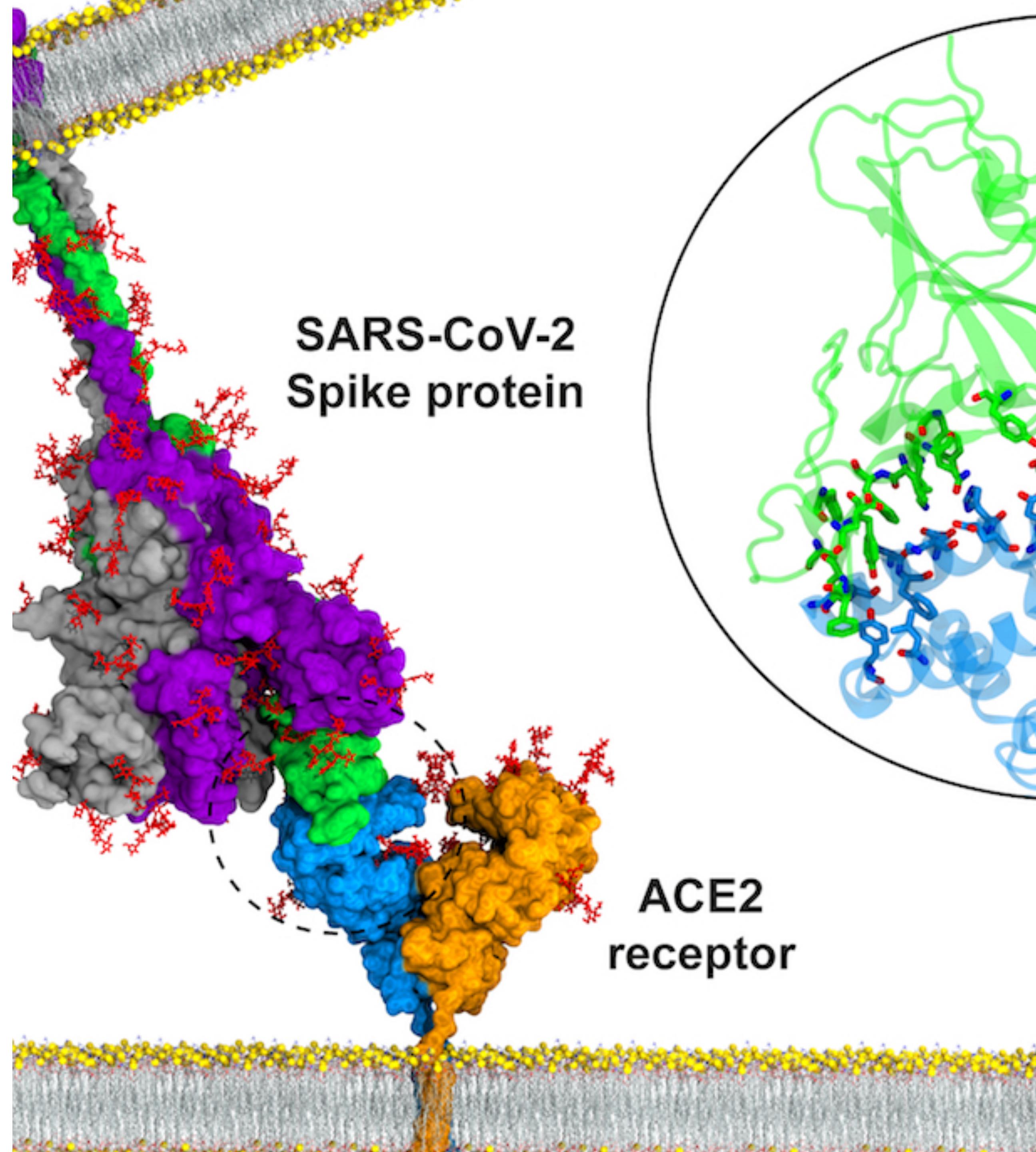


Structures predictions and
molecular docking

Structural Bioinformatics

Outline

- Structure prediction: the origins
- Structure prediction: key advances
- State of the art and AI approaches
- Protein Design
- Challenges: conformational changes, effect of mutations, binding, quaternary structure



Motivations

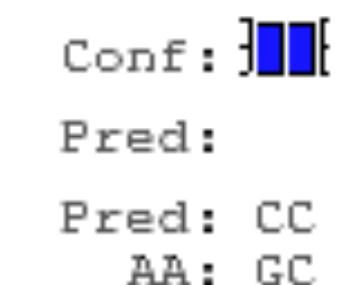
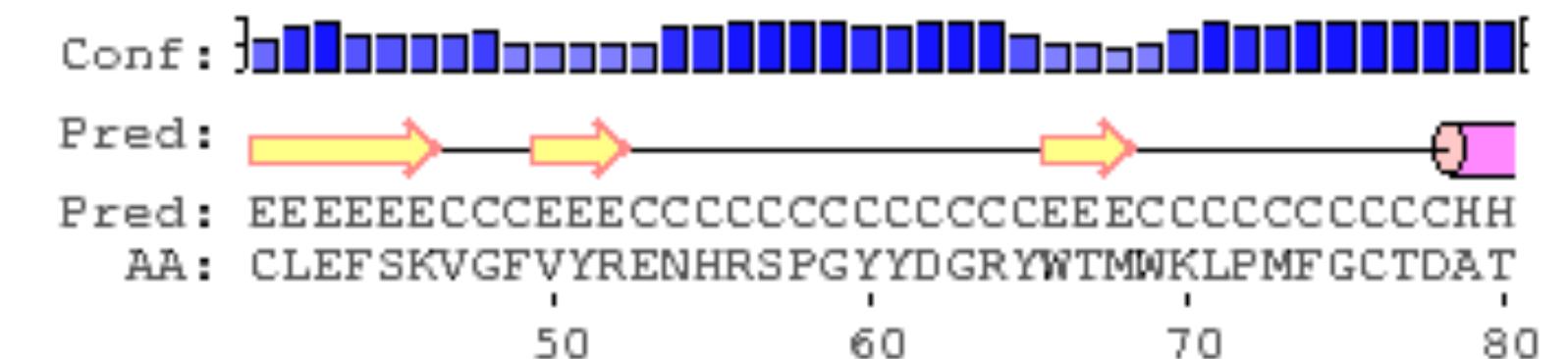
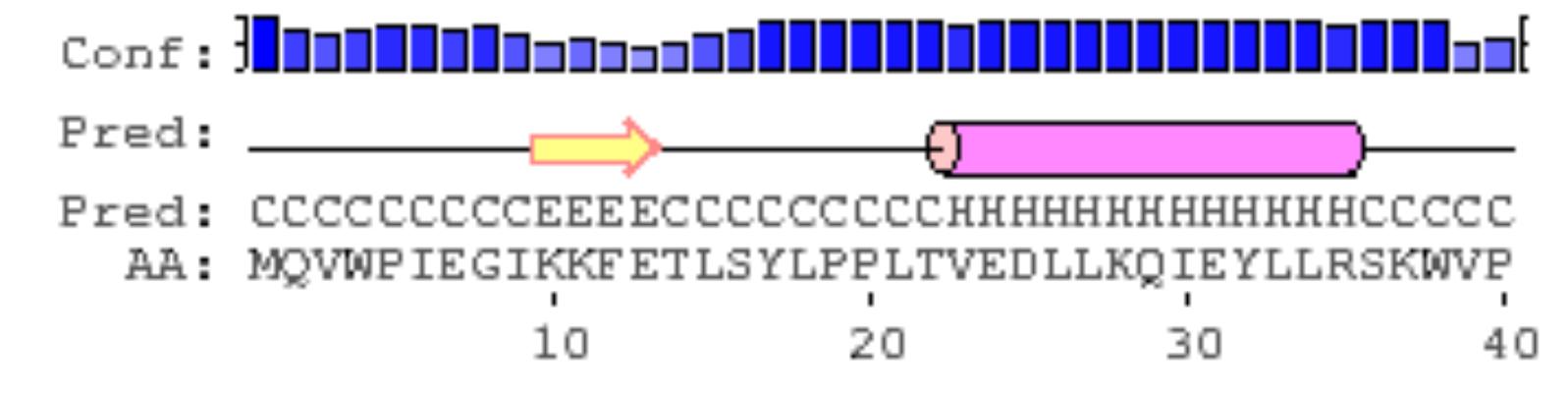
- * High-throughput sequencing technology have greatly exacerbated the gap between the number of known sequences ($\sim 10^8$) and the number of experimentally determined protein structures ($\sim 10^5$).
- * Experimental structure determination is a long and expensive process (1+ year and k\$ per structure)
- * Protein structures help drug design and development
- * Protein structures are used to understand biological functions and dysfunctions (mechanisms, effect of mutations, post translational modifications, interactions with other partners, ...)
- * Protein structures are used to understand evolution (from long times scales to viruses)
- * ...



Let's visualise the structure prediction problems

Secondary Structure:

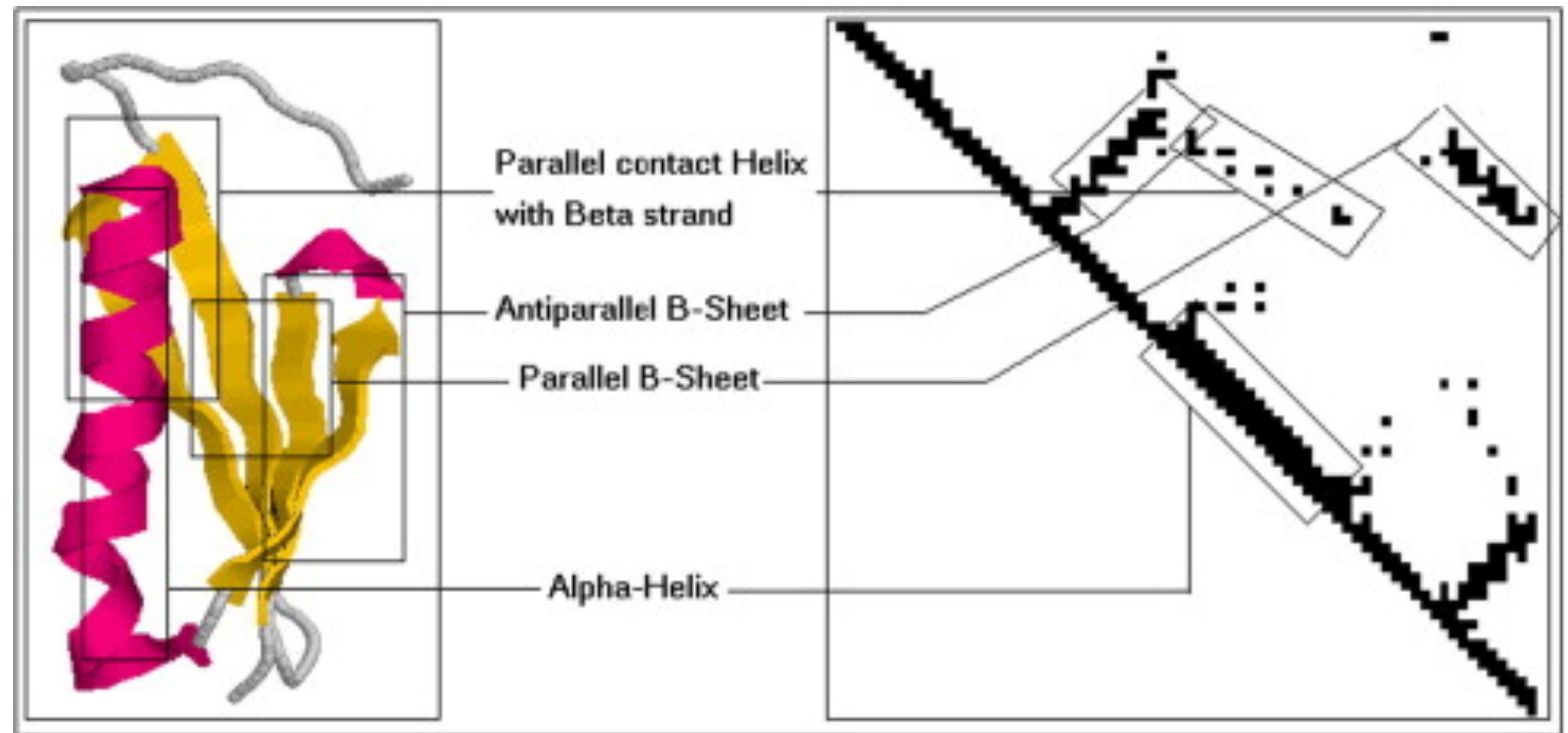
From a sequence of letters (protein sequence) to
a new sequence of letters (secondary structure
assignment)



