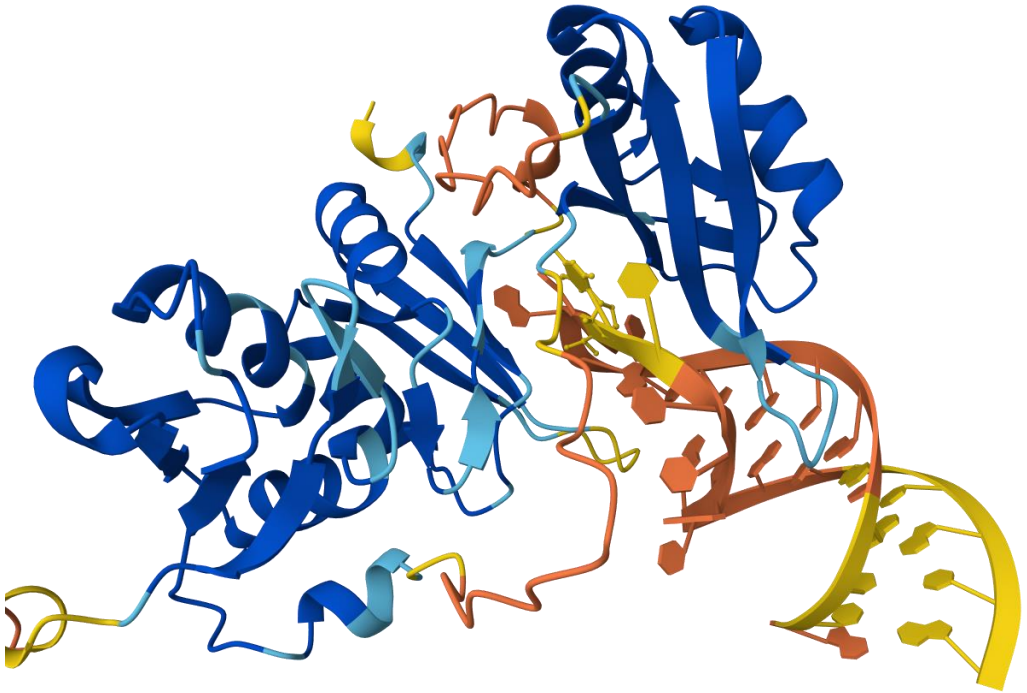
- The colour at (x, y) indicates AlphaFold's expected position error at residue x if the predicted and true structures were aligned on residue y.
- PAE within a structural core – domain will be low (dark green)
- If the PAE is generally low for residue pairs x, y from two different domains, it indicates that AlphaFold predicts well-defined relative positions and orientations for them.
- If the PAE is generally high for residue pairs x, y from two different domains, then the relative positions and/or orientations of these domains in the 3D structure are uncertain and should not be interpreted.
- The PAE x-y is not the same as y-x

