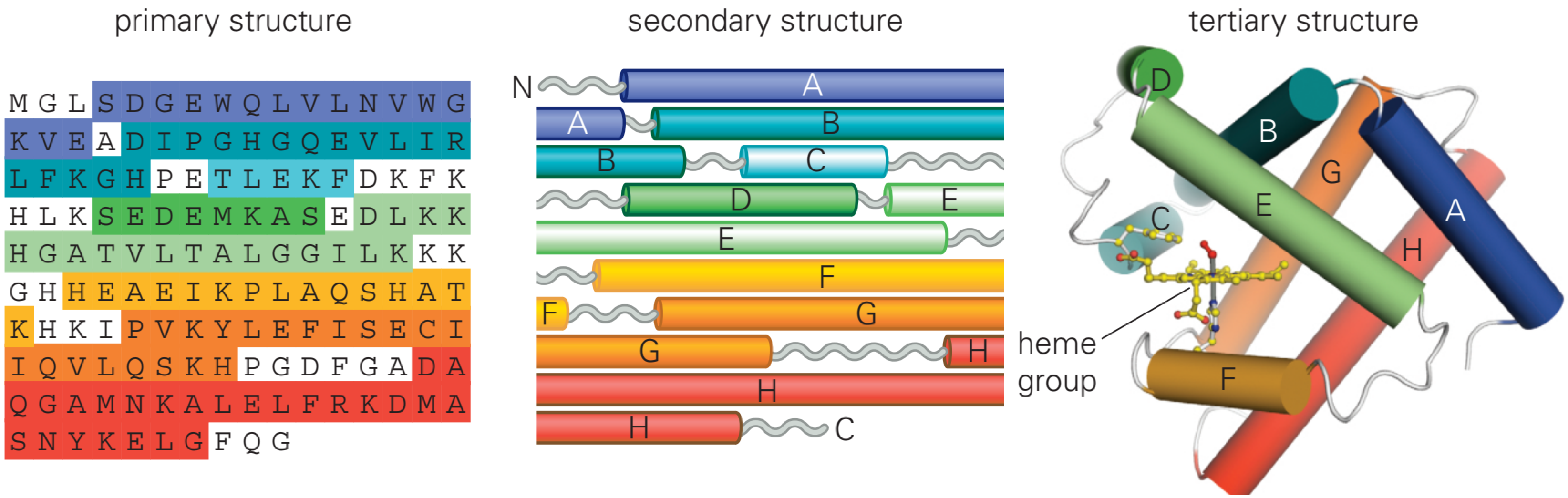
The amino-acid sequence defines the 3D-structure of proteins



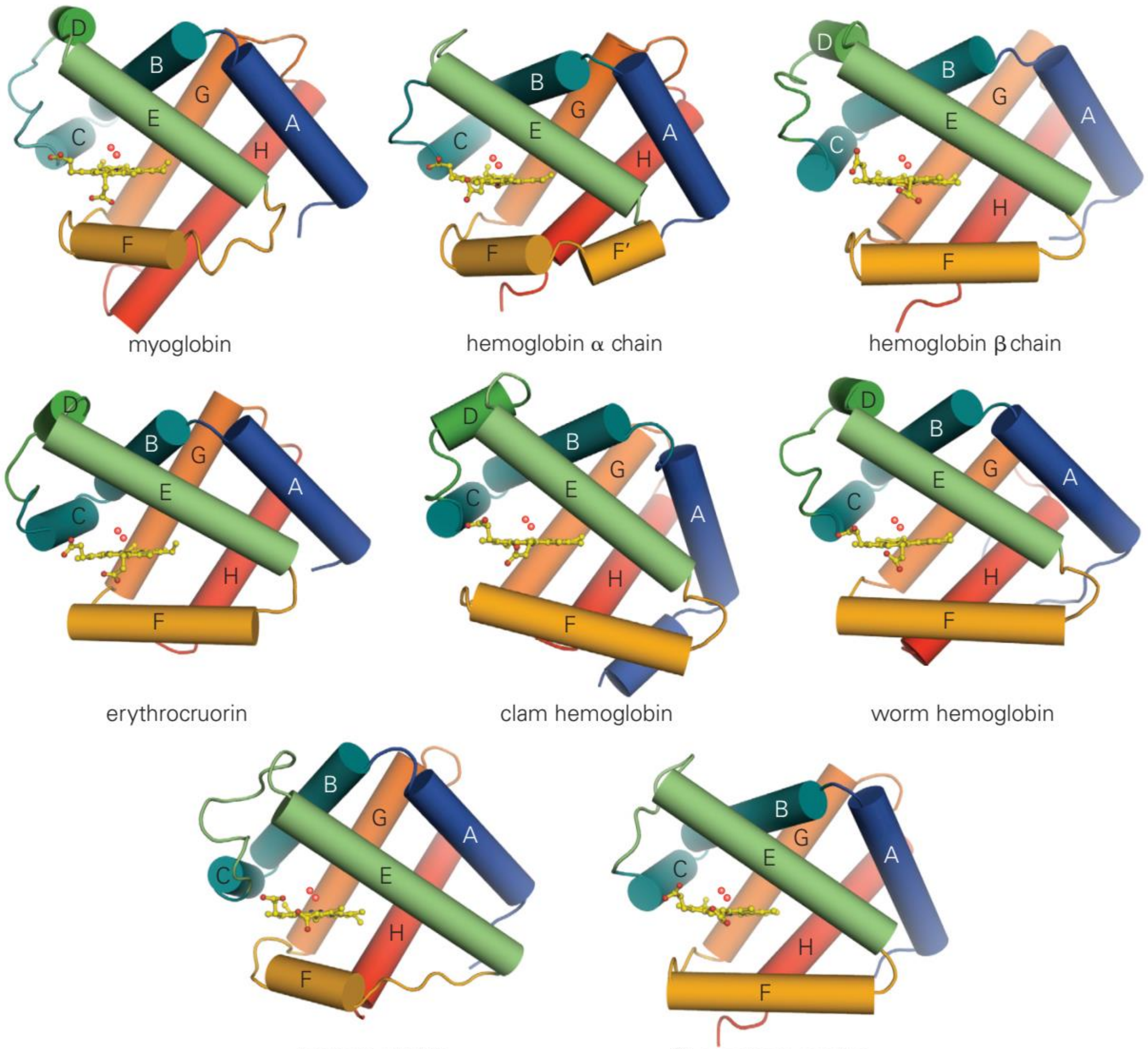
primary structure

MGLSDGEWQLVLNVWG
KVEADIPGHGQEVLIR
LFKGHPETLEKFDKFK
HLKSEDEMKASEDLKK
HGATVLTALGGILKKK
GHHEAEIKPLAQSHAT
KHKIPVKYLEFISECI
IQVLQSKHPGDFGADA
QGAMNKALELFRKDMA
SNYKELGFQG

secondary structure

tertiary structure

Similar sequences have similar folds



myoglobin

hemoglobin α chain

hemoglobin β chain

erythrocruorin

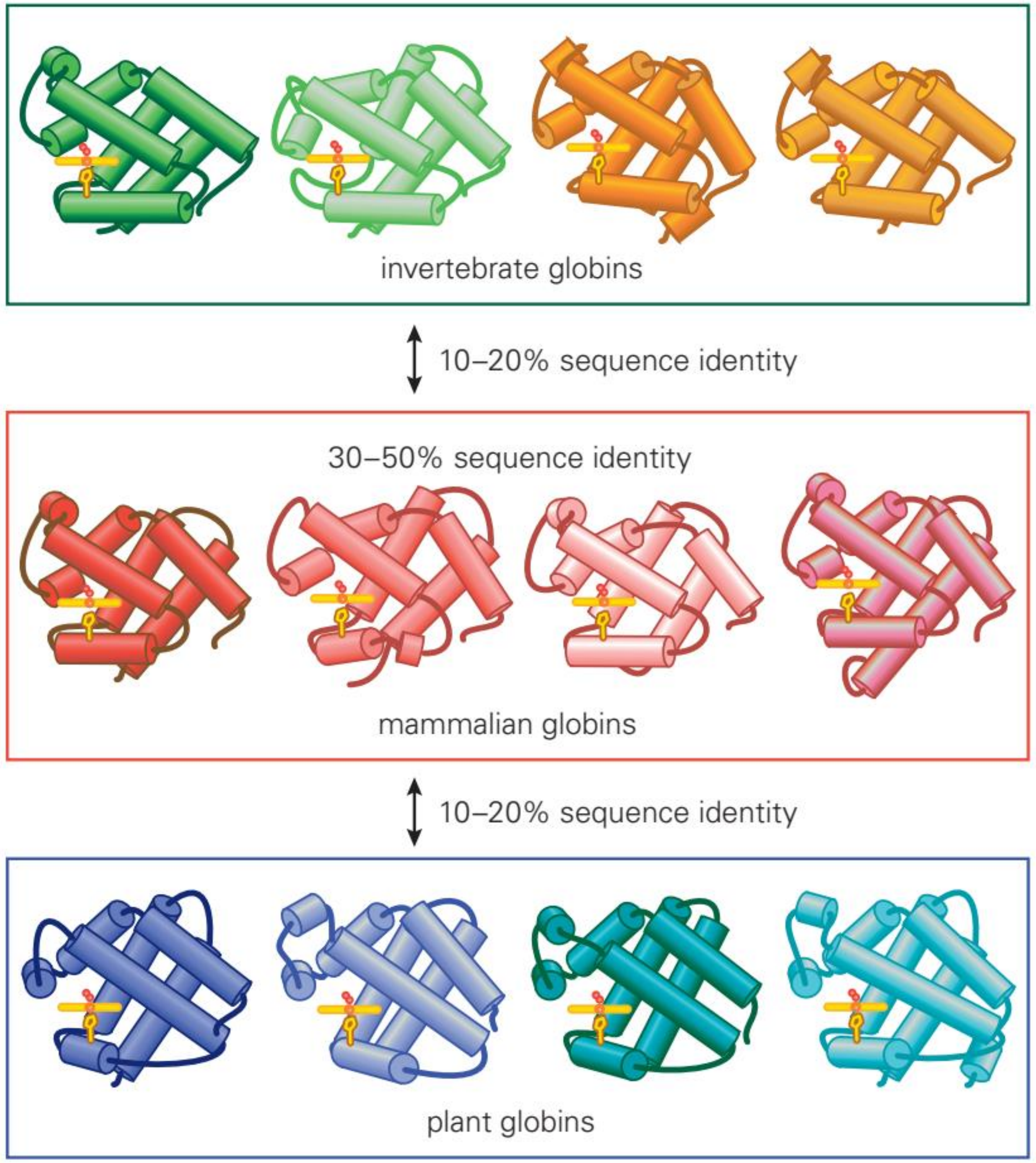
clam hemoglobin

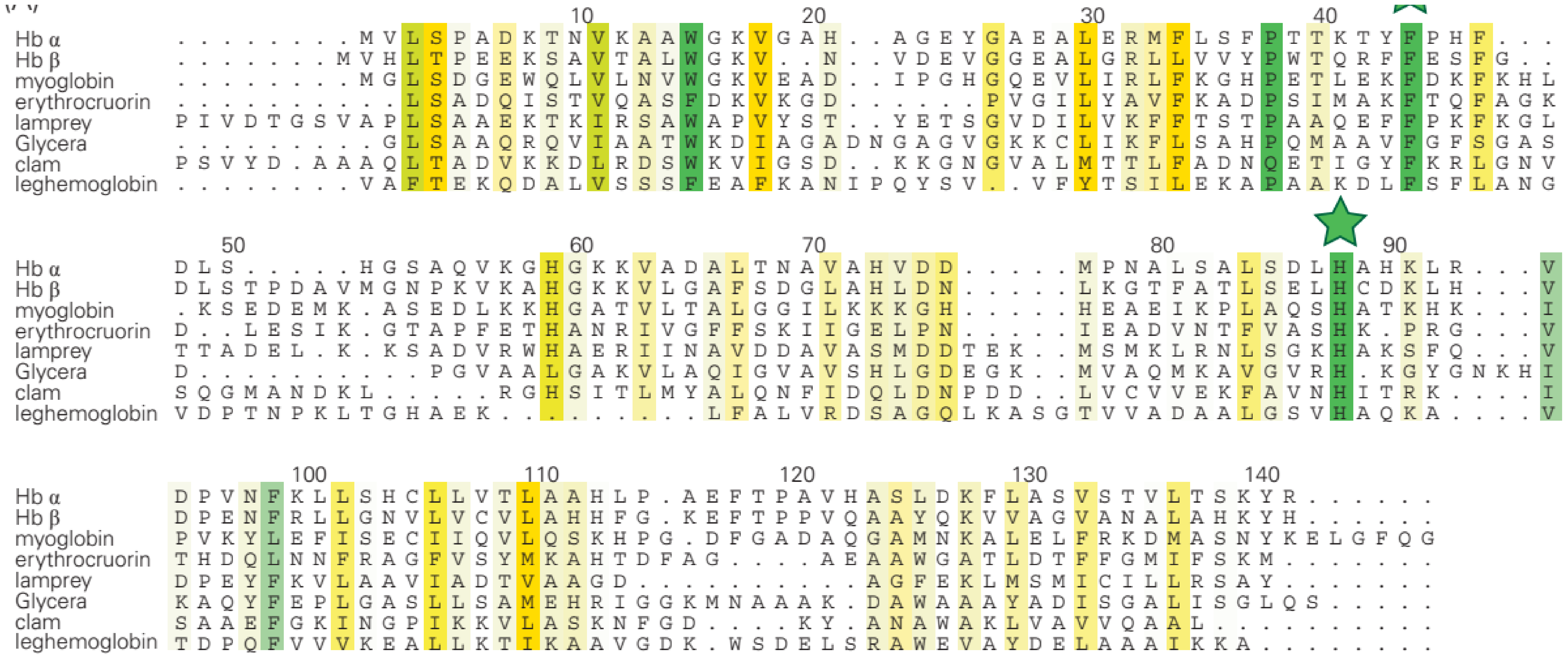
worm hemoglobin

leghemoglobin

Glycera hemoglobin

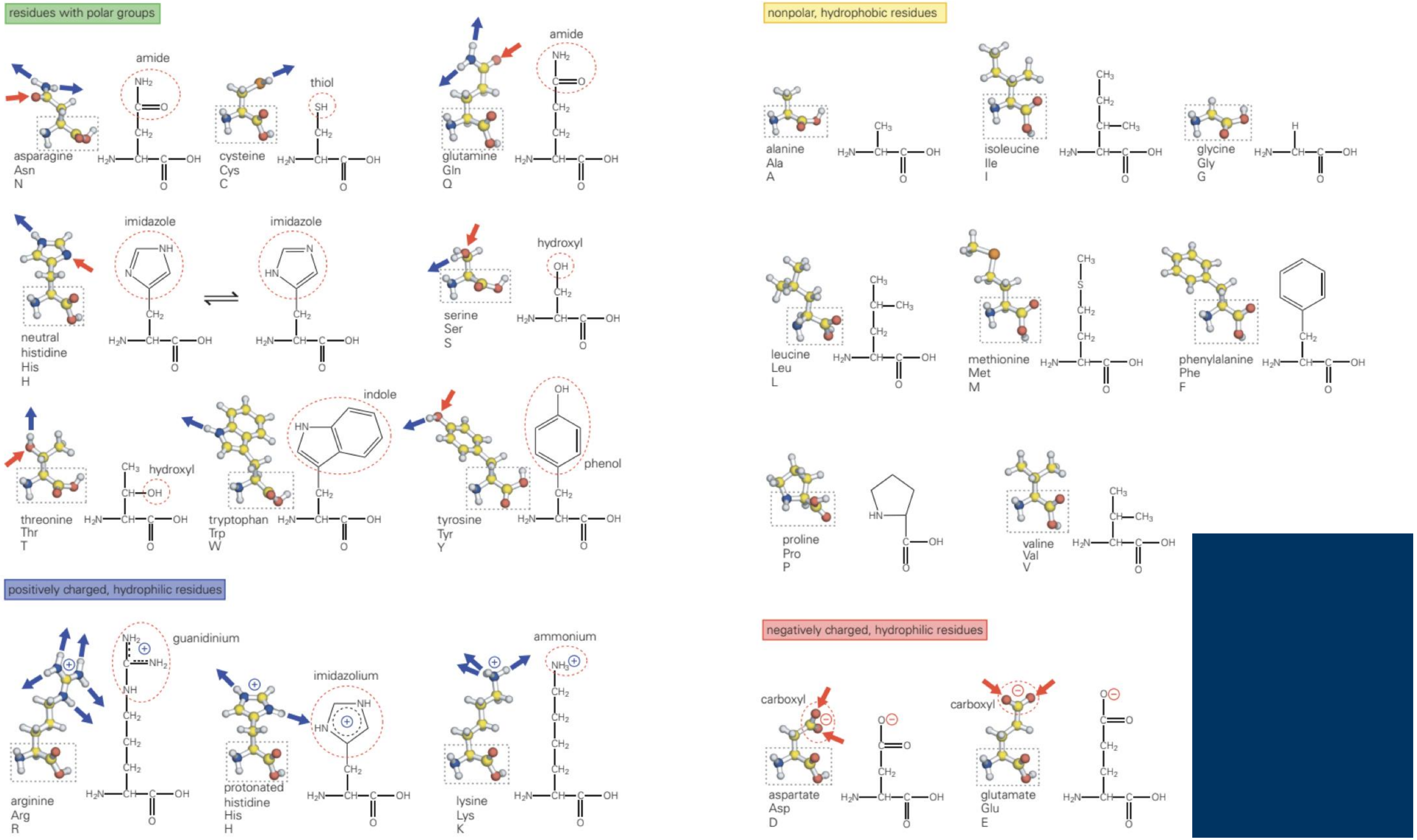
Importance of sequence similarity





```
                                                  10                  20                            30                      40
Hb α           . . . . . . . . M V L S P A D K T N V K A A W G K V G A H . . A G E Y G A E A L E R M F L S F P T T K T Y F P H F . . .
Hb β           . . . . . . . M V H L T P E E K S A V T A L W G K V . . N . . V D E V G G E A L G R L L V V Y P W T Q R F F E S F G . .
myoglobin      . . . . . . . . M G L S D G E W Q L V L N V W G K V E A D . . I P G H G Q E V L I R L F K G H P E T L E K F D K F K H L
erythrocruorin . . . . . . . . . . L S A D Q I S T V Q A S F D K V K G D . . . . . . P V G I L Y A V F K A D P S I M A K F T Q F A G K
lamprey        P I V D T G S V A P L S A A E K T K I R S A W A P V Y S T . . Y E T S G V D I L V K F F T S T P A A Q E F F P K F K G L
Glycera        . . . . . . . . . G L S A A Q R Q V I A A T W K D I A G A D N G A G V G K K C L I K F L S A H P Q M A A V F G F S G A S
clam           P S V Y D . A A A Q L T A D V K K D L R D S W K V I G S D . . K K G N G V A L M T T L F A D N Q E T I G Y F K R L G N V
leghemoglobin  . . . . . . . . V A F T E K Q D A L V S S S F E A F K A N I P Q Y S V . . V F Y T S I L E K A P A A K D L F S F L A N G

                  50                  60                  70                            80                      90
Hb α           D L S . . . . . H G S A Q V K G H G K K V A D A L T N A V A H V D D . . . . . M P N A L S A L S D L H A H K L R . . . V
Hb β           D L S T P D A V M G N P K V K A H G K K V L G A F S D G L A H L D N . . . . . L K G T F A T L S E L H C D K L H . . . V
myoglobin      . K S E D E M K . A S E D L K K H G A T V L T A L G G I L K K K G H . . . . . H E A E I K P L A Q S H A T K H K . . . I
erythrocruorin D . . L E S I K . G T A P F E T H A N R I V G F F S K I I G E L P N . . . . . I E A D V N T F V A S H K . P R G . . . V
lamprey        T T A D E L . K . K S A D V R W H A E R I I N A V D D A V A S M D D T E K . . M S M K L R N L S G K H A K S F Q . . . V
Glycera        D . . . . . . . . . . . . . P G V A A L G A K V L A Q I G V A V S H L G D E G K . . M V A Q M K A V G V R H . K G Y G N K H I
clam           S Q G M A N D K L . . . . . R G H S I T L M Y A L Q N F I D Q L D N P D D . . L V C V V E K F A V N H I T R K . . . . I
leghemoglobin  V D P T N P K L T G H A E K . . . . . . . . . L F A L V R D S A G Q L K A S G T V V A D A A L G S V H A Q K A . . . . V

                         100                 110                 120                 130                 140
Hb α           D P V N F K L L S H C L L V T L A A H L P . A E F T P A V H A S L D K F L A S V S T V L T S K Y R . . . . . .
Hb β           D P E N F R L L G N V L V C V L A H H F G . K E F T P P V Q A A Y Q K V V A G V A N A L A H K Y H . . . . . .
myoglobin      P V K Y L E F I S E C I I Q V L Q S K H P G . D F G A D A Q G A M N K A L E L F R K D M A S N Y K E L G F Q G
erythrocruorin T H D Q L N N F R A G F V S Y M K A H T D F A G . . . . A E A A W G A T L D T F F G M I F S K M . . . . . . .
lamprey        D P E Y F K V L A A V I A D T V A A G D . . . . . . . . . . A G F E K L M S M I C I L L R S A Y . . . . . . .
Glycera        K A Q Y F E P L G A S L L S A M E H R I G G K M N A A A K . D A W A A A Y A D I S G A L I S G L Q S . . . . .
clam           S A A E F G K I N G P I K K V L A S K N F G D . . . . K Y . A N A W A K L V A V V Q A A L . . . . . . . . . .
leghemoglobin  T D P Q F V V V K E A L L K T I K A A V G D K . W S D E L S R A W E V A Y D E L A A A I K K A . . . . . . . .
```

How do we “define” similarity?

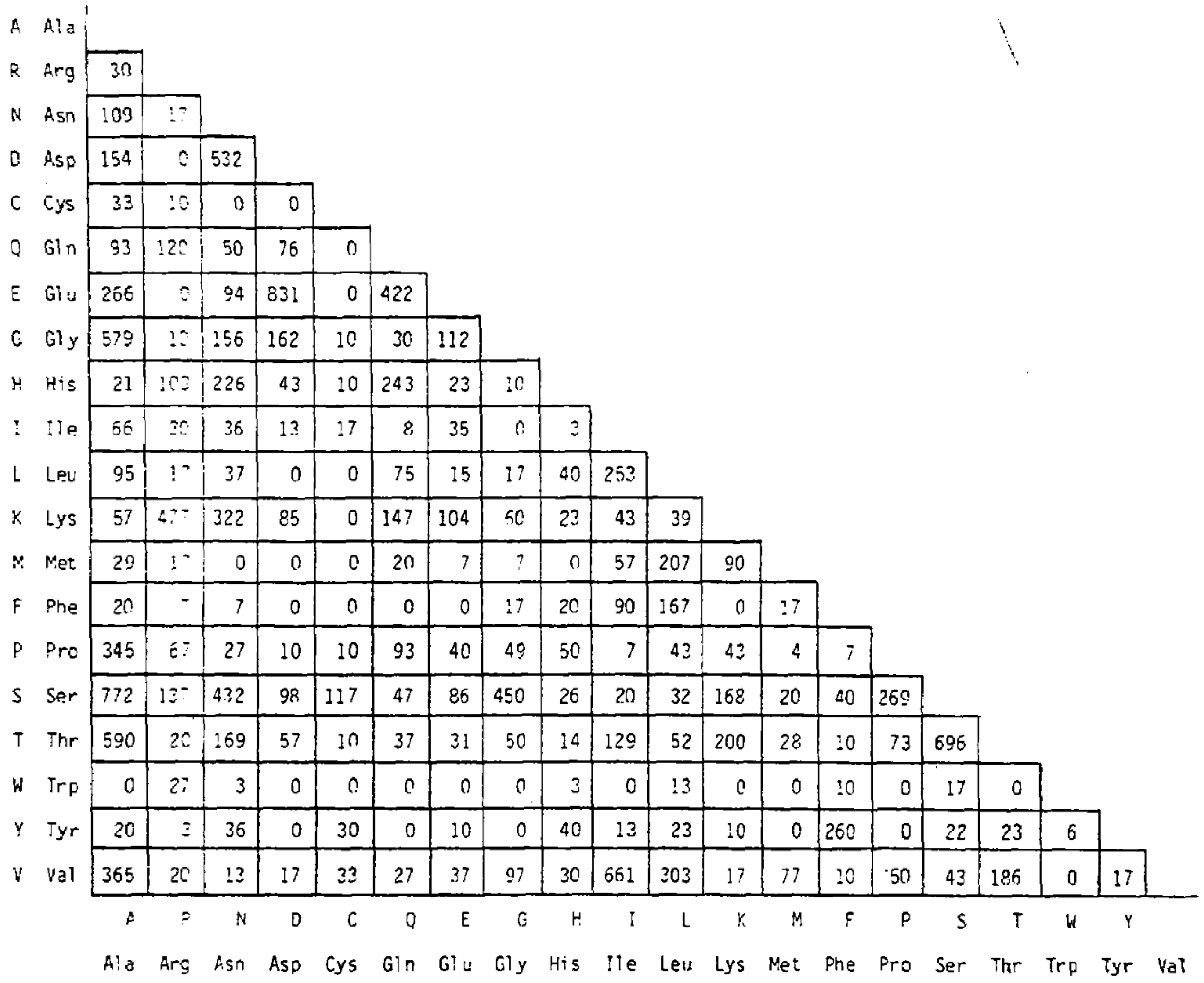


Amino acid substitution matrix

- 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

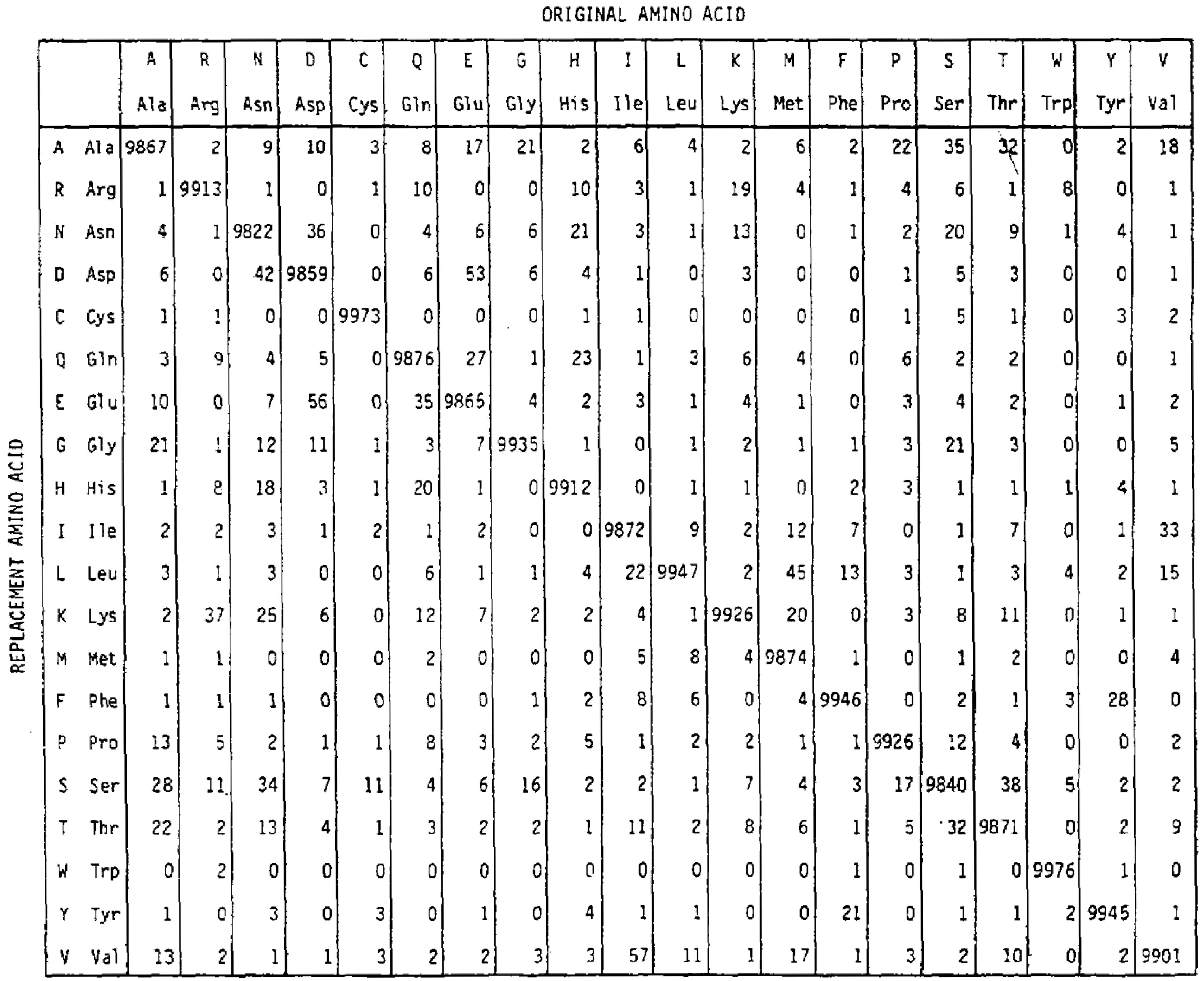


		A Ala	R Arg	N Asn	D Asp	C Cys	Q Gln	E Glu	G Gly	H His	I Ile	L Leu	K Lys	M Met	F Phe	P Pro	S Ser	T Thr	W Trp	Y Tyr	Val
A	Ala																				
R	Arg	30																			
N	Asn	109	17																		
D	Asp	154	0	532																	
C	Cys	33	10	0	0																
Q	Gln	93	120	50	76	0															
E	Glu	266	0	94	831	0	422														
G	Gly	579	10	156	162	10	30	112													
H	His	21	103	226	43	10	243	23	10												
I	Ile	66	30	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	

Figure 80. Numbers of accepted point mutations (X 10) 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 ≠ 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

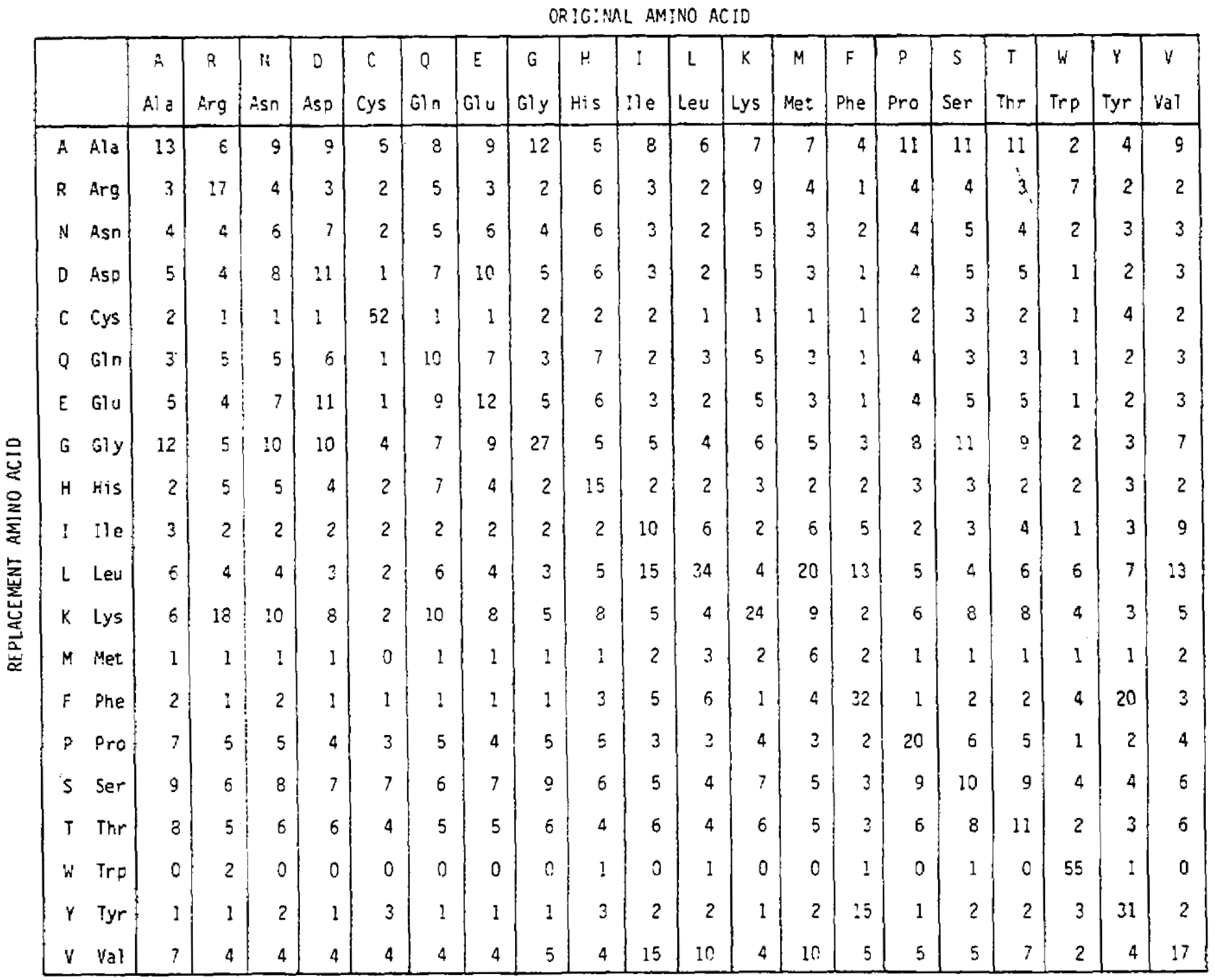
REPLACEMENT AMINO ACID

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
		Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	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	8	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