Let's visualise the structure prediction problems

Contact prediction:

From a sequence of N letters (protein sequence)
to a matrix NxN with 1 and 0s:



Let's visualise the structure prediction problems

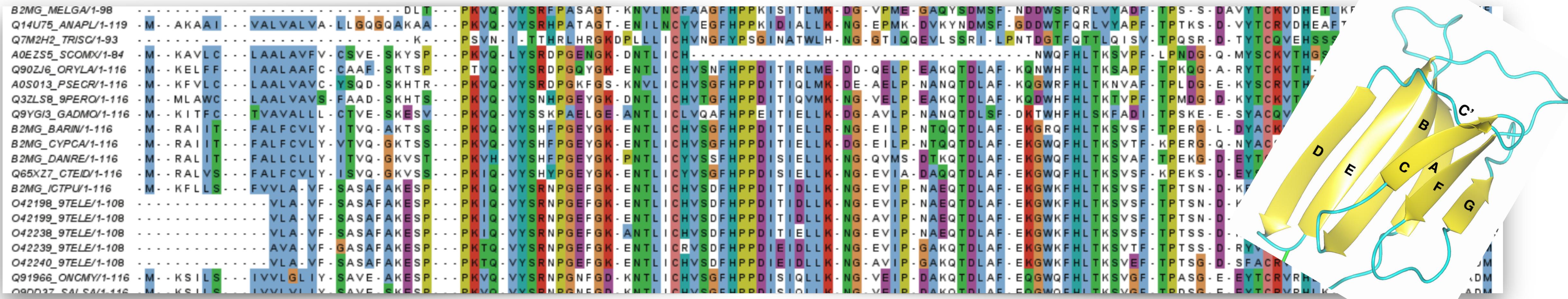
Structure prediction:

>1PGB_1|Chain A|PROTEIN G|Streptococcus sp. GX7805 (1325)
MTYKLILNGKTLKGETTTEAVDAATAEKVFKQYANDNGVDGEWTYDDATKTFTVTE

From a sequence of N
letters (protein sequence)
to a table of $3 \times N_{\text{atoms}}$
numbers (x,y,z)
coordinates of each atom

ATOM	1	N	MET	A	1	12.969	18.506	30.954	1.00	15.93	N
ATOM	2	CA	MET	A	1	13.935	18.529	29.843	1.00	17.40	C
ATOM	3	C	MET	A	1	13.138	18.692	28.517	1.00	14.65	C
ATOM	4	O	MET	A	1	12.007	18.222	28.397	1.00	13.04	C
ATOM	5	CB	MET	A	1	14.733	17.216	29.882	1.00	20.72	C
ATOM	6	CG	MET	A	1	15.742	16.983	28.738	1.00	23.81	C
ATOM	7	SD	MET	A	1	17.378	17.025	29.359	1.00	28.11	C
ATOM	8	CE	MET	A	1	17.166	16.055	30.819	1.00	27.51	C
ATOM	9	N	THR	A	2	13.719	19.413	27.573	1.00	12.63	C
ATOM	10	CA	THR	A	2	13.088	19.661	26.283	1.00	12.68	C
ATOM	11	C	THR	A	2	13.561	18.631	25.300	1.00	12.02	C
ATOM	12	O	THR	A	2	14.763	18.432	25.121	1.00	13.07	C
ATOM	13	CB	THR	A	2	13.527	20.980	25.667	1.00	14.62	C
ATOM	14	OG1	THR	A	2	13.307	22.020	26.627	1.00	15.31	C
ATOM	15	CG2	THR	A	2	12.704	21.284	24.409	1.00	14.47	C
ATOM	16	N	TYR	A	3	12.574	18.048	24.642	1.00	11.17	C
ATOM	17	CA	TYR	A	3	12.726	17.033	23.612	1.00	10.11	C
ATOM	18	C	TYR	A	3	12.109	17.637	22.316	1.00	10.52	C
ATOM	19	O	TYR	A	3	11.165	18.449	22.364	1.00	9.38	C
ATOM	20	CB	TYR	A	3	11.907	15.809	24.042	1.00	10.96	C
ATOM	21	CG	TYR	A	3	12.497	15.093	25.196	1.00	10.60	C
ATOM	22	CD1	TYR	A	3	13.560	14.276	25.012	1.00	12.20	C
ATOM	23	CD2	TYR	A	3	12.045	15.324	26.492	1.00	11.77	C
ATOM	24	CE1	TYR	A	3	14.205	13.693	26.058	1.00	13.25	C
ATOM	25	CE2	TYR	A	3	12.663	14.737	27.567	1.00	12.45	C
ATOM	26	CZ	TYR	A	3	13.772	13.910	27.323	1.00	11.39	C
ATOM	27	OH	TYR	A	3	14.476	13.300	28.344	1.00	14.48	C
ATOM	28	N	LYS	A	4	12.633	17.222	21.175	1.00	9.63	C
ATOM	29	CA	LYS	A	4	12.179	17.659	19.887	1.00	9.41	C
ATOM	30	C	LYS	A	4	11.677	16.470	19.087	1.00	9.49	C
ATOM	31	O	LYS	A	4	12.151	15.336	19.237	1.00	8.55	C
ATOM	32	CB	LYS	A	4	13.376	18.247	19.100	1.00	12.36	C
ATOM	33	CG	LYS	A	4	12.954	19.035	17.857	1.00	17.46	C
ATOM	34	CD	LYS	A	4	14.119	19.494	16.982	1.00	20.77	C
ATOM	35	CE	LYS	A	4	14.184	21.056	16.755	1.00	24.12	C
ATOM	36	NZ	LYS	A	4	12.929	21.820	16.303	1.00	25.14	C
ATOM	37	N	LEU	A	5	10.771	16.761	18.157	1.00	8.76	C
ATOM	38	CA	LEU	A	5	10.253	15.790	17.221	1.00	7.91	C
ATOM	39	C	LEU	A	5	10.360	16.415	15.781	1.00	9.03	C
ATOM	40	O	LEU	A	5	9.916	17.539	15.553	1.00	6.35	C
ATOM	41	CB	LEU	A	5	8.765	15.468	17.506	1.00	8.63	C
ATOM	42	CG	LEU	A	5	8.058	14.607	16.411	1.00	8.98	C
ATOM	43	CD1	LEU	A	5	8.626	13.160	16.373	1.00	8.38	C
ATOM	44	CD2	LEU	A	5	6.577	14.522	16.660	1.00	9.13	C
ATOM	45	N	ILE	A	6	10.995	15.689	14.856	1.00	7.05	C
ATOM	46	CA	ILE	A	6	11.082	16.103	13.475	1.00	9.67	C
ATOM	47	C	ILE	A	6	10.046	15.228	12.753	1.00	8.83	C
ATOM	48	O	ILE	A	6	10.068	14.016	12.892	1.00	7.29	C
ATOM	49	CB	ILE	A	6	12.484	15.880	12.922	1.00	9.90	C
ATOM	50	CG1	ILE	A	6	13.453	16.788	13.678	1.00	13.88	C
ATOM	51	CG2	ILE	A	6	12.520	16.275	11.428	1.00	9.38	C
ATOM	52	CD1	ILE	A	6	14.844	16.502	13.276	1.00	15.90	C
ATOM	53	N	LEU	A	7	9.085	15.850	12.087	1.00	9.18	C
ATOM	54	CA	LEU	A	7	8.009	15.163	11.389	1.00	7.98	C
ATOM	55	C	LEU	A	7	8.312	15.181	9.902	1.00	9.81	C
ATOM	56	O	LEU	A	7	8.580	16.234	9.291	1.00	7.52	C
ATOM	57	CB	LEU	A	7	6.690	15.903	11.602	1.00	11.14	C
ATOM	58	CG	LEU	A	7	6.147	16.007	13.030	1.00	13.03	C
ATOM	59	CD1	LEU	A	7	5.647	17.410	13.283	1.00	14.31	C
ATOM	60	CD2	LEU	A	7	5.037	14.952	13.247	1.00	12.67	C
ATOM	61	N	ASN	A	8	8.383	14.018	9.323	1.00	10.71	C
ATOM	62	CA	ASN	A	8	8.628	13.975	7.913	1.00	13.03	C
ATOM	63	C	ASN	A	8	7.575	13.039	7.320	1.00	12.38	C
ATOM	64	O	ASN	A	8	7.885	11.921	6.936	1.00	11.56	C
ATOM	65	CB	ASN	A	8	10.025	13.485	7.676	1.00	15.38	C
ATOM	66	CG	ASN	A	8	10.270	13.214	6.226	1.00	19.45	C
ATOM	67	OD1	ASN	A	8	10.134	14.101	5.386	1.00	19.46	C
ATOM	68	ND2	ASN	A	8	10.497	11.955	5.900	1.00	22.31	C
ATOM	69	N	GLY	A	9	6.309	13.415	7.484	1.00	10.44	C
ATOM	70	CA	GLY	A	9	5.213	12.642	6.966	1.00	12.52	C
ATOM	71	C	GLY	A	9	4.914	12.990	5.506	1.00	12.67	C
ATOM	72	O	GLY	A	9	5.571	13.842	4.929	1.00	14.71	C
ATOM	73	N	LYS	A	10	3.922	12.342	4.909	1.00	14.25	C
ATOM	74	CA	LYS	A	10	3.589	12.601	3.497	1.00	15.77	C
ATOM	75	C	LYS	A	10	2.910					

Proteins structures are more conserved than sequences..



Evolution works at both the sequence (randomness) and function (selection) levels

Random mutations (or insertions or deletions) of the sequence are selected by their effect on the function. This determines the exploration of the sequence space under the restraint of conserving/improving the function, that often corresponds to conserving the structure (the idea is that the cell is not happy to have useless proteins, so in most of the cases the fold is conserved and mutations are accommodated)

If two protein sequences are similar it is ‘likely’ that they will fold into very similar structures.



Homologs: The Sequence - Structure - Function Paradigm

Sequence alignment -> Structure alignment

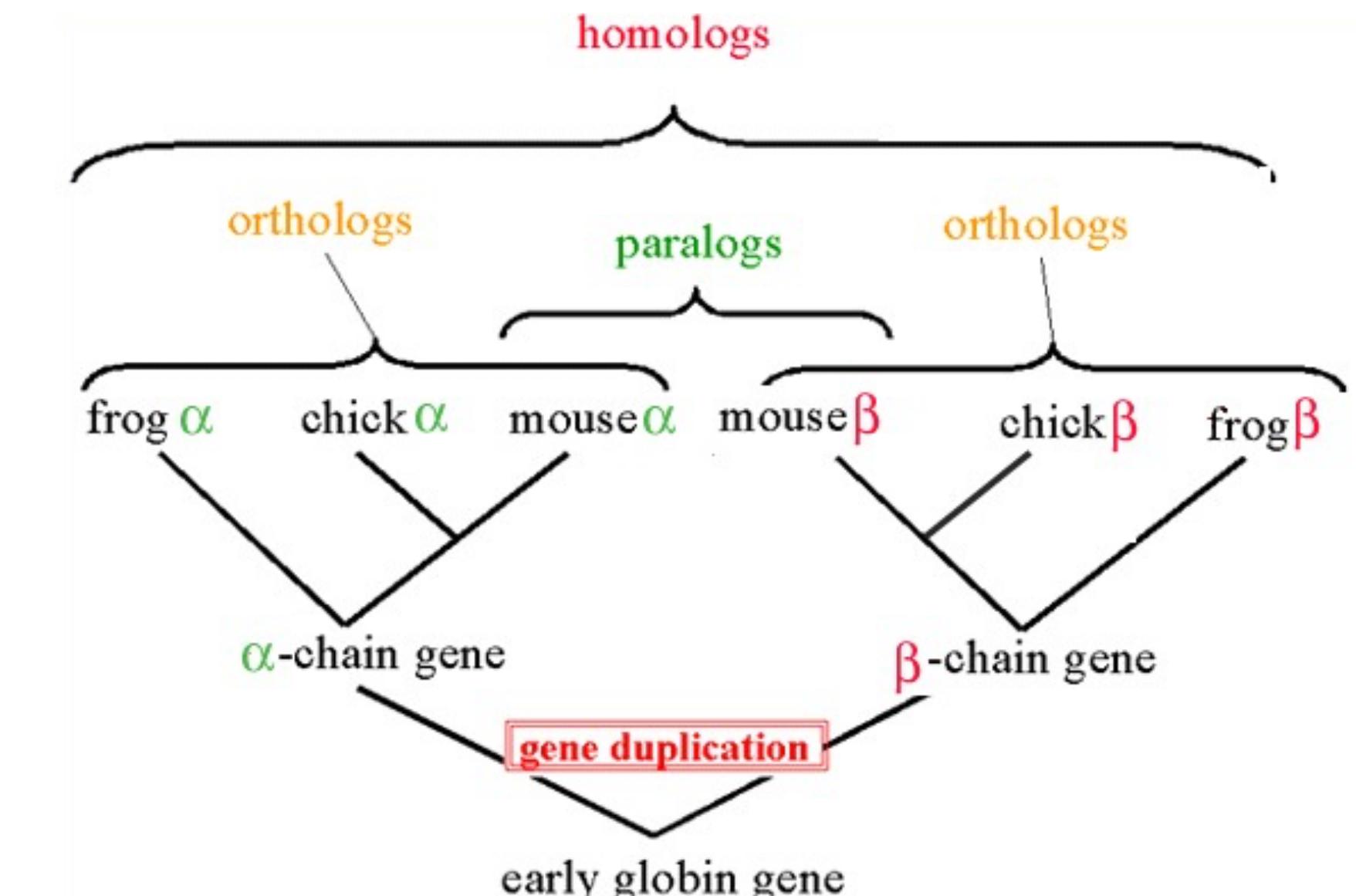
Protein sequence alignment can be used as guide for structural alignments, thus helping in identifying structural patterns.