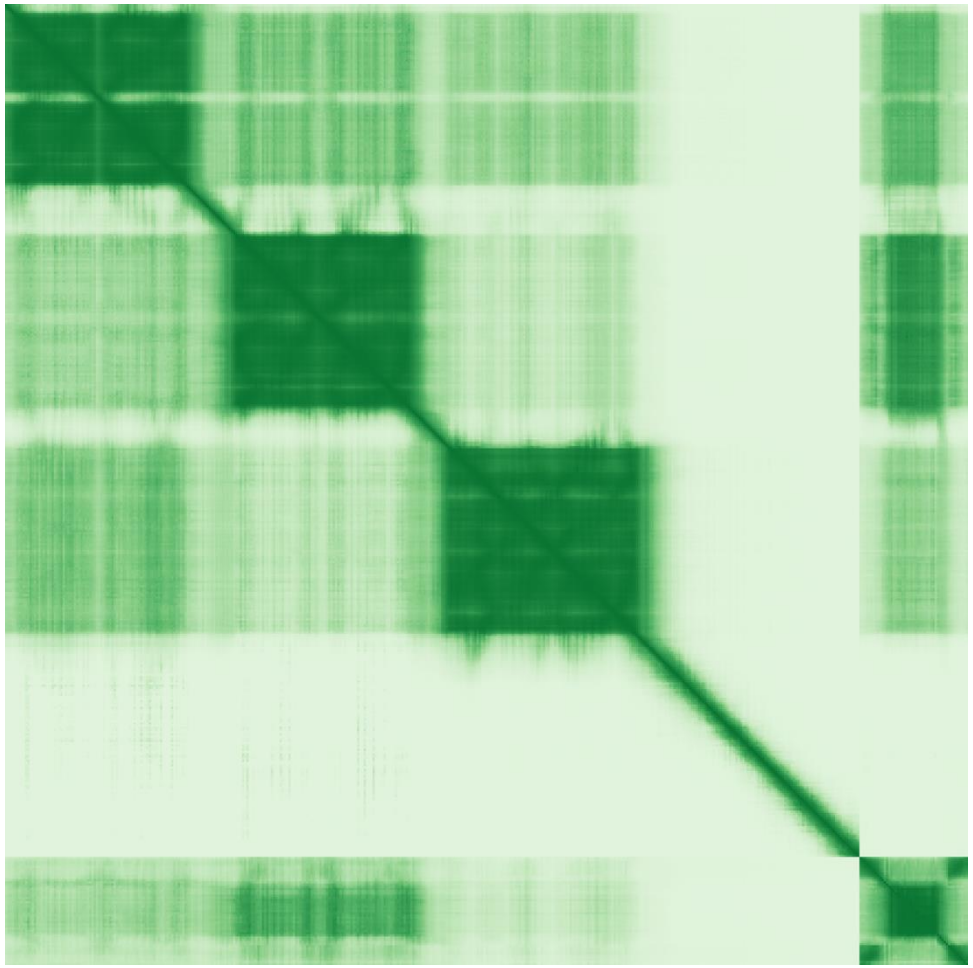
A membrane protein



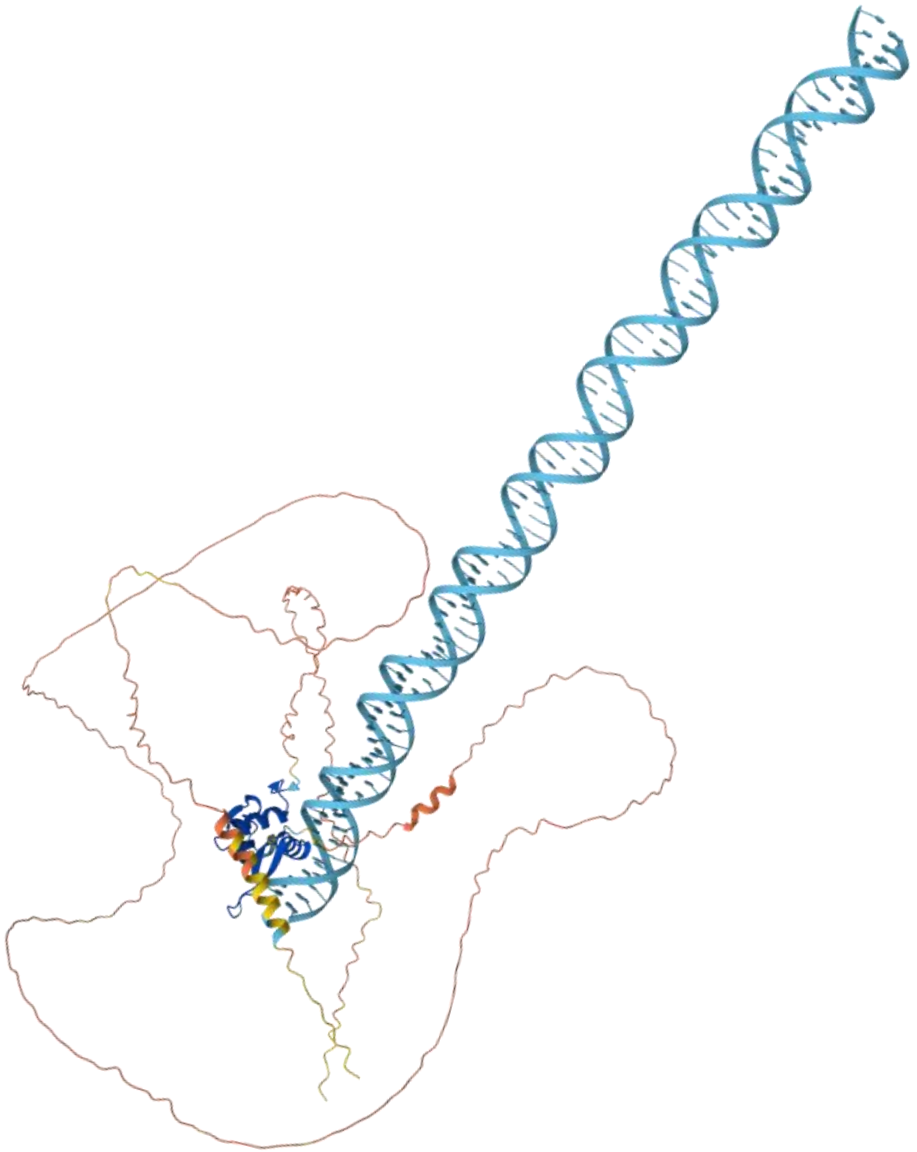


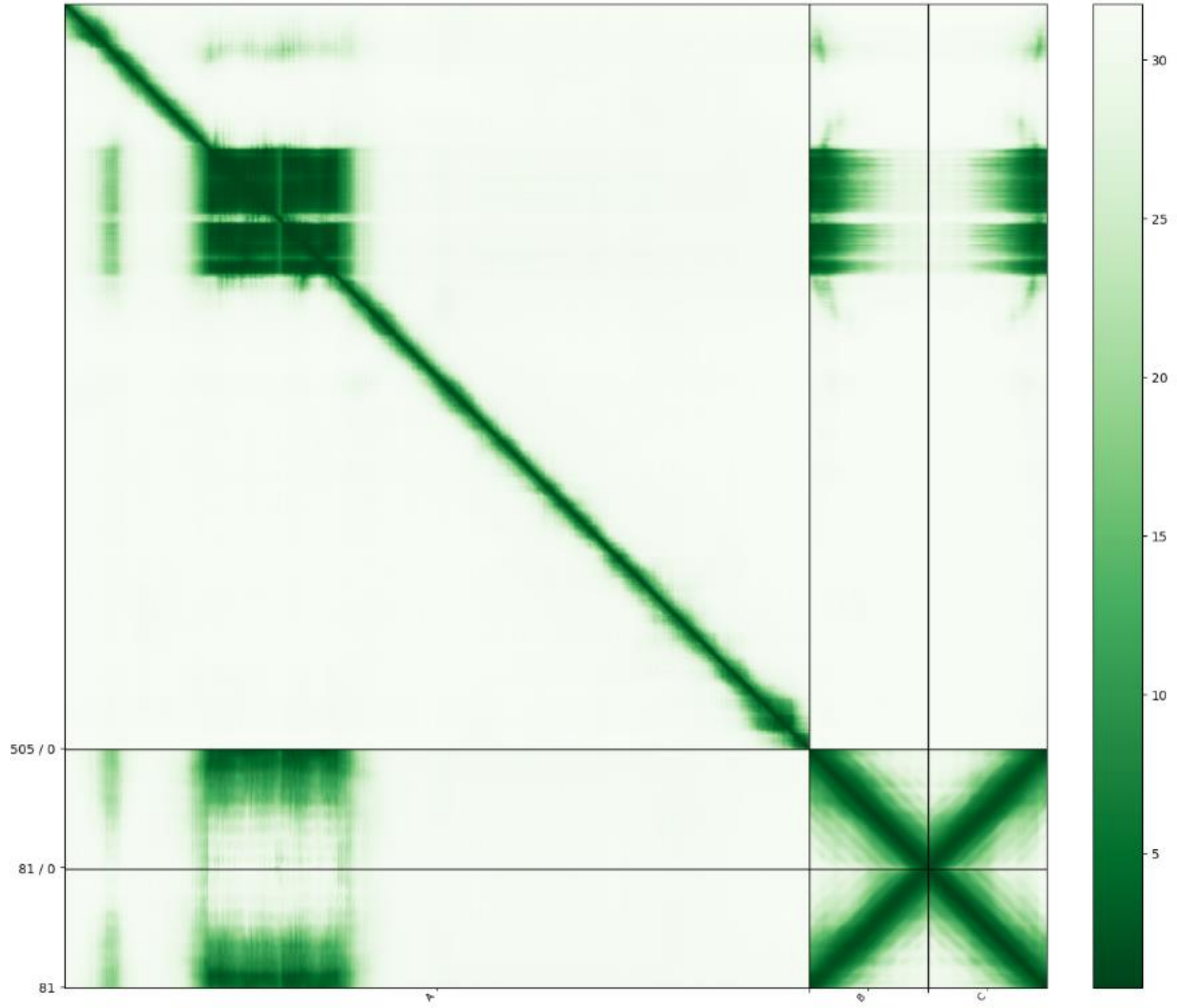
Protein-RNA interaction





Protein-DNA interaction





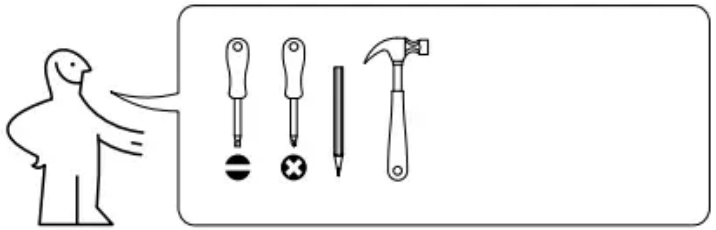
On the asymmetry of the PAE matrix

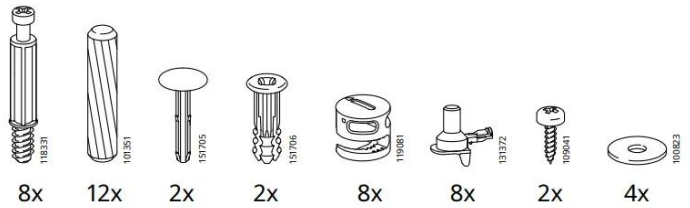
- From the perspective of the lighthouse the position of the boat cannot be accurately determined
- From the perspective of the boat the position of the lighthouse is clear.

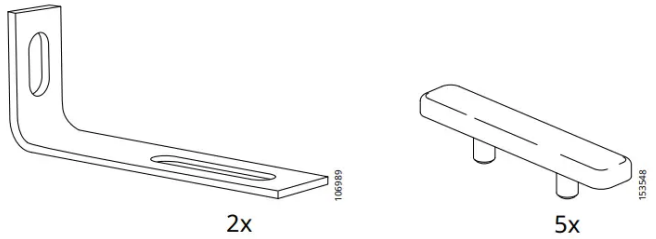
MAKING SENSE OF AI

In macromolecular structure prediction

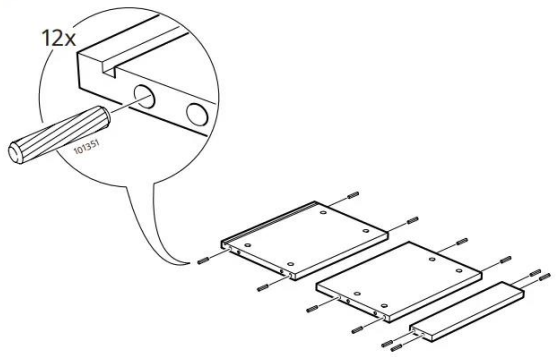
how is a cell assembled?



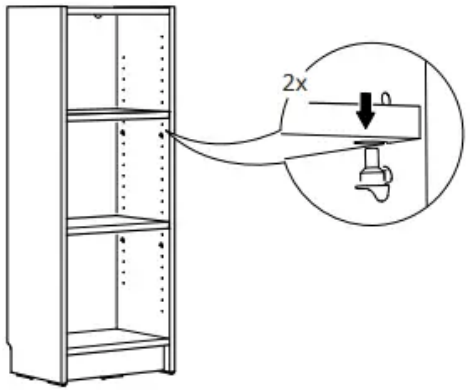




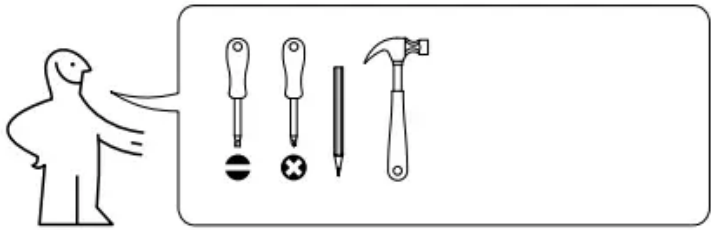
1

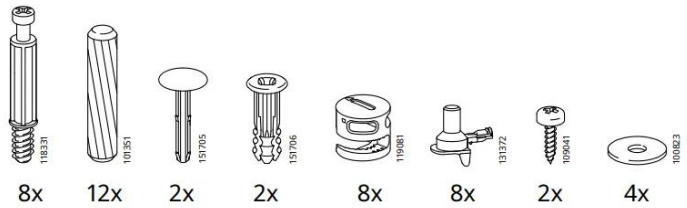


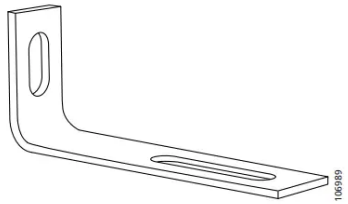
15

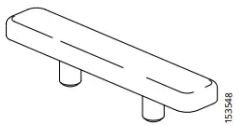


how is a cell assembled?