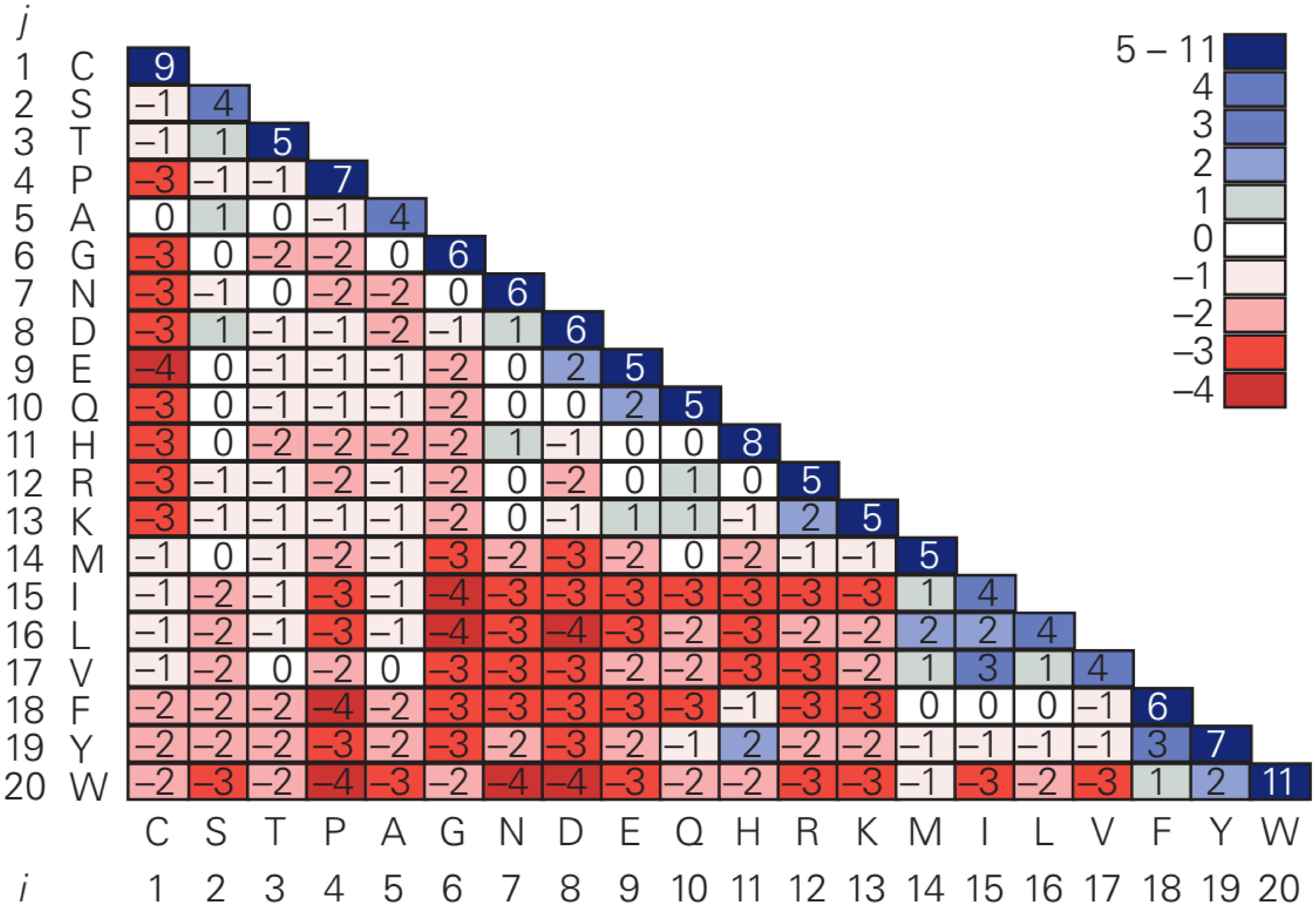
	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
A Ala	13	6	9	9	5	8	9	12	5	8	6	7	7	4	11	11	11	2	4	9
R Arg	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N Asn	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D Asp	5	4	8	11	1	7	10	5	6	3	2	5	3	1	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	3	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	5	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

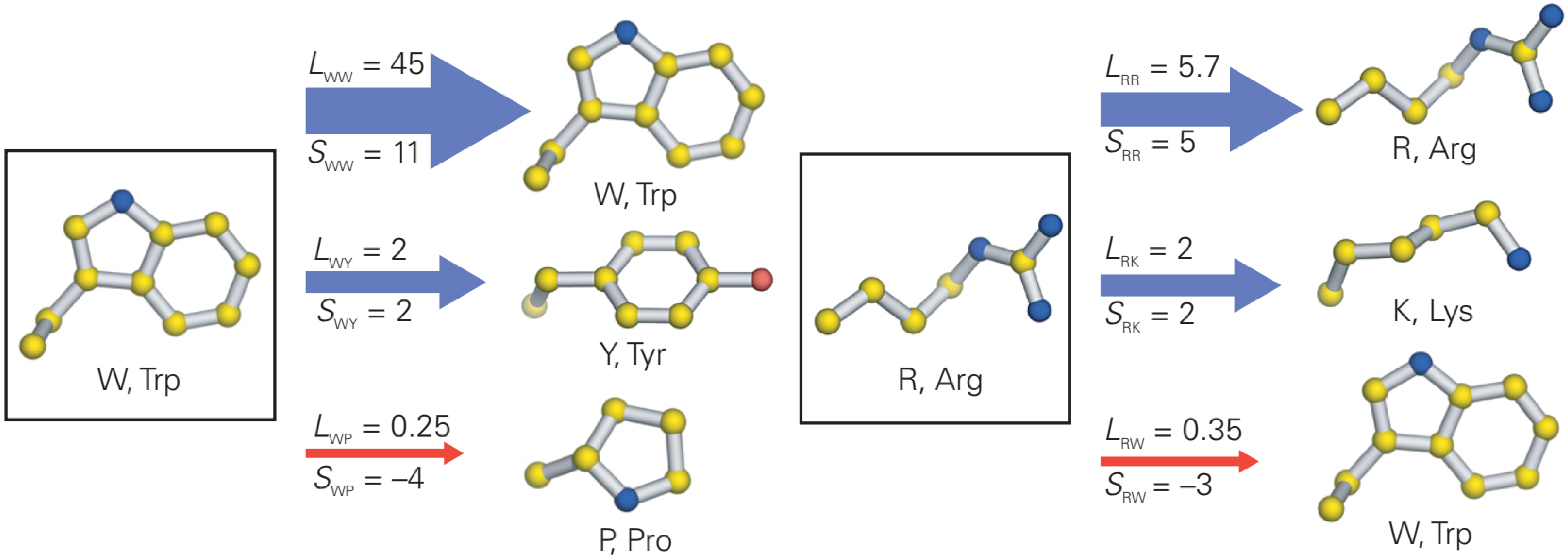
- BLOcks SUbstitution Matrix
- BLOSUM looks directly at mutations in motifs of related sequences while PAM's extrapolate evolutionary information based on closely related sequences.
- BLOSUM matrices were built based on pre-existing local alignments and optimised iteratively.
- The BLOSUM substitution 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



j																					
1	C	9																			
2	S	–1	4																		
3	T	–1	1	5																	
4	P	–3	–1	–1	7																
5	A	0	1	0	–1	4															
6	G	–3	0	–2	–2	0	6														
7	N	–3	–1	0	–2	–2	0	6													
8	D	–3	1	–1	–1	–2	–1	1	6												
9	E	–4	0	–1	–1	–1	–2	0	2	5											
10	Q	–3	0	–1	–1	–1	–2	0	0	2	5										
11	H	–3	0	–2	–2	–2	–2	1	–1	0	0	8									
12	R	–3	–1	–1	–2	–1	–2	0	–2	0	1	0	5								
13	K	–3	–1	–1	–1	–1	–2	0	–1	1	1	–1	2	5							
14	M	–1	0	–1	–2	–1	–3	–2	–3	–2	0	–2	–1	–1	5						
15	I	–1	–2	–1	–3	–1	–4	–3	–3	–3	–3	–3	–3	–3	1	4					
16	L	–1	–2	–1	–3	–1	–4	–3	–4	–3	–2	–3	–2	–2	2	2	4				
17	V	–1	–2	0	–2	0	–3	–3	–3	–2	–2	–3	–3	–2	1	3	1	4			
18	F	–2	–2	–2	–4	–2	–3	–3	–3	–3	–3	–1	–3	–3	0	0	0	–1	6		
19	Y	–2	–2	–2	–3	–2	–3	–2	–3	–2	–1	2	–2	–2	–1	–1	–1	–1	3	7	
20	W	–2	–3	–2	–4	–3	–2	–4	–4	–3	–2	–2	–3	–3	–1	–3	–2	–3	1	2	11
		C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
i		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20

Substitution scores reflect the chemical properties of the amino acids



W, Trp

$L_{WW} = 45$

$S_{WW} = 11$

W, Trp

$L_{WY} = 2$

$S_{WY} = 2$

Y, Tyr

$L_{WP} = 0.25$

$S_{WP} = -4$

P, Pro

R, Arg

$L_{RR} = 5.7$

$S_{RR} = 5$

R, Arg

$L_{RK} = 2$

$S_{RK} = 2$

K, Lys

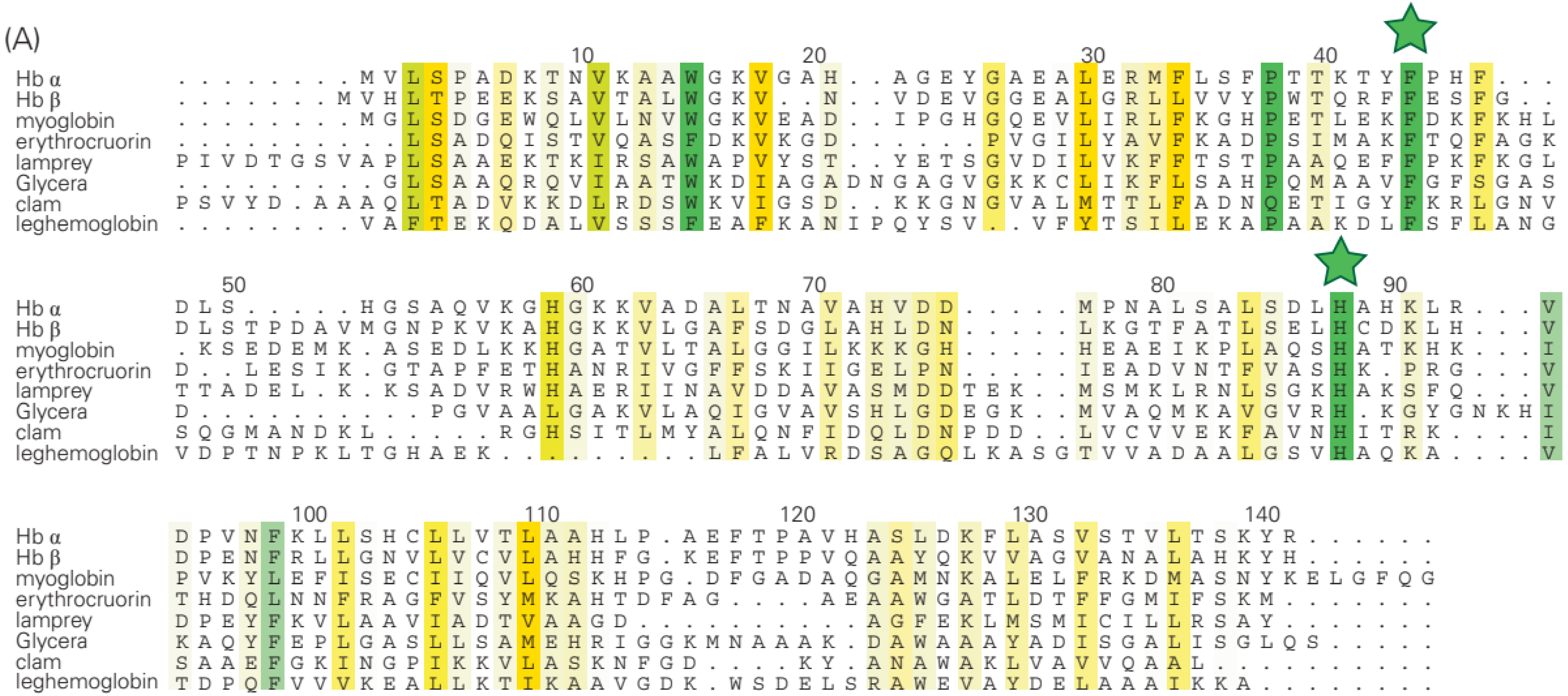
$L_{RW} = 0.35$

$S_{RW} = -3$

W, Trp

Multiple Sequence Alignments

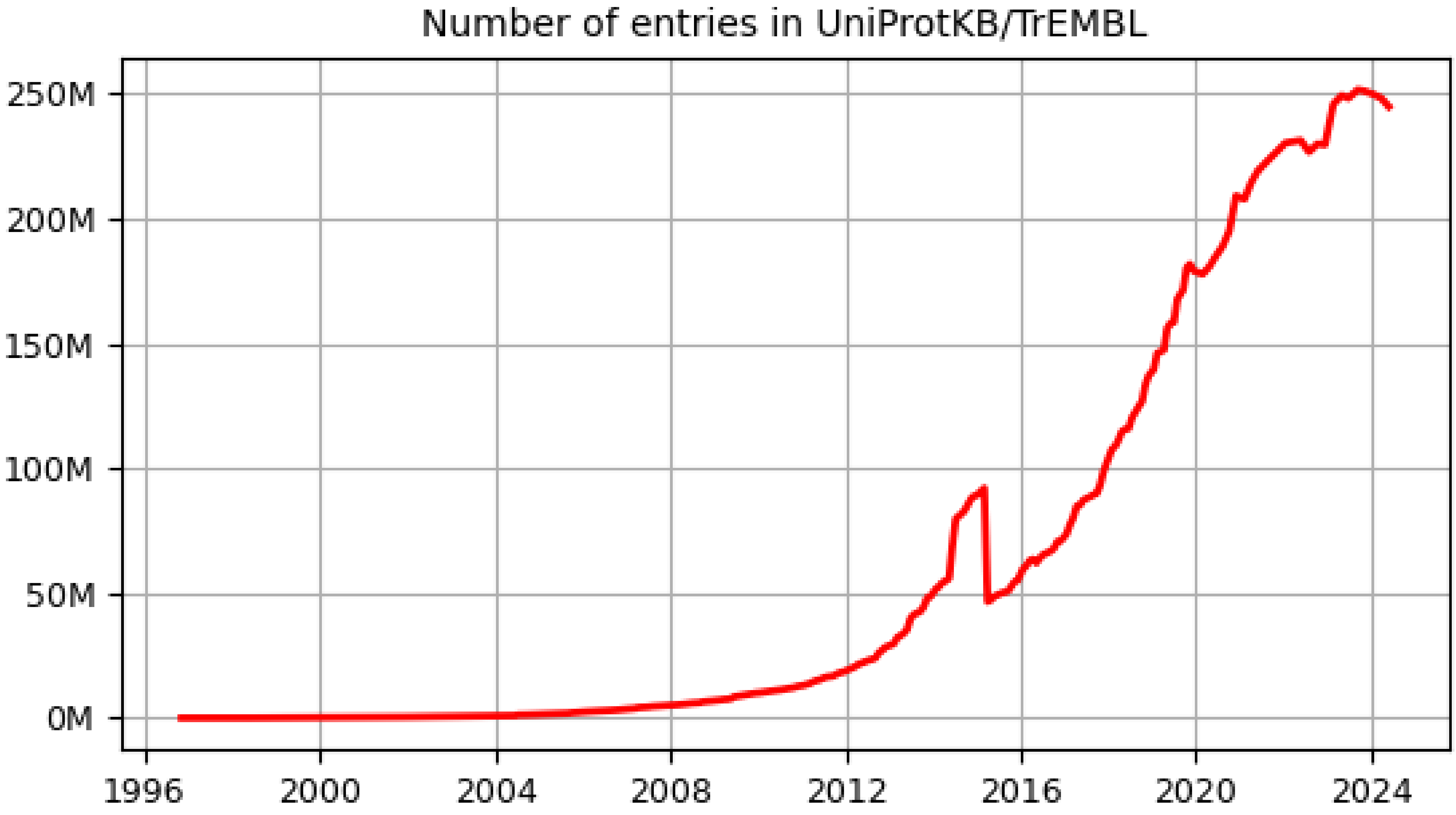
- One of the most important tools to understand structure function and structure
- Important to align many (all) avilable sequences to understand (
- (



(A)

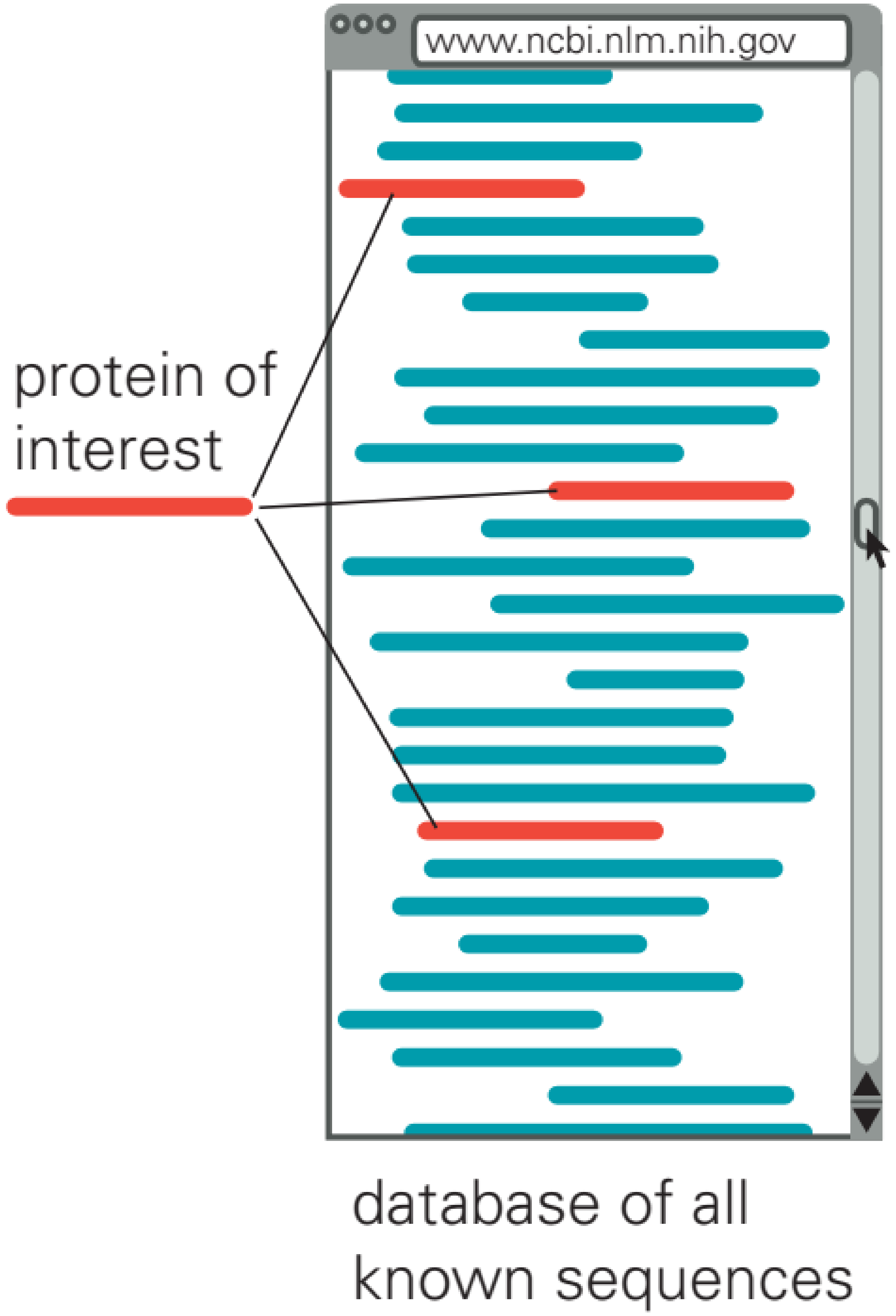
```
                                    10                  20                        30                  40
Hb α            . . . . . . . . M V L S P A D K T N V K A A W G K V G A H . . A G E Y G A E A L E R M F L S F P T T K T Y F P H F . . .
Hb β            . . . . . . . M V H L T P E E K S A V T A L W G K V . . N . . V D E V G G E A L G R L L V V Y P W T Q R F F E S F G . .
myoglobin       . . . . . . . . M G L S D G E W Q L V L N V W G K V E A D . . I P G H G Q E V L I R L F K G H P E T L E K F D K F K H L
erythrocruorin  . . . . . . . . . . L S A D Q I S T V Q A S F D K V K G D . . . . . . P V G I L Y A V F K A D P S I M A K F T Q F A G K
lamprey         P I V D T G S V A P L S A A E K T K I R S A W A P V Y S T . . Y E T S G V D I L V K F F T S T P A A Q E F F P K F K G L
Glycera         . . . . . . . . . G L S A A Q R Q V I A A T W K D I A G A D N G A G V G K K C L I K F L S A H P Q M A A V F G F S G A S
clam            P S V Y D . A A A Q L T A D V K K D L R D S W K V I G S D . . K K G N G V A L M T T L F A D N Q E T I G Y F K R L G N V
leghemoglobin   . . . . . . . . V A F T E K Q D A L V S S S F E A F K A N I P Q Y S V . . V F Y T S I L E K A P A A K D L F S F L A N G

                     50                  60                  70                            80                  90
Hb α            D L S . . . . . H G S A Q V K G H G K K V A D A L T N A V A H V D D . . . . . M P N A L S A L S D L H A H K L R . . . V
Hb β            D L S T P D A V M G N P K V K A H G K K V L G A F S D G L A H L D N . . . . . L K G T F A T L S E L H C D K L H . . . V
myoglobin       . K S E D E M K . A S E D L K K H G A T V L T A L G G I L K K K G H . . . . . H E A E I K P L A Q S H A T K H K . . . I
erythrocruorin  D . . L E S I K . G T A P F E T H A N R I V G F F S K I I G E L P N . . . . . I E A D V N T F V A S H K . P R G . . . V
lamprey         T T A D E L . K . K S A D V R W H A E R I I N A V D D A V A S M D D T E K . . M S M K L R N L S G K H A K S F Q . . . V
Glycera         D . . . . . . . . . . P G V A A L G A K V L A Q I G V A V S H L G D E G K . . M V A Q M K A V G V R H . K G Y G N K H I
clam            S Q G M A N D K L . . . . . R G H S I T L M Y A L Q N F I D Q L D N P D D . . L V C V V E K F A V N H I T R K . . . . I
leghemoglobin   V D P T N P K L T G H A E K . . . . . . . . . L F A L V R D S A G Q L K A S G T V V A D A A L G S V H A Q K A . . . . V

                          100                 110                 120                 130                 140
Hb α            D P V N F K L L S H C L L V T L A A H L P . A E F T P A V H A S L D K F L A S V S T V L T S K Y R . . . . . .
Hb β            D P E N F R L L G N V L V C V L A H H F G . K E F T P P V Q A A Y Q K V V A G V A N A L A H K Y H . . . . . .
myoglobin       P V K Y L E F I S E C I I Q V L Q S K H P G . D F G A D A Q G A M N K A L E L F R K D M A S N Y K E L G F Q G
erythrocruorin  T H D Q L N N F R A G F V S Y M K A H T D F A G . . . . A E A A W G A T L D T F F G M I F S K M . . . . . . .
lamprey         D P E Y F K V L A A V I A D T V A A G D . . . . . . . . . . A G F E K L M S M I C I L L R S A Y . . . . . . .
Glycera         K A Q Y F E P L G A S L L S A M E H R I G G K M N A A A K . D A W A A A Y A D I S G A L I S G L Q S . . . . .
clam            S A A E F G K I N G P I K K V L A S K N F G D . . . . K Y . A N A W A K L V A V V Q A A L . . . . . . . . . .
leghemoglobin   T D P Q F V V V K E A L L K T I K A A V G D K . W S D E L S R A W E V A Y D E L A A A I K K A . . . . . . . .
```

Number of known protein sequences



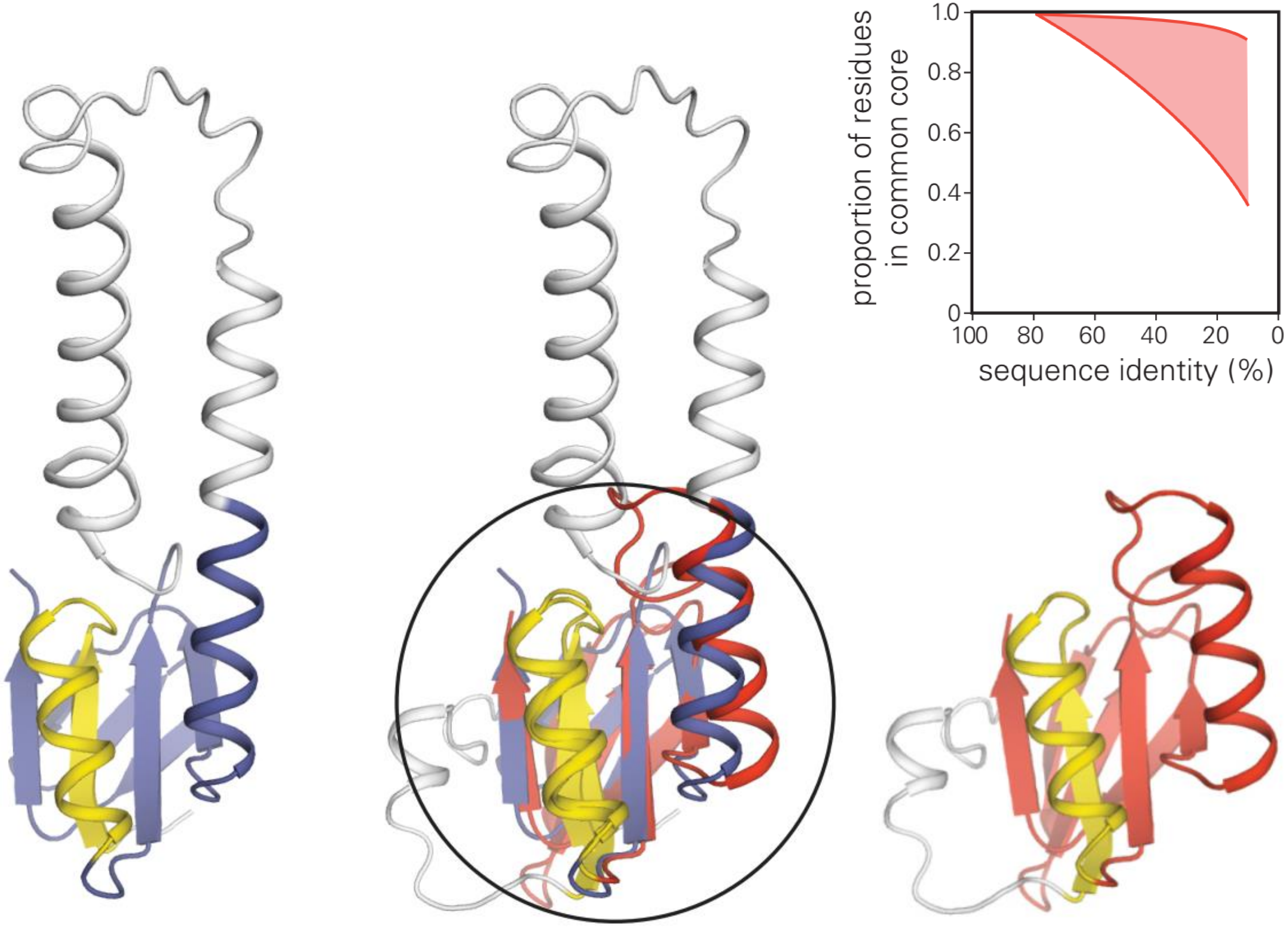
Structure variation is related to sequence variation