Structure alignment -> Sequence alignment (Structural genomics and phylogenetic)

When sequence overlap is very poor, structural alignment can produce aligned sequences.

A 25% sequence similarity is often enough to share the same fold, then:

- A) Either predicting a protein fold should be not too difficult because the details of the sequence don't matter
- B) Or most proteins have common ancestors and there wasn't enough time to decorrelate



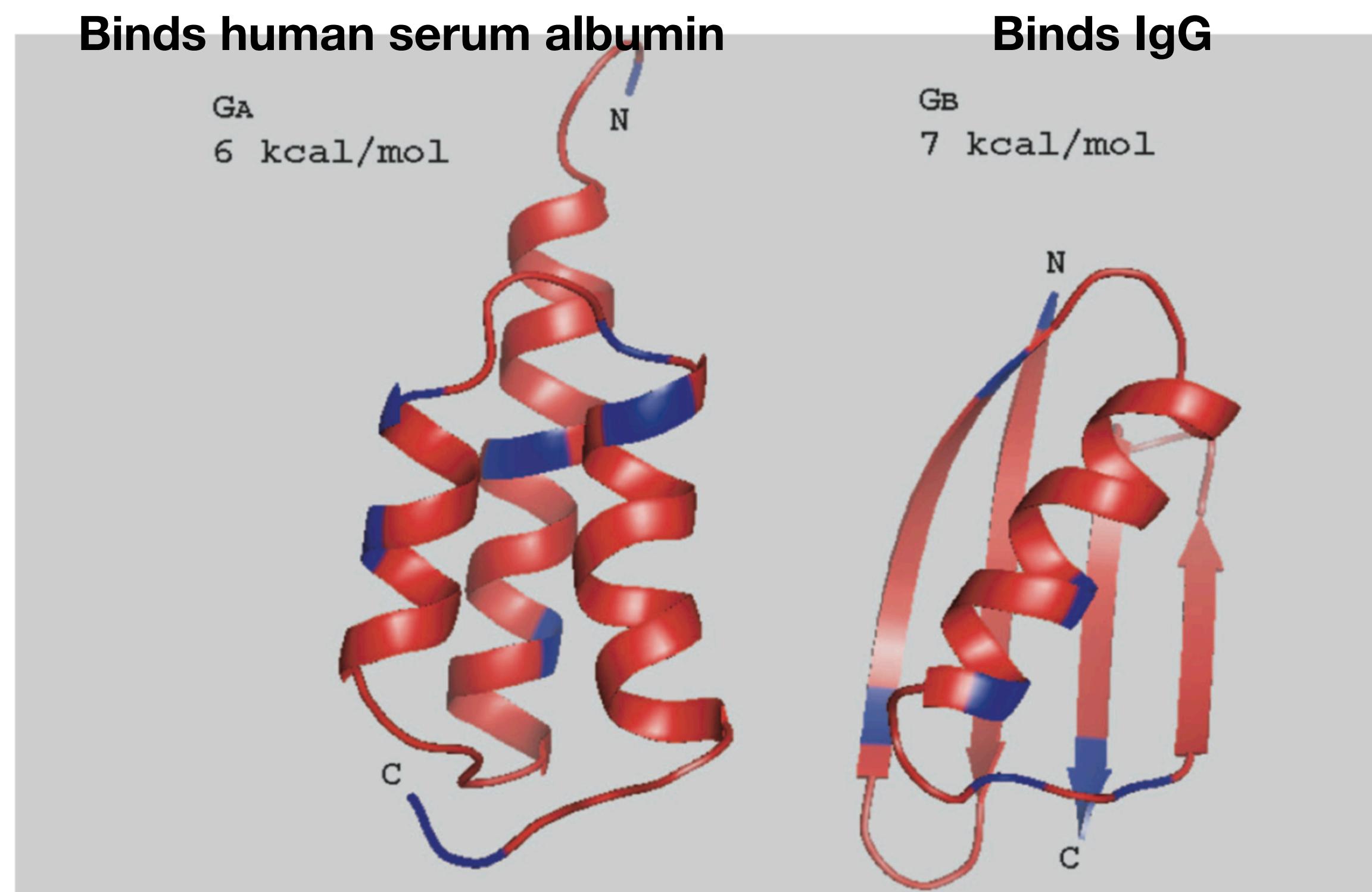
PARALOGS: Homologous biological components (genes, proteins, structures) within a single species that arose by gene duplication.

ORTHOLOGS: Homologous biological components (genes, proteins, structures) in different species that arose from a single component present in the common ancestor of the species; orthologs may or may not have a similar function.

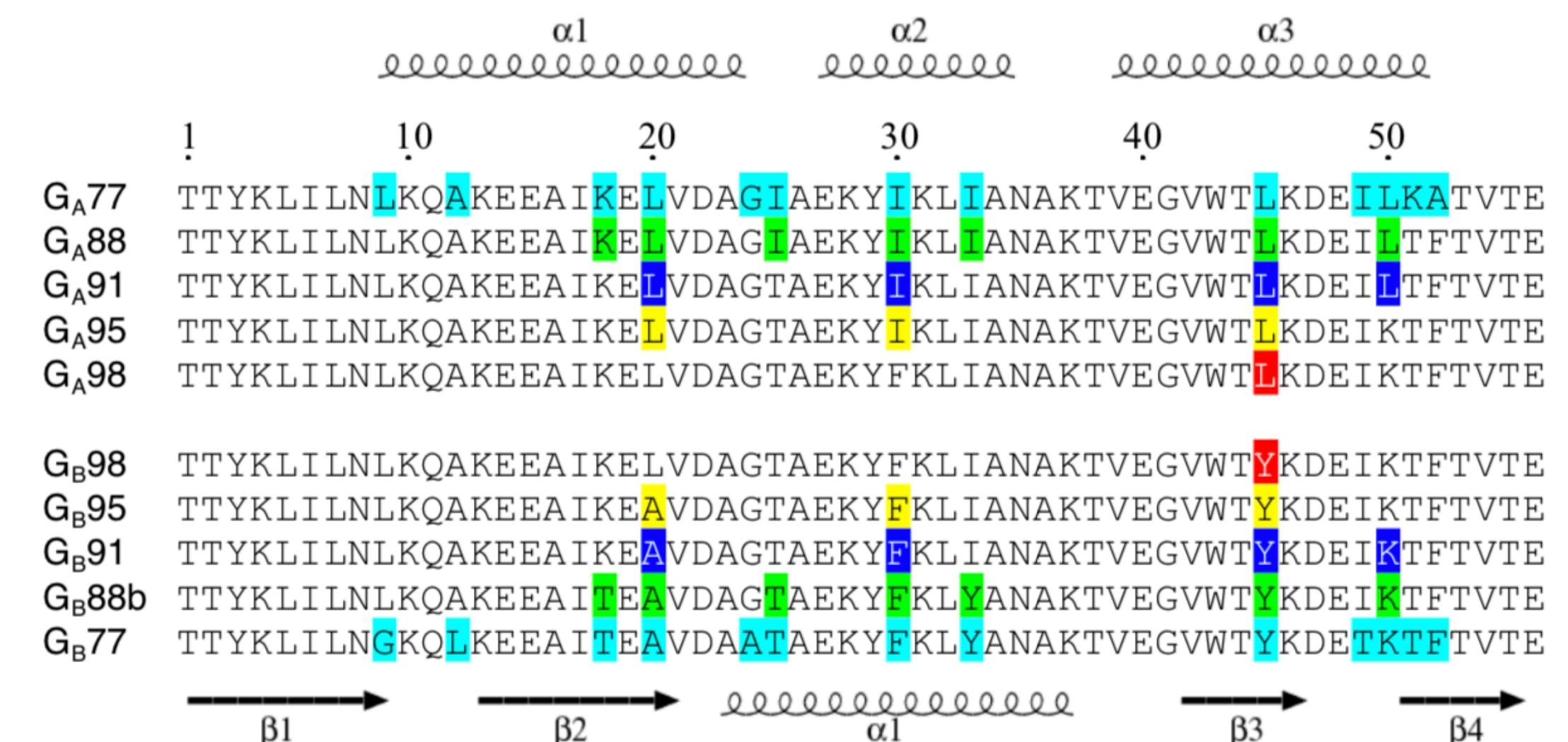


...yet small sequence variations may lead to dramatic differences

Very often a mutation of a single key aminoacid (usually very conserved upon evolution) is enough to go from a protein to a disordered polymer. But it can be even worse:



1.Alexander, P. A., He, Y., Chen, Y., Orban, J. & Bryan, P. N. A minimal sequence code for switching protein structure and function. *Proceedings of the National Academy of Sciences of the United States of America* 106, 21149 (2009).



1AA is enough to determine the fold!
And the function?

GA98 exhibits diminished affinity for HSA but has acquired affinity for IgG. GB98 binds tightly to IgG but not HSA.