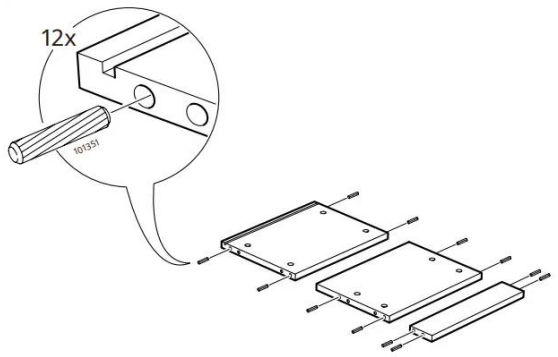






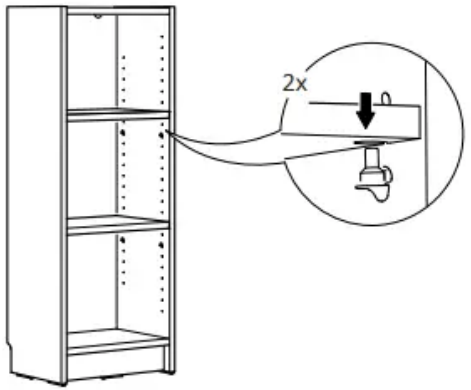


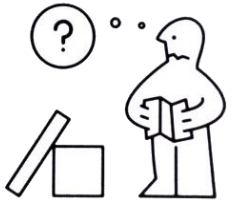
1

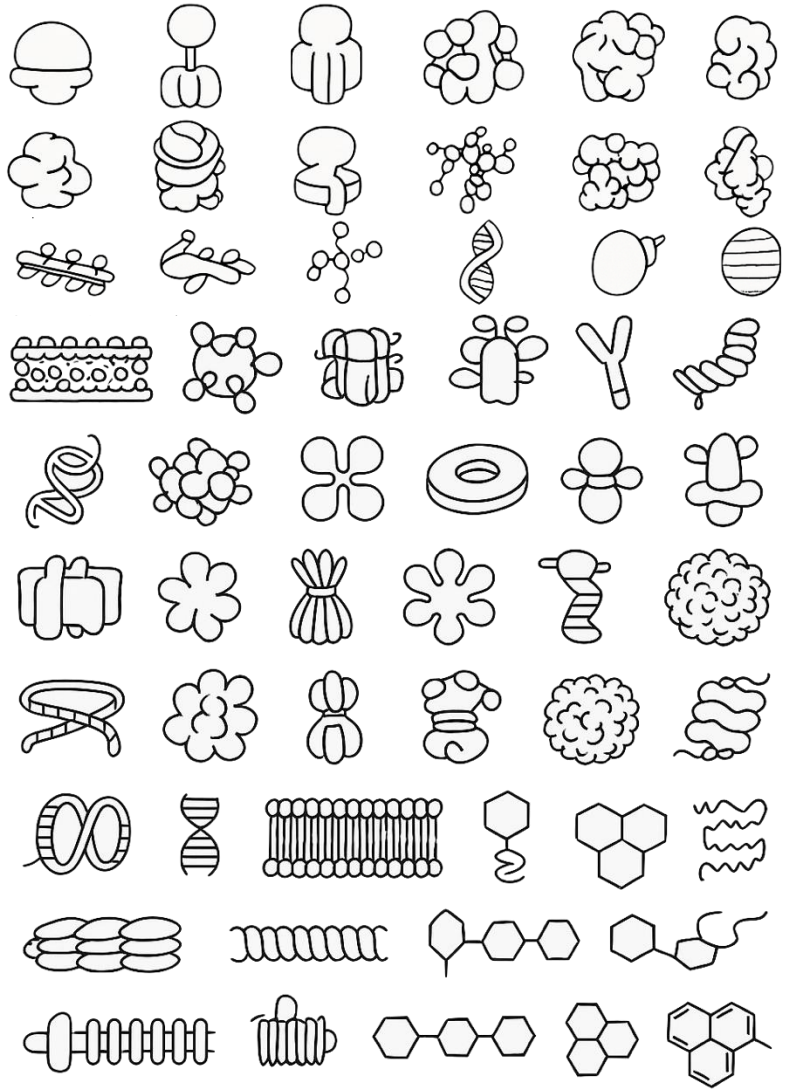


...