- 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)*



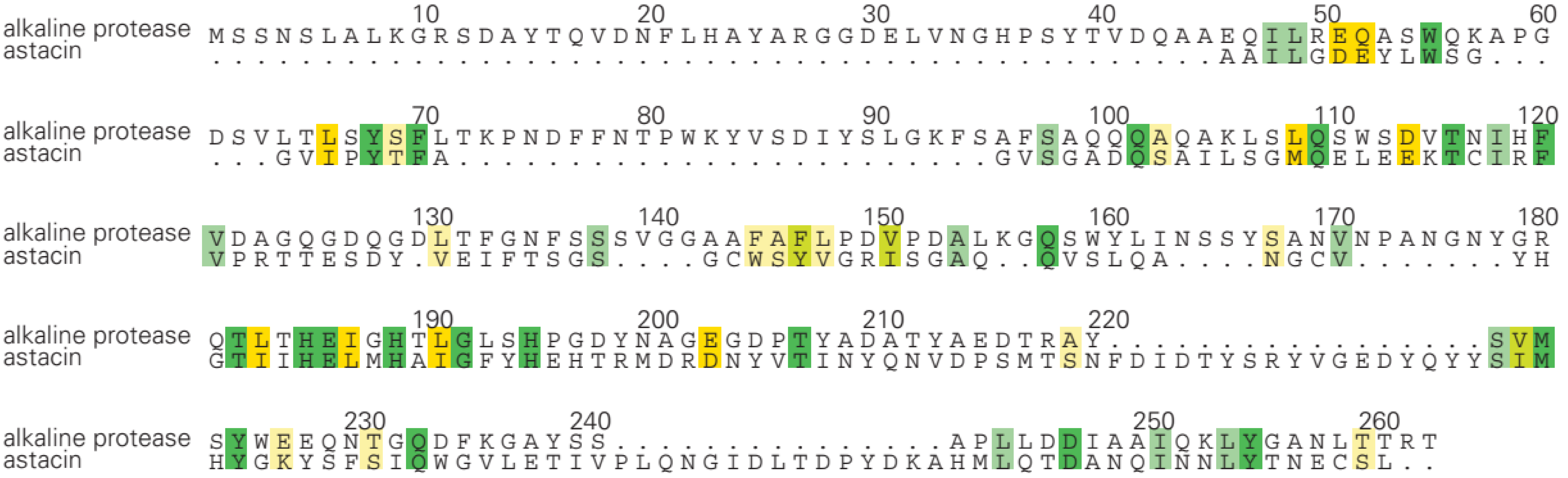
Common structural core

- 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



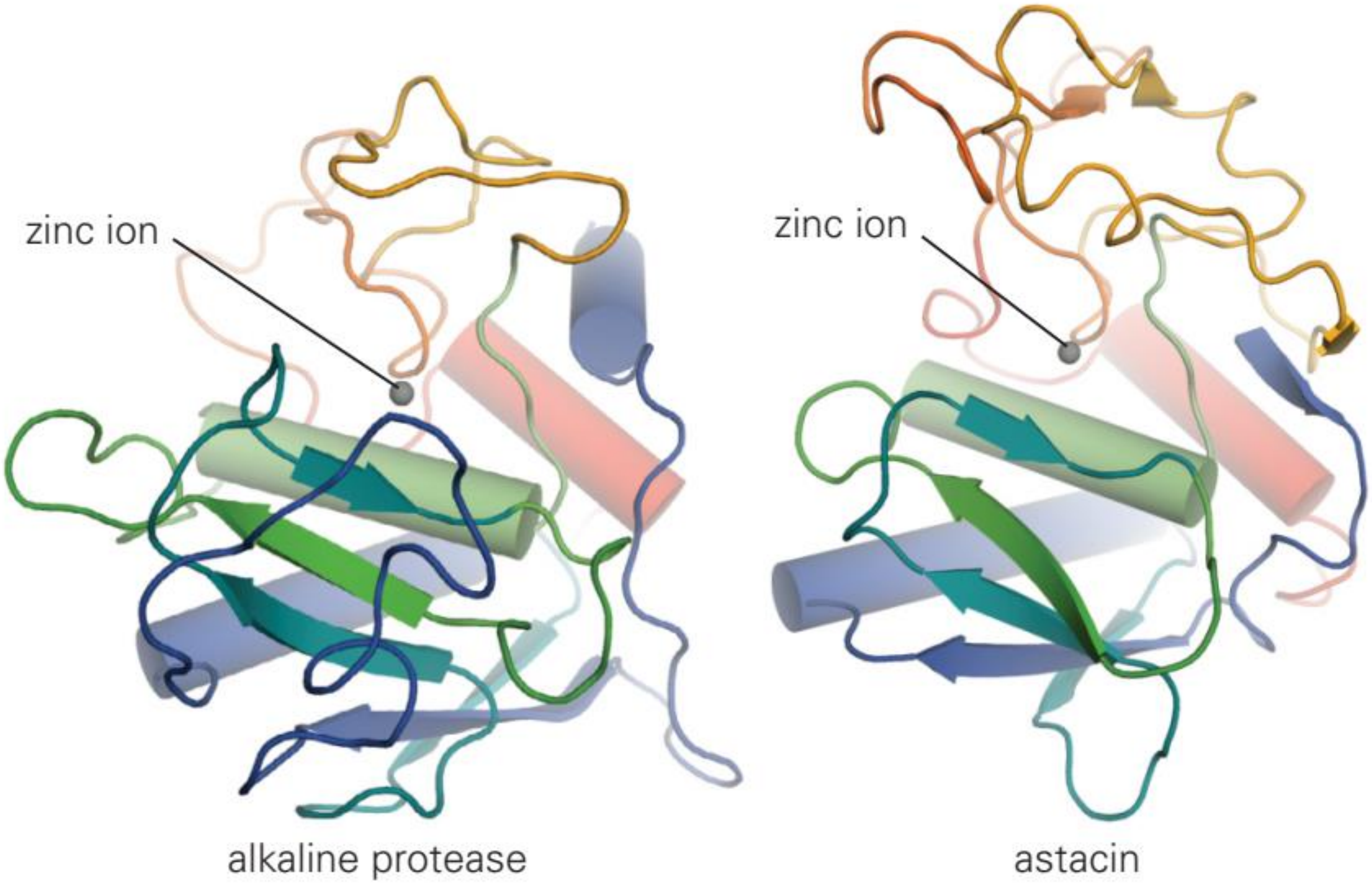
common core

Conservation along the whole sequence can be missleading

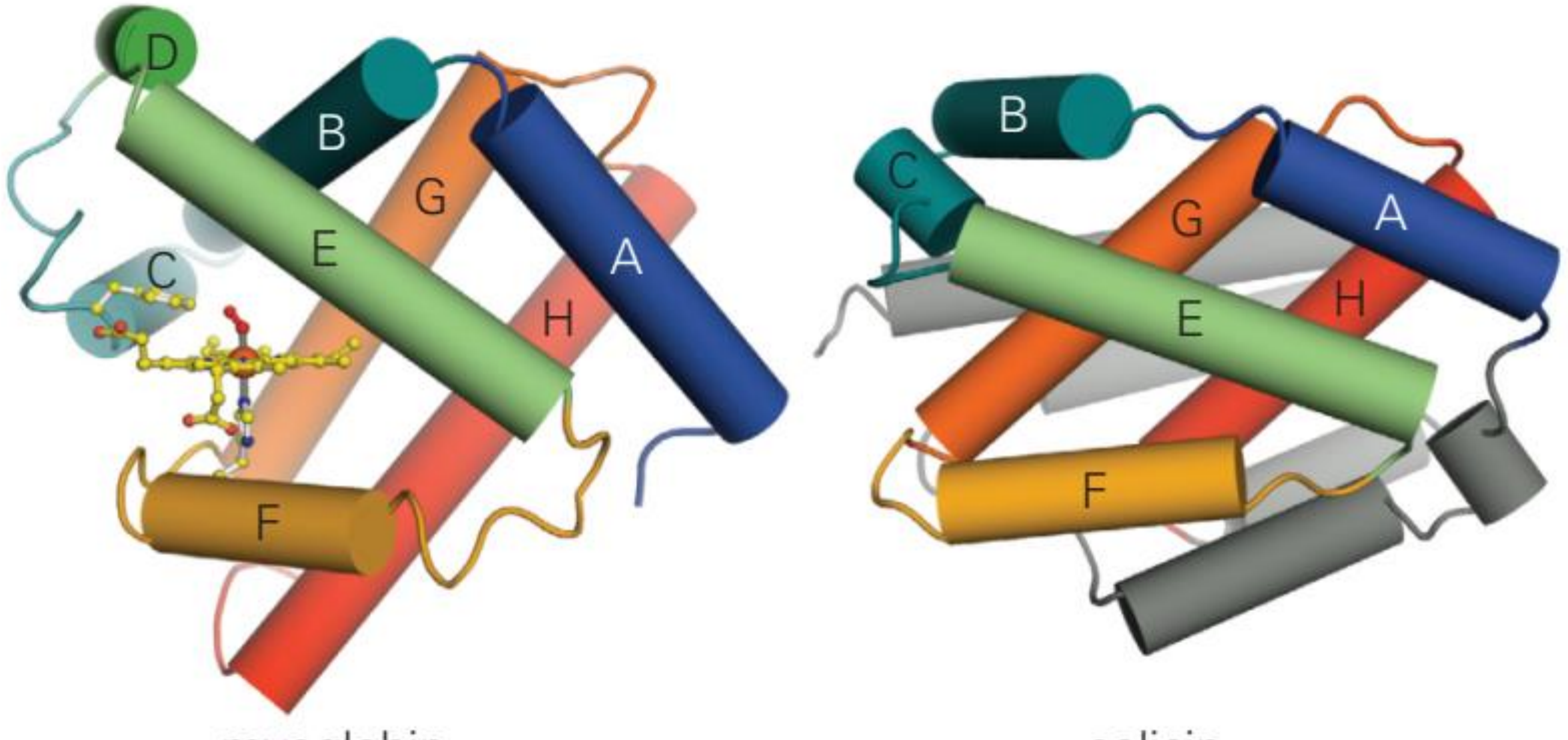


```
                  10        20        30        40        50        60
alkaline protease MSSNSLALKGRSDAYTQVDNFLHAYARGGDELVNGHPSYTVDQAAEQILREQASWQKAPG
astacin           .............................................AAILGDEYLWSG...

                  70        80        90        100       110       120
alkaline protease DSVLTLSYSFLTKPNDFFNTPWKYVSDIYSLGKFSAFSAQQQAQAKLSLQSWSDVTNIHF
astacin           ...GVIPYTFA........................GVSGADQSAILSGMQELEEKTCIRF

                  130       140       150       160       170       180
alkaline protease VDAGQGDQGDLTFGNFSSSVGGAAFAFLPDVPDALKGQSWYLINSSYSANVNPANGNYGR
astacin           VPRTTESDY.VEIFTSGS....GCWSYVGRISGAQ..QVSLQA....NGCV.......YH

                  190       200       210       220
alkaline protease QTLTHEIGHTLGLSHPGDYNAGEGDPTYADATYAEDTRAY.................SVM
astacin           GTIIHELMHAIGFYHEHTRMDRDNYVTINYQNVDPSMTSNFDIDTYSRYVGEDYQYYSIM

                  230       240       250       260
alkaline protease SYWEEQNTGQDFKGAYSS...............APLLDDIAAIQKLYGANLTTRT
astacin           HYGKYSFSIQWGVLETIVPLQNGIDLTDPYDKAHMLQTDANQINNLYTNECSL..
```

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

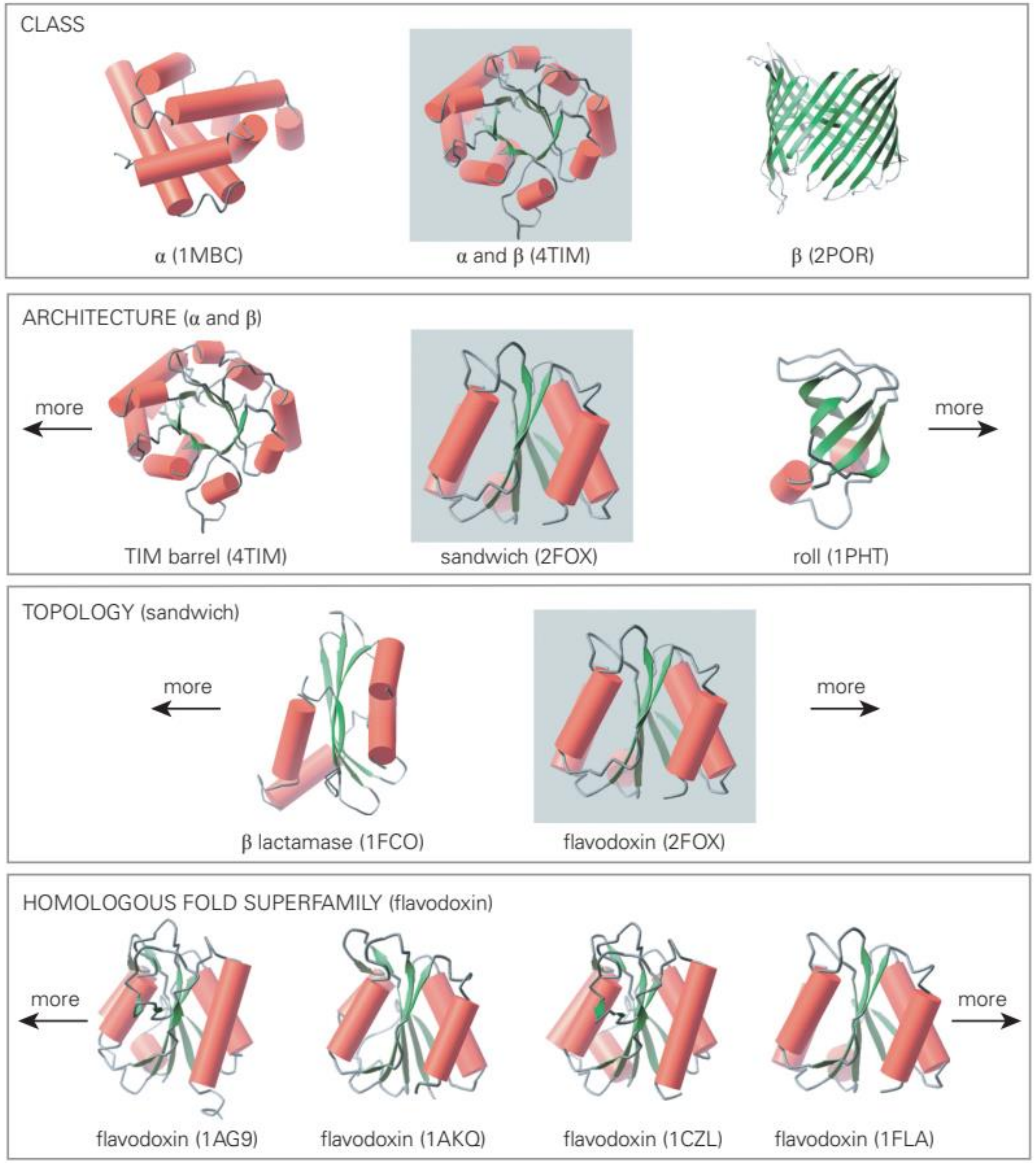


Conserved cores are often reused for different functions

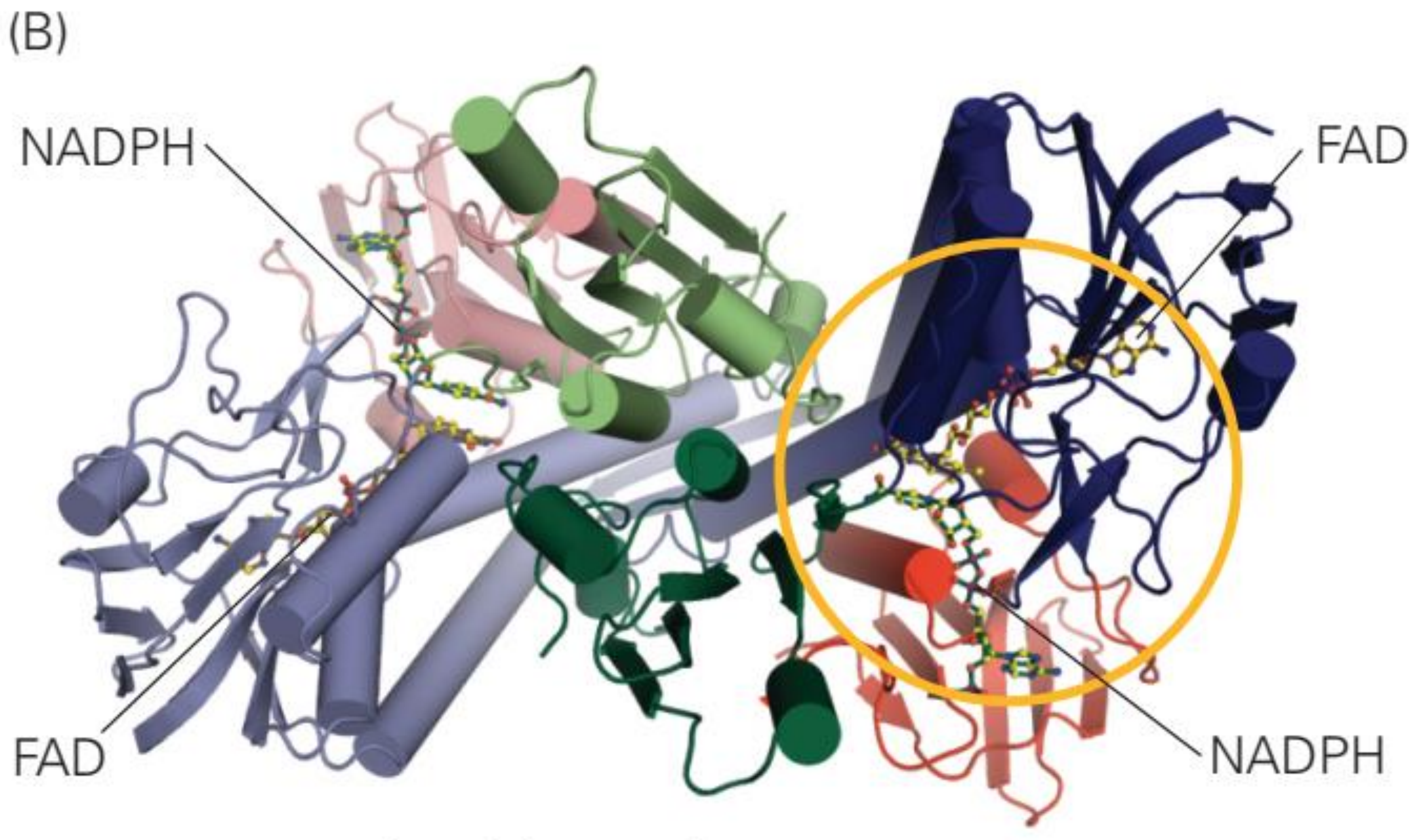


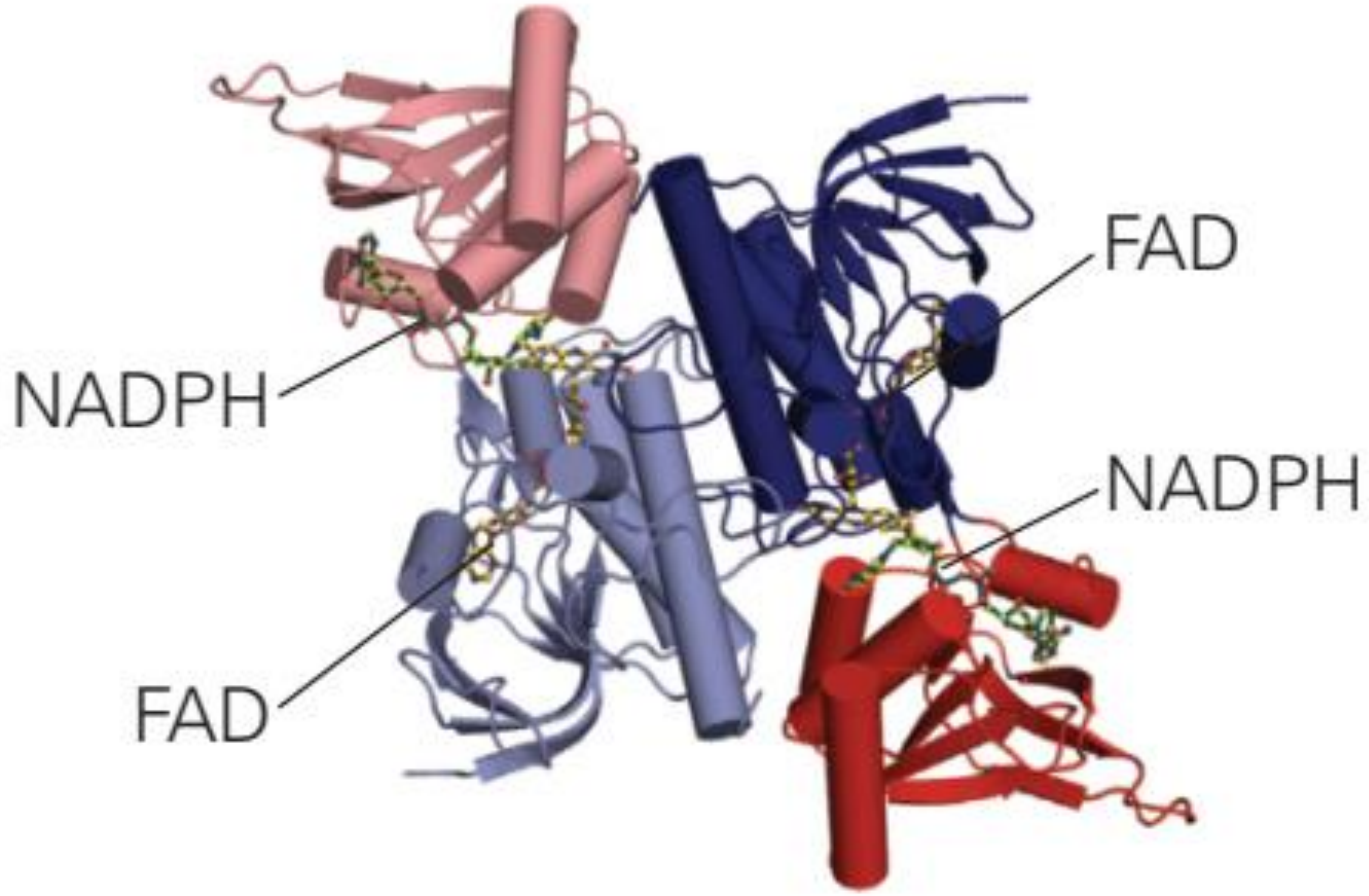
myoglobin

colicin

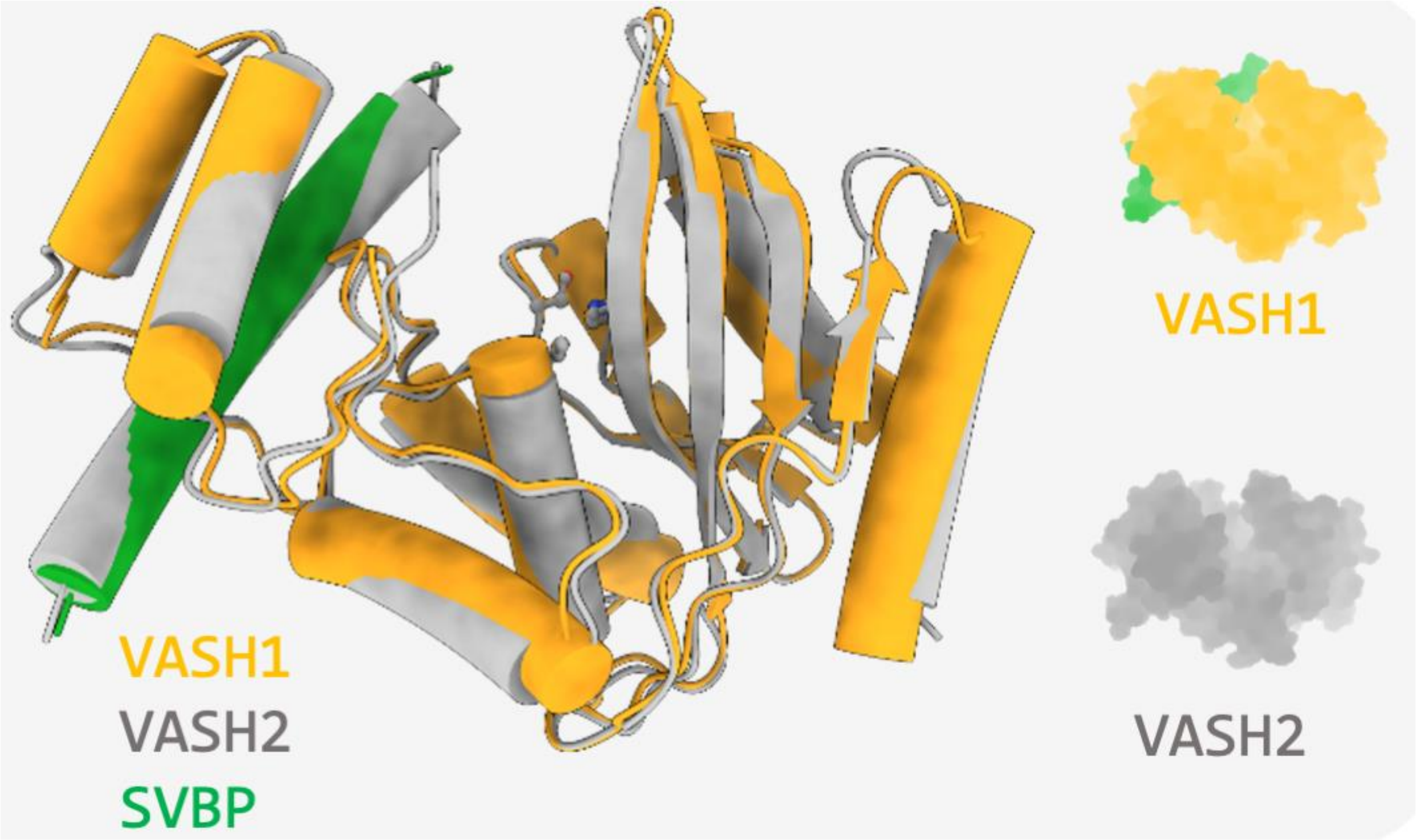


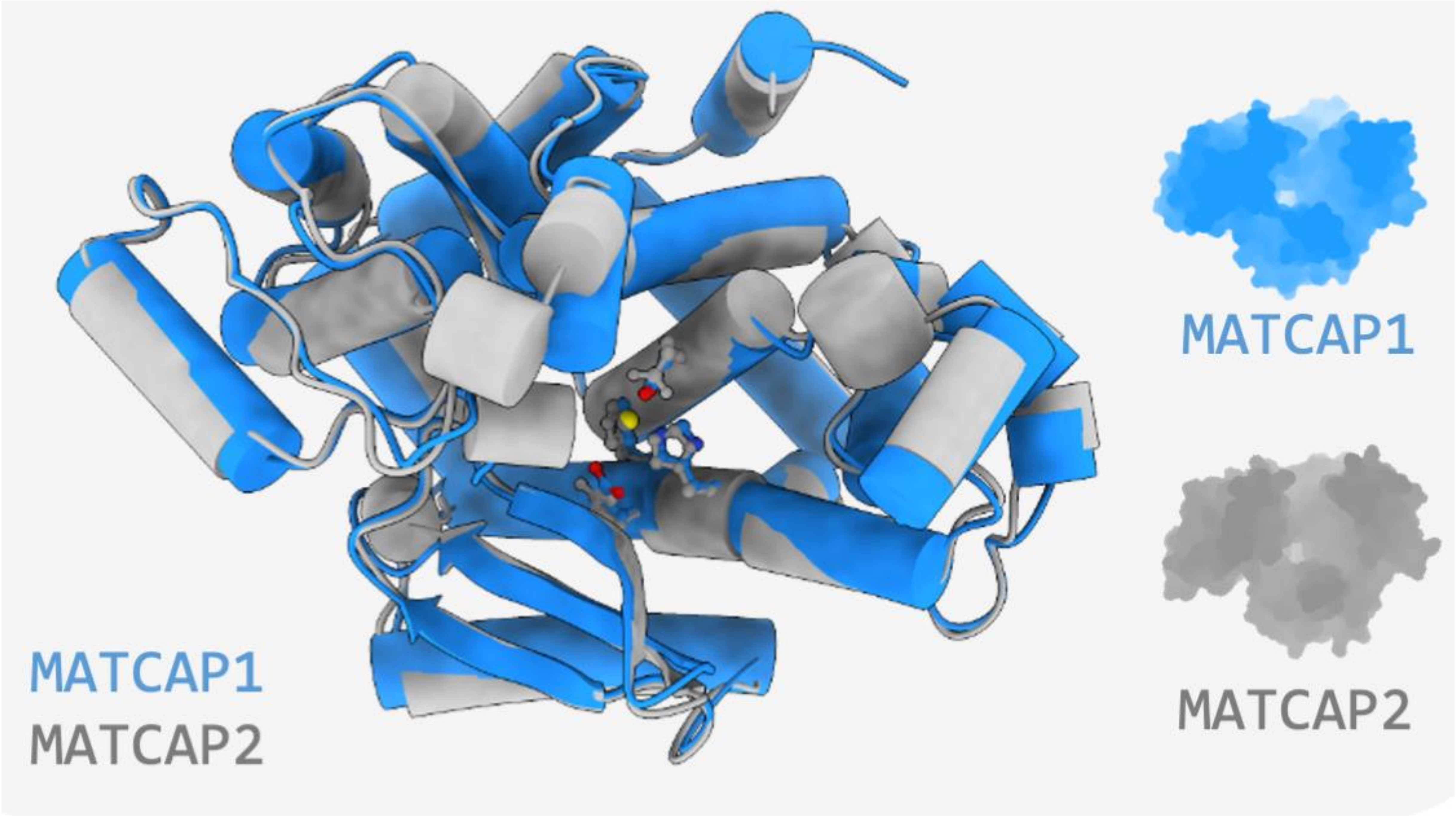
The same function can be served by different folds: convergent evolution





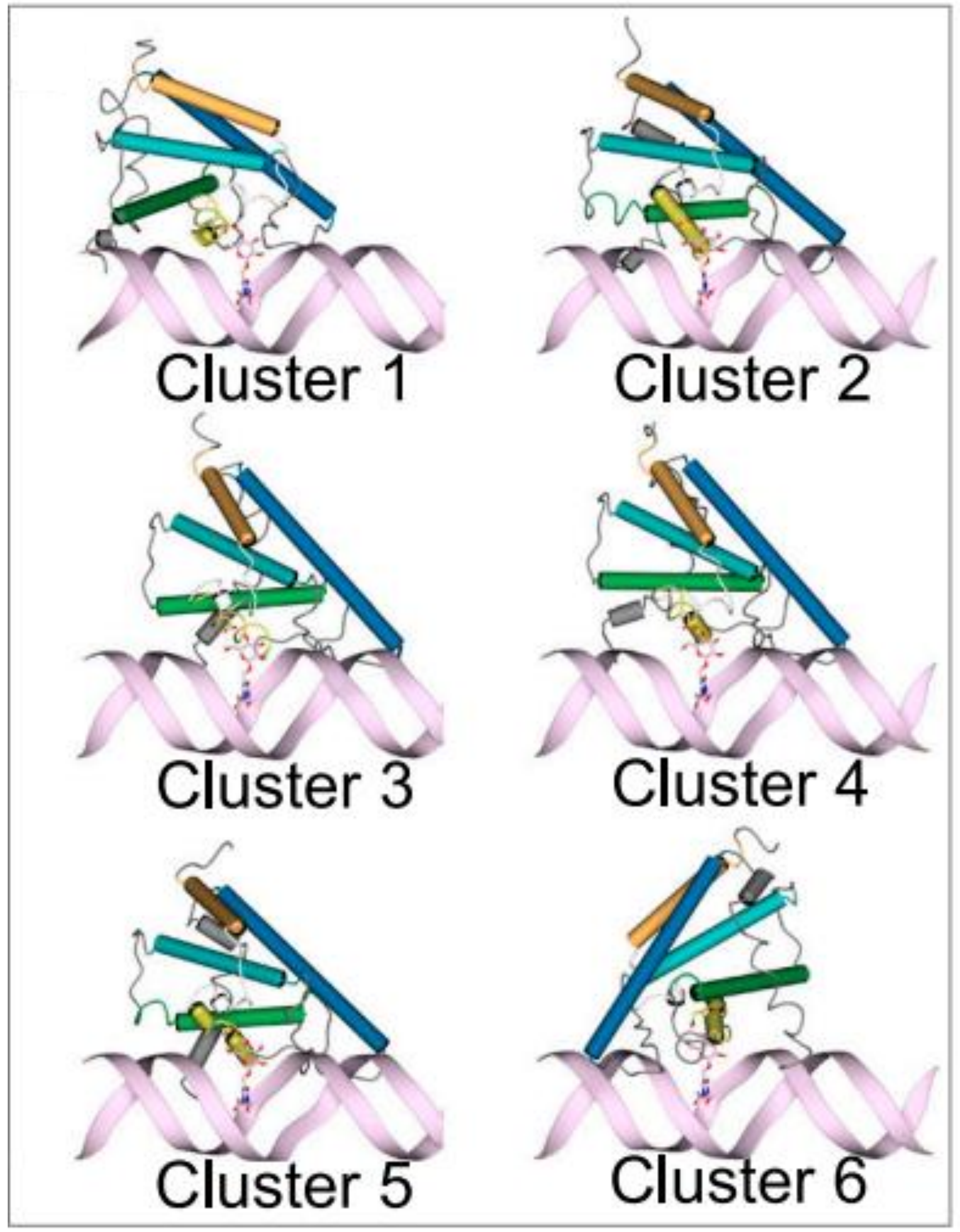
The same function can be served by different folds: convergent evolution