Secondary structure prediction is relatively easy but secondary structure can depend on tertiary and quaternary structure

Useful as:

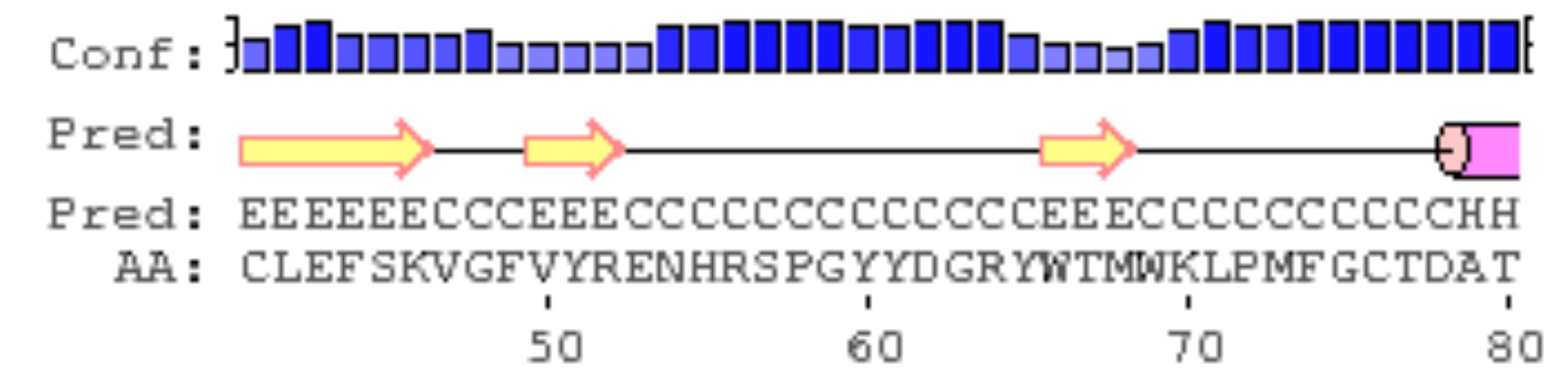
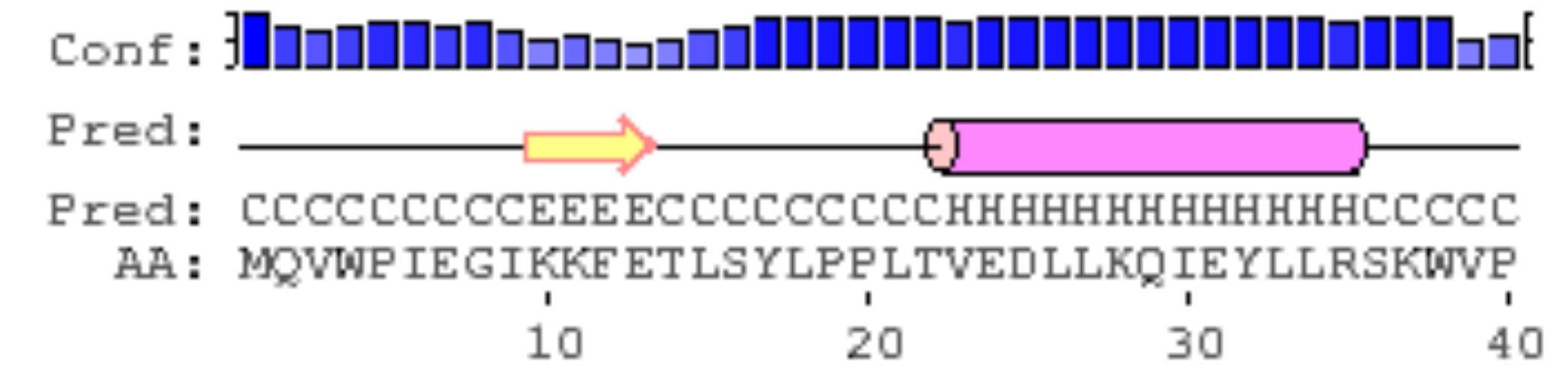
- a step towards tertiary structure prediction
- Sequence alignment
- Structure determination at intermediate resolution
- Prediction of other properties (aggregation, trans-membrane, disorder)

First generation of predictors (~1970) were based on single residues frequencies, second generation (~1990) took into account patterns for segments. Both biased by the small size of the PDB and the lack of systematic assignment.

Accuracy ~60%

Sequence alignment is performed (psi-blast) and neural networks are trained to assign secondary structures (from one letter sequence to one letter secondary structure)

Accuracy ~81%



Conf: [blue bars]
Pred:
Pred: CC
AA: GC

<http://bioinf.cs.ucl.ac.uk/psipred/>
<http://www.compbio.dundee.ac.uk/jpred>

Can we use a physico-chemical approach? molecular dynamics simulations have size, time scale and accuracy limits.

Classical molecular dynamics simulations of a protein in explicit environment with an energy function that tries to approximate the interactions (Force Field):

$$V(r) = \sum_{bonds} k_b(b - b_0)^2 + \sum_{angles} k_\theta(\theta - \theta_0)^2 + \sum_{torsions} k_\phi[\cos(n\phi + \delta) + 1] \\ + \sum_{nonbond\ pairs} \left[\frac{q_i q_j}{r_{ij}} + \frac{A_{ij}}{r_{ij}^{12}} - \frac{C_{ij}}{r_{ij}^6} \right]$$

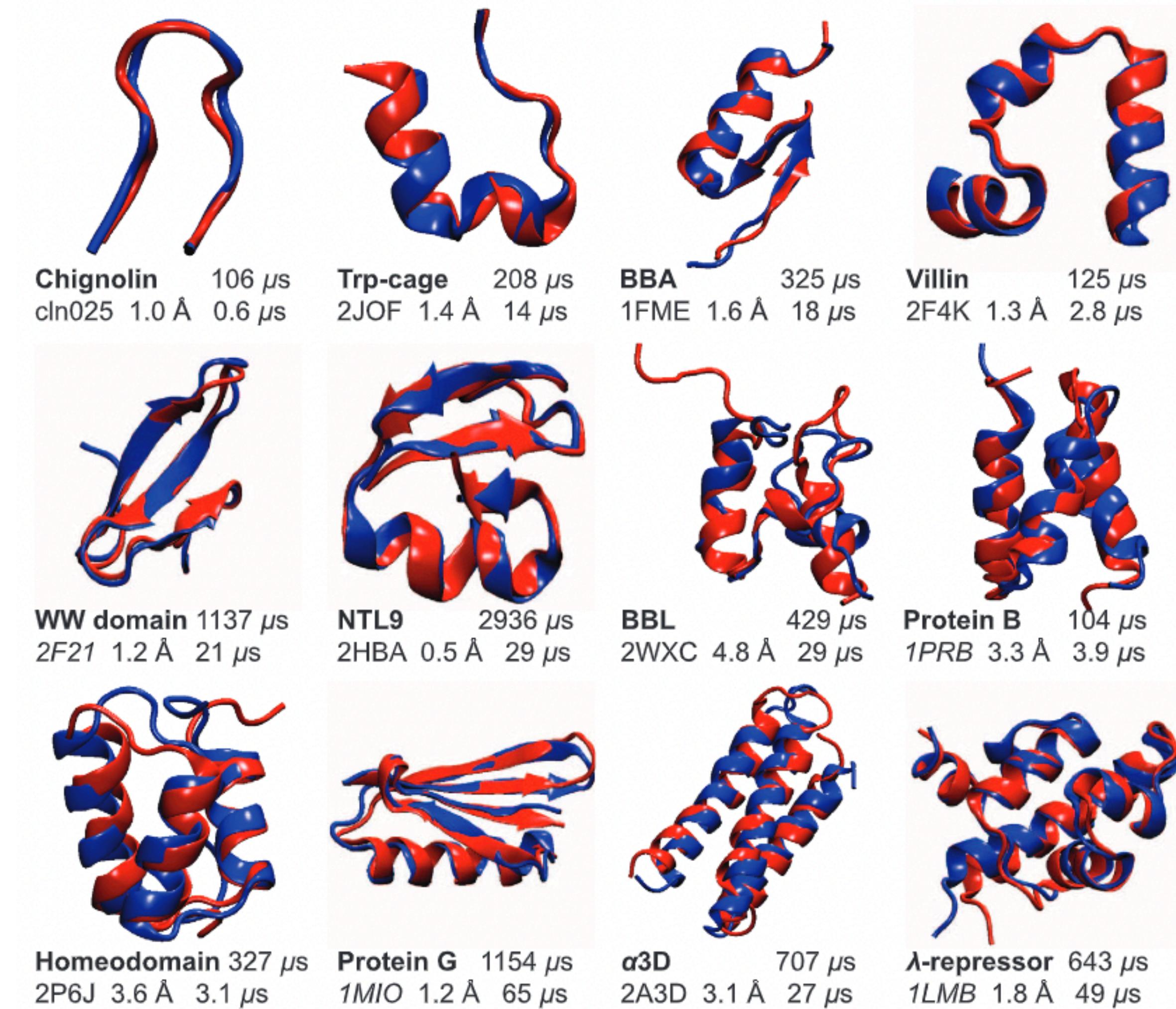
first approx for vibrations
(anharmonic potential can be used for more accurate vibrations)

geometrical consideration
pi-bonds, etc

point charge Coulomb

excluded volume

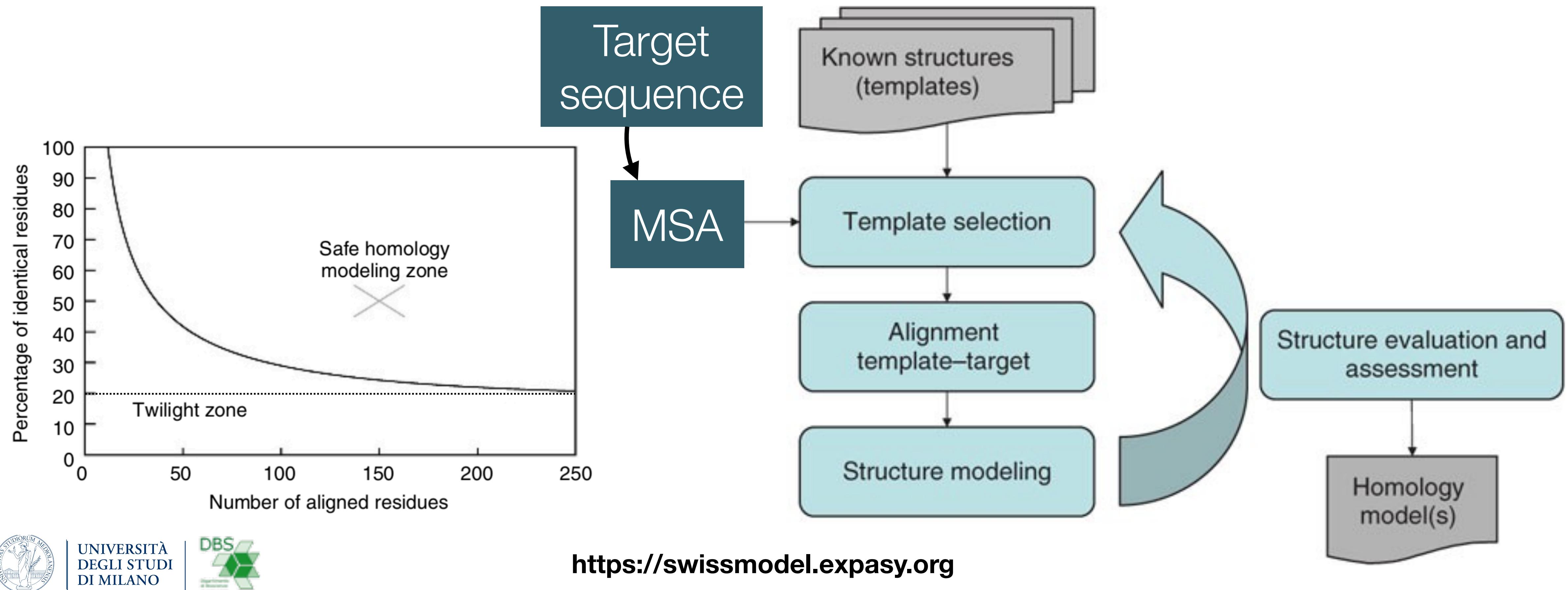
Dispersions (interactions of neutral molecules)



How fast-folding proteins fold. *Sci New York N Y* 334, 517–20 (2011).