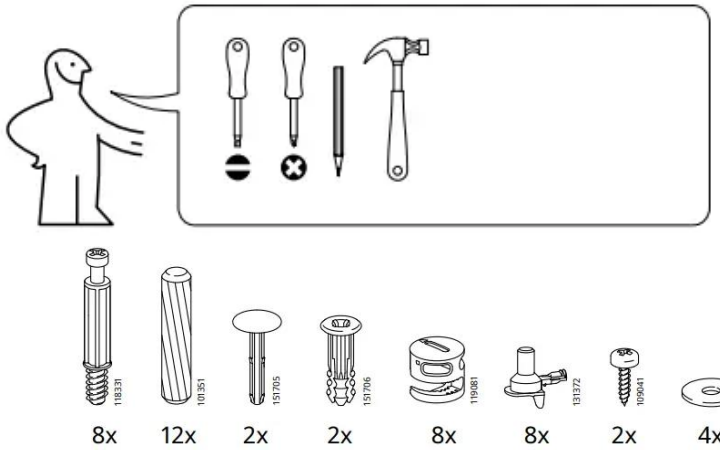
15

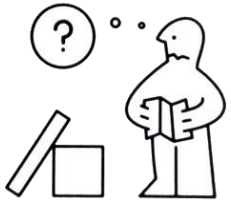


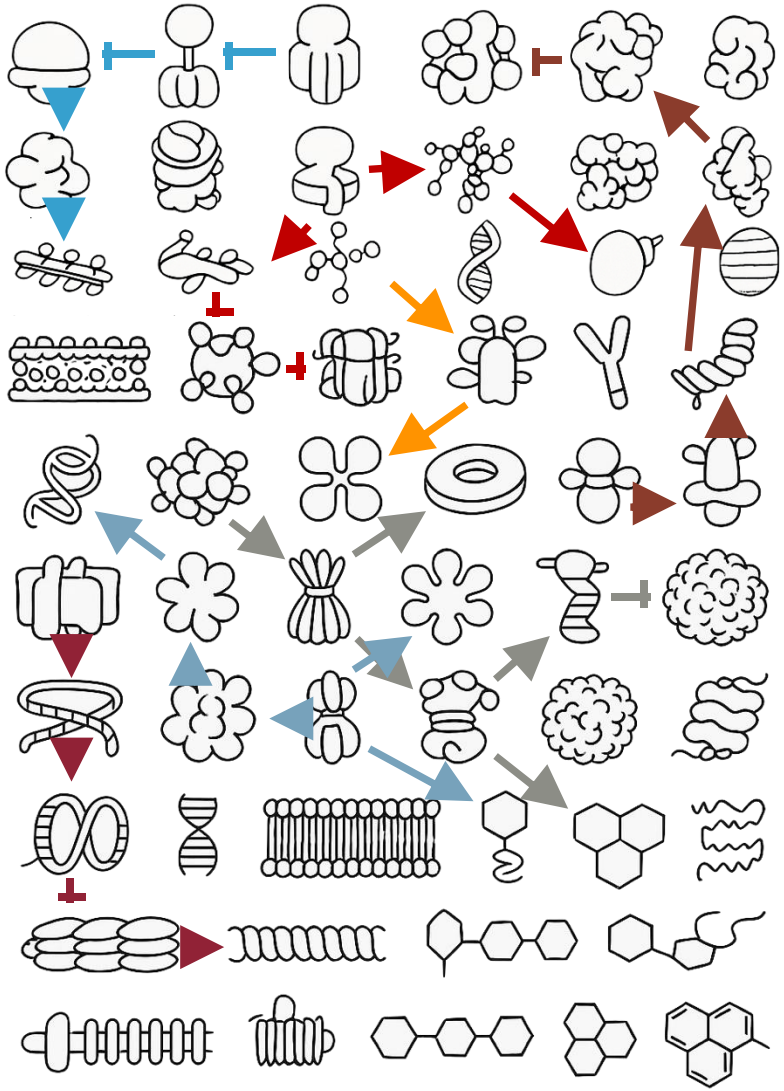


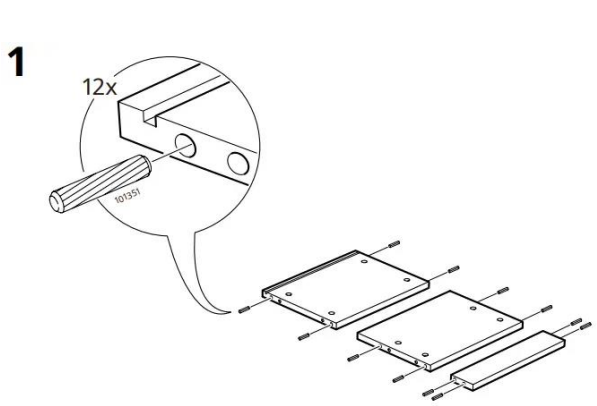


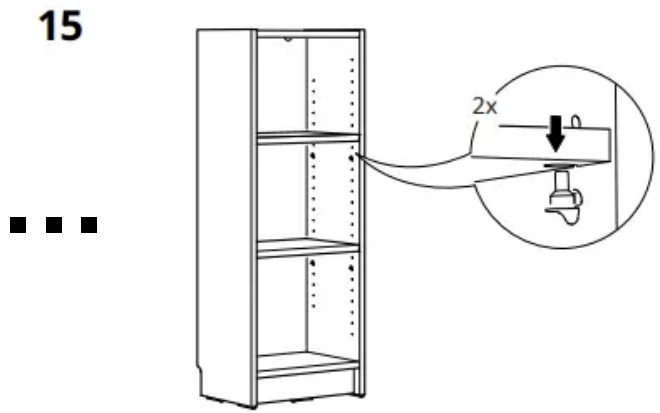
how is a cell assembled?





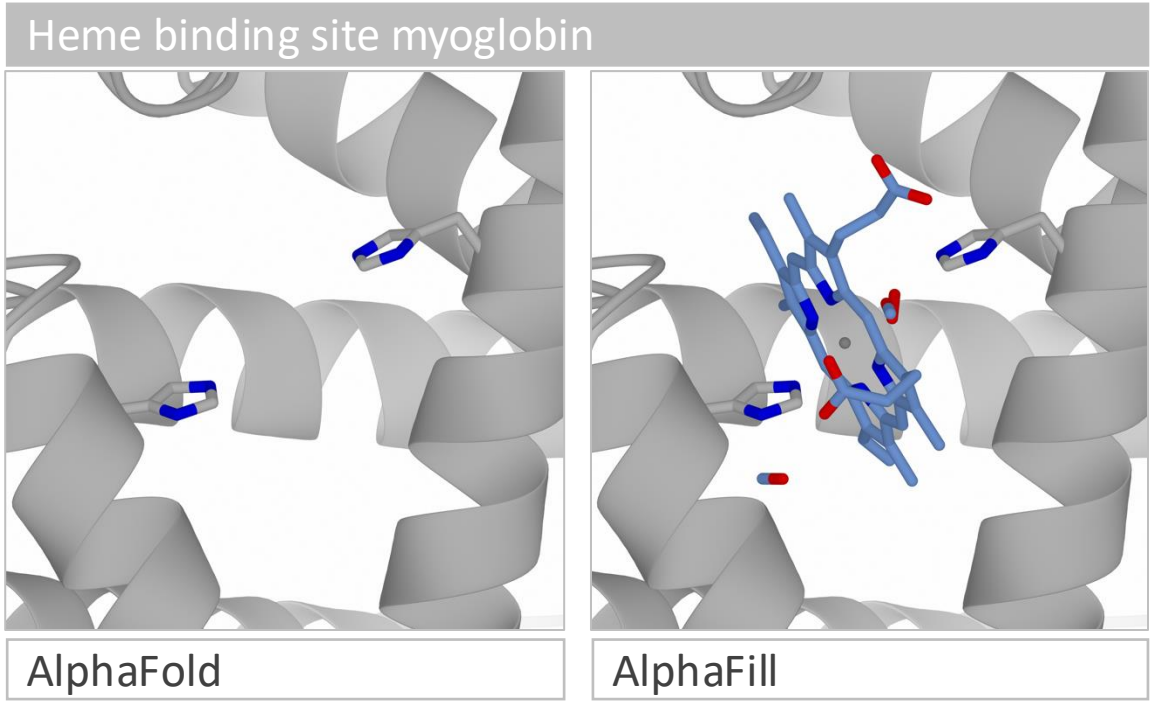






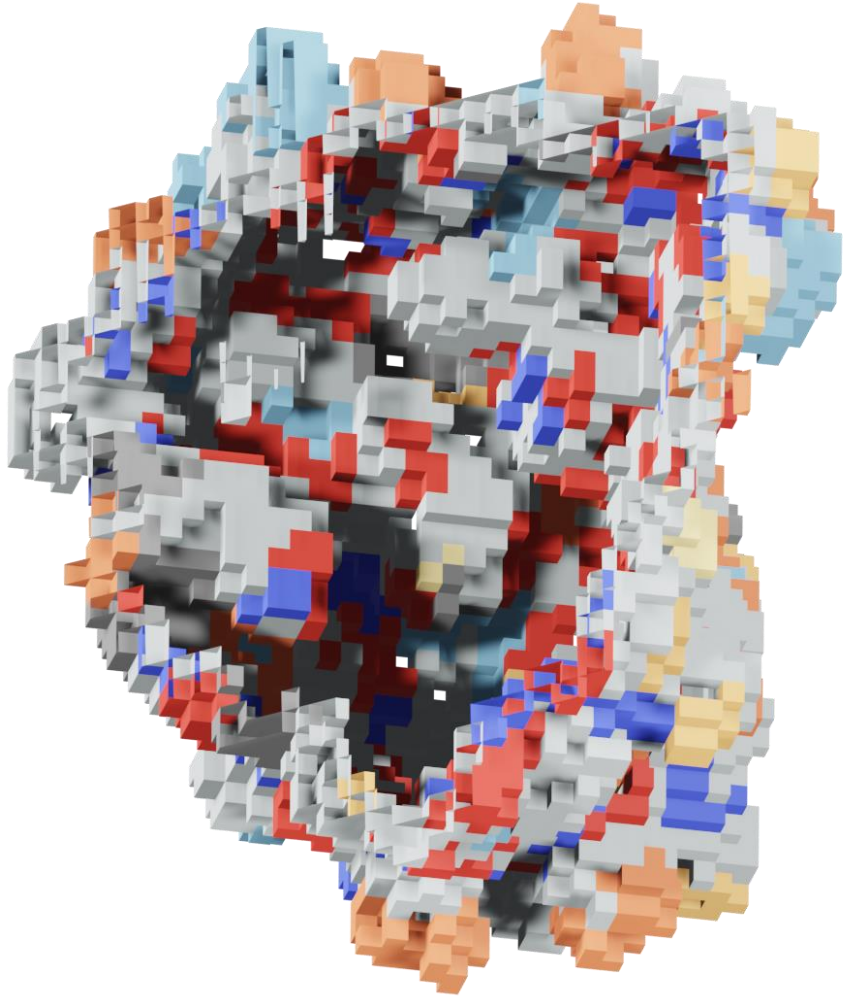
Proteins have their PARTNERS: AlPHA-FILL

- AlphaFill “transplants” missing “partners” from homologous experimental structures to predicted models
  - *co-factors (e.g heme, acetyl-CoA, etc)*
  - *ions*
  - *natural ligands.*
- The AlphaFill resource contains
  - *all model organism proteomes proteomes relevant to global health (e.g. viruses)*
  - *User-initiated AlphaFill transplants*
  - *±12 million transplanted compounds and ±1 million proteins*