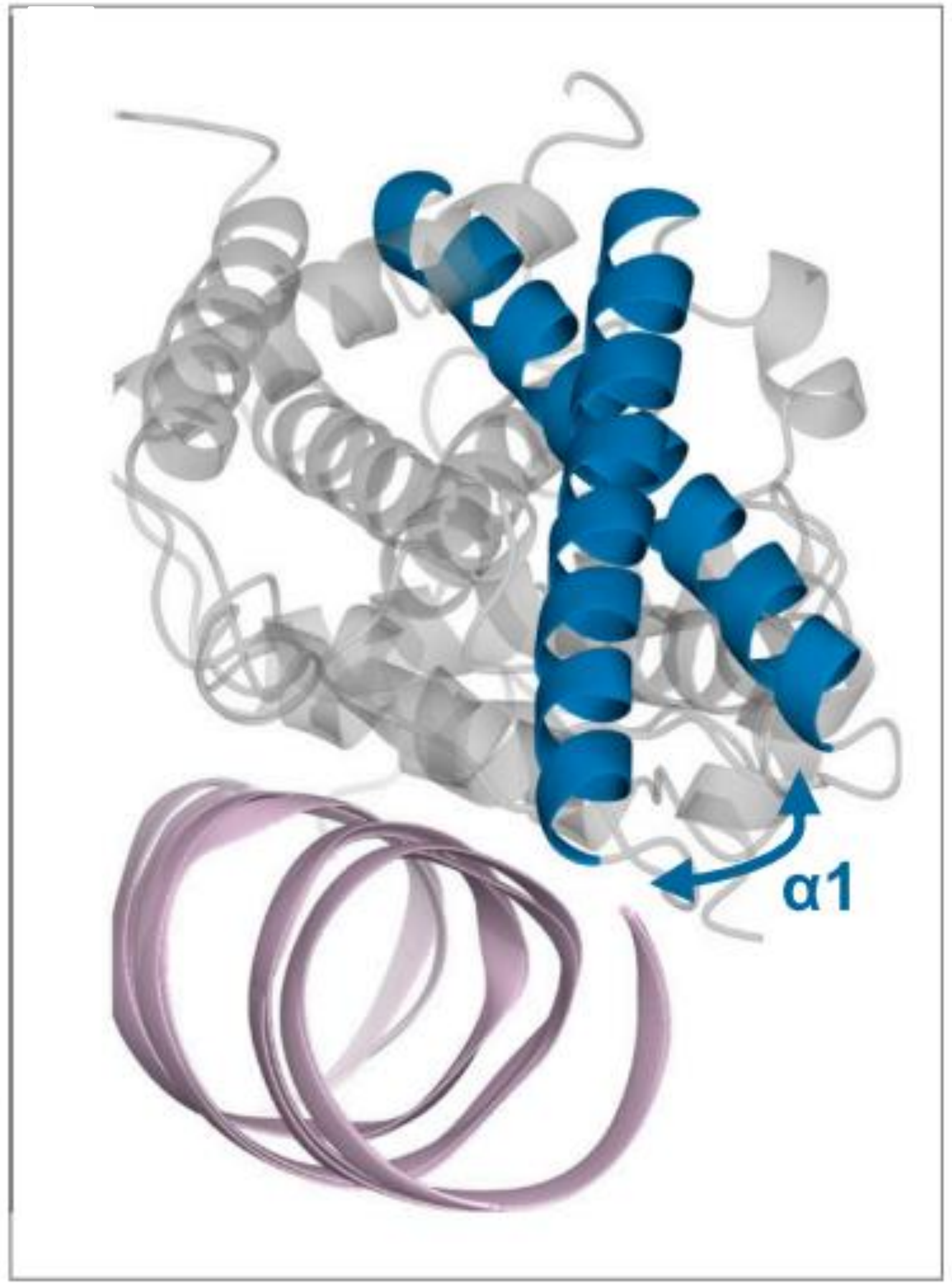




Aligning (superposing) structures

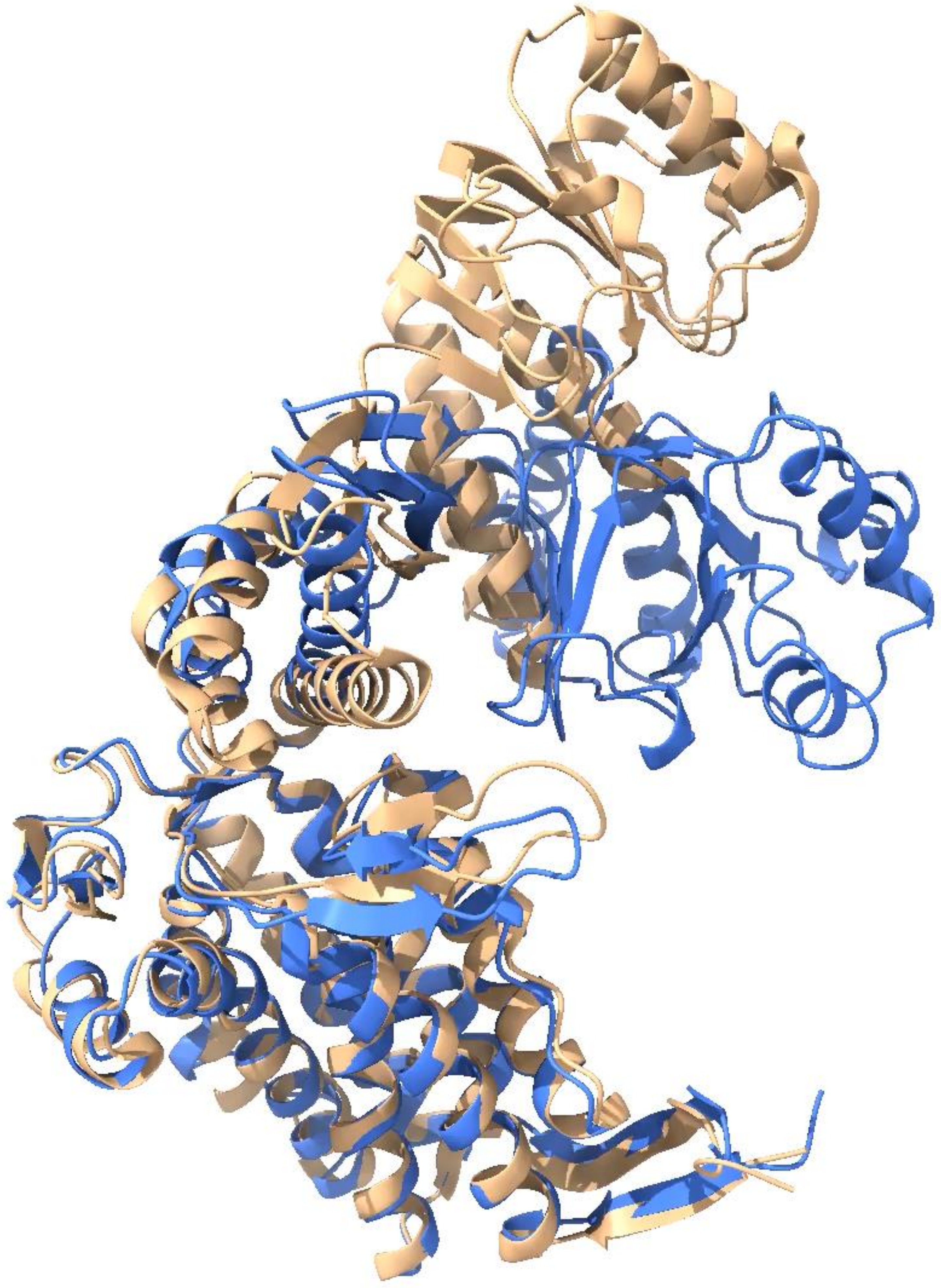
- Structure alignemnt is not identical to sequence alignment.
- Which one is preferred depends to the case study.
- Both can be informative.
- Automated structure alignments (e.g. Chimera-X “matchmaker” or Coot “SSM superpose”) will align automatically two structures.
- You can show superposed structures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





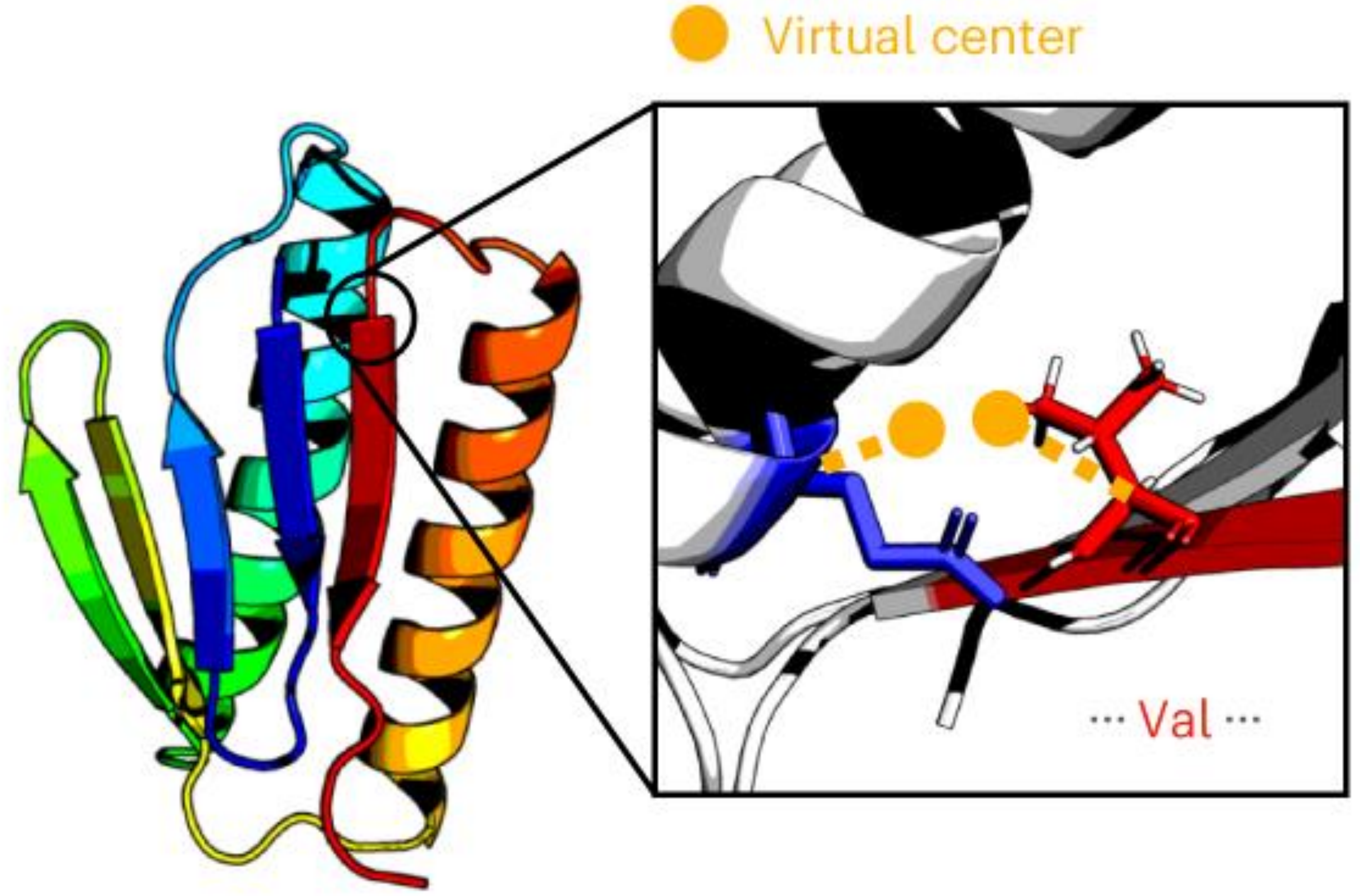
Aligning (superposing) multidomain proteins

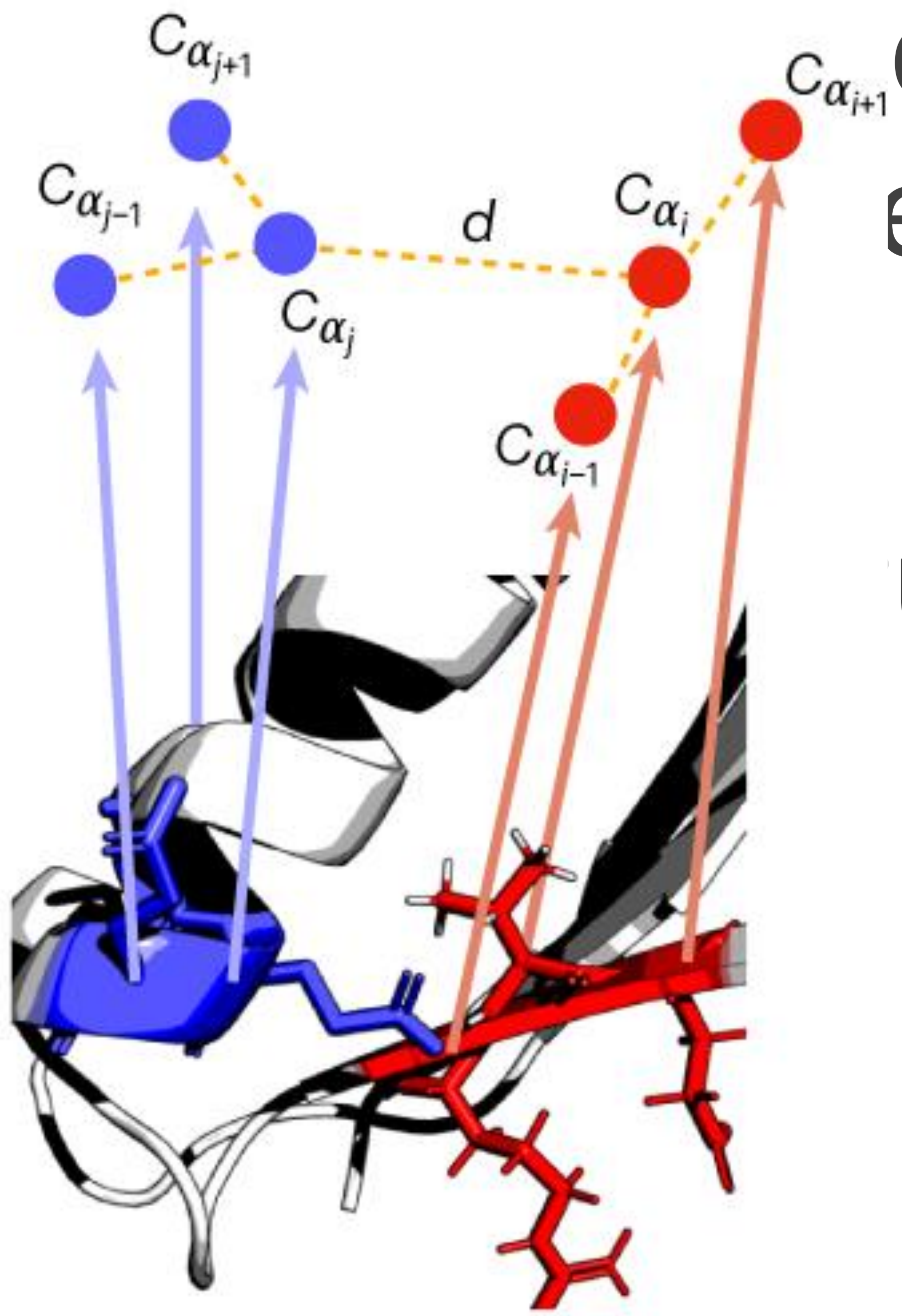
- Proteins are often put together by stiching together different domains.
- Domain movement needs to be taken into account during a structural alignment.
- Tools like RAPIDO, THESEUS, or [*your brain*] can be used to decide how to superpose.
- Superposition-independent structure comparisons are sometimes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

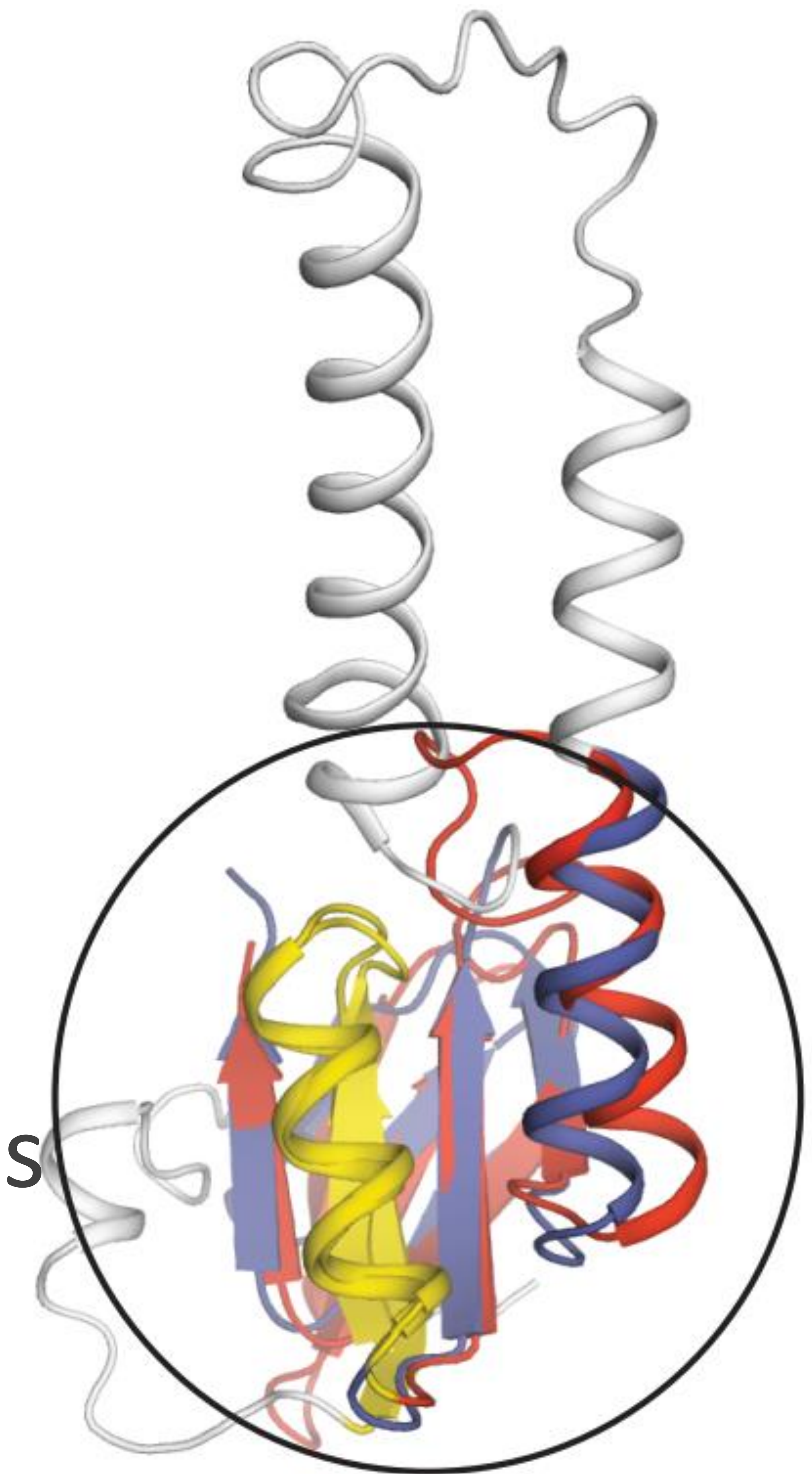


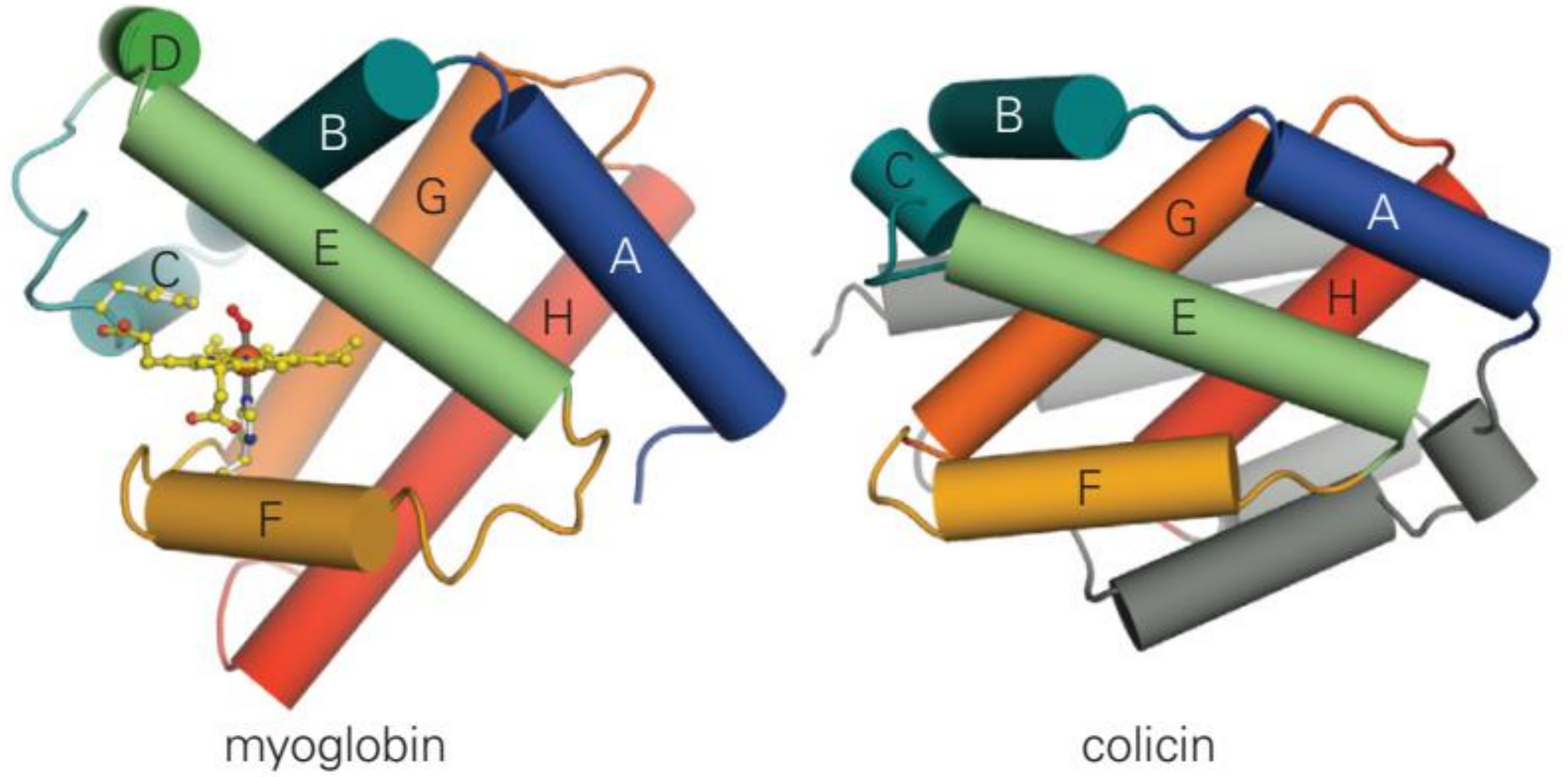
Searching for structure similarity

- 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.
- Cl conformations in ence al
- Fo uctu al









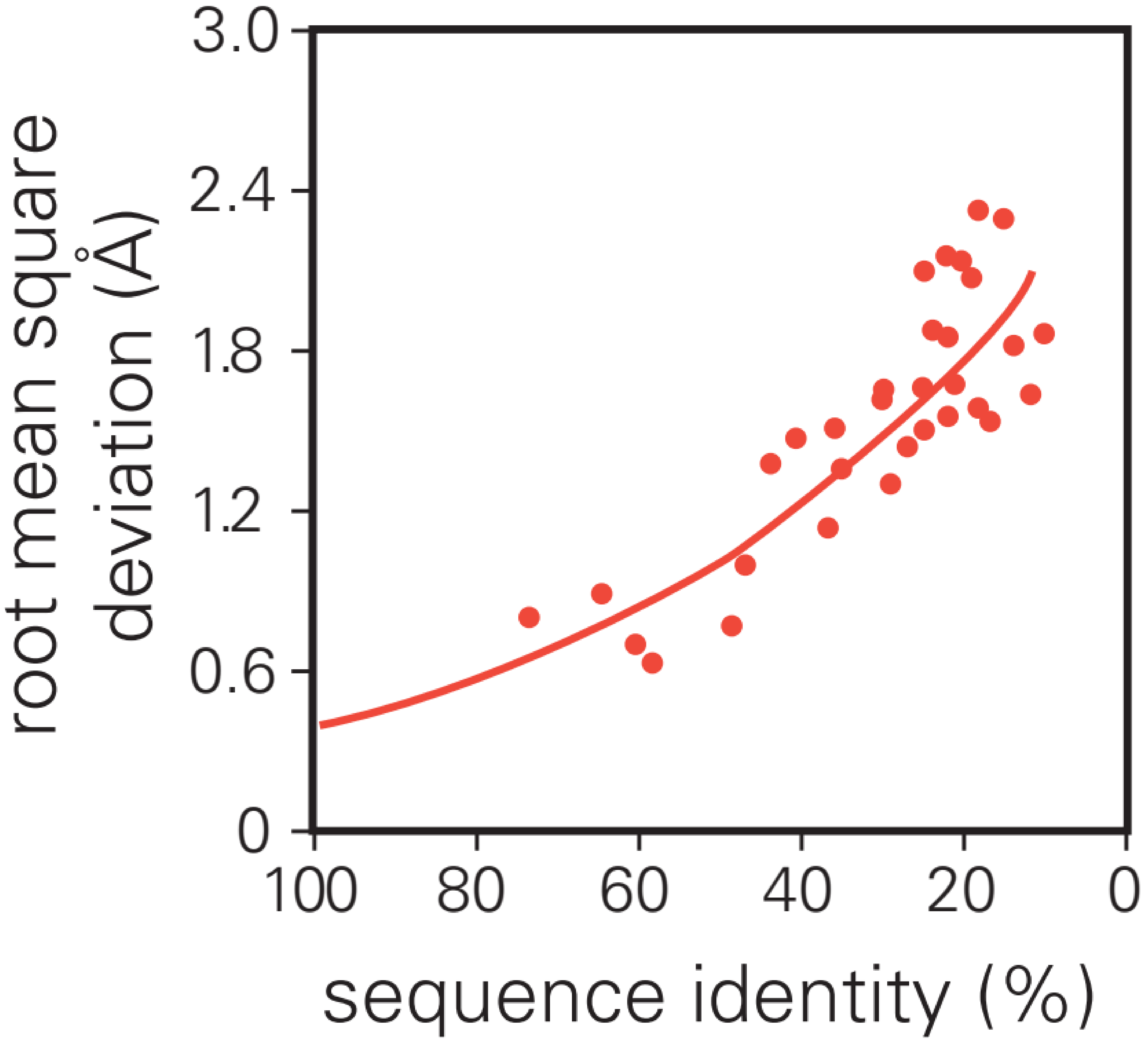
Sequence and structure divergence

- 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.

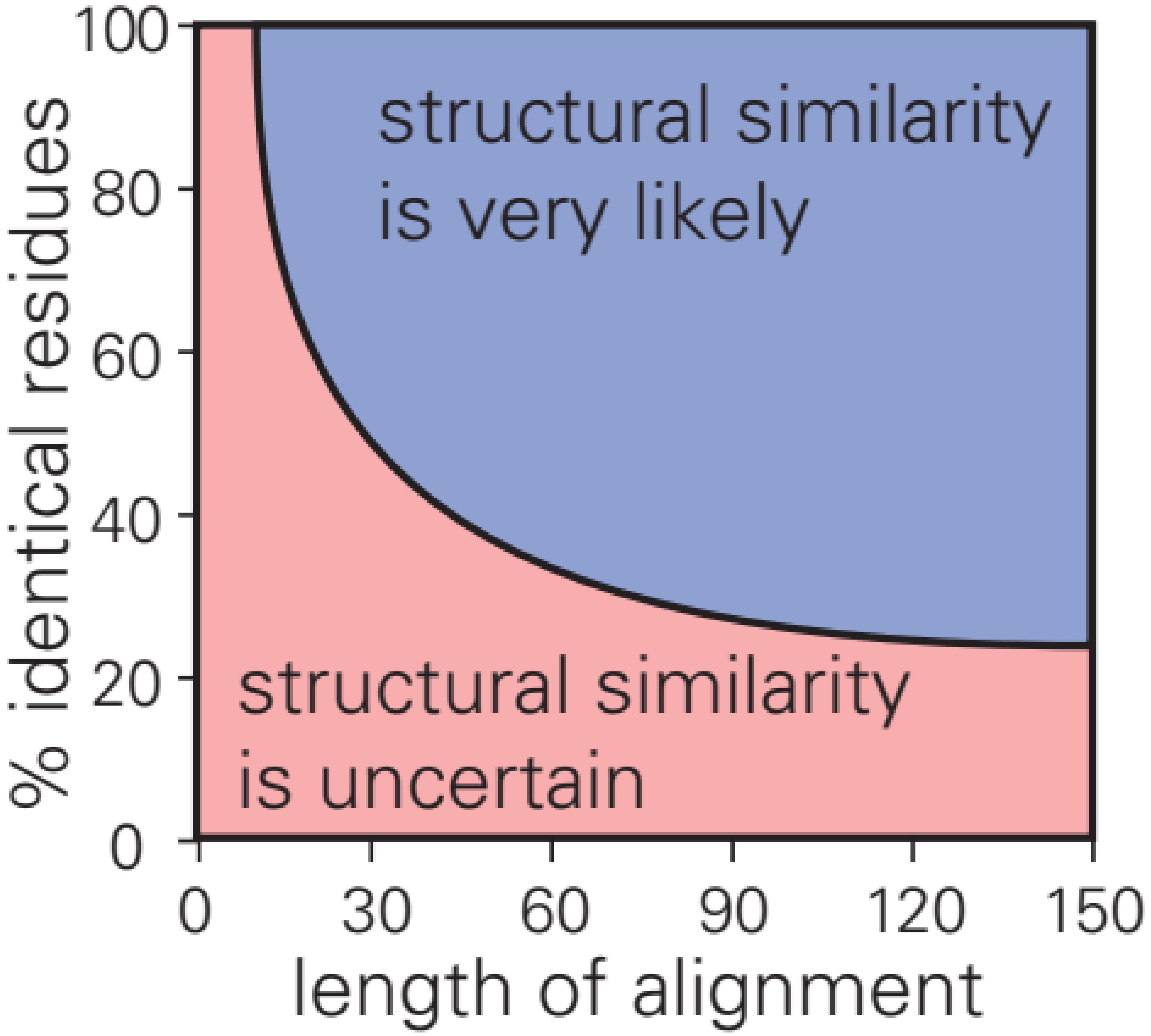
- At what degree does sequence similarity guarantee structure similarity?

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, tc be meaningful.
- Proteins that are related by ~50% or more sequence identity have common cores tha are structurally very similar, with rms deviations that 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.