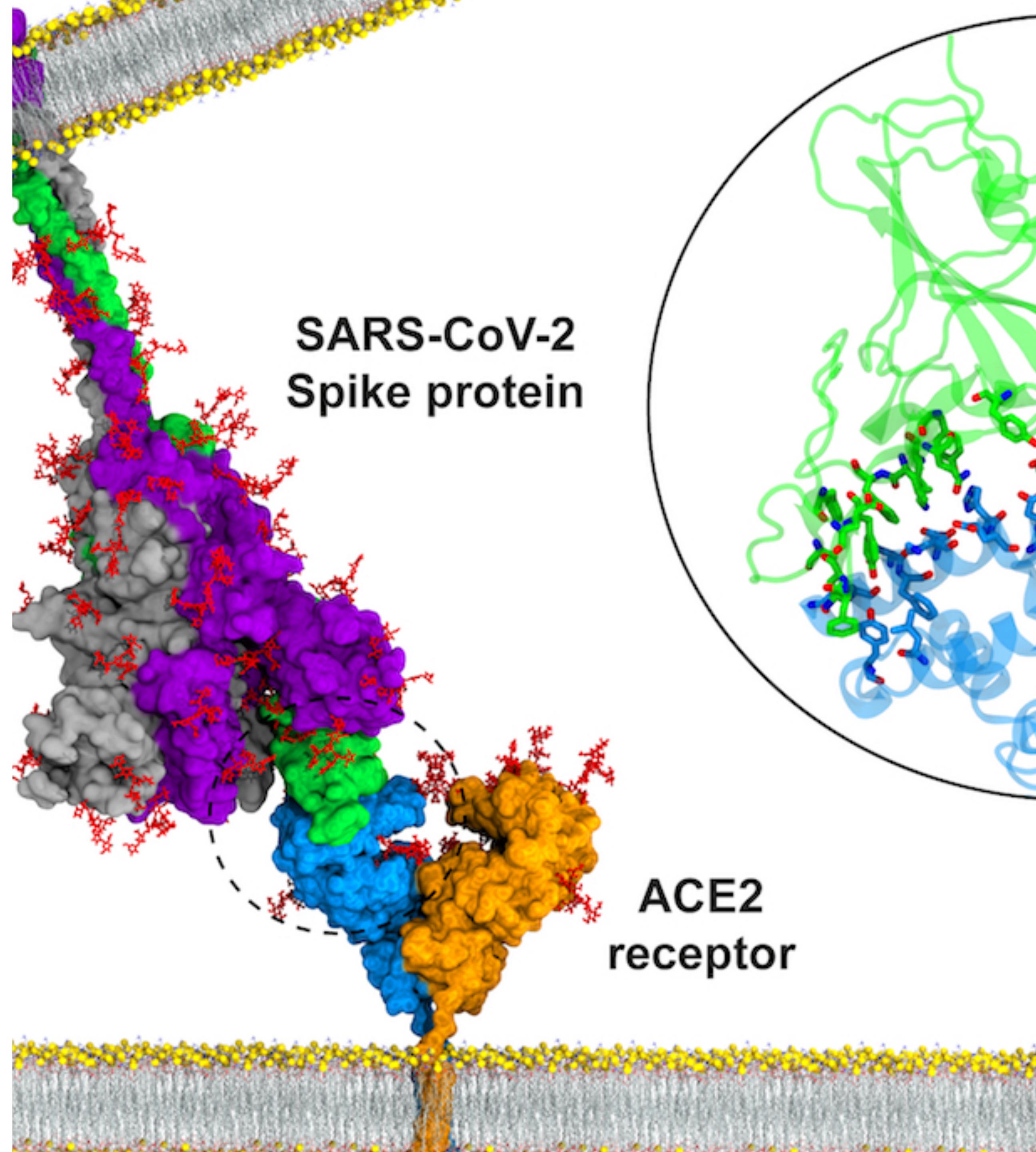
Evolutionary based bioinformatic approach: homology modelling aka template based modelling (TBM)

Given a sequence (target) we want to make use of sequence similarity and existing structures (templates) to model our target structure.



Outline

- Structure prediction: the origins
- **Structure prediction: key advances**
- State of the art and AI approaches
- Protein Design
- Challenges: conformational changes, effect of mutations, binding, quaternary structure

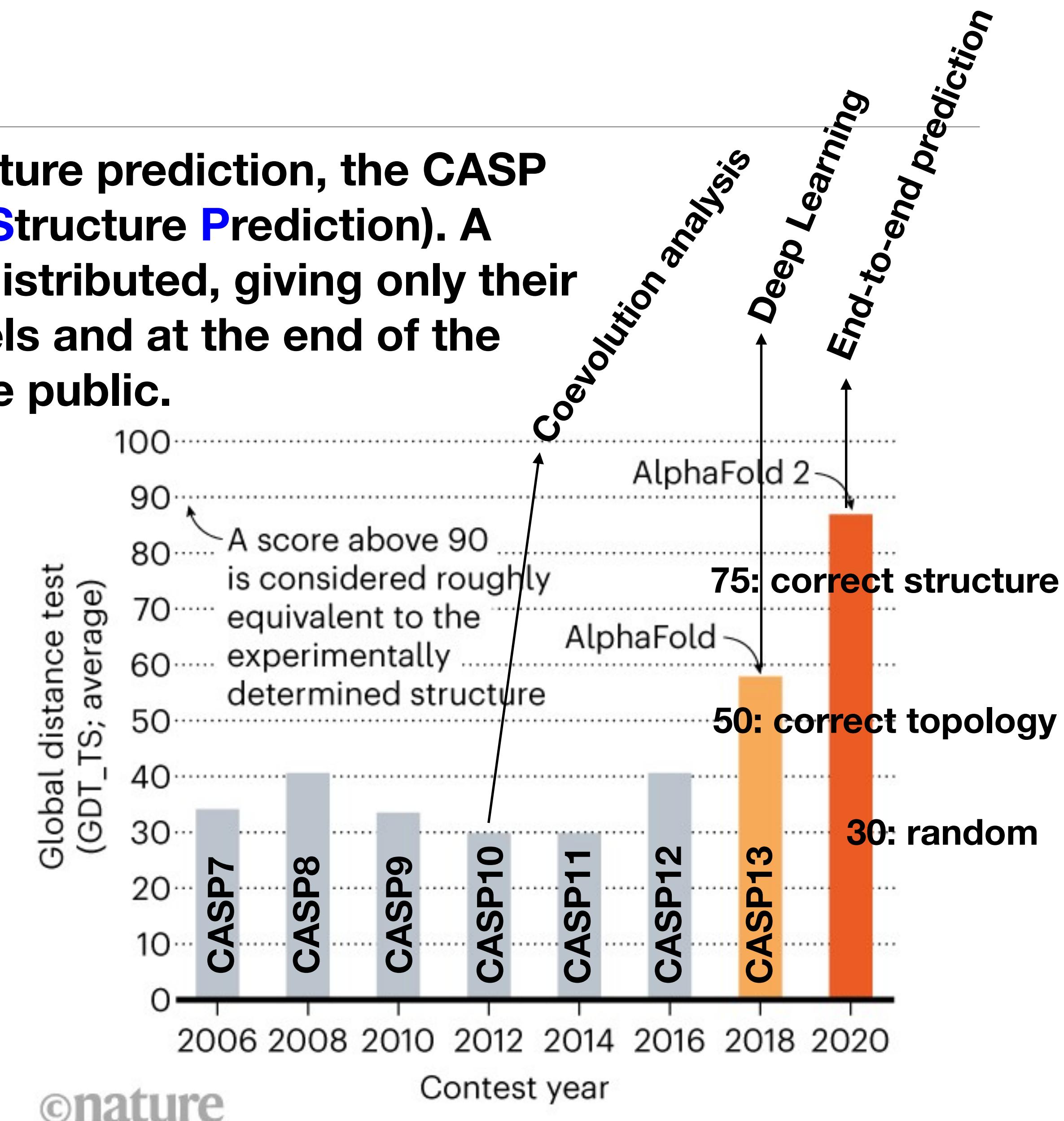


Ab-initio protein structure predictions

From 1994 there is a competition for protein structure prediction, the CASP (Critical Assessment of techniques for protein Structure Prediction). A number of structures recently determined are not distributed, giving only their sequences. Participants can deposit their models and at the end of the competition the results are made public.

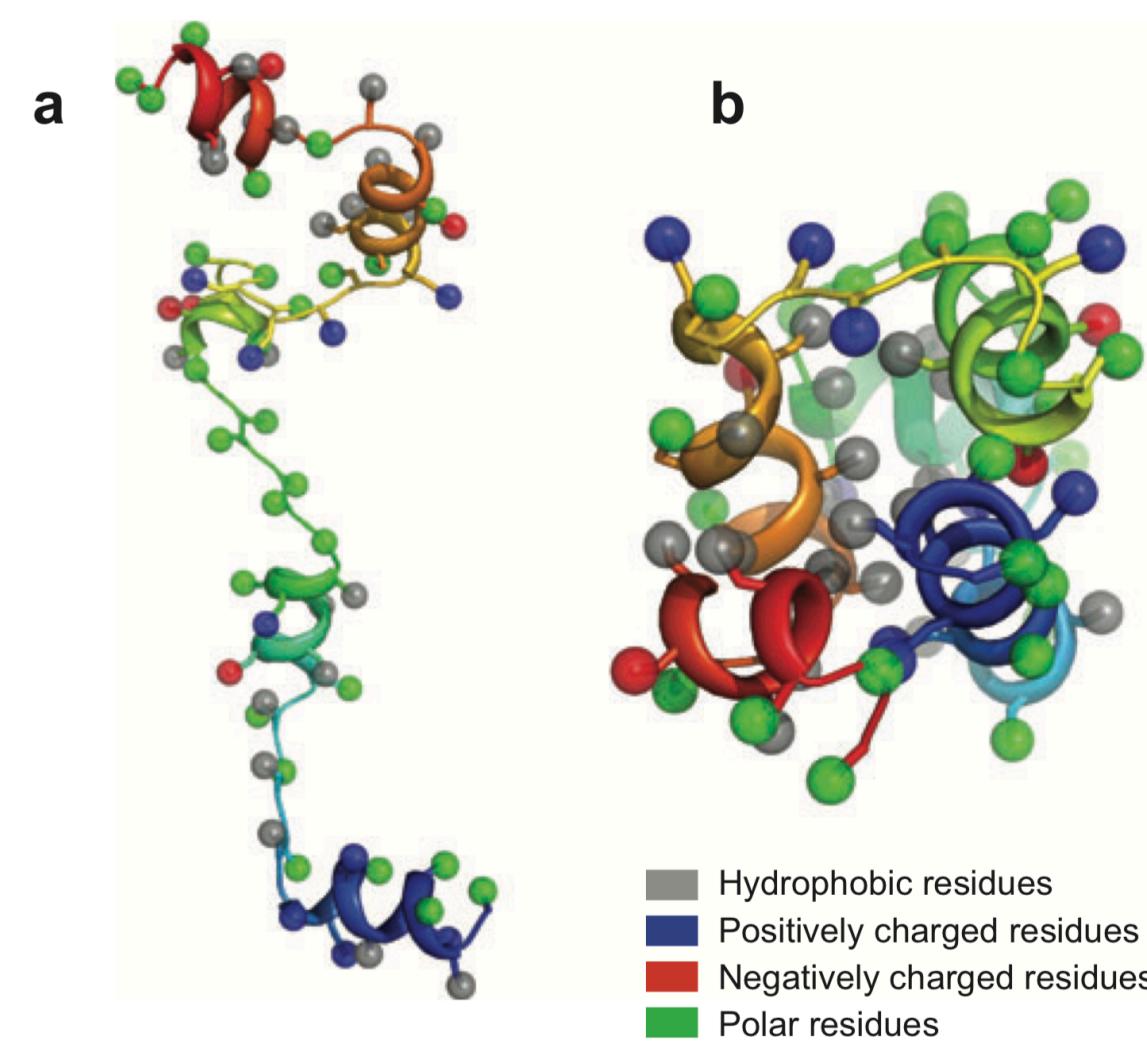
From the early 2000 the way to predict the structure of a protein has been to follow a “fragment replacement” strategy:

1. Generate many configuration using local information like secondary structure prediction as well as contact restraint using physico-chemical principles.
2. Refine the final structure at atomic resolution and use some fancy “scoring function”

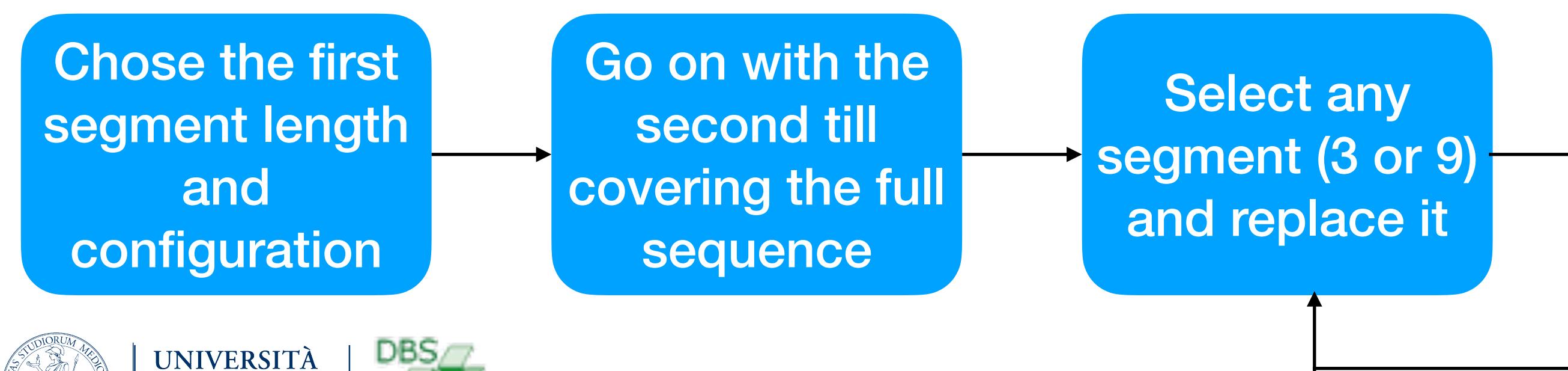


Before 2012: Fragment Replacement

A common strategy is that of splitting the problem in two. First the fold space is sampled with a very simplified model.