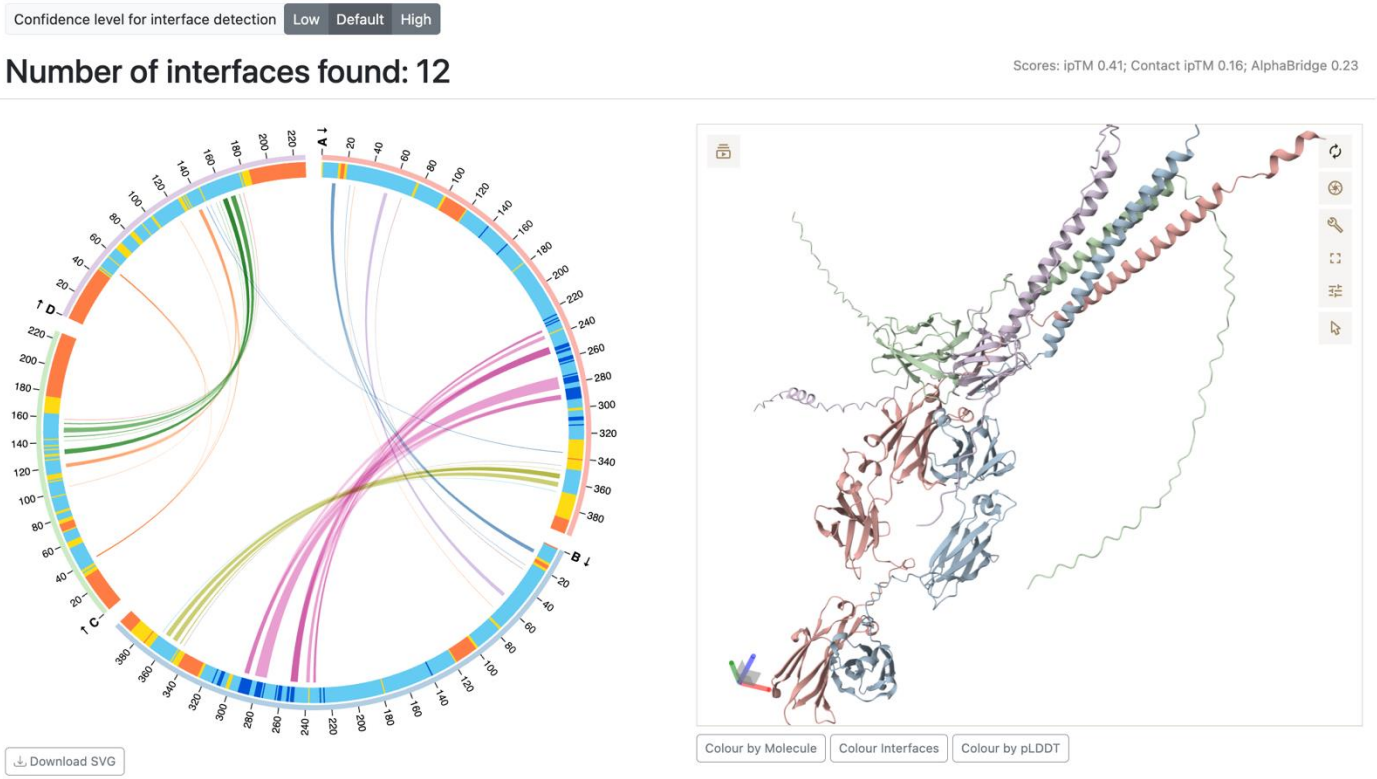
LEGO-FILL: finding new partners for proteins

- Lego-Fill goes beyonf homology
- Translates proteins to “3D images”
- Uses classic image recognition deep learning methods (eUNET++)
- Can predict ions, ligands binding sites to any structure
- Used to assign functions to metagenome movel folds



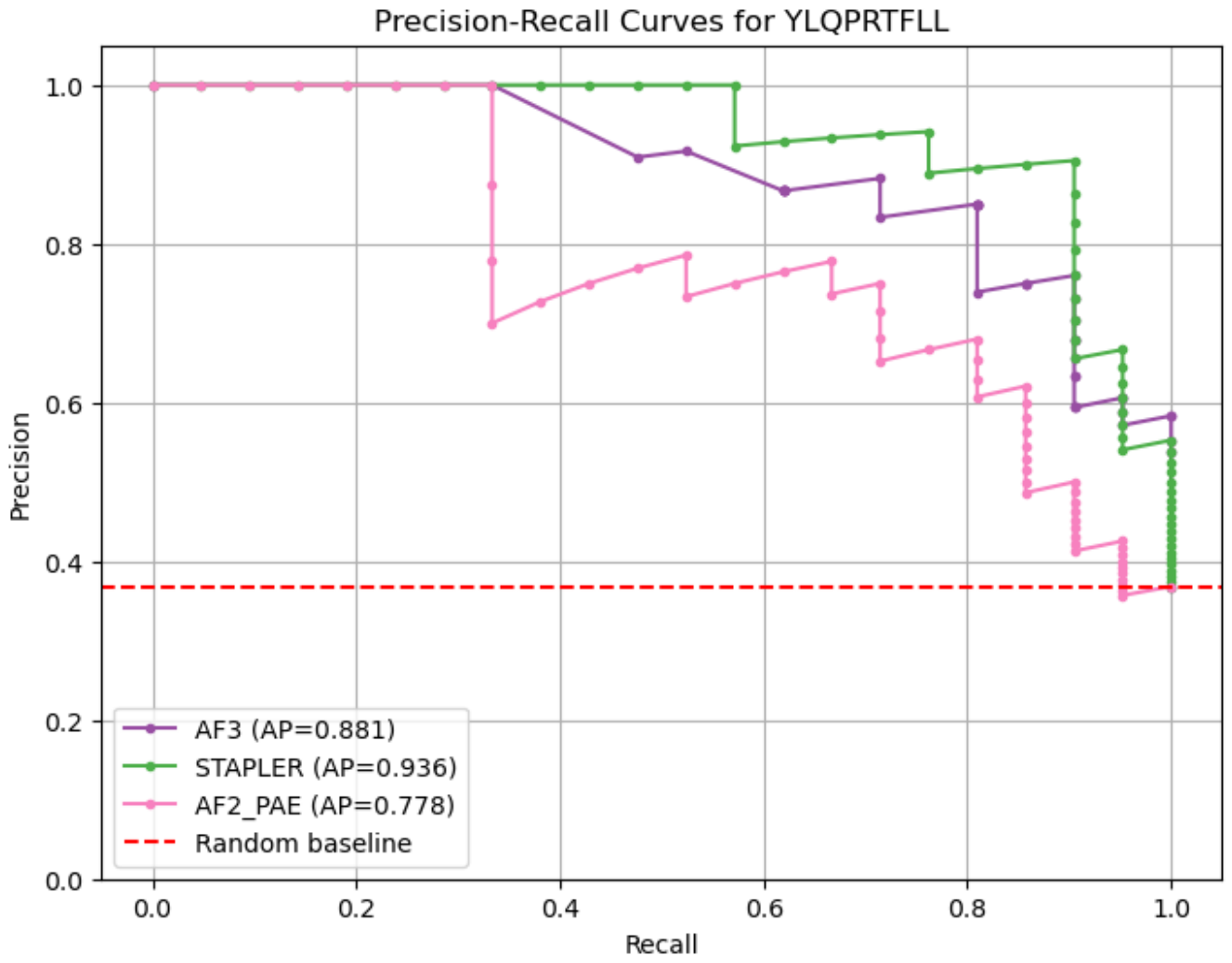
Alpha-bridge: understanding partnerships

- Structure prediction is increasingly used to examine if two proteins interact or not.
- How can we validate if the prediction of an ineraction we do not know if it occurs is correct?
- Use reliability scores and protein atructure knowledge together with clustering and image processing algorithms to post-process AF3 predictions