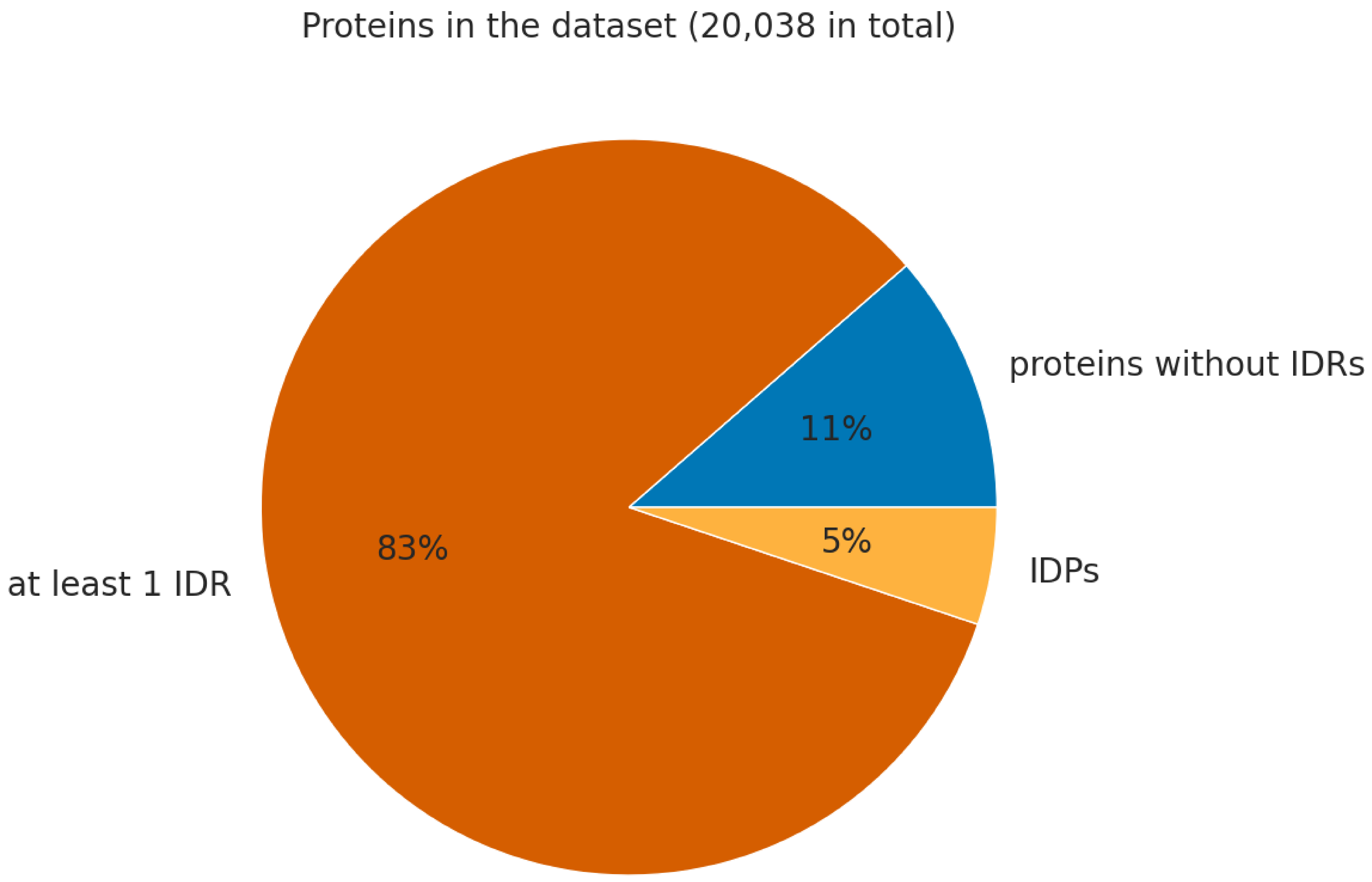


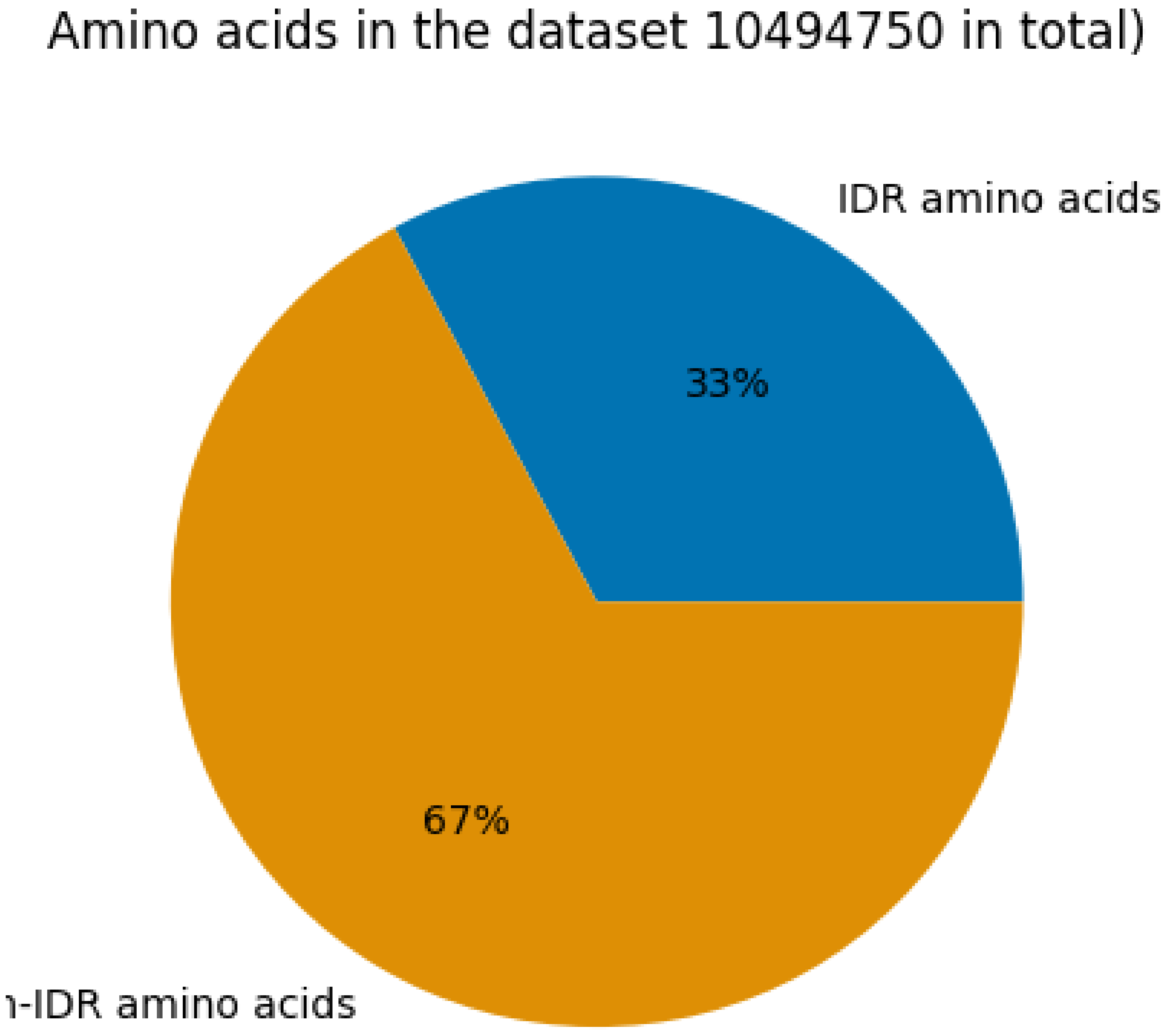
How much can we tell about structure from sequence?



- 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





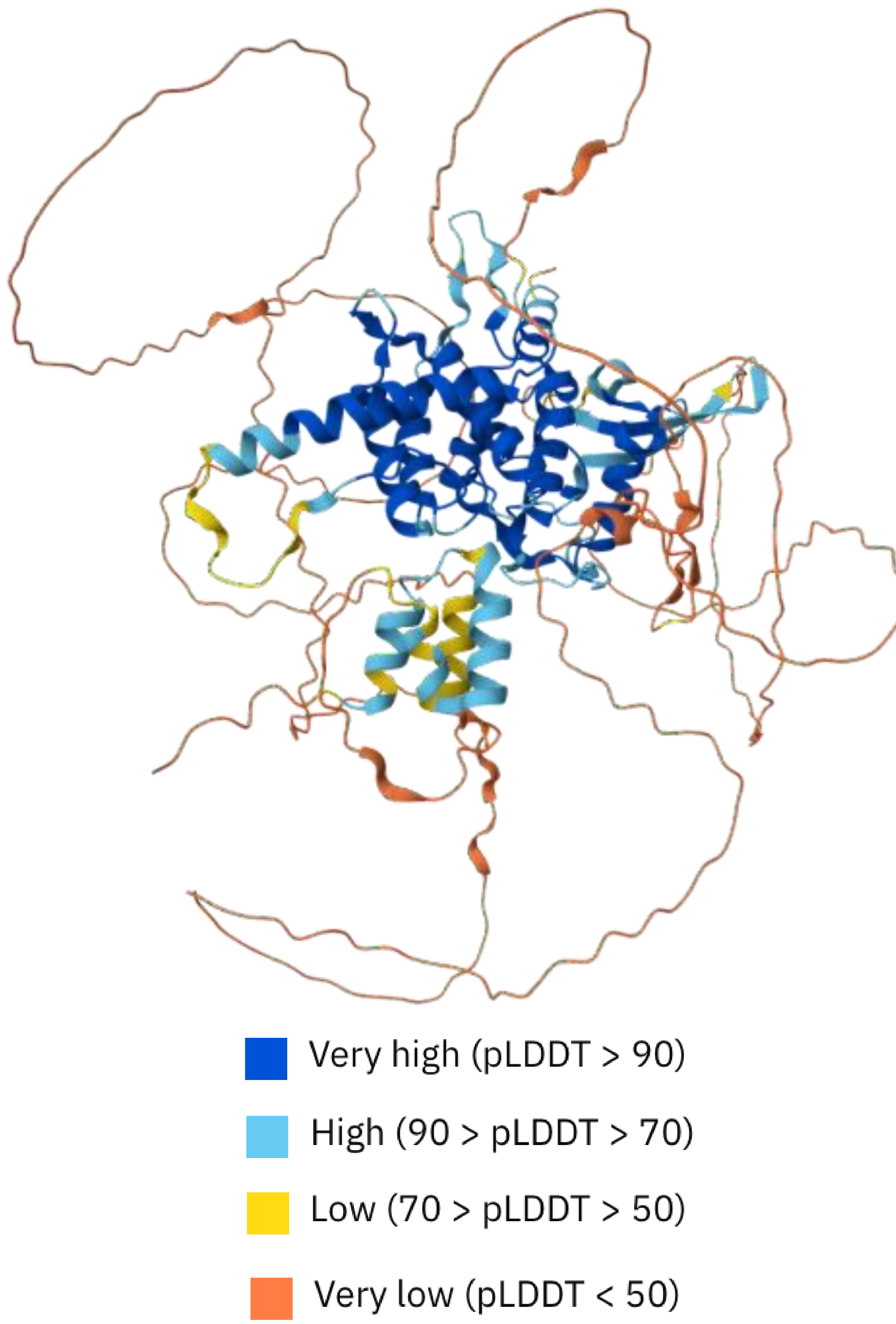
AI redefines the limits on the information that can be extracted from sequences.

- 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.

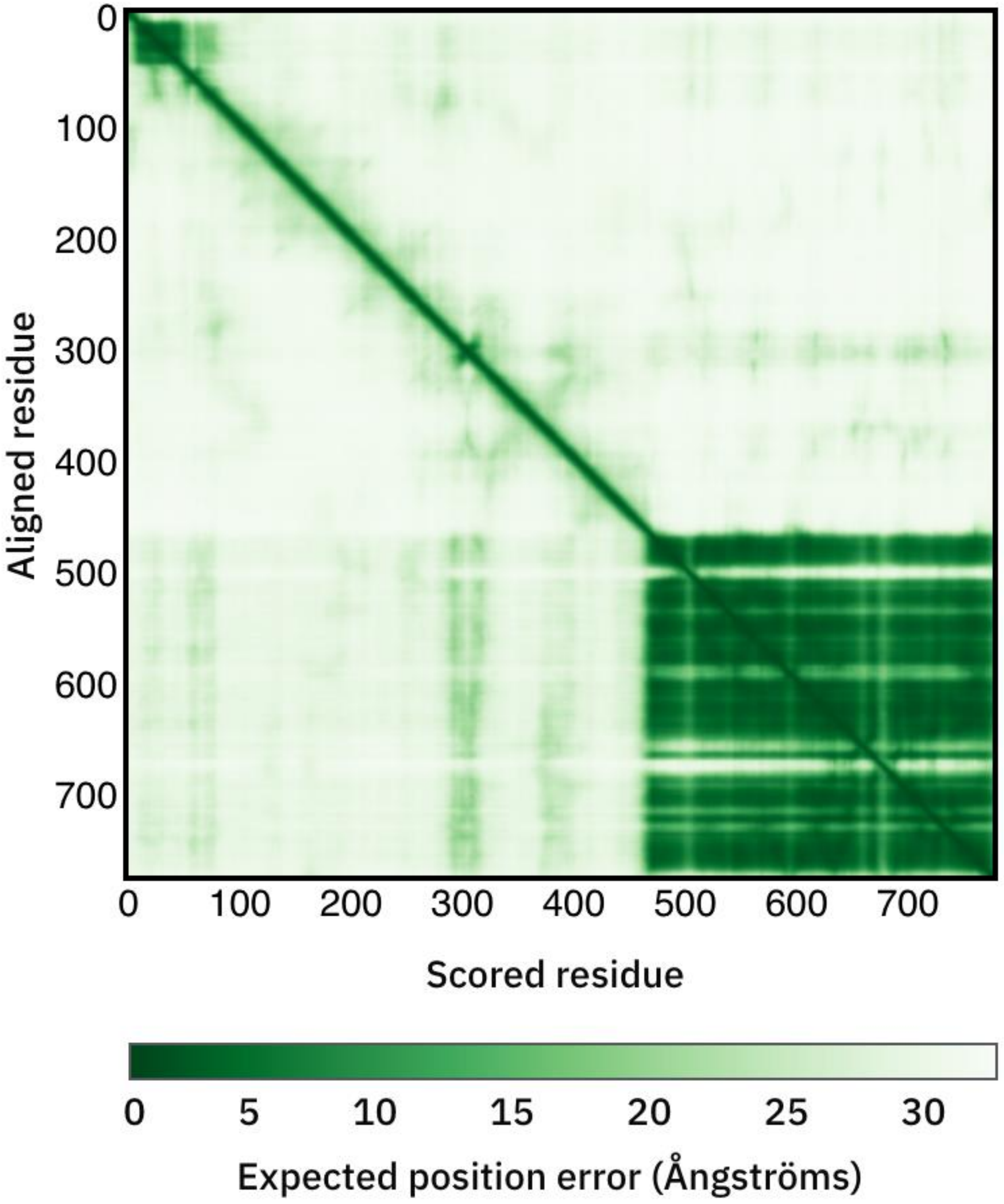
- AlphaFold3 provides excellent validation criteria.

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
- AlphaFold 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



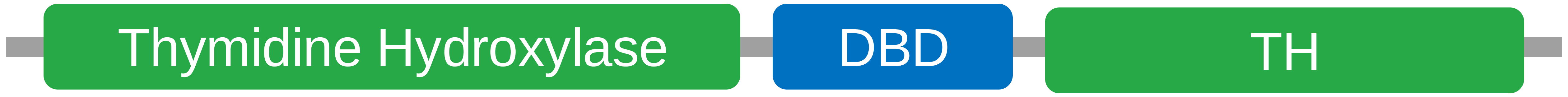
- The colour at (x, y) indicates AlphaFold's expected position error at residue x if the predicted and true structures were aligned on residue y.
- PAE within a structural core – domain will be low (dark green)
- If the PAE is generally low for residue pairs x, y from two different domains, it indicates 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

On the asymmetry of the PAE matrix

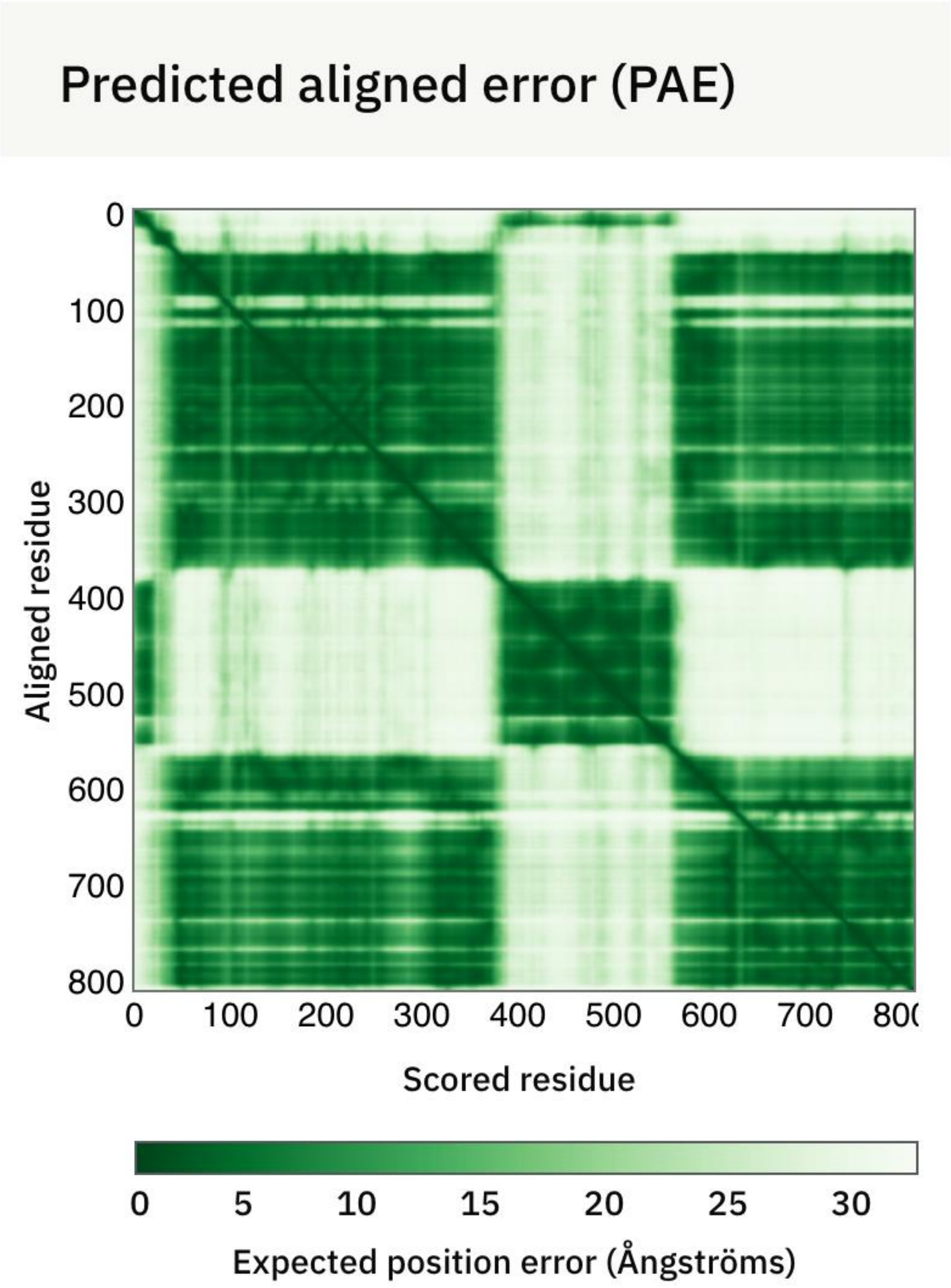
- From the perspective of the lighthouse the position of the boat cannot be accurately determined
- From the perspective of the boat the position of the lighthouse is clear.

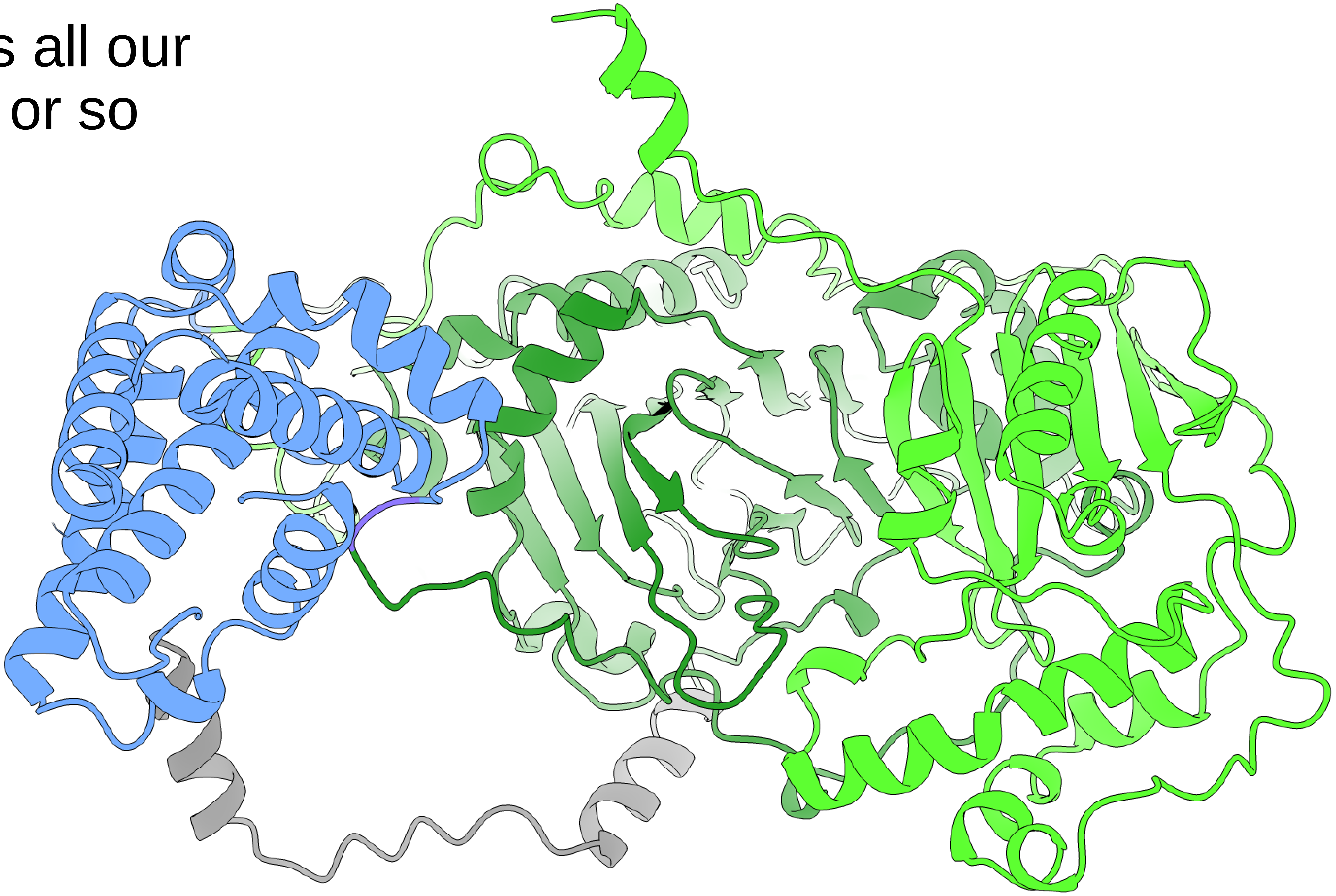
What does the model tell us?

The importance of PAE plots



- AlphaFold model confirms all our work for the last 15 years or so

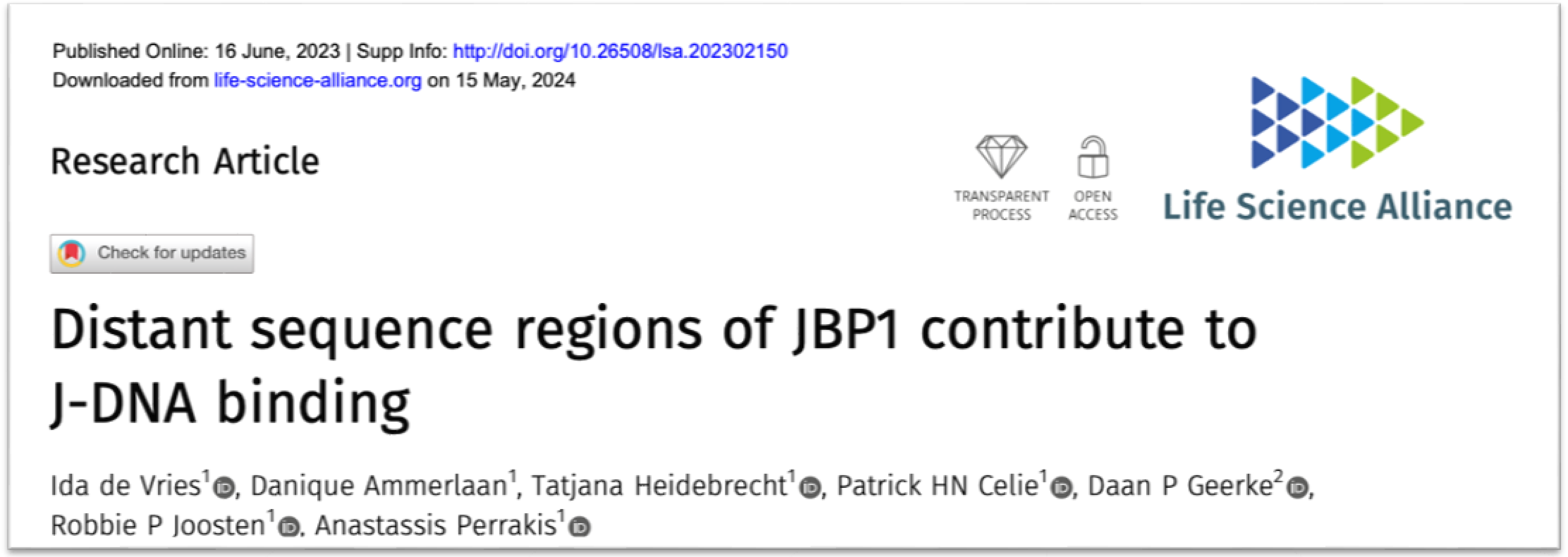

Predicted aligned error (PAE)



Are predictions enough?

No. You still need experiments

JBP1 is an Fe-dependent thymidine hydroxylase.
Where is the iron?



Published Online: 16 June, 2023 | Supp Info: http://doi.org/10.26508/lsa.202302150
Downloaded from life-science-alliance.org on 15 May, 2024

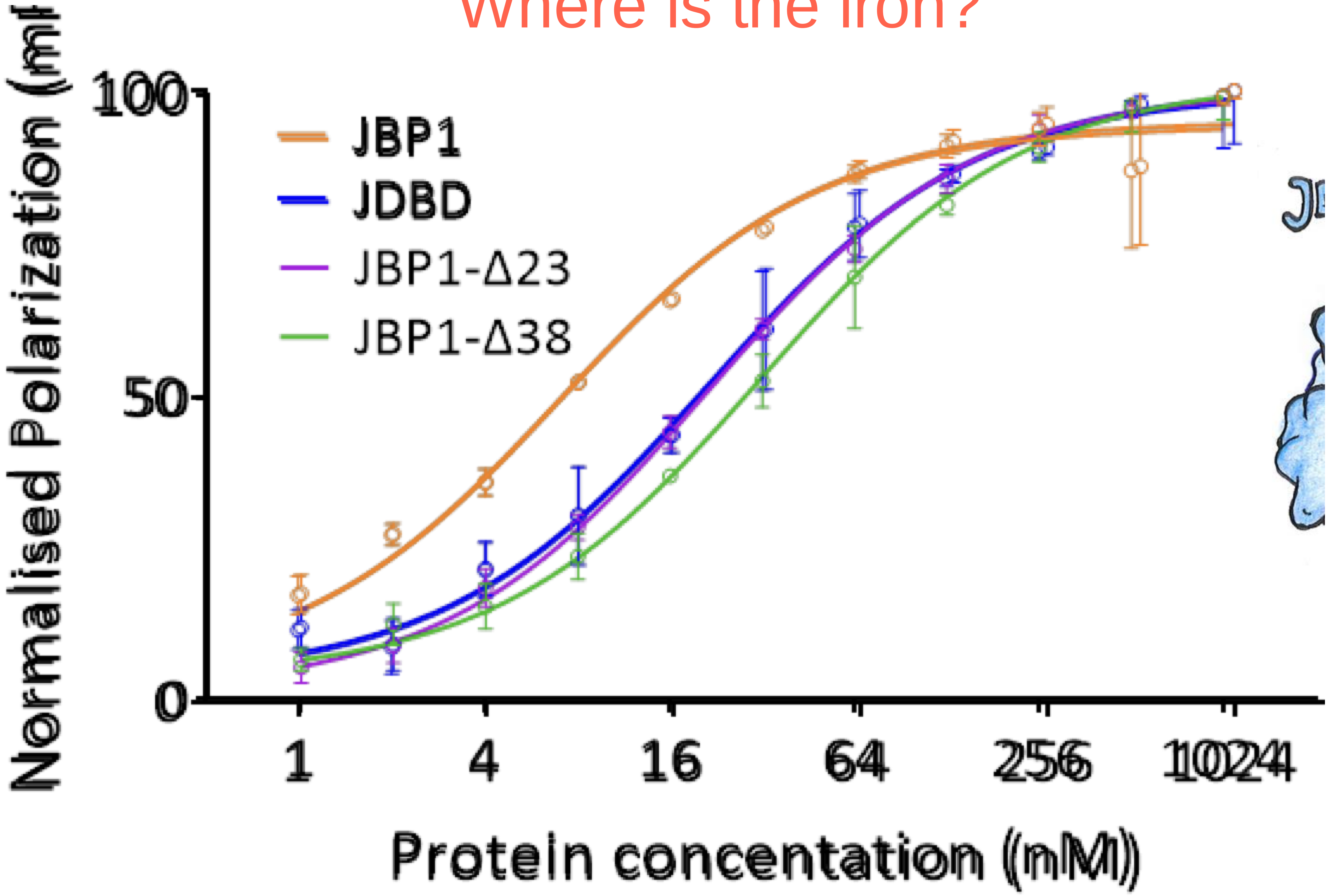
Research Article

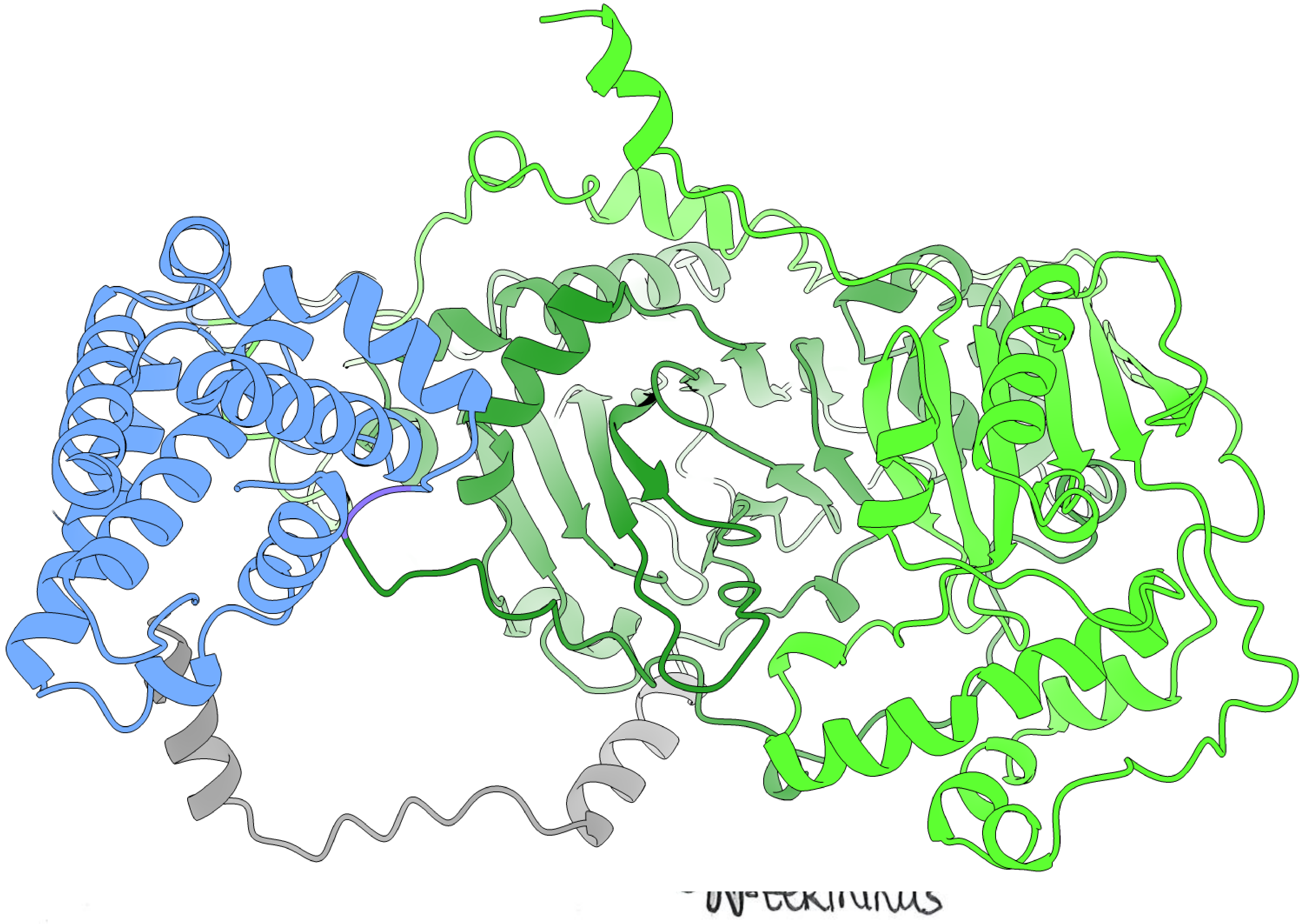
TRANSPARENT PROCESS OPEN ACCESS Life Science Alliance

Check for updates

Distant sequence regions of JBP1 contribute to J-DNA binding

Ida de Vries[1], Danique Ammerlaan[1], Tatjana Heidebrecht[1], Patrick HN Celie[1], Daan P Geerke[2], Robbie P Joosten[1], Anastassis Perrakis[1]