Fragment replacement:



1. Split the sequence of your protein in segments (nine and three residues long)
2. For each segment make an ensemble of configurations found in the PDB
3. Make the coarse-grain representation of the fragments
4. Monte-Carlo Assembly:

Monte Carlo sampling

In principle one could generate configurations at random, evaluate their scoring function, and keep the lowest energy one. The limit of this approach is that most of the time will be spent evaluating scoring functions for bad scoring configurations.

Calculating $P(x)$ would mean calculating it for all possible configurations, but the probability is simply proportional to:

$$f(x) = \exp\left[\frac{-U(x)}{k_B T}\right]$$

$$f(x) \propto P(x)$$

Starting from a configuration we can calculate its energy or score, then we can generate a second configuration and get its score and we need to decide if to keep it or not. How? Their relative probability is:

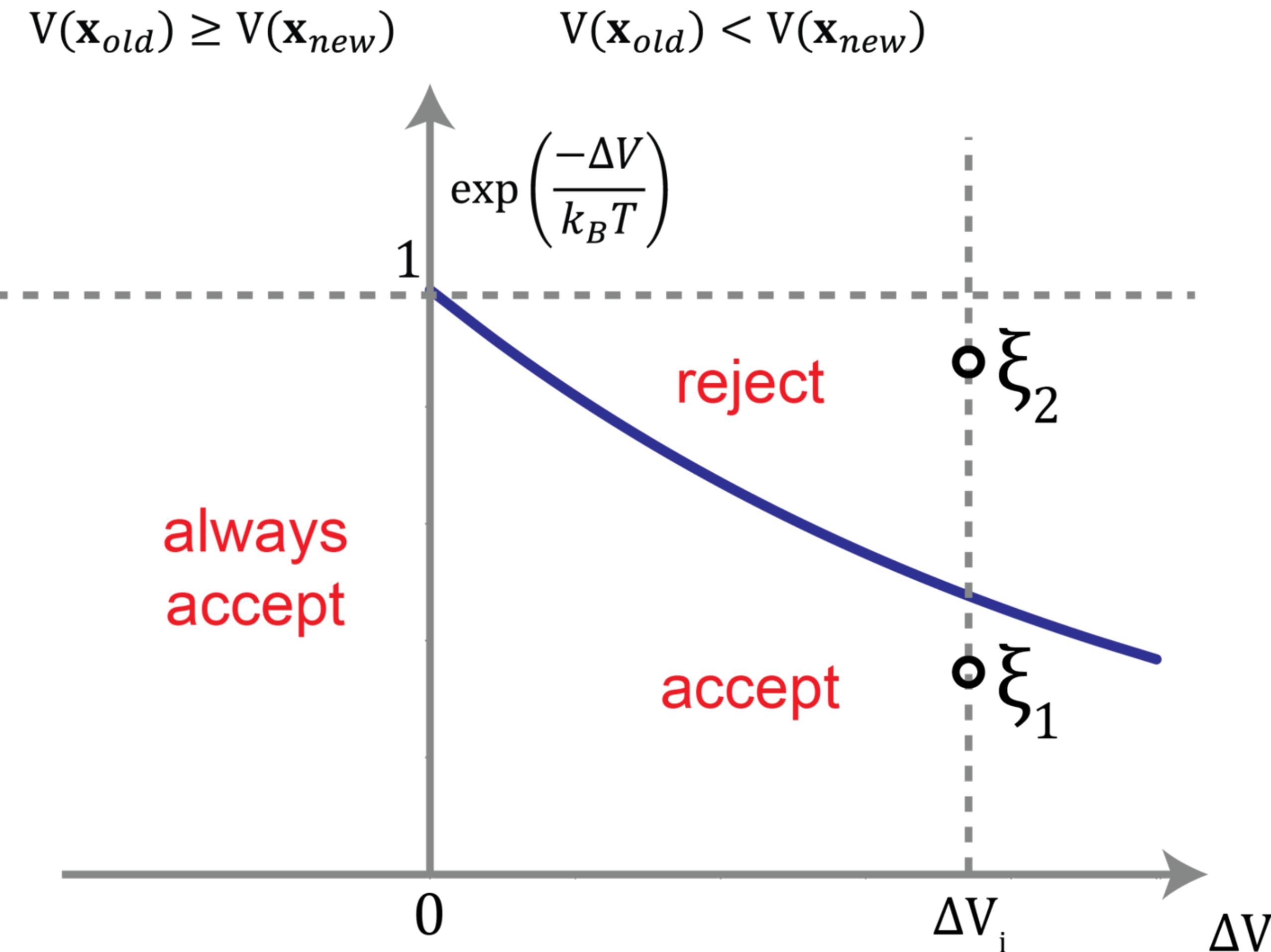
$$\frac{f(x_{new})}{f(x)} = \exp\left(\frac{-U(x_{new}) + U(x)}{k_B T}\right)$$

Do we keep the new one or not? We generate a random number (0..1), if it is smaller than the above ratio we keep the new configuration, otherwise the old.



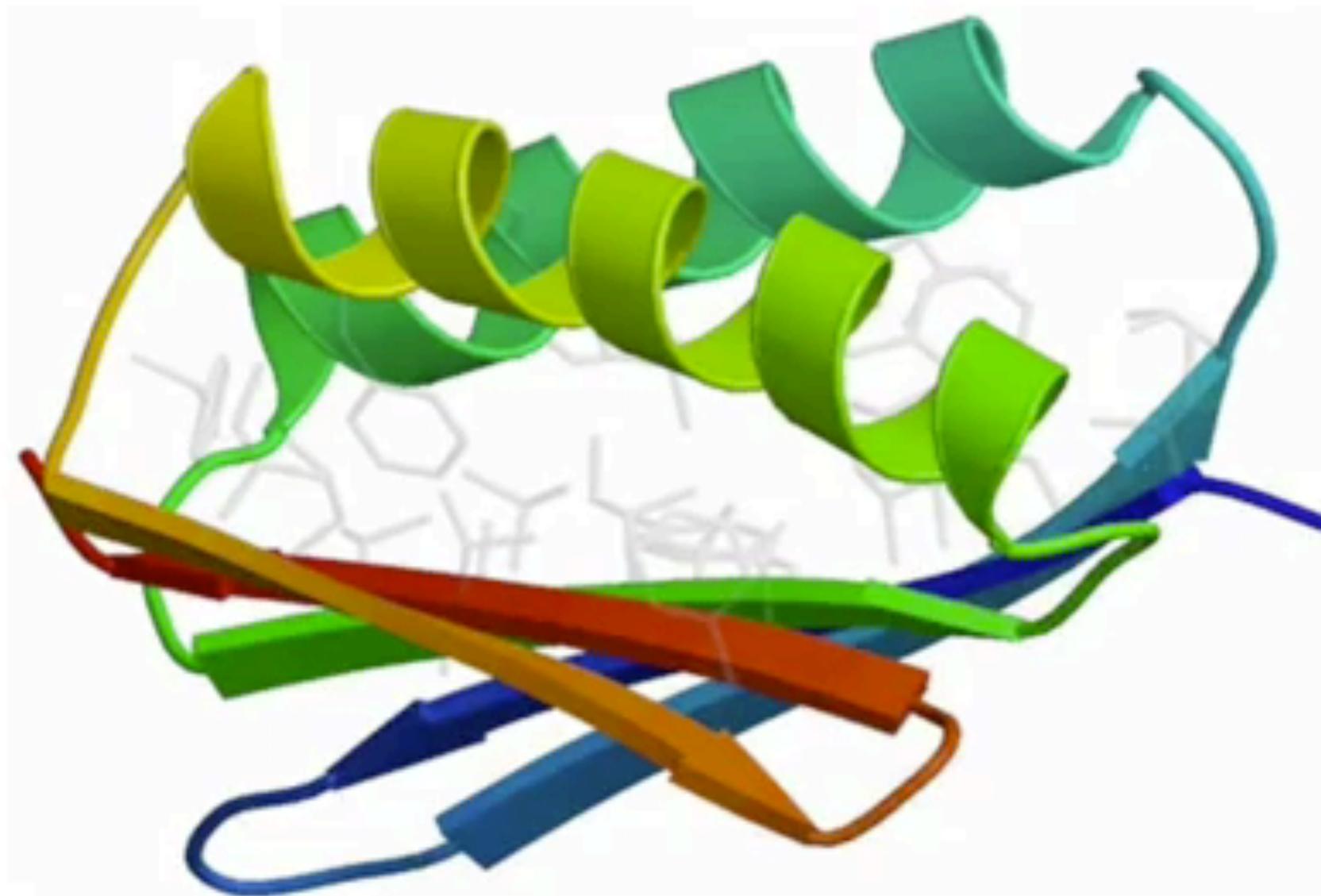
Monte-Carlo sampling

1. Given a starting configuration x_0 , suggest a new configuration x_1
2. Calculate $\text{score}(x_0)$ and $\text{score}(x_1)$
3. Calculate $a = \exp[(\text{score}(x_0) - \text{score}(x_1))/\text{optim_rate}]$
4. Generate a random number u between 0..1
5. If $u \leq a$ accept the configuration x_1 and go on
6. If $u > a$ discard x_1 and keep x_0
7. Keep track of the lowest energy found till now