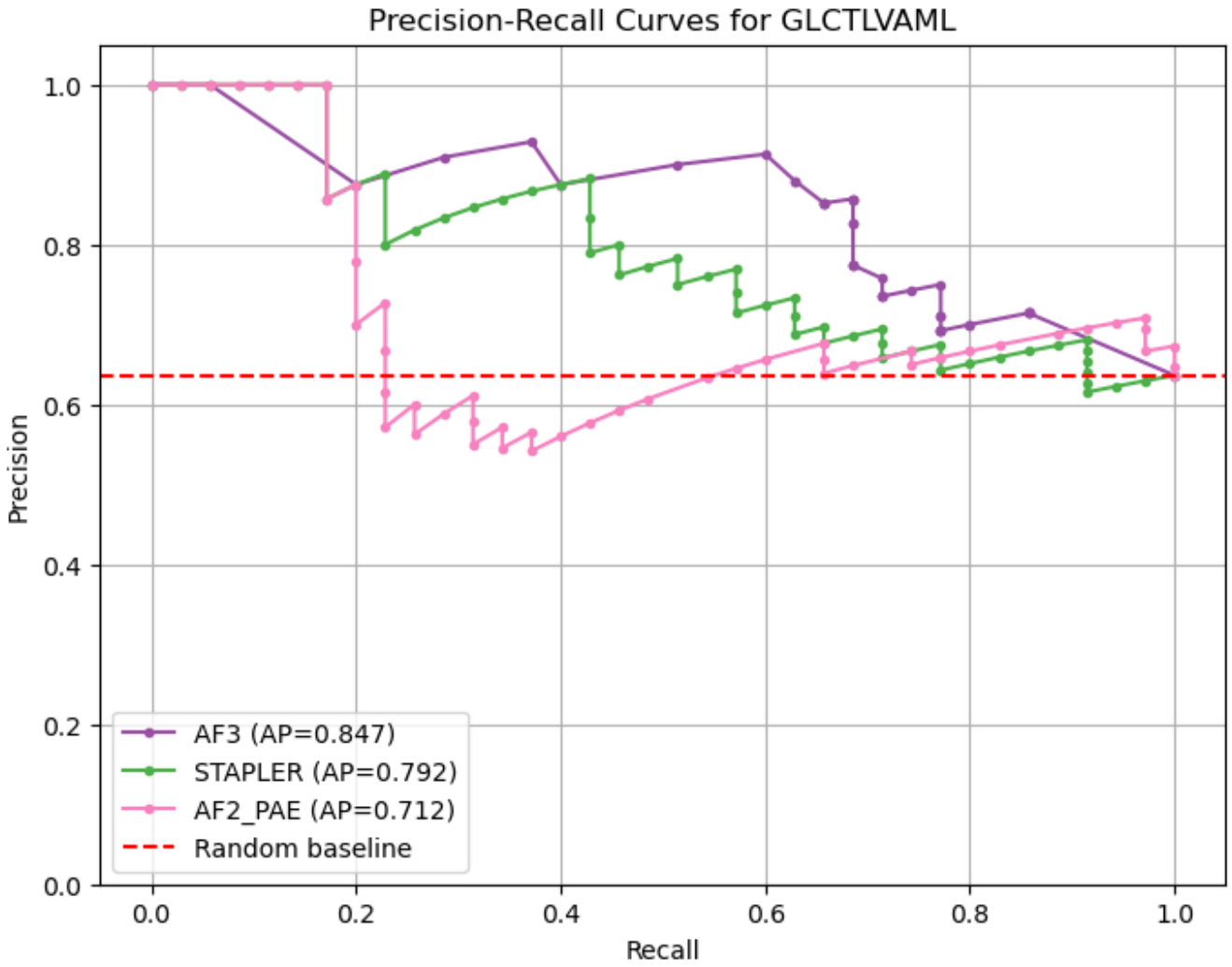
AlphaBridge and Immune response

- Can we predict which TCRs are likely to respond to specific (neo)antigens?
- Algorithm based on confidence of the prediction
- Already pretty good
- We can “learn” from bigger dataset as it accumulates



AlphaBridge and Immune response

- Can we predict which TCRs are likely to respond to specific (neo)antigens?
- Algorithm based on confidence of the prediction
- Already pretty good
  - *Also for "new" peptides*
- We can "learn" from bigger dataset as it accumulates

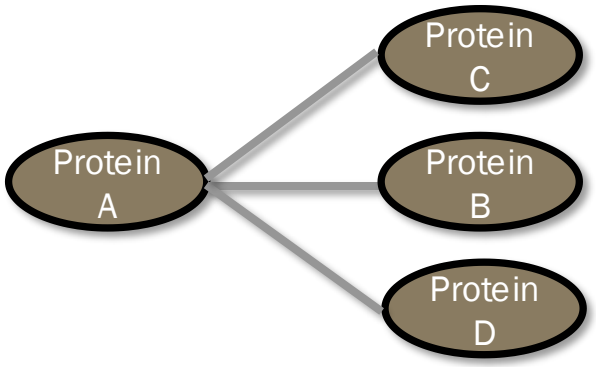


How can we put the pieces together ?

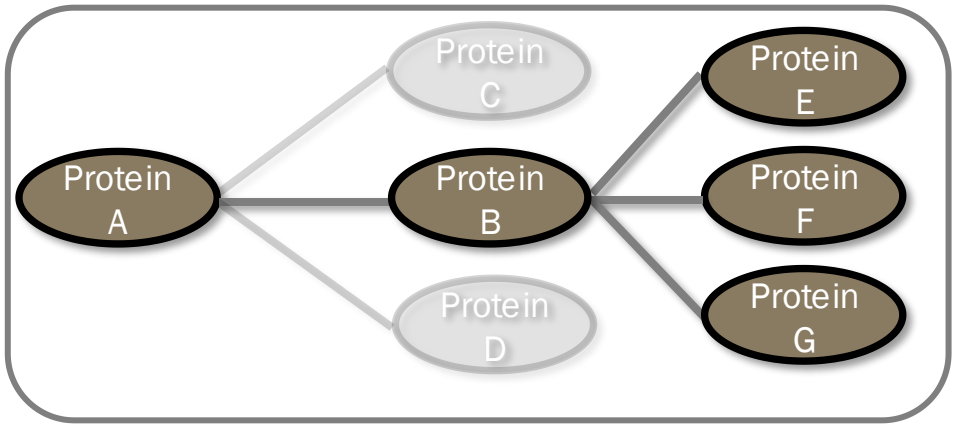
- Proteins do not act alone in the cell – they for dynamic complexes.
- Protein complexes are some of the core functional units of a cell.
- Can we learn about function studying the way proteins interact?
- Structural data: binary protein interactions across the human genome
  - *Zhang et al., Science, 2025*
- Functional data: DepMap CRISPR gene effect estimates
  - *DepMap, Broad (2025). Post-Chronos Gene effect estimates.*

Combing orthogonal data: structures first

Binary interaction



Level 2 interactions



Potential Complex

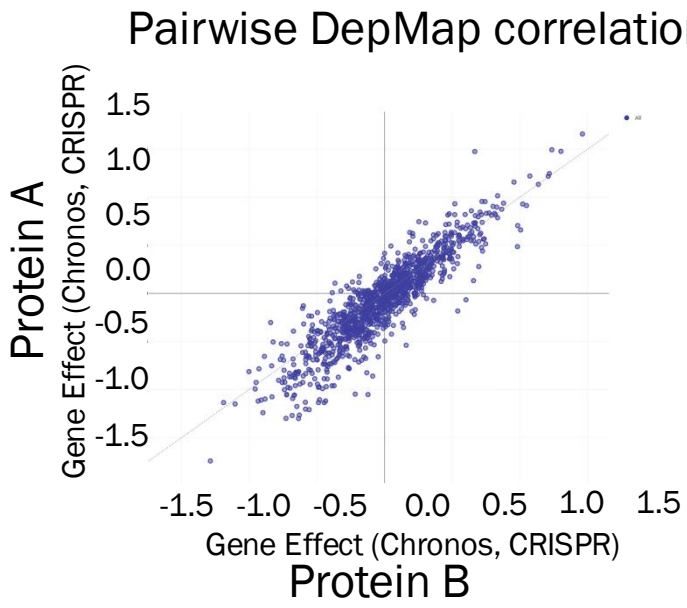
Protein A
Protein B
Protein C
Protein D
Protein E
Protein F
Protein G
Protein H