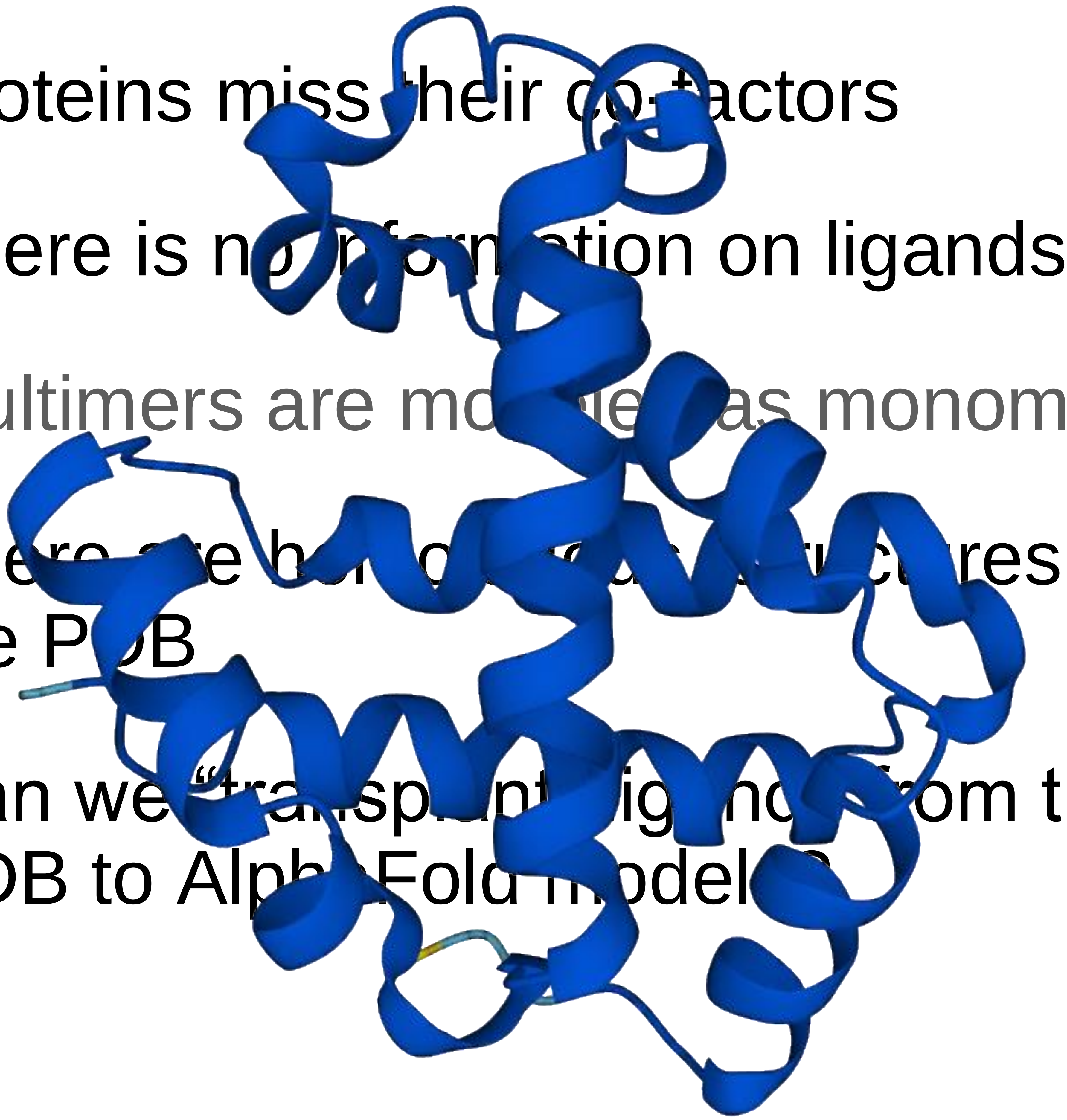


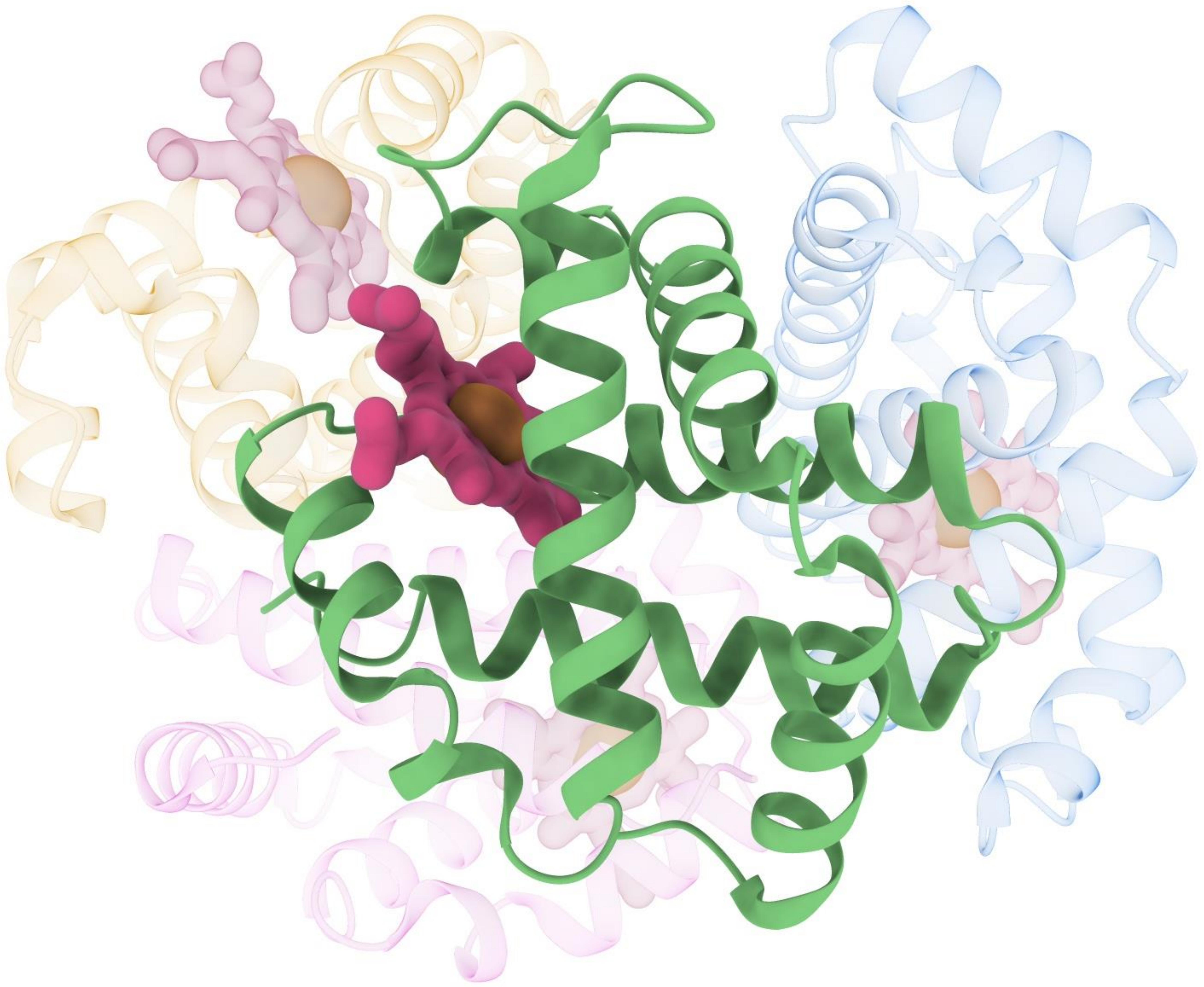


Things are missing from protein structure predictions?

… how do we enrich these predictions with such information?

- Proteins miss their co-factors
- There is no information on ligands
- Multimers are modelled as monomers
- There are homologous structures in the PDB
- Can we "transplant" ligands from the PDB to AlphaFold models?

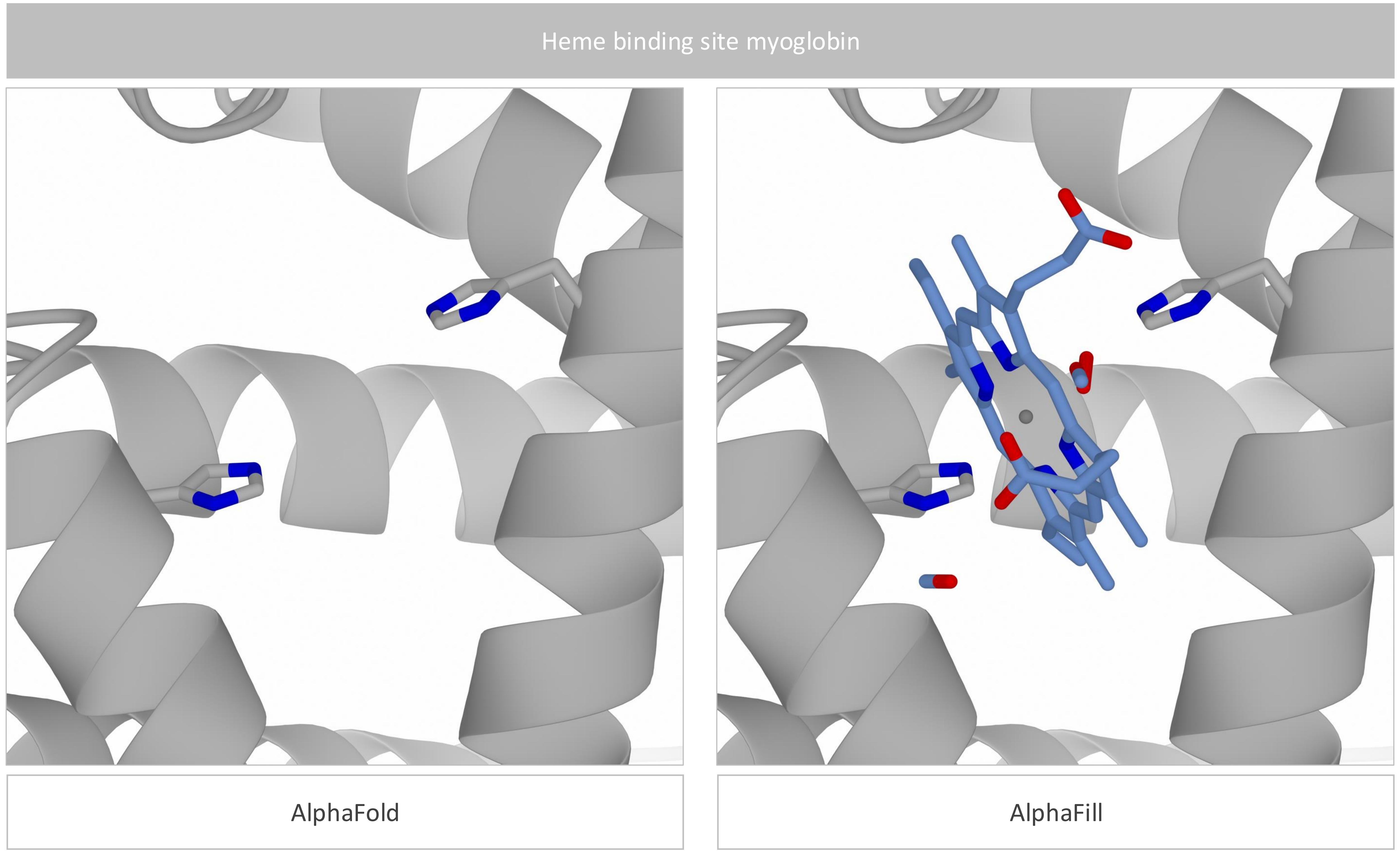




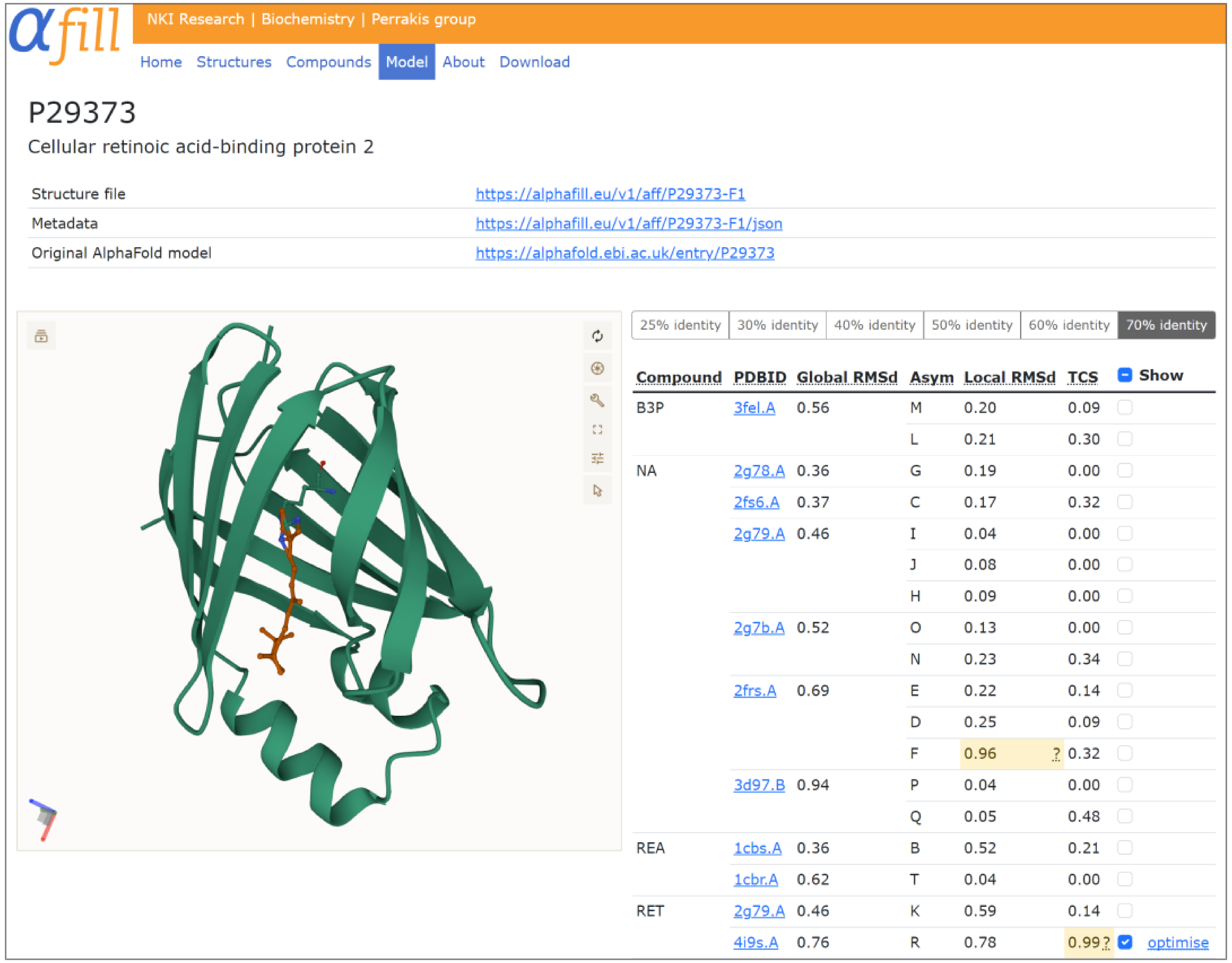
A summary of the AlphaFill resource

Using AlphaFold DataBase June 2022

- Apply to the AlphaFold database: 995k predicted protein structures
  - Model organism proteomes
  - Proteomes relevant to global health
  - Swiss-Prot
- AlphaFill:
  - 586k AlphaFold models with transplants
  - 12 million transplanted compounds
- All the recently added AlphaFold models will be generated “on the fly”



The AlphaFill databank



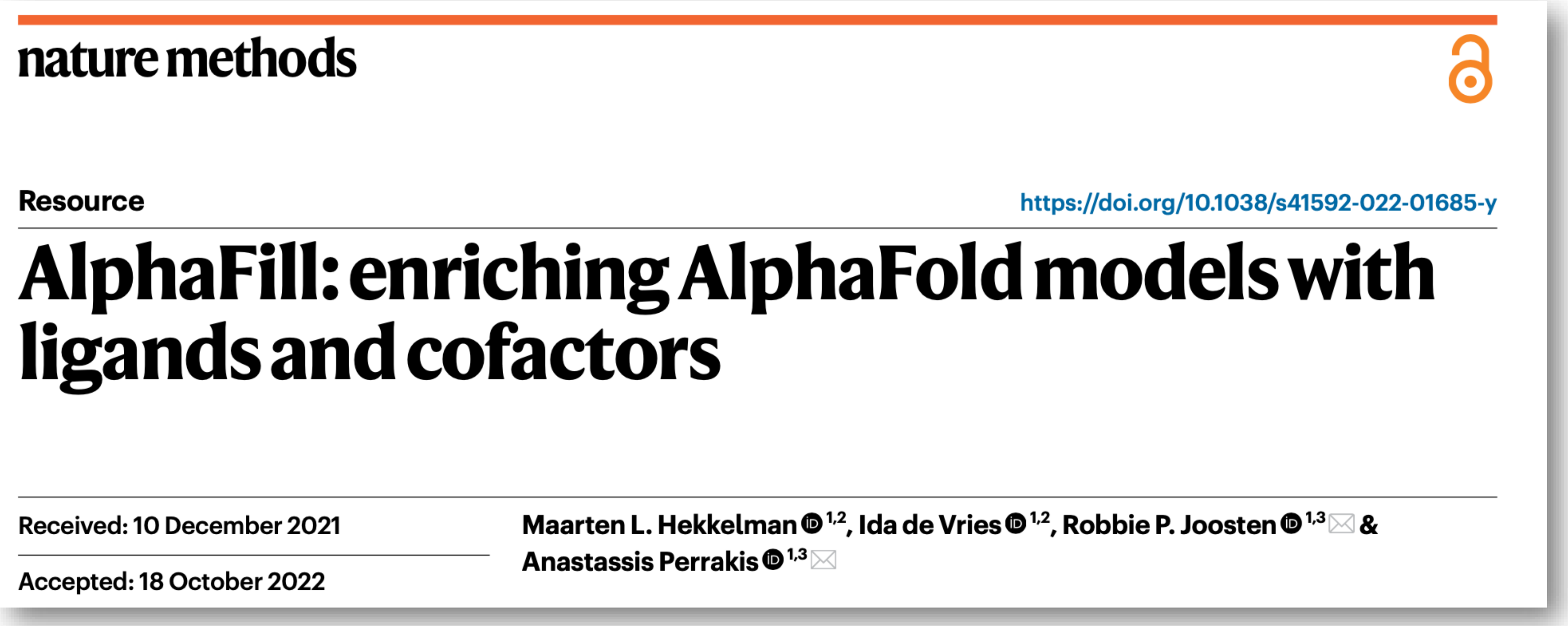
P29373

Cellular retinoic acid-binding protein 2

Structure file	https://alphafill.eu/v1/aff/P29373-F1
Metadata	https://alphafill.eu/v1/aff/P29373-F1/json
Original AlphaFold model	https://alphafold.ebi.ac.uk/entry/P29373

25% identity | 30% identity | 40% identity | 50% identity | 60% identity | 70% identity

Compound	PDBID	Global RMSd	Asym	Local RMSd	TCS	Show	
B3P	3fel.A	0.56	M	0.20	0.09	□	
			L	0.21	0.30	□	
NA	2g78.A	0.36	G	0.19	0.00	□	
	2fs6.A	0.37	C	0.17	0.32	□	
	2g79.A	0.46	I	0.04	0.00	□	
			J	0.08	0.00	□	
			H	0.09	0.00	□	
	2g7b.A	0.52	O	0.13	0.00	□	
			N	0.23	0.34	□	
	2frs.A	0.69	E	0.22	0.14	□	
			D	0.25	0.09	□	
			F	0.96 ?	0.32	□	
	3d97.B	0.94	P	0.04	0.00	□	
			Q	0.05	0.48	□	
REA	1cbs.A	0.36	B	0.52	0.21	□	
	1cbr.A	0.62	T	0.04	0.00	□	
RET	2g79.A	0.46	K	0.59	0.14	□	
	4i9s.A	0.76	R	0.78	0.99 ?	☑	optimise



nature methods

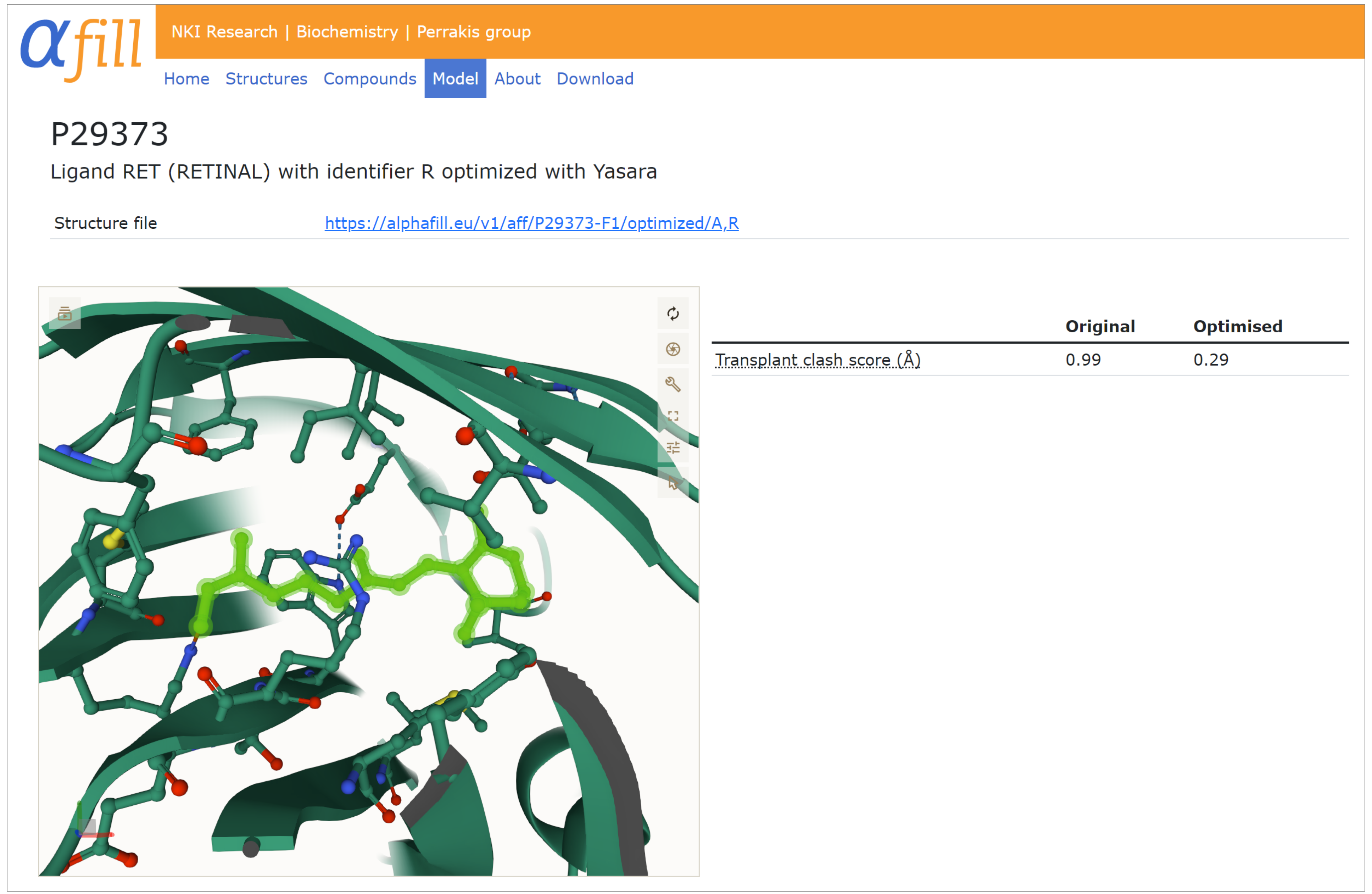
Resource https://doi.org/10.1038/s41592-022-01685-y

AlphaFill: enriching AlphaFold models with ligands and cofactors

Received: 10 December 2021

Accepted: 18 October 2022

Maarten L. Hekkelman [1,2], Ida de Vries [1,2], Robbie P. Joosten [1,3] & Anastassis Perrakis [1,3]



P29373

Ligand RET (RETINAL) with identifier R optimized with Yasara

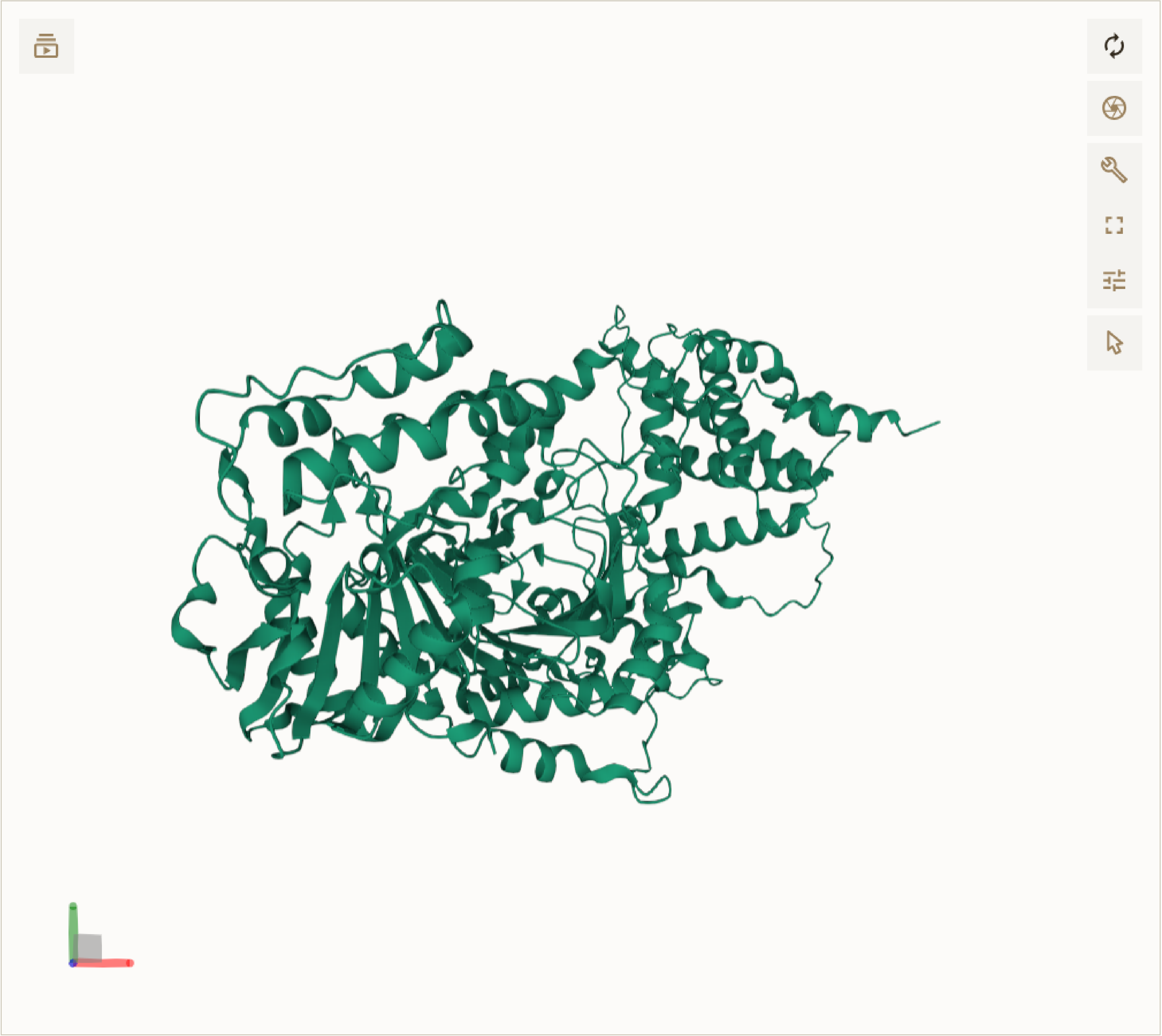
Structure file	https://alphafill.eu/v1/aff/P29373-F1/optimized/A,R

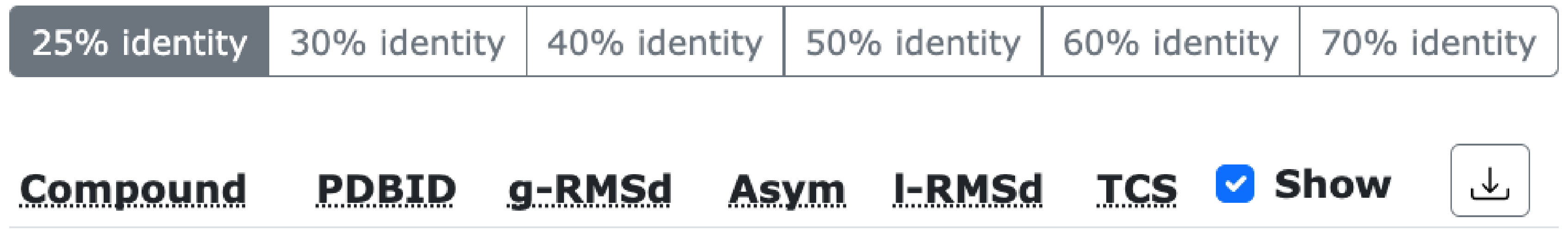
	Original	Optimised
Transplant clash score (Å)	0.99	0.29

Does JBP1 now have its Fe?

No.

- AlphaFill depends on the existence of homologous structures.





25% identity | 30% identity | 40% identity | 50% identity | 60% identity | 70% identity

Compound	PDBID	g-RMSd	Asym	l-RMSd	TCS	☑ Show

AlphaFill is based on homologous structures

Can we go beyond that?

- We have thousand of ligand-binding events in experimental structures.
- And AlphaFill offers even more such reliable observations.
- Can we train an AI to learn what defines binding of specific ligands?
- … at least for an easy case, e.g. for metals
  - no rotational and conformational freedom
- … and while doing that think of what defines ligand binding?

How should we describe a protein to an AI?

Why not as an image?