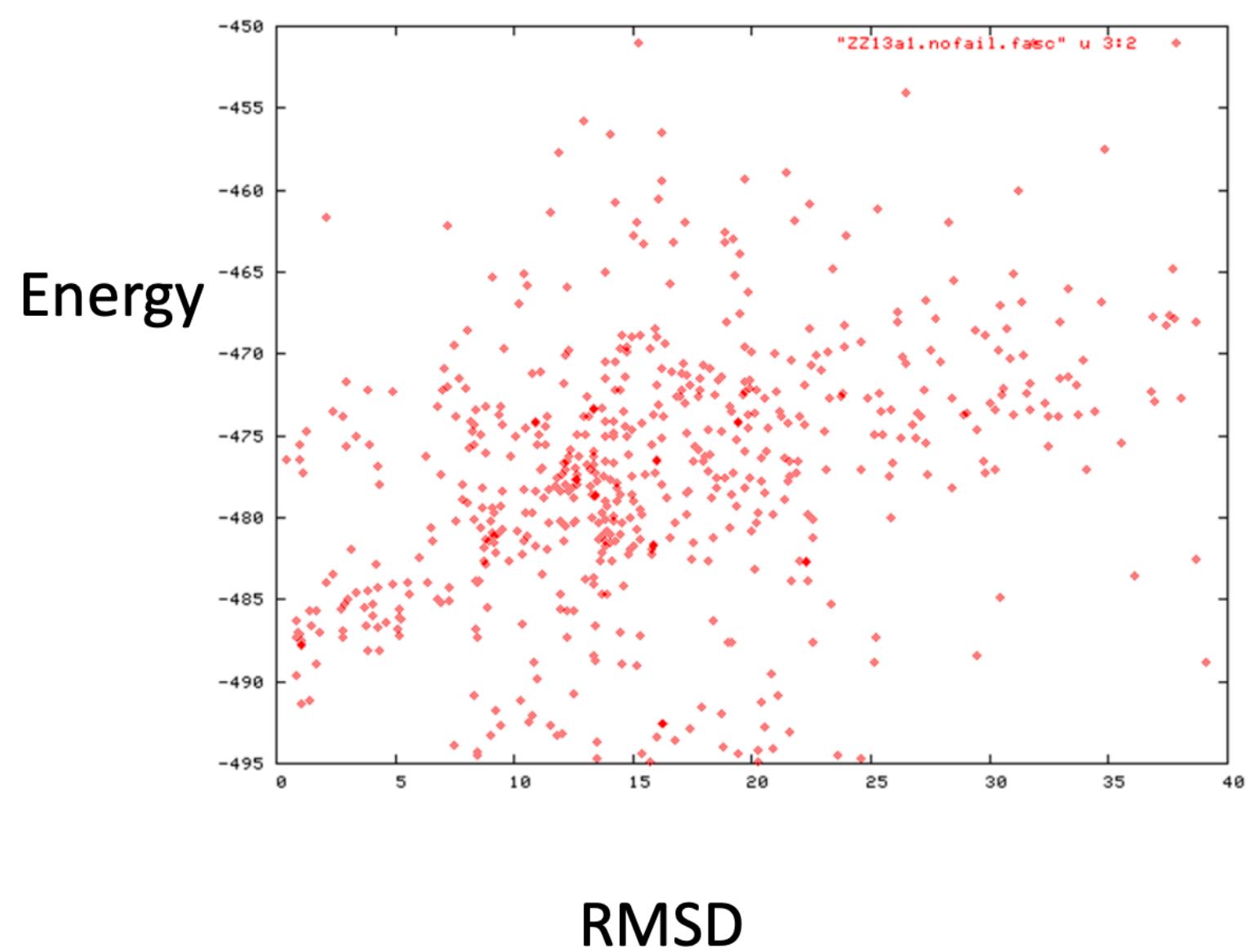


Convergence of a prediction is checked by scores distributions

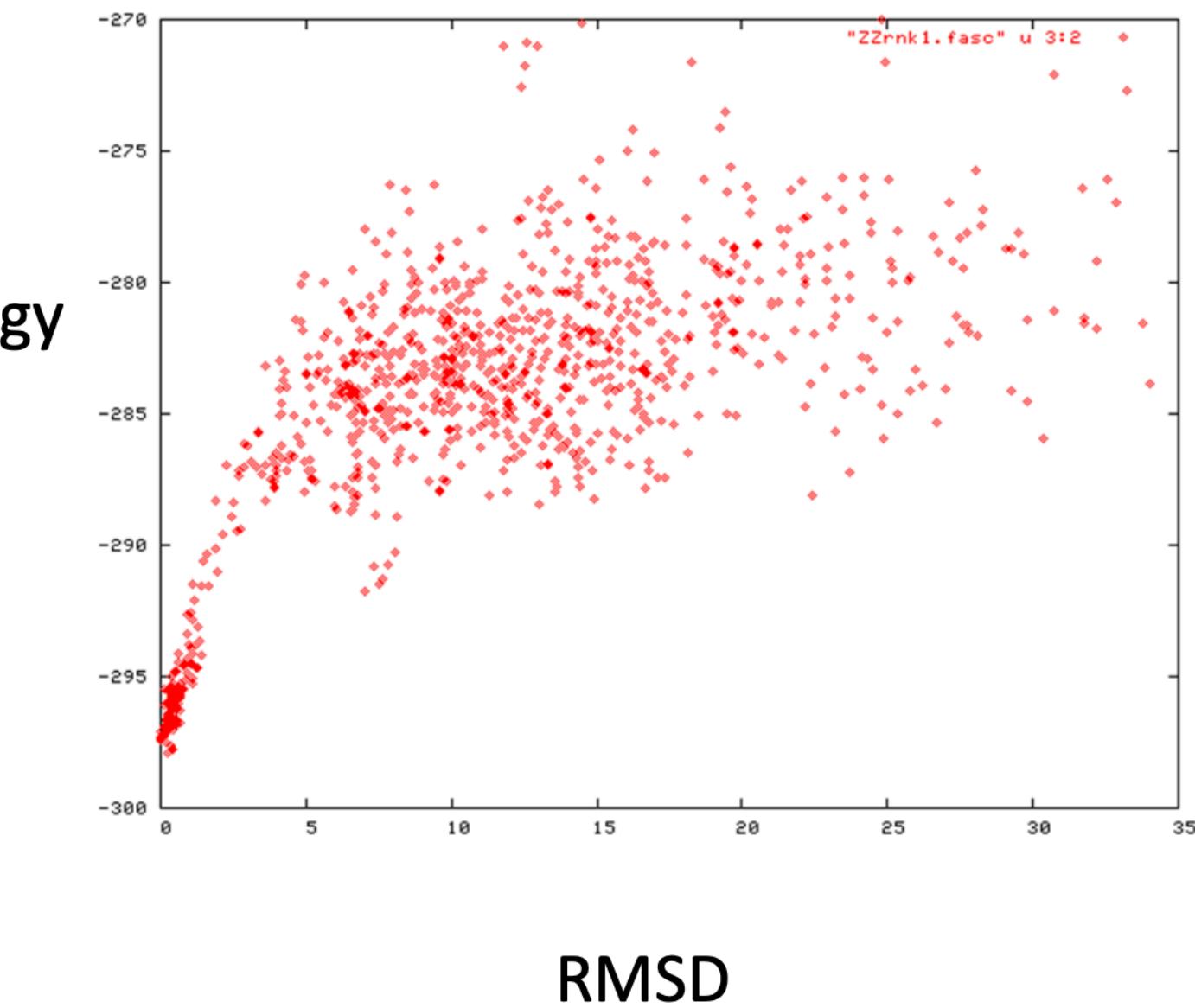
Then the best possible folds are further sampled at full atomistic resolution (without water)



Tens of thousands of structures are then scored:



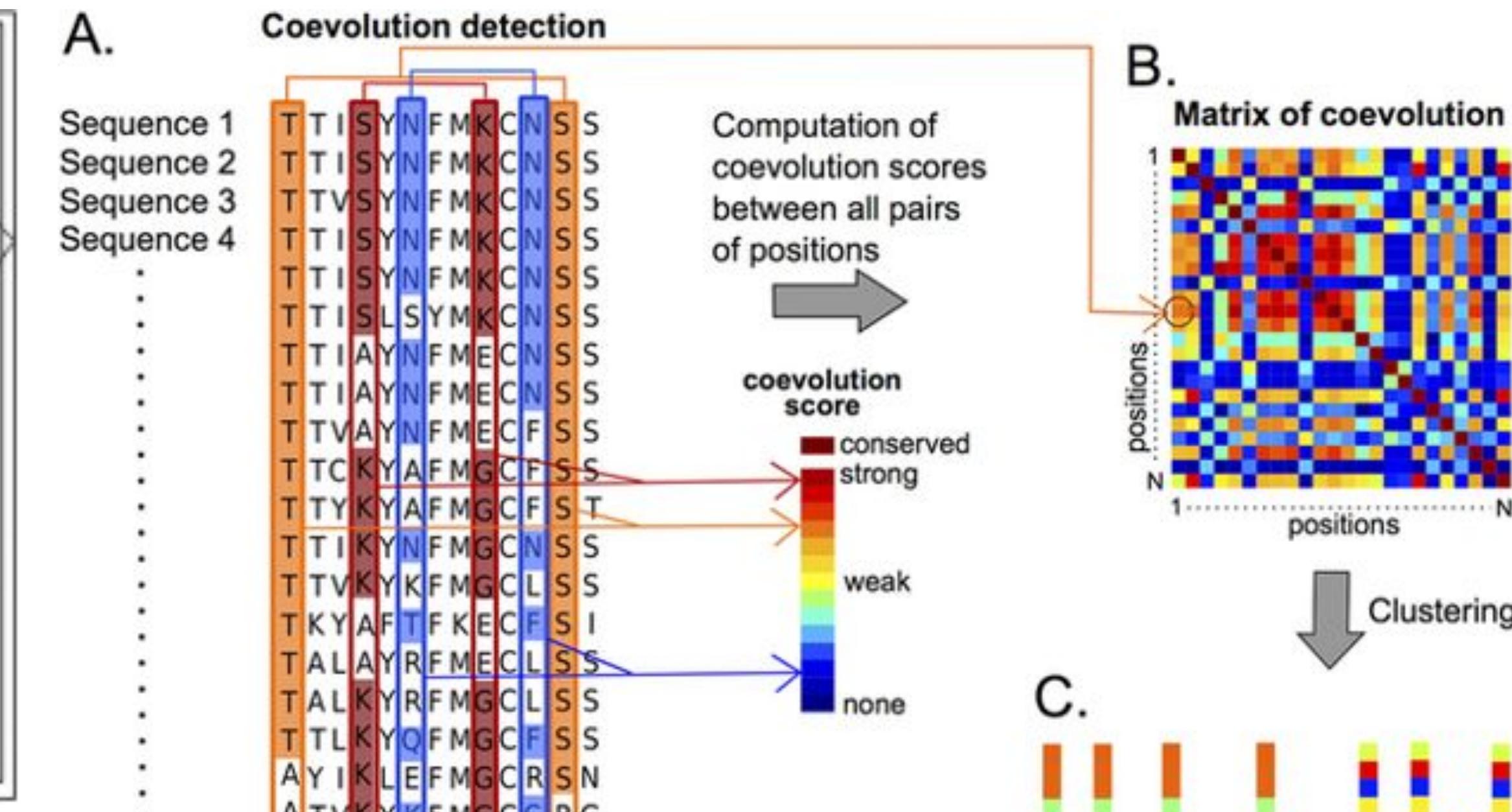
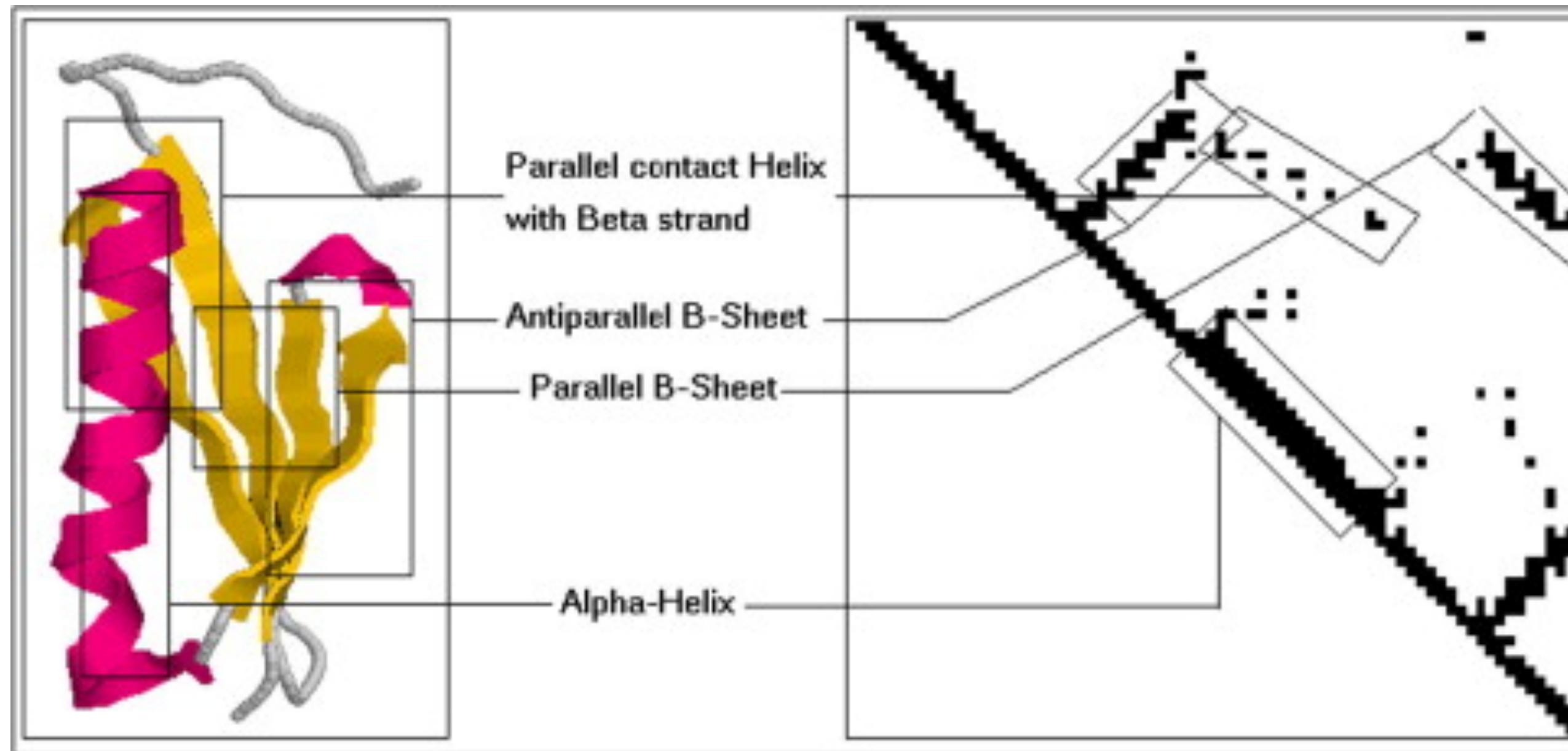
Bad