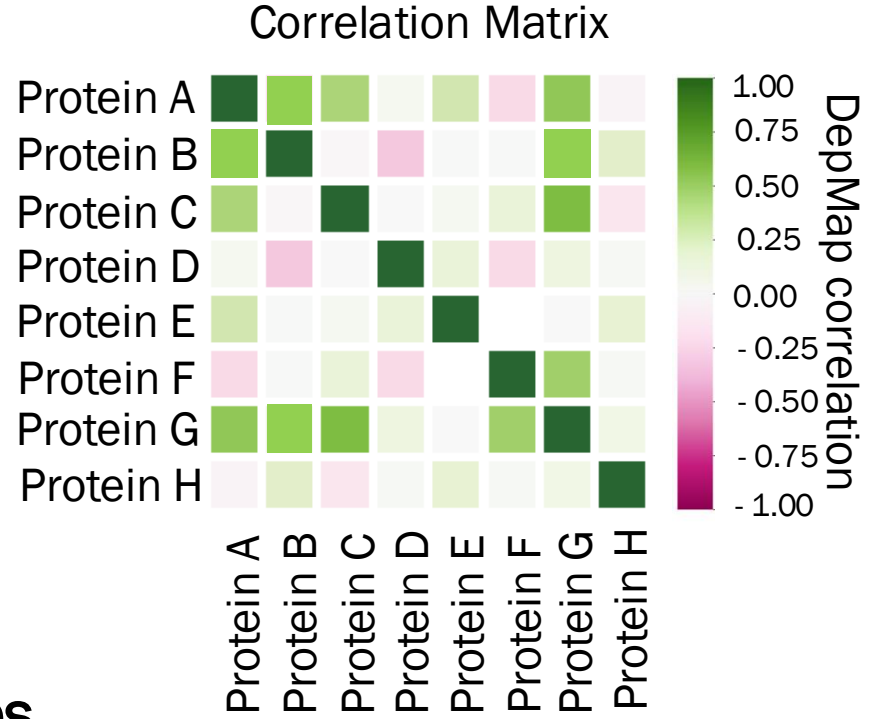
9768 protein sets

Combing orthogonal data: cell function

Potential Complex

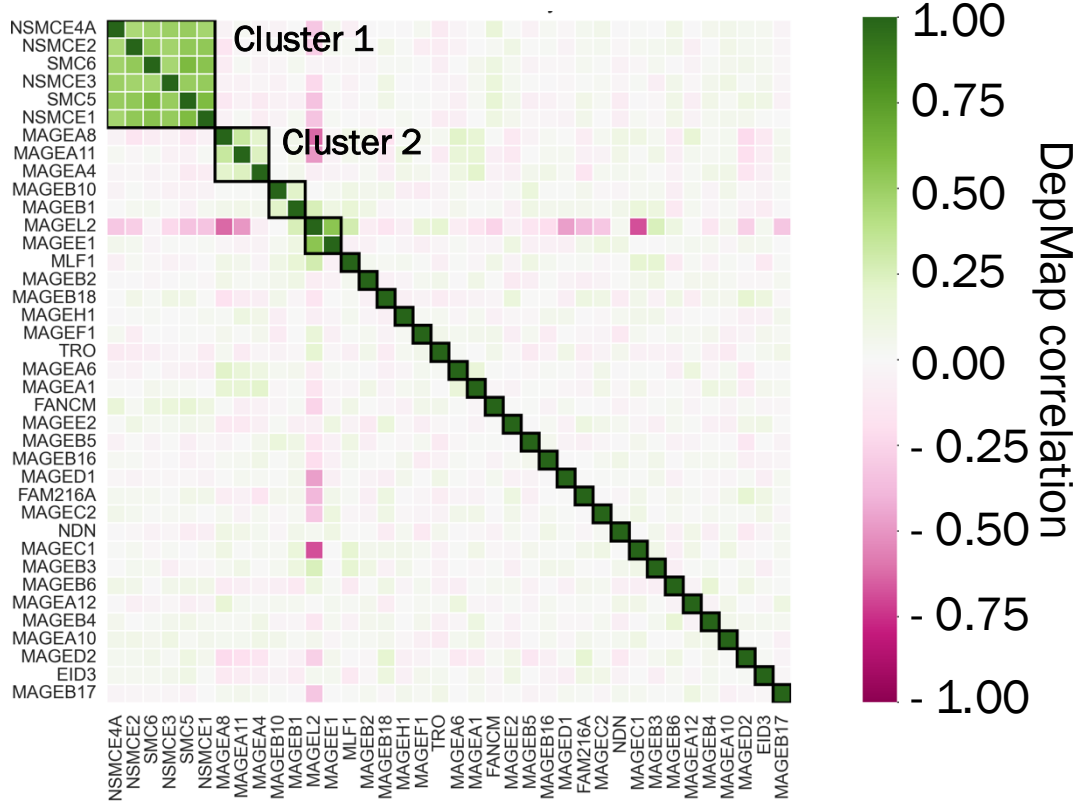
Protein A
Protein B
Protein C
Protein D
Protein E
Protein F
Protein G
Protein H

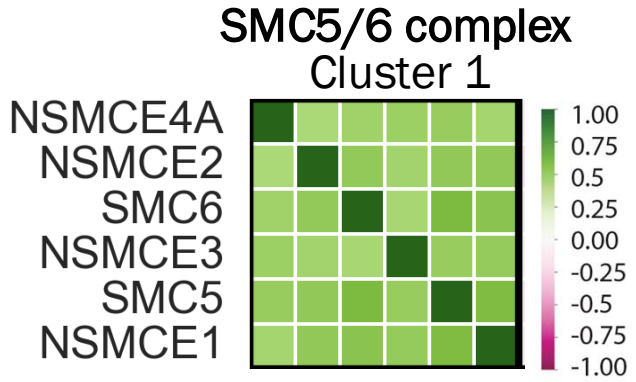


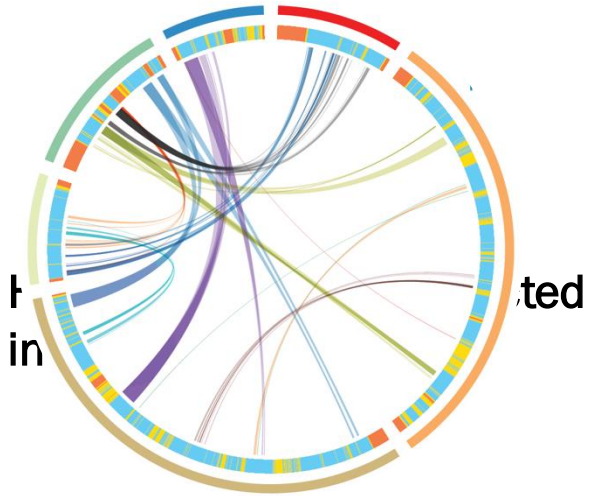


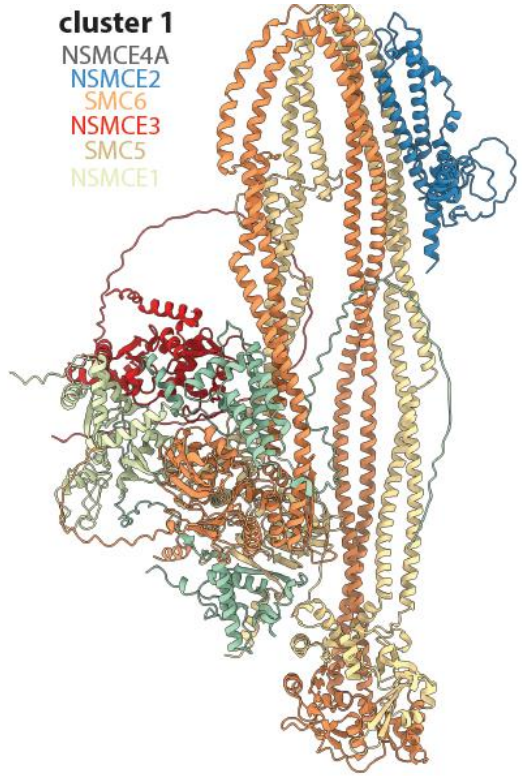
9768 correlation matrices

Clustering and validation









After clustering and filtering:
873 protein clusters for AlphaFold3 structure prediction

Adding biological sense

GO Feature Matrix
(353 x 4846)