Carbon

Nitrogen in histidine

Positively charged nitrogen atoms

Other nitrogen atoms

Sulfur in methionine

Hydroxylic oxygen

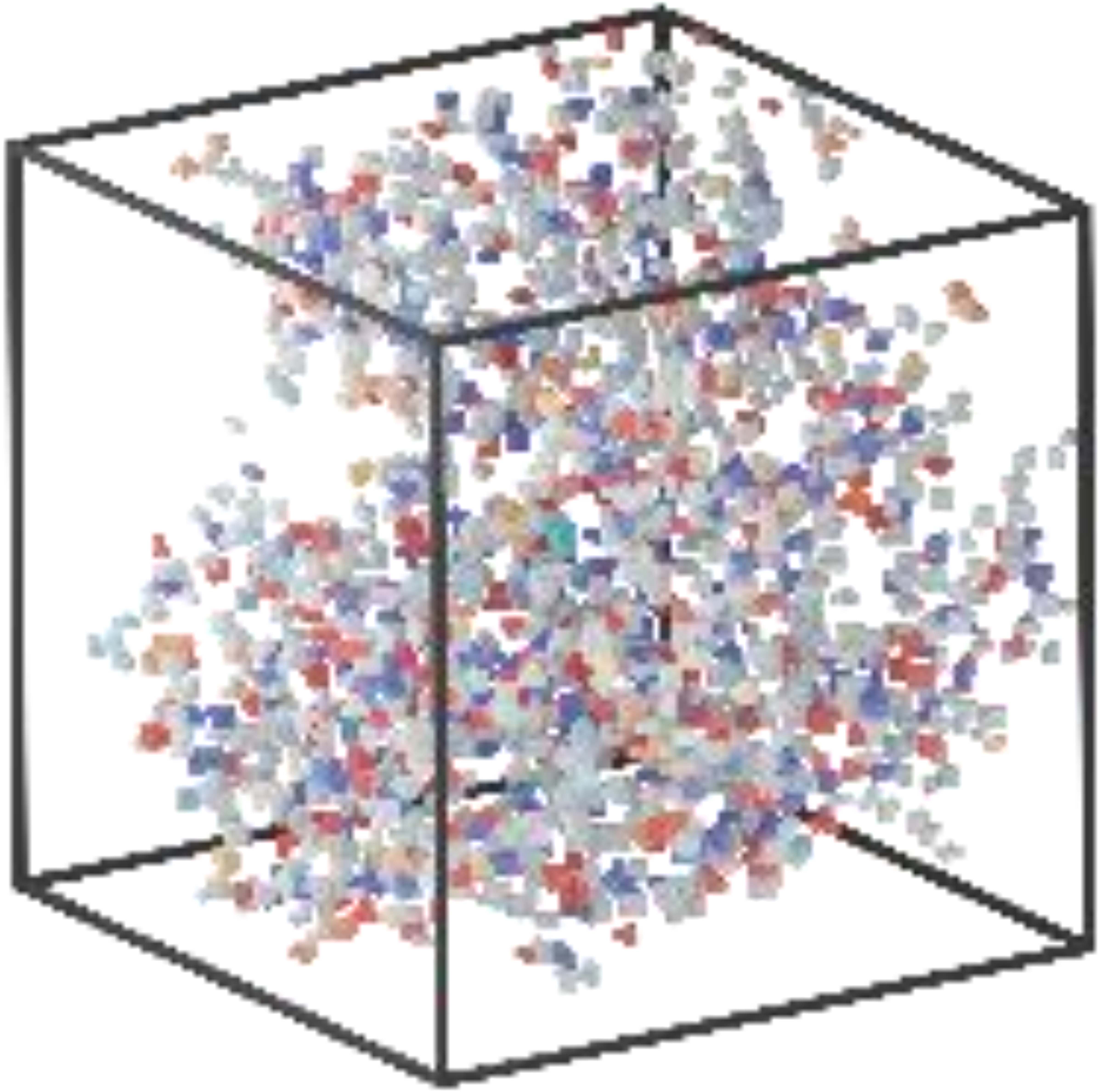
Carboxylic oxygen

Carbonyl oxygen

Sulfur in cysteine

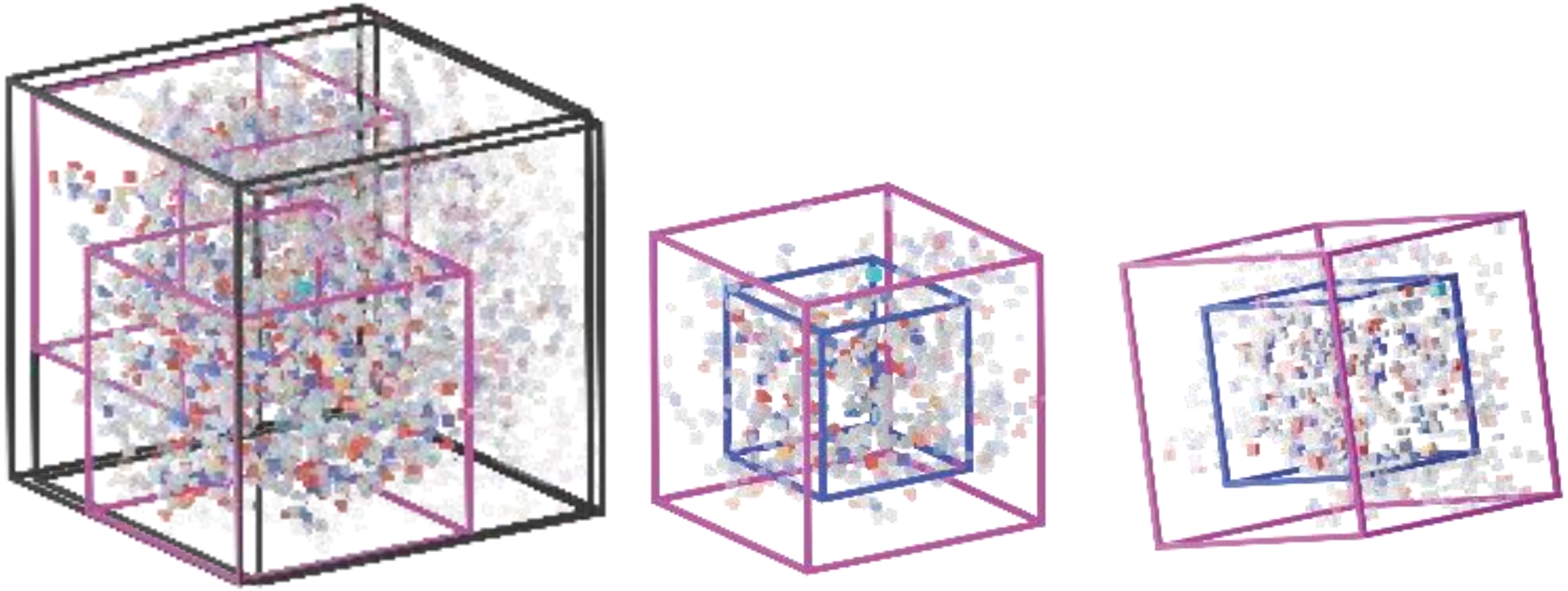
NH
N
R
HN
HN
+
NH
R
H_3N^+–R
O
R
H_2N
NH
R
R
NH
H–S
R
H–O
R
O
O
–
R
O
R
R
H_3C–S
R
O
NH_2
H
N
R_1
R_2
O

A whole protein as a 3D “color”-image



Tricks and tips to learn in 3D

Pretty standard in image recognition AI



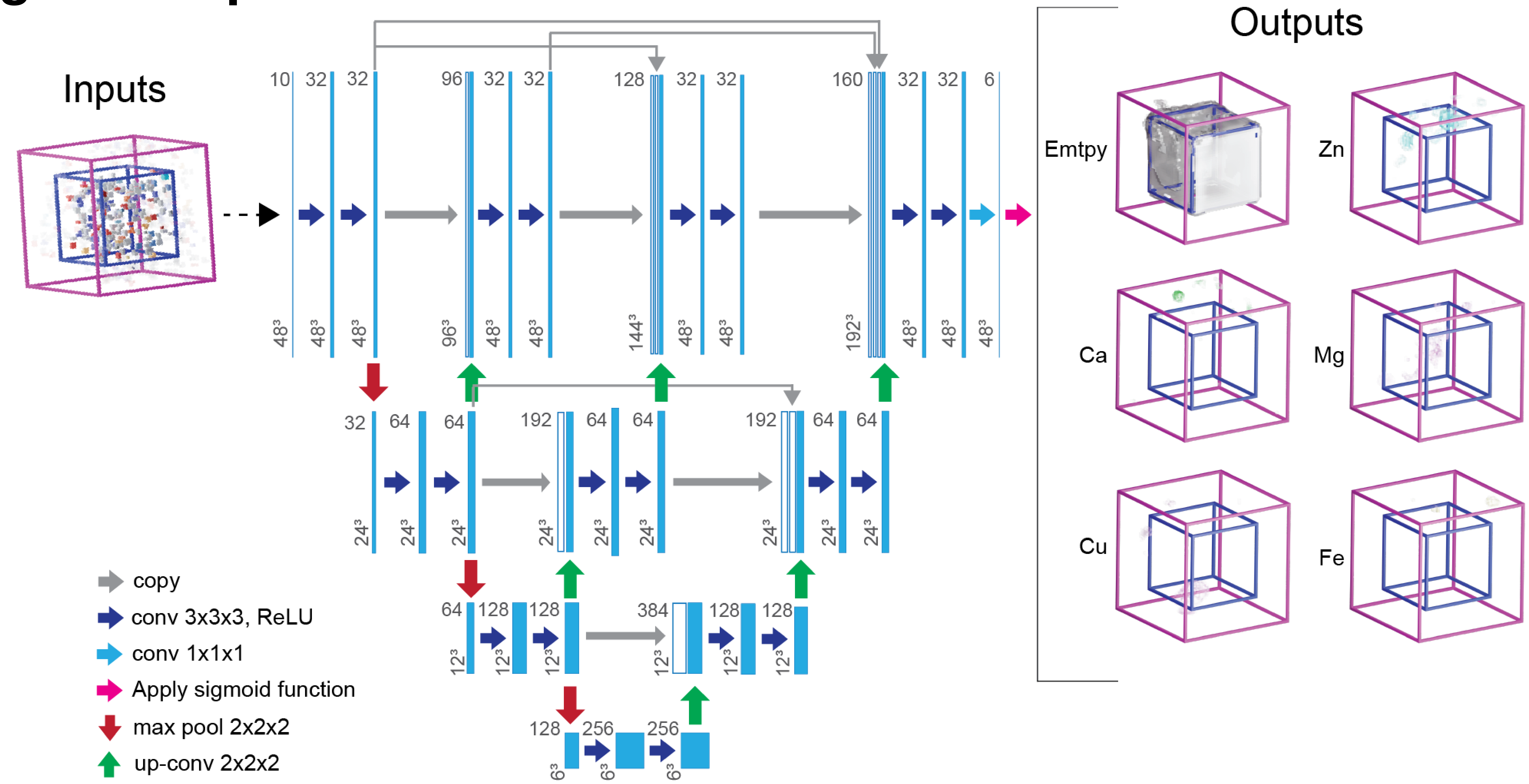
Container Cube

Sampling Cube

Learning Cube

Learning to recognize metals in protein images

Using an adapted UNET++

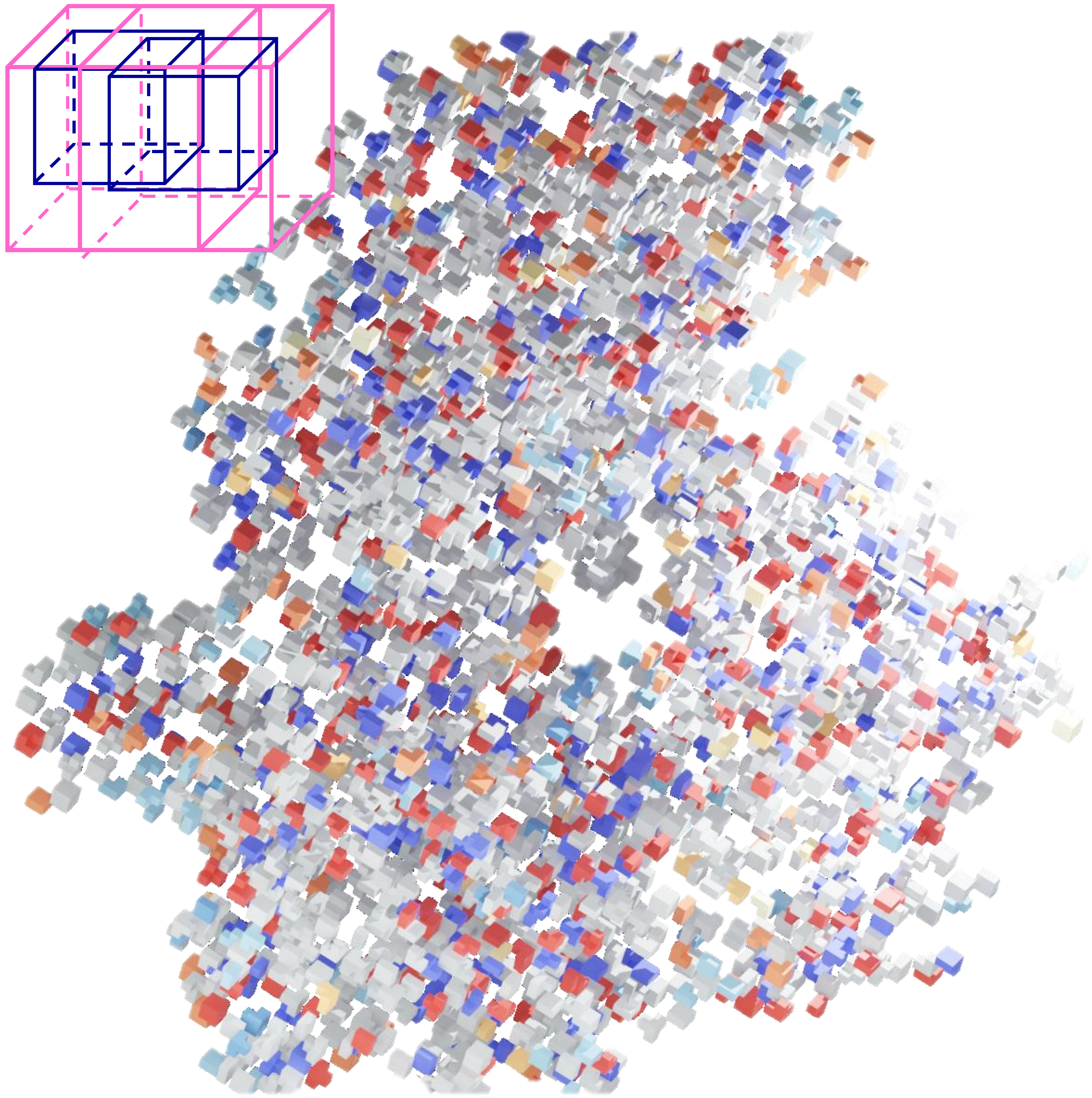


- copy
- conv 3x3x3, ReLU
- conv 1x1x1
- Apply sigmoid function
- max pool 2x2x2
- up-conv 2x2x2

Metal prediction using AI

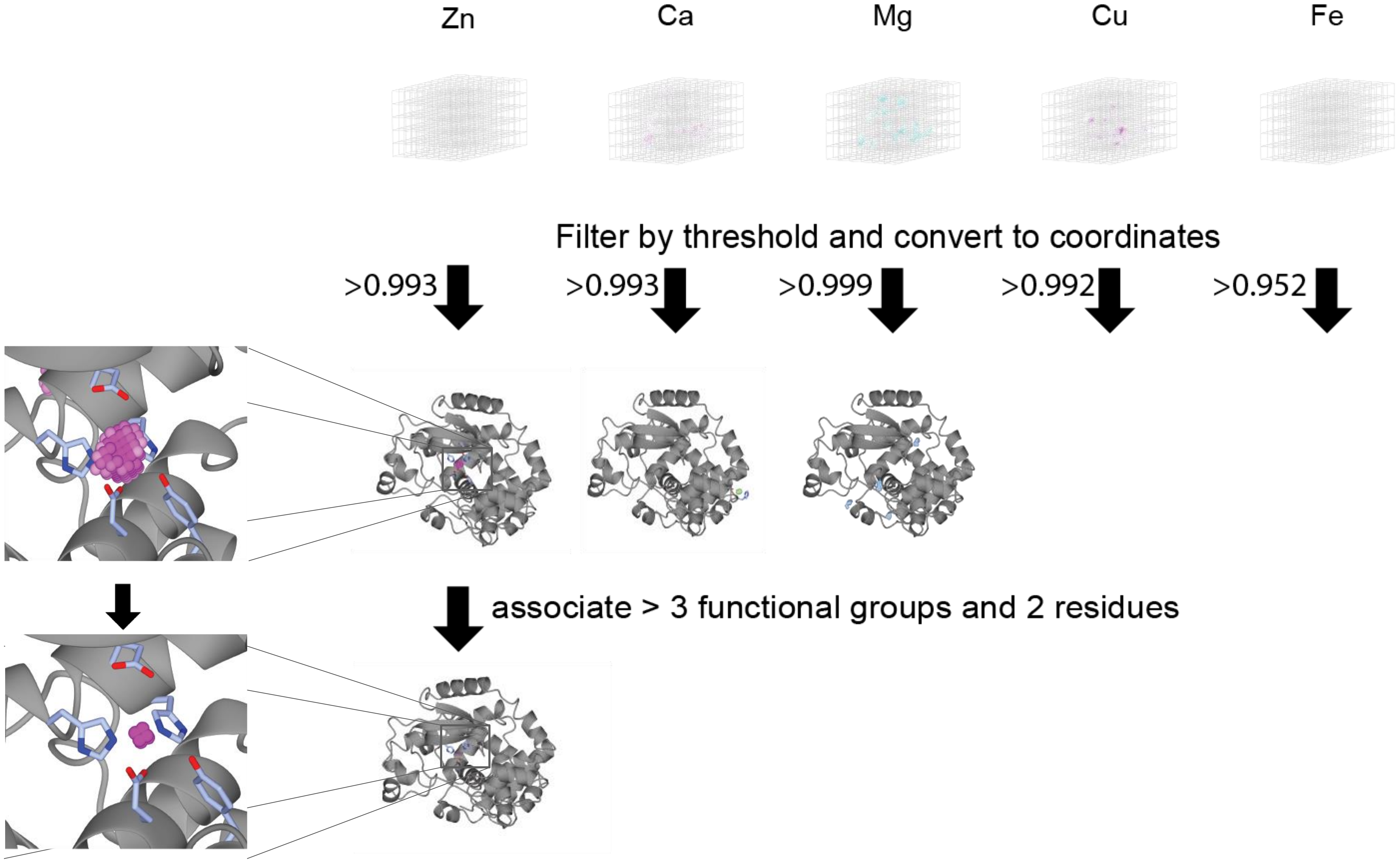
Scanning the target

- Start with simplest case: metal ions
  - Most common: Zn, Mg, Ca, Fe, Cu
- Provide chemistry to the computer
- Learn what a metal binding site looks like
- **Find metal binding sites in proteins**



Post-processing by probability

Filter results by probabilities and surroundings

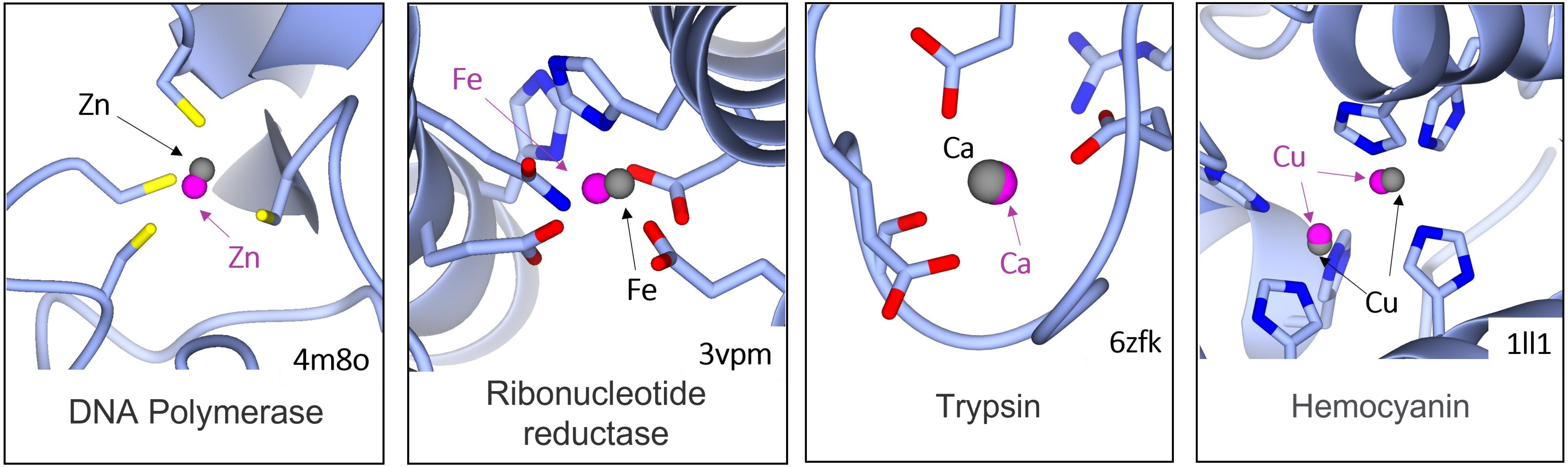


Filter by threshold and convert to coordinates

associate > 3 functional groups and 2 residues

Finding Metals

In PDB structures of the “left-out” validation set



DNA Polymerase

Ribonucleotide reductase

Trypsin

Hemocyanin

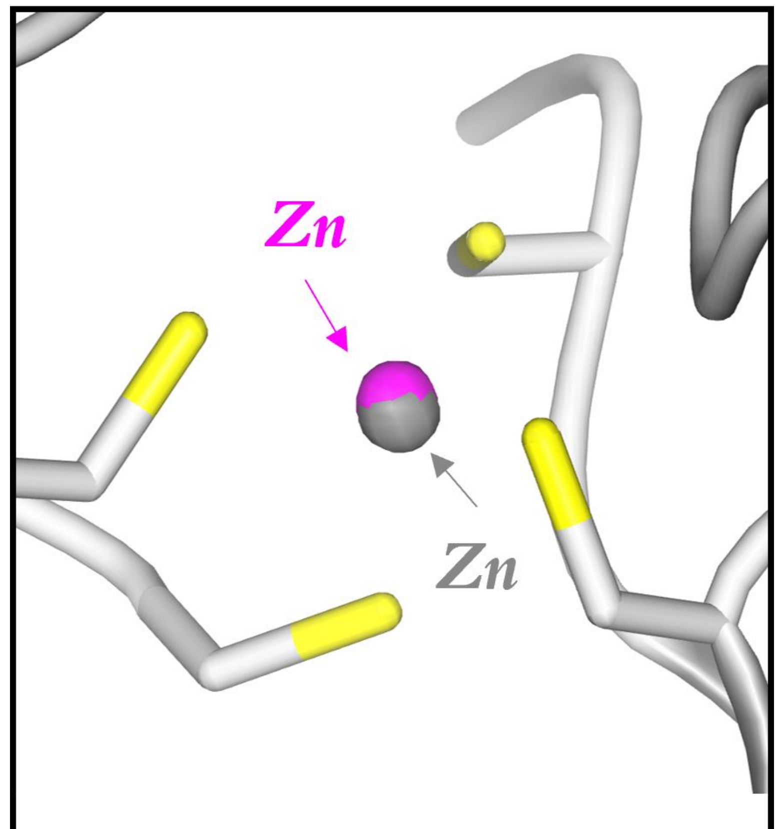
● Metal ion in PDB-REDO

● Metal ion predicted by neural network

Overall performance

	Zn	Ca	Mg	Cu	Fe
Known metal binding sites	27	13	19	20	12
Re-discovered	26	12	16	19	10
Not discovered	1	1	3	1	2
Newly discovered	5	4	14	2	3

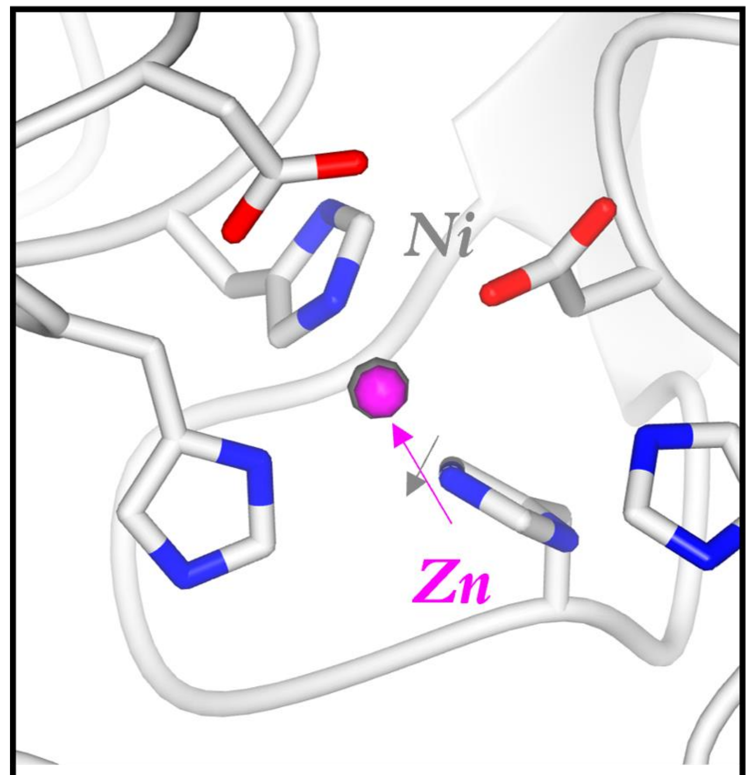
Re-discovered



DNA Polymerase

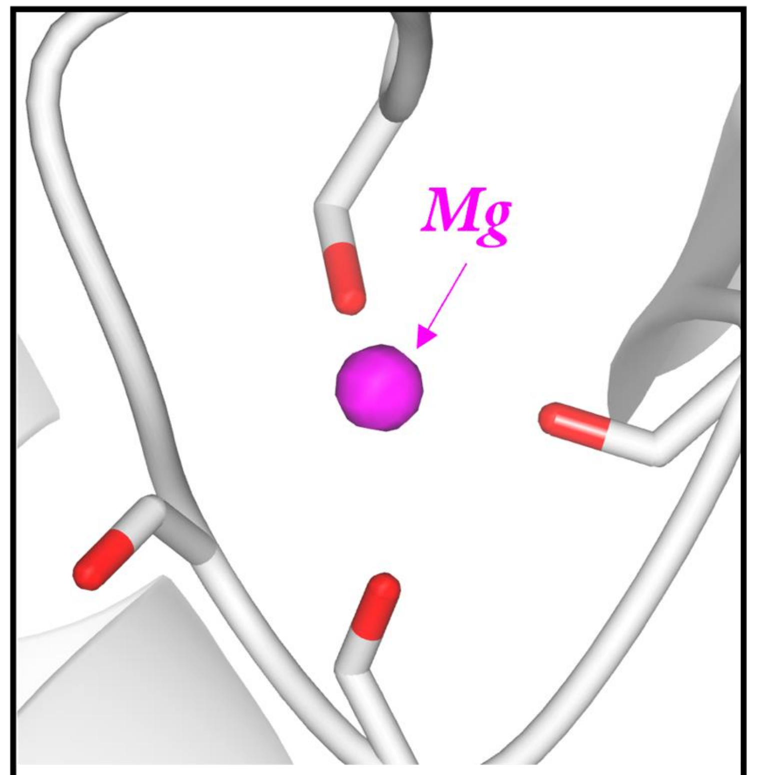
Newly discovered

Miss-identified



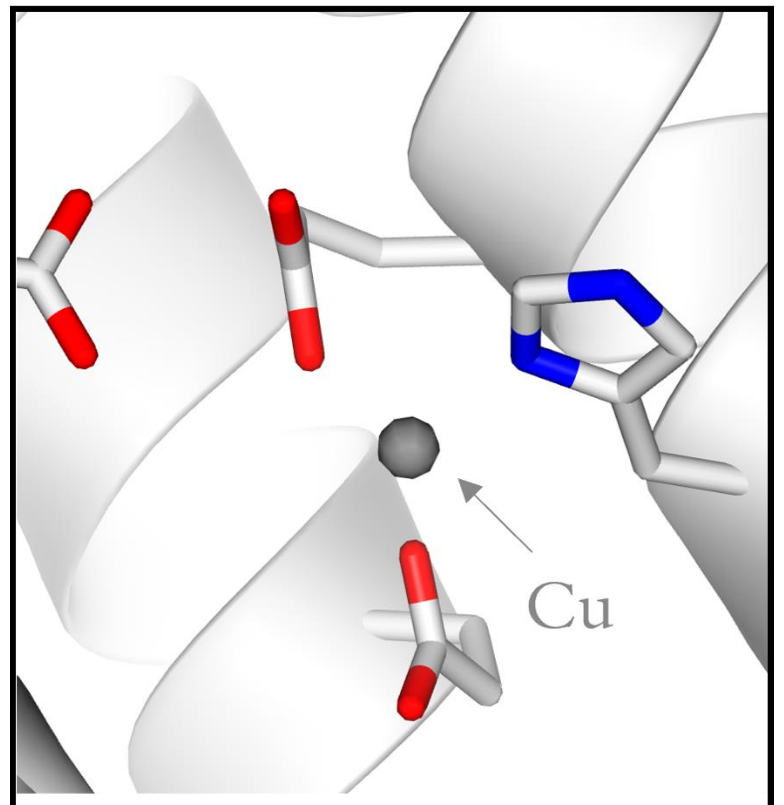
Oxidase

False positive



5’ Exonuclease Apollo

Not discovered



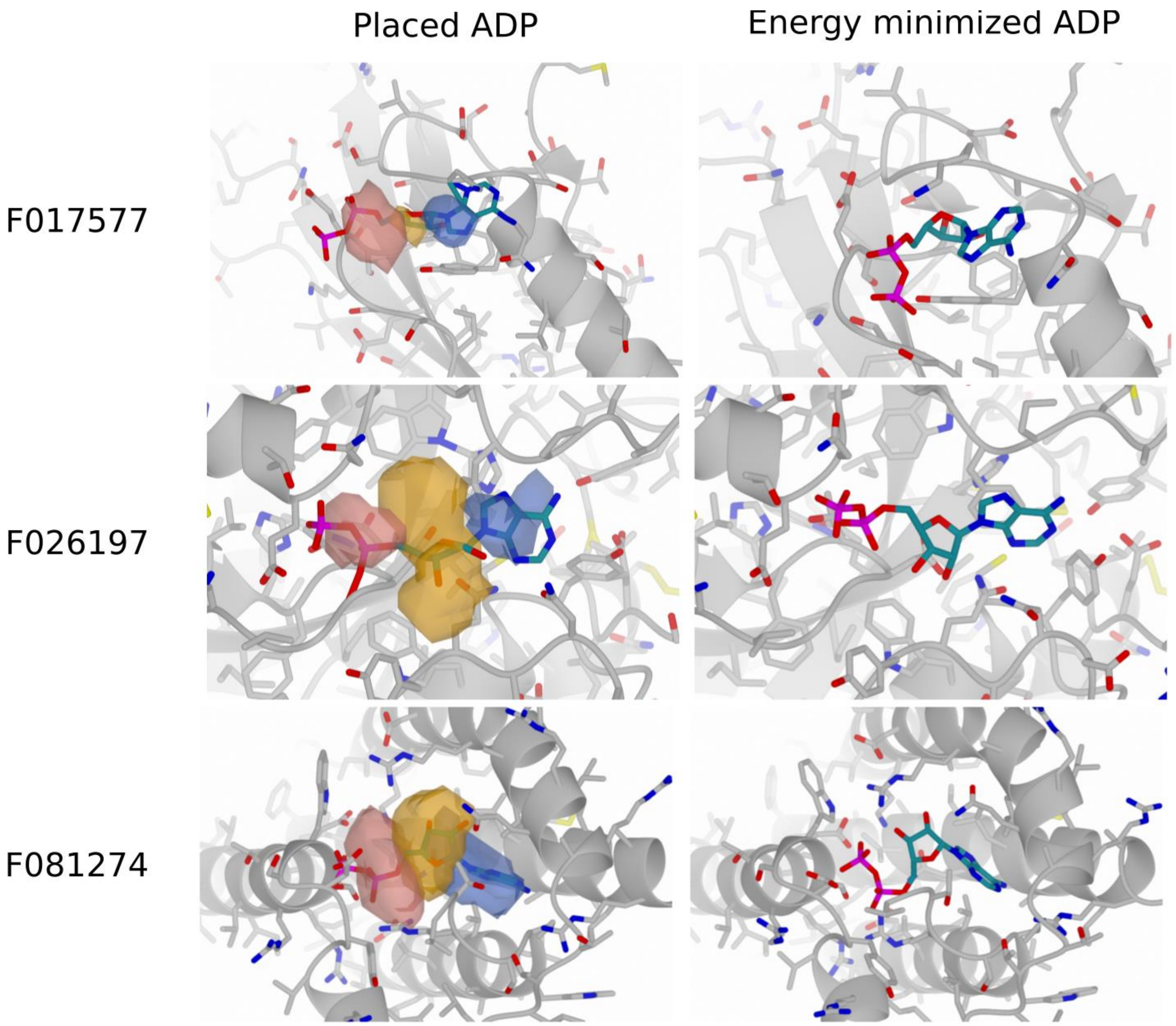
Ferritin

● Metal ion in PDB-REDO

● Metal ion predicted by neural network

Finding more complex blocks

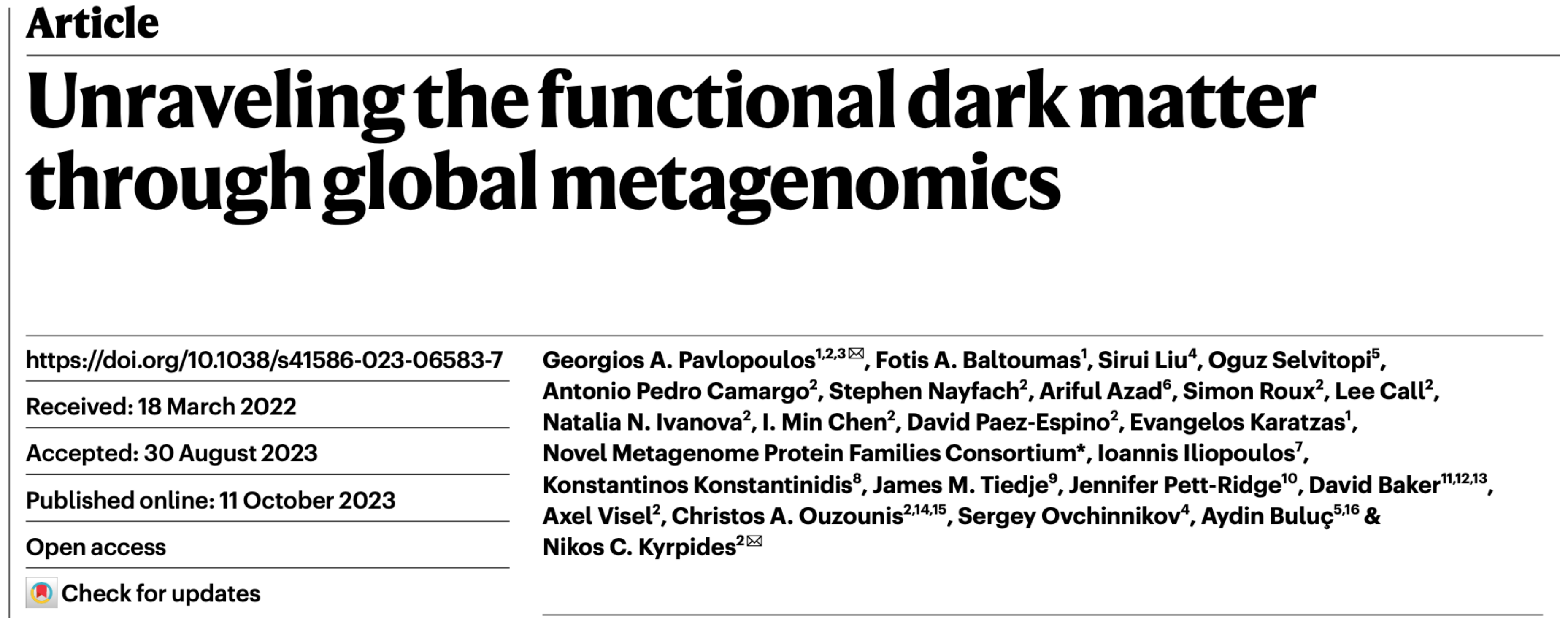
Phosphates, sugars, nucleobases (to build nucleotide)



	Nucleotide prediction
Total tested	1365
Re-discovered	710
Not discovered	655
Newly discovered	157

Can we find new things?