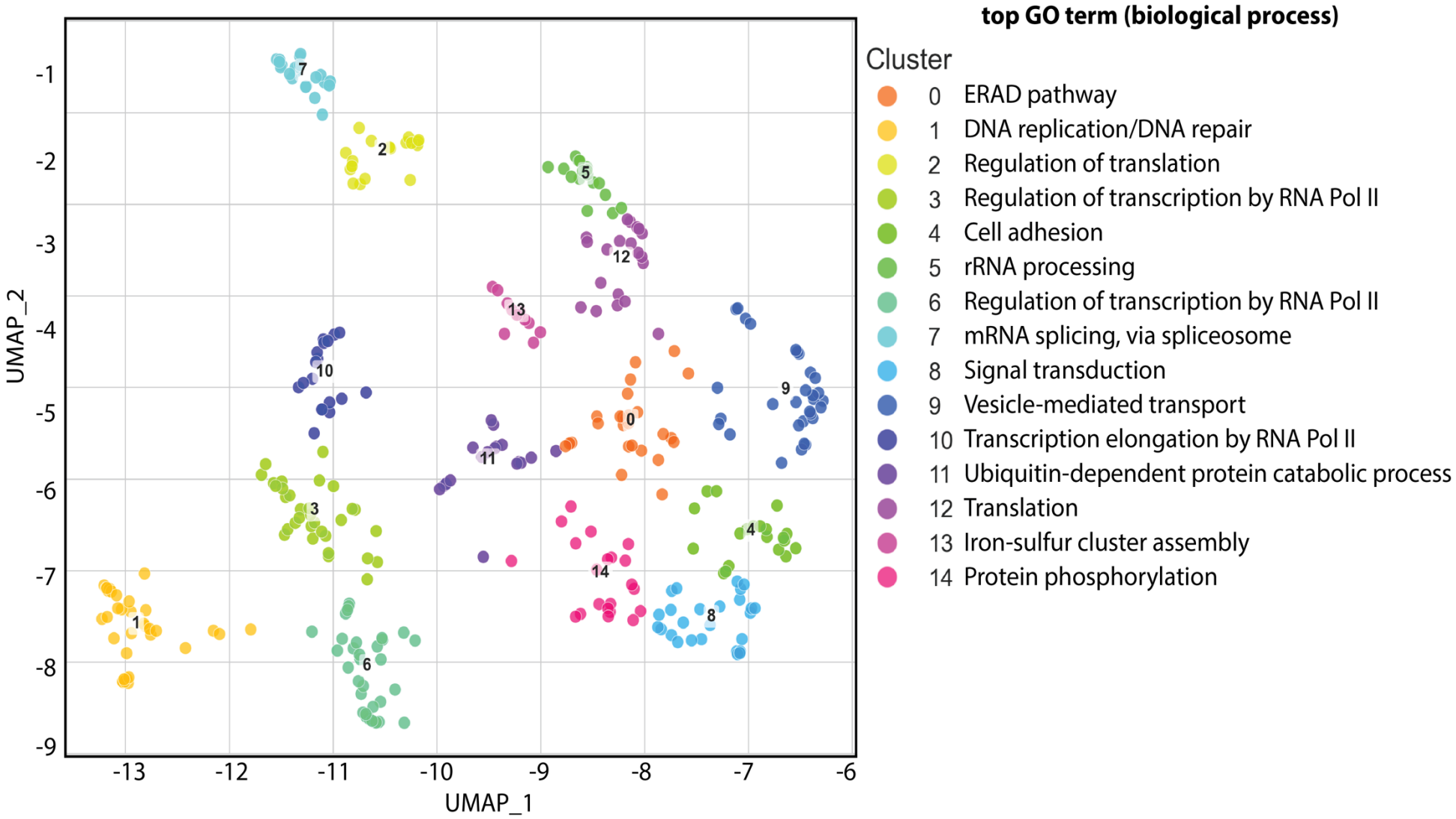
	GO-Term 1	GO-Term 2	GO-Term 3	...	GO-Term 4845	GO-Term 4846
Complex (1)	0	1	0	...	0	1
Complex (2)	0	1	1	...	1	1
⋮				⋮		
Complex (i)	0	0	0	...	1	0
⋮				⋮		
Complex (353)	1	0	0	...	0	0

String Feature Matrix
(353 x 14969)

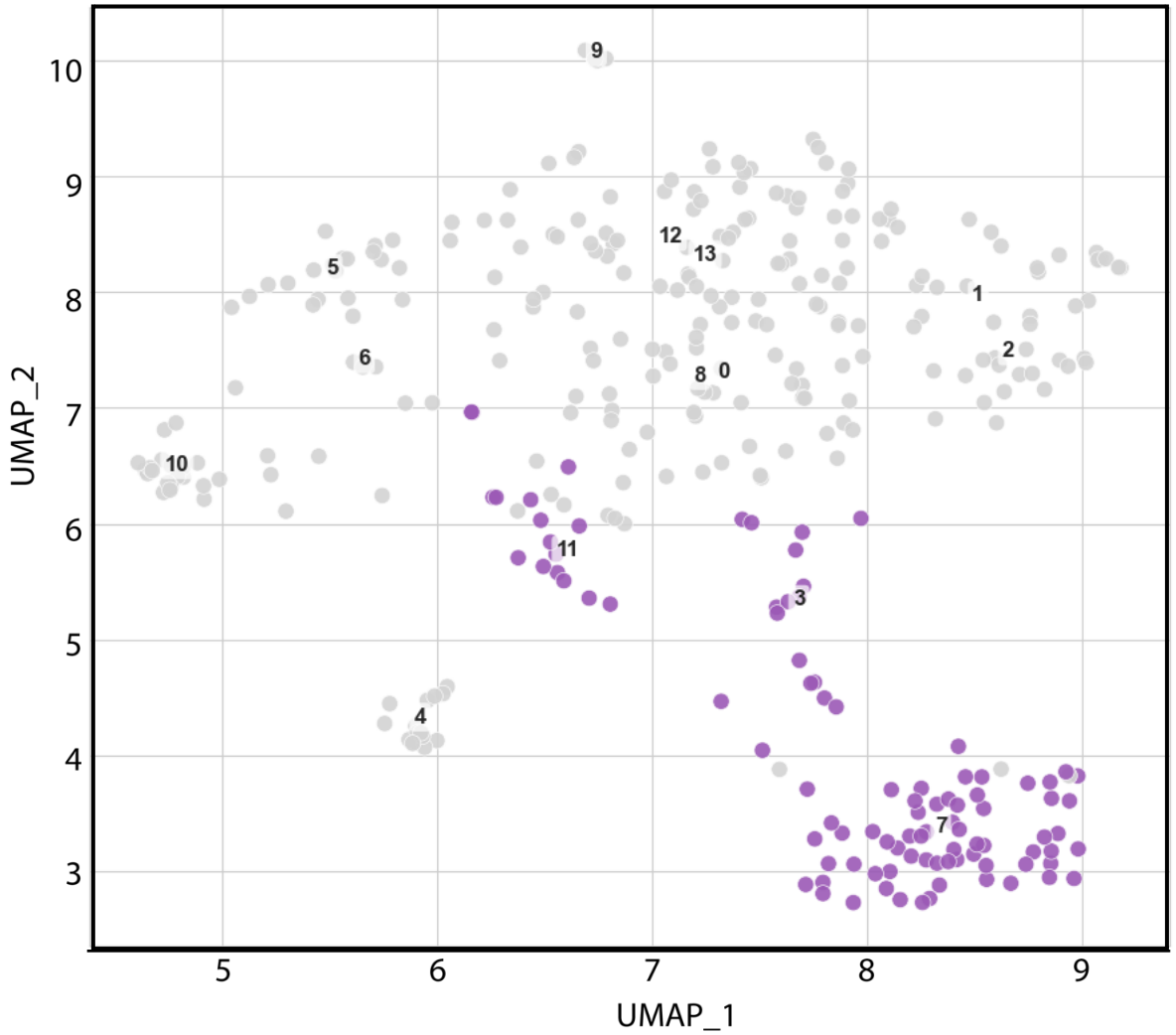
	interactor 1	interactor 2	interactor 3	...	interactor 14968	interactor 14969
Complex (1)	0.0	0.85	0.45	...	0.0	0.78
Complex (2)	0.0	0.41	0.0	...	0.0	0.0
⋮				⋮		
Complex (i)	0.98	0.0	0.0	...	0.88	0.0
⋮				⋮		
Complex (353)	0.98	0.0	0.0	...	0.88	0.0

Vector representation allows similarity comparison

Functional Clustering



Clustering based on disease association



3 Multiple cancers

7 Multiple cancers

11 Multiple cancers

Take home messages

- The three-dimensional structures of proteins are determined by their amino acid sequences.
- Proteins that share very little sequence similarity can have the same three-dimensional fold.
- The relationship between sequence and structure is not symmetrical: while one sequence corresponds to one fold, that fold is compatible with a large number of sequences.
- The detection of similarities in distantly related proteins relies on the use of amino acid substitution scores, e.g. PAM or BLOSUM.
- The number of protein structures that have been determined experimentally is far smaller than the number of proteins that have had their genes sequenced.
- Very distantly related proteins can have the same fold, but the relationship between them may not be detectable by methods that align amino acid sequences.
- Similarities between folds can be detected even for sequence similarity that cannot be detected.
- Most proteins of unknown structure are related in their folds to proteins whose structures are already known.
- Modern AI software can predict the structure of proteins even if alignments fail to relate them to known structures.