Test in metagenome database of new folds



Article

Unraveling the functional dark matter through global metagenomics

https://doi.org/10.1038/s41586-023-06583-7

Received: 18 March 2022

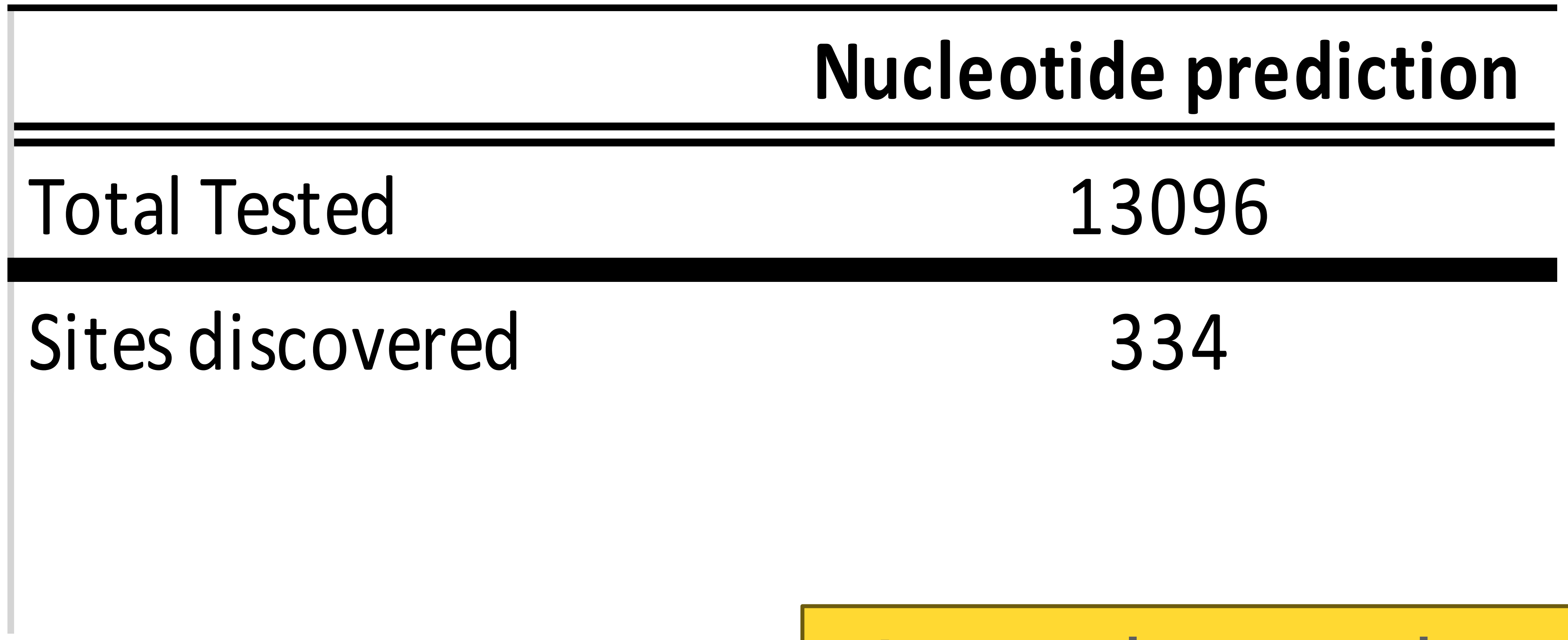
Accepted: 30 August 2023

Published online: 11 October 2023

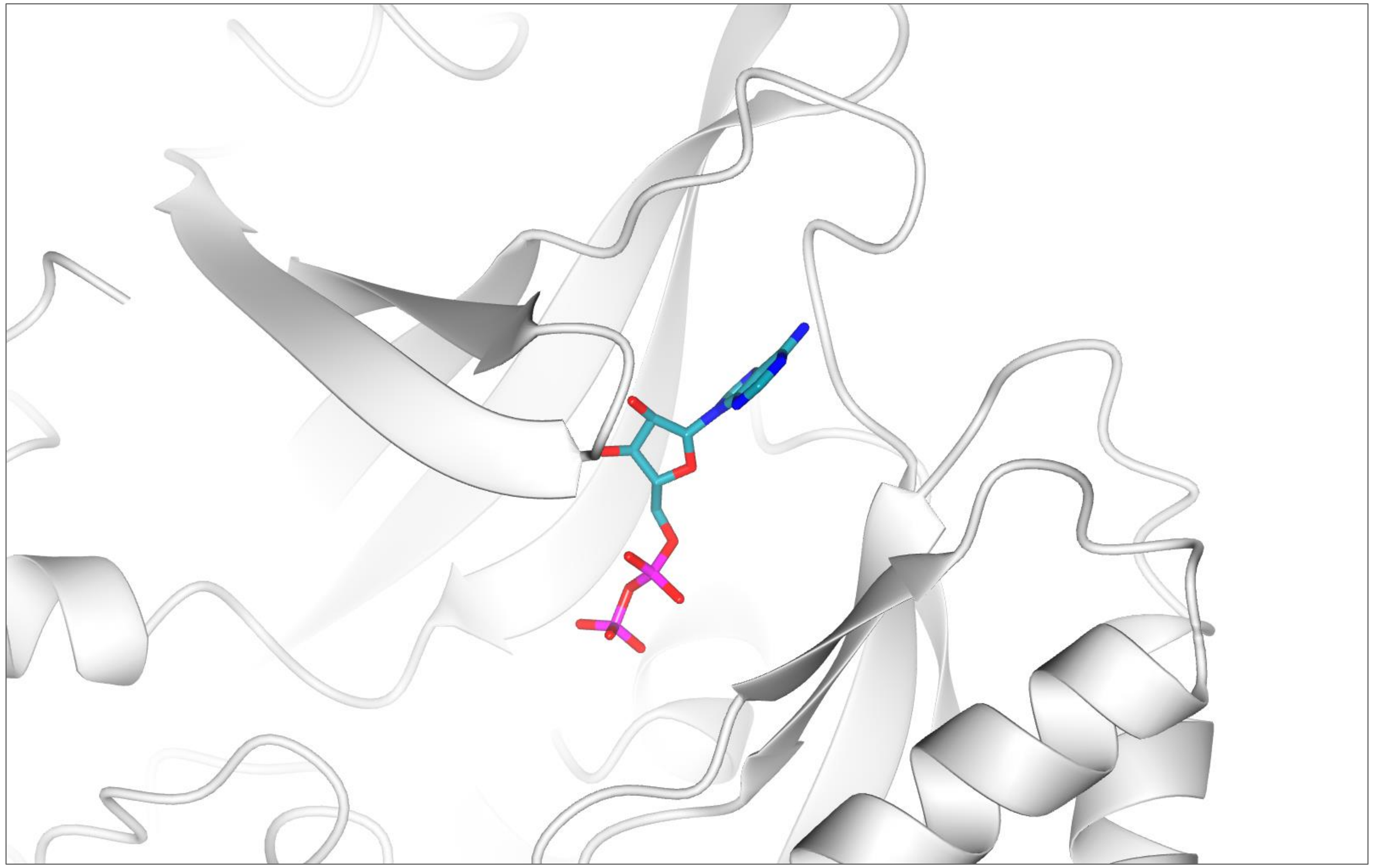
Open access

Check for updates

Georgios A. Pavlopoulos[1,2,3]✉, Fotis A. Baltoumas[1], Sirui Liu[4], Oguz Selvitopi[5], Antonio Pedro Camargo[2], Stephen Nayfach[2], Ariful Azad[6], Simon Roux[2], Lee Call[2], Natalia N. Ivanova[2], I. Min Chen[2], David Paez-Espino[2], Evangelos Karatzas[1], Novel Metagenome Protein Families Consortium*, Ioannis Iliopoulos[7], Konstantinos Konstantinidis[8], James M. Tiedje[9], Jennifer Pett-Ridge[10], David Baker[11,12,13], Axel Visel[2], Christos A. Ouzounis[2,14,15], Sergey Ovchinnikov[4], Aydin Buluç[5,16] & Nikos C. Kyrpides[2]✉



	Nucleotide prediction
Total Tested	13096
Sites discovered	334

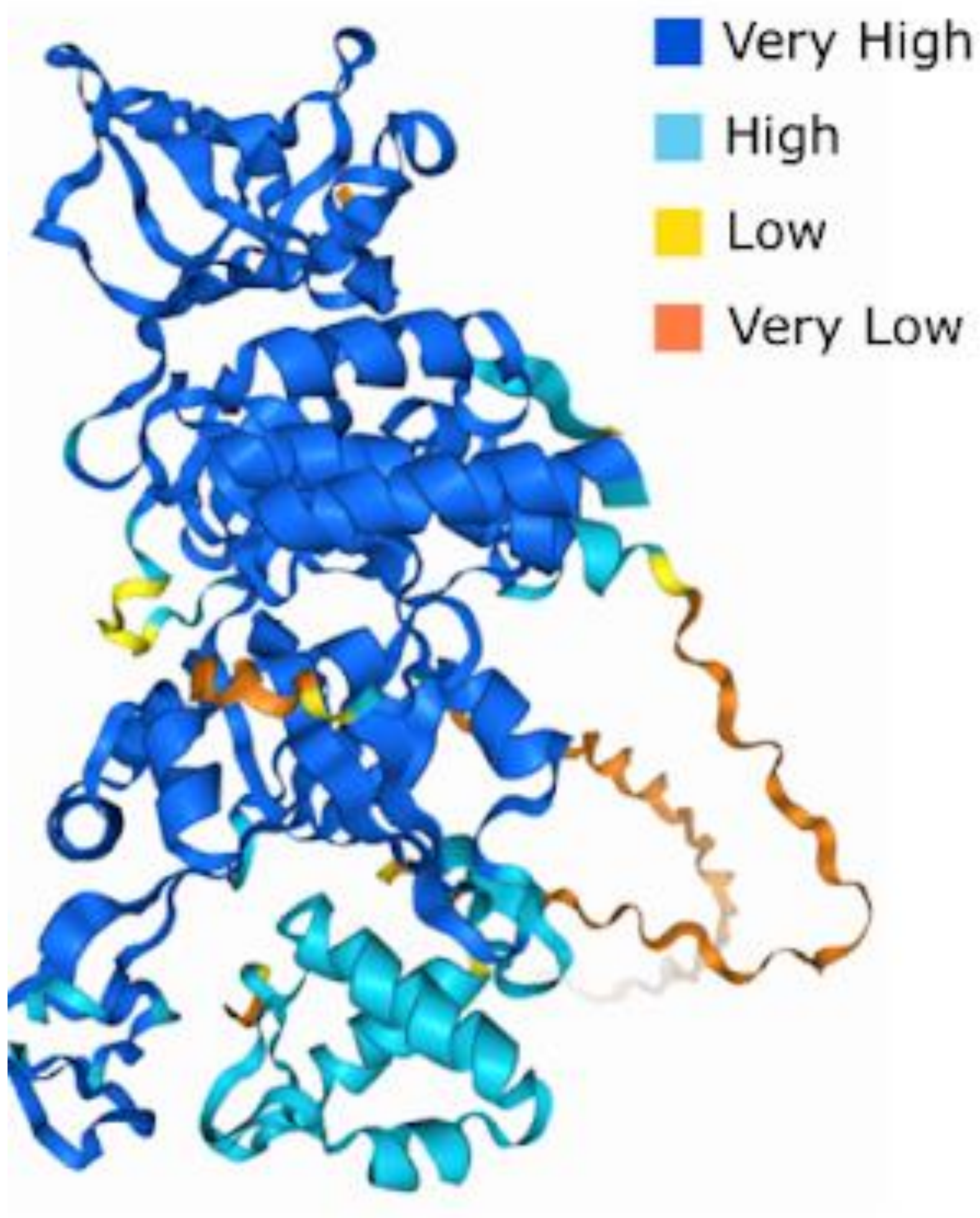


4 proteins produced, testing now ATP/ADP binding

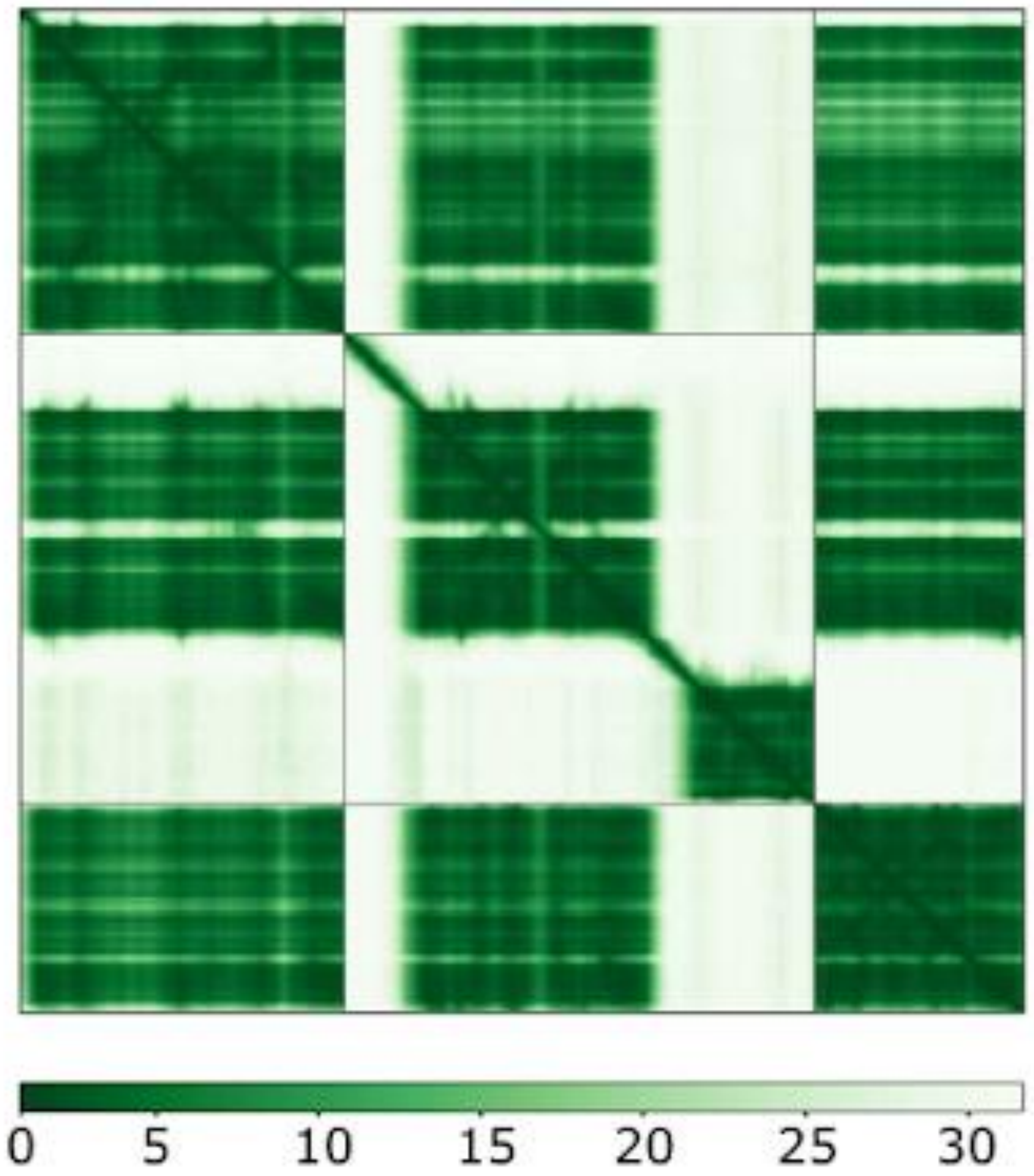
TAMRA-ADP binds in all 4

What you see in the AlphaFold Server

pLDDT



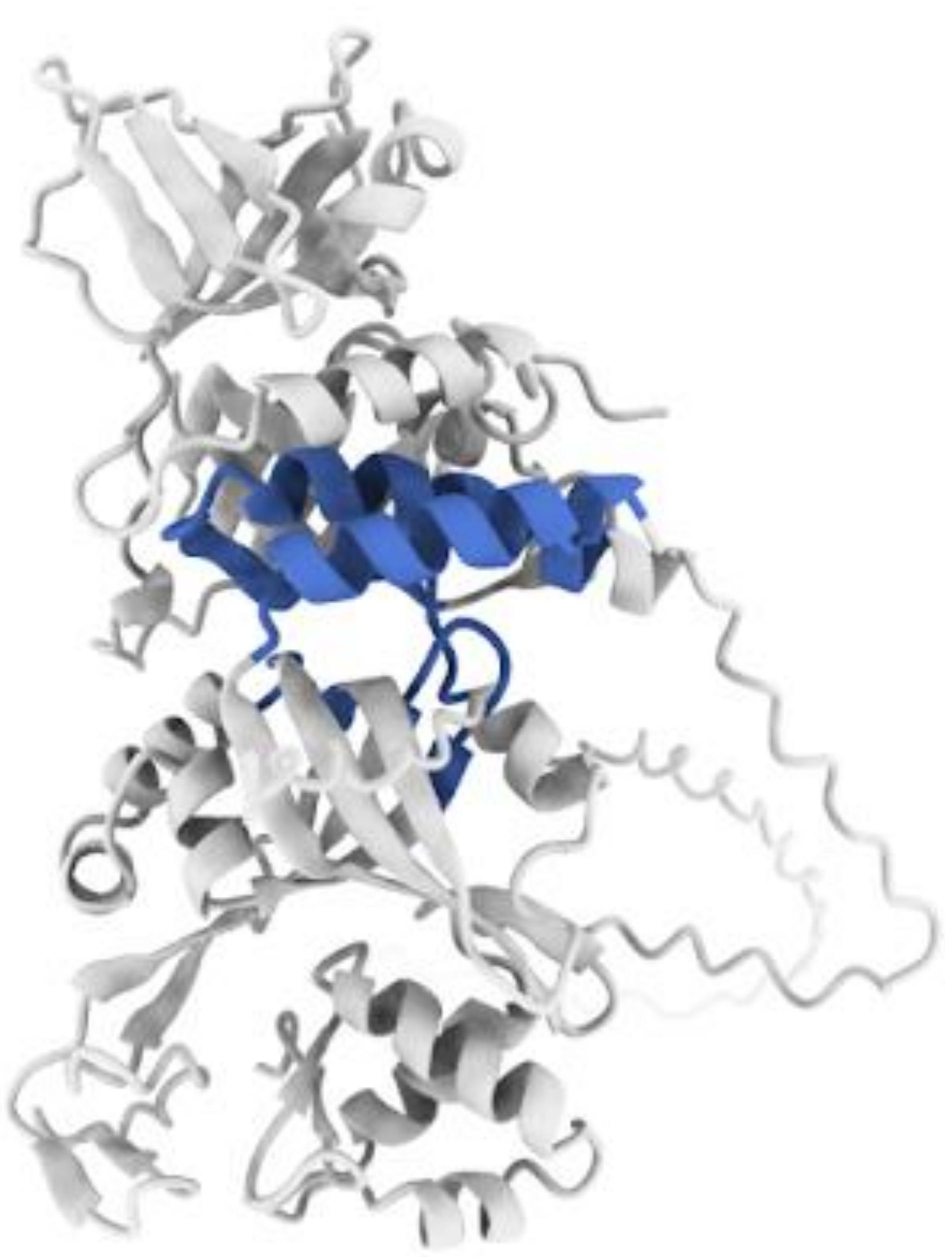
PAE



Plotting interfaces in 2D for easy inspection

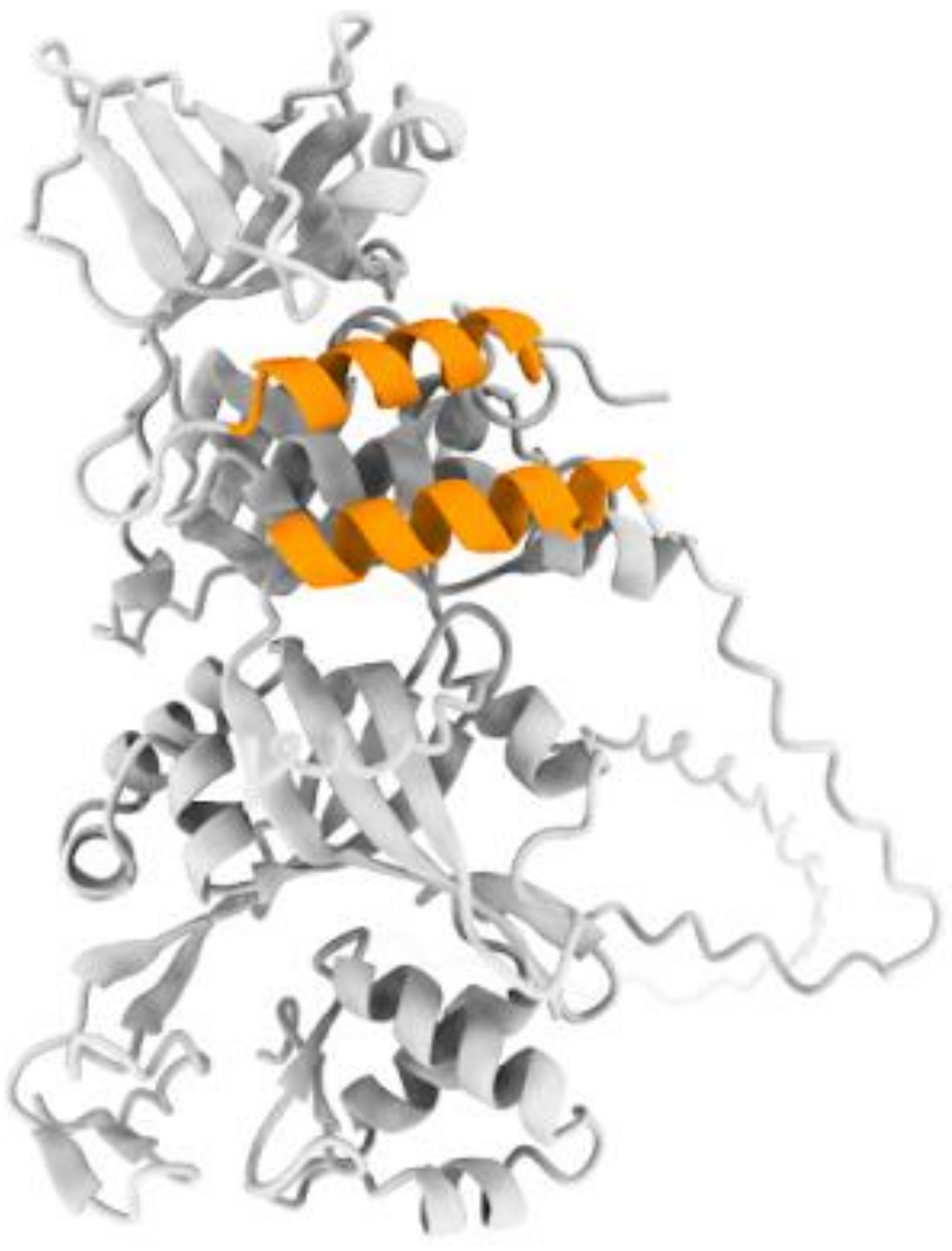
Interface 1

RFA1 - RFA2



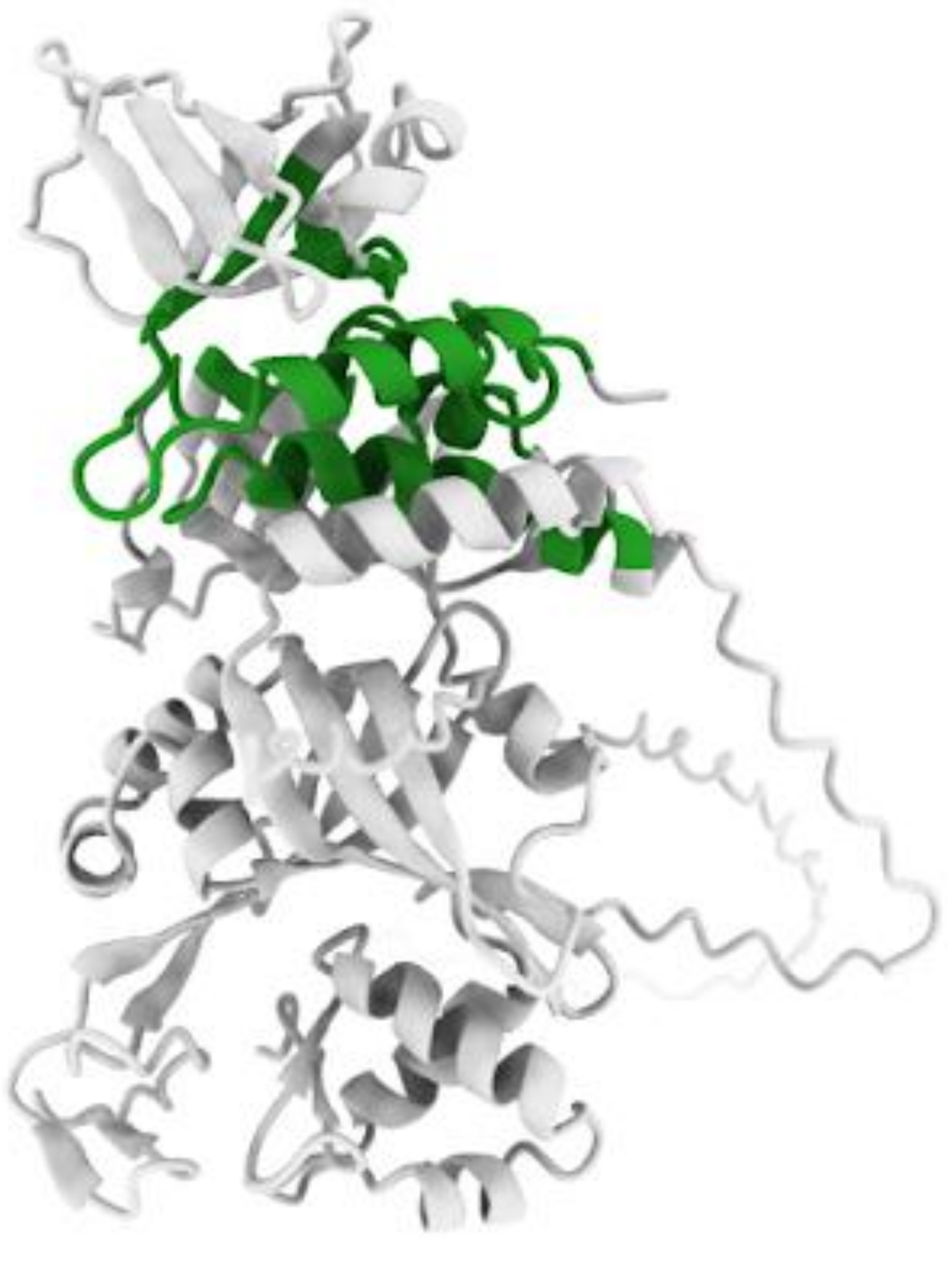
Interface 2

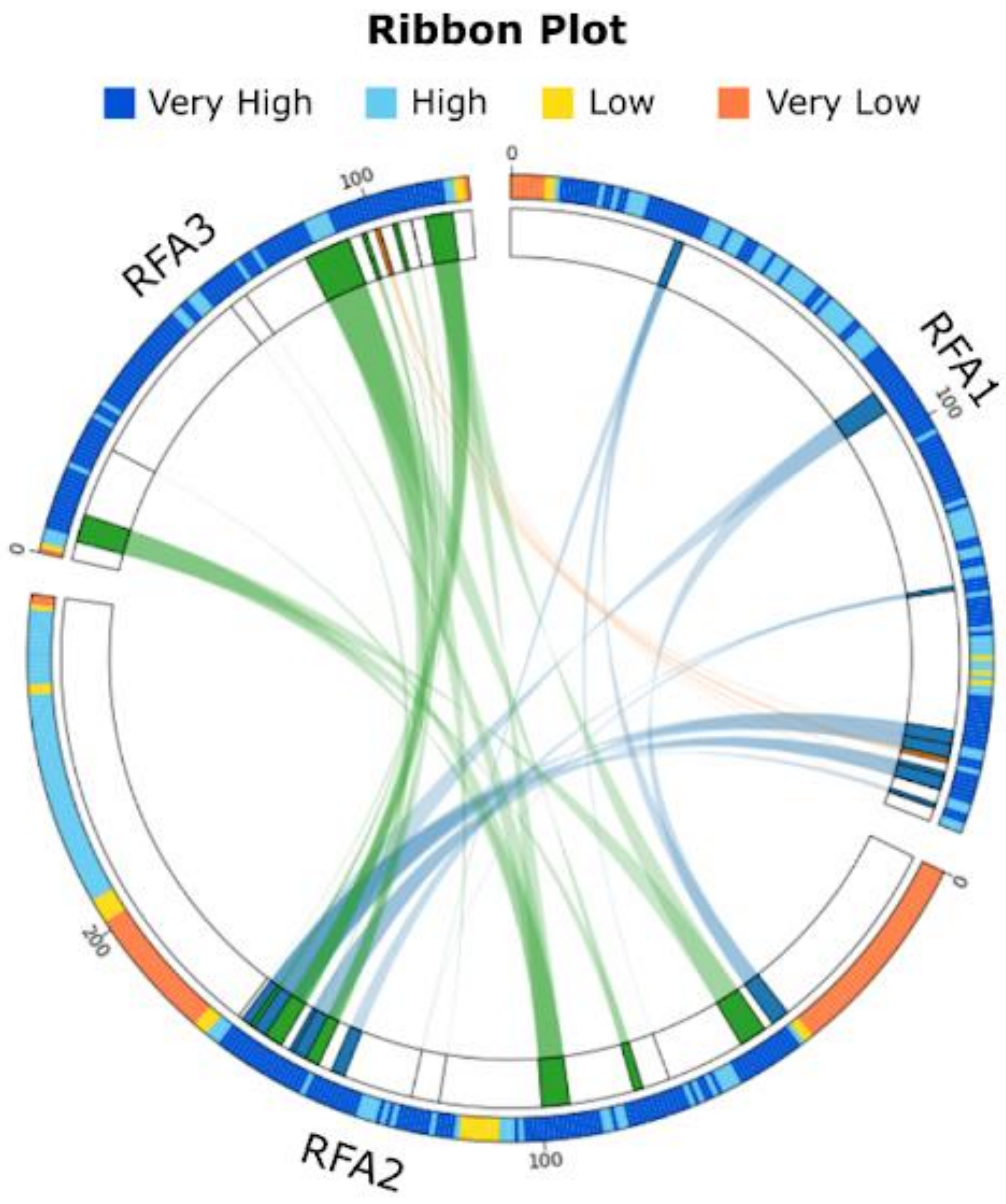
RFA1 - RFA3



Interface 3

RFA2 - RFA3





Alphabridge
Home
A
20
40
60
80
100
120
140
160
180
200
220
240
260
280
300
320
340
360
380
B
C
D

Found 8 interfaces