Good



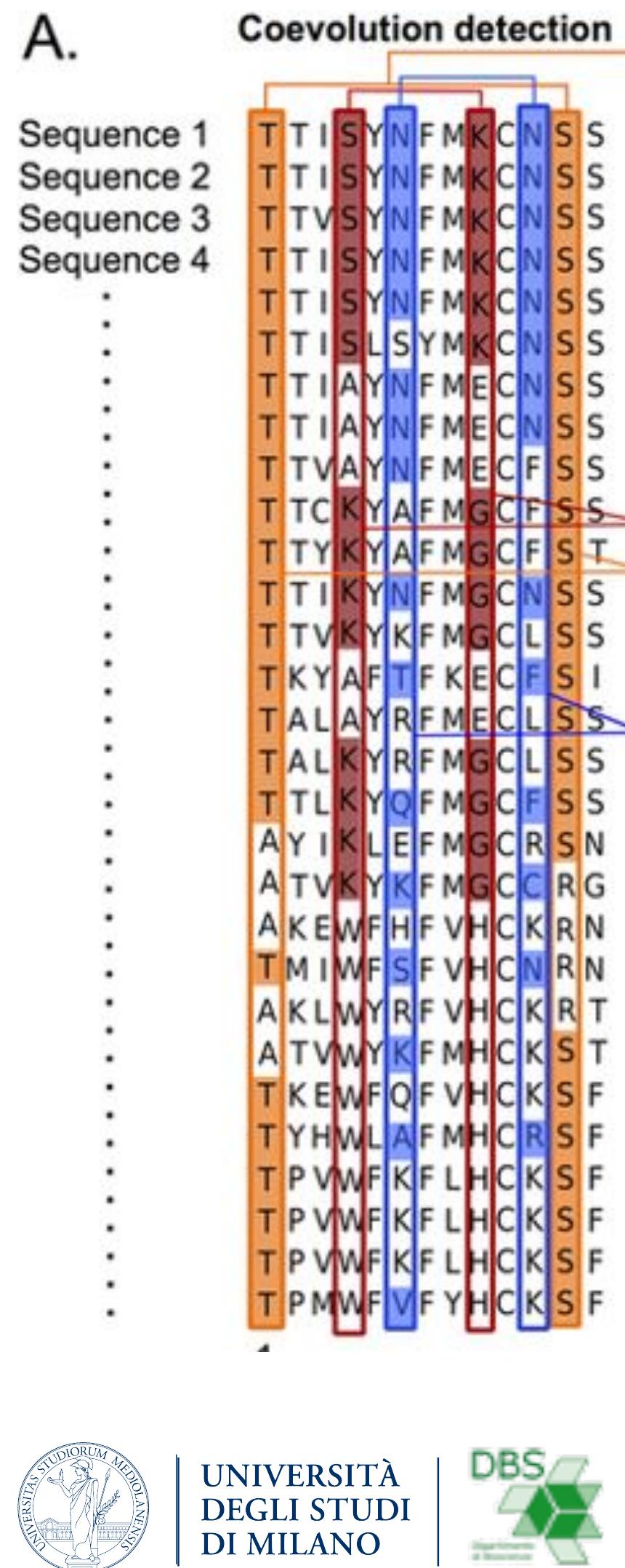
From 2012: adding inter-residues contact restraints predicted from sequences co-evolution analysis



The contact map is a binary 2D representation of a protein structure, it may be simpler to go from 1D sequence to 2D CMAP as an intermediate step.

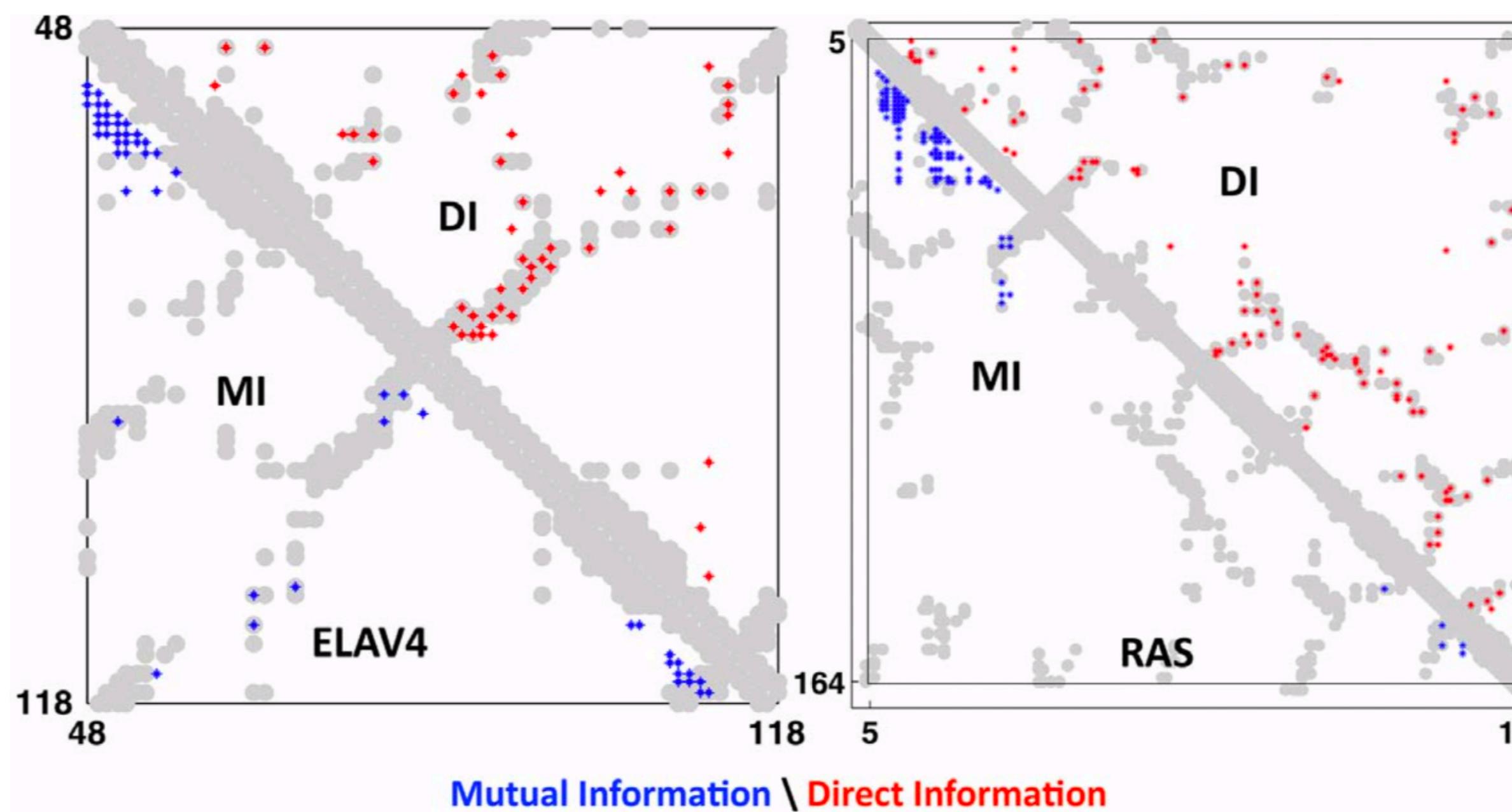
The key idea is that correlations among mutations as observed in MSA may indicate amino acid that coevolved together to maintain the structure

From 2012: adding inter-residues contact restraints predicted from sequences co-evolution analysis



$$f_i(A) = \frac{1}{M} \sum_{a=1}^M \delta_{A, A_i^a}; \quad f_{ij}(A, B) = \frac{1}{M} \sum_{a=1}^M \delta_{A, A_i^a} \delta_{B, A_j^a},$$

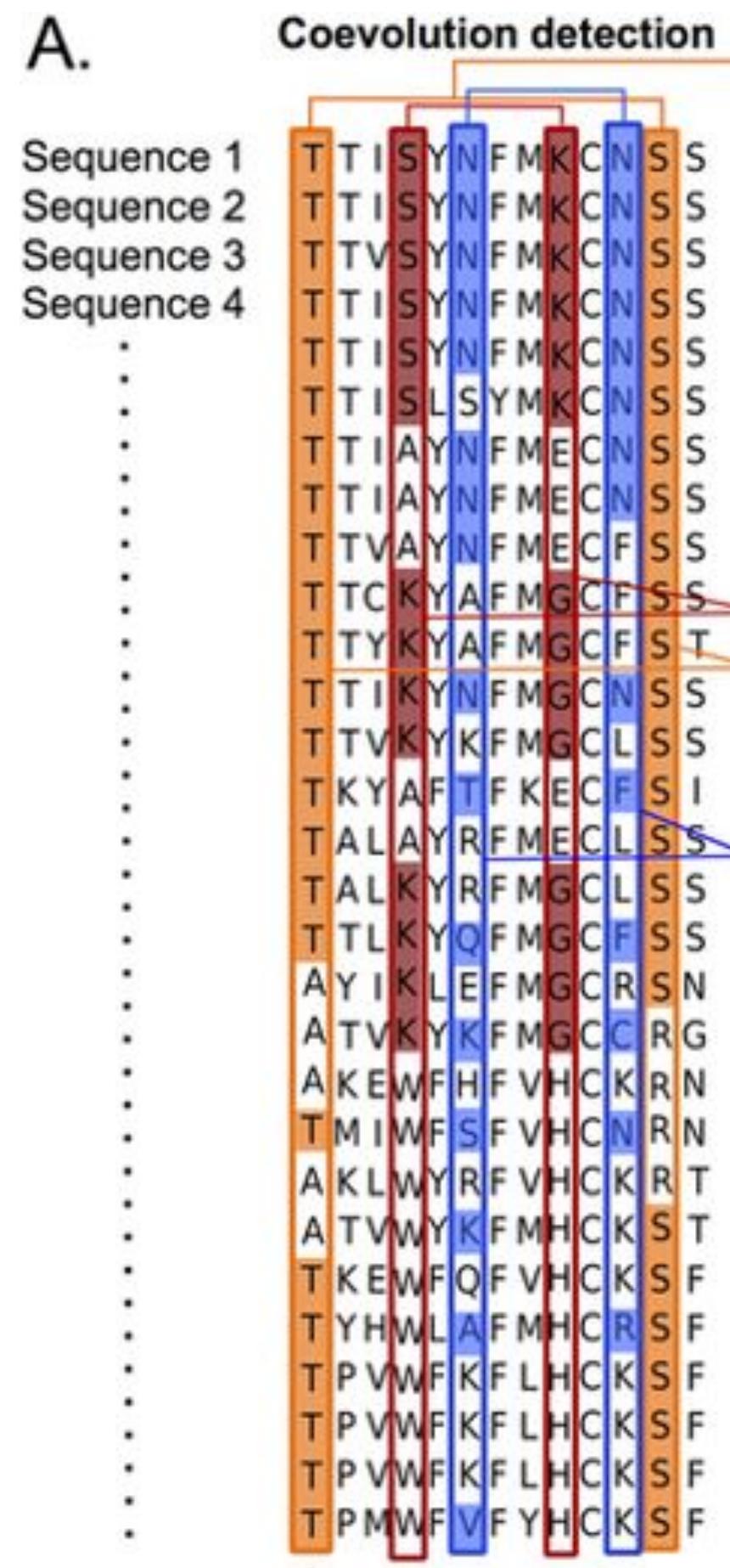
$$MI_{ij} = \sum_{A, B} f_{ij}(A, B) \ln \frac{f_{ij}(A, B)}{f_i(A)f_j(B)}. \quad \textbf{Mutual Information}$$



Mutual information does not capture residue-residue contacts in 3D structures, why?

2012: From Mutual to Direct Information

1.Morcos, F. et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *P Natl Acad Sci Usa* 108, E1293-301 (2011).



The problem is that MSA do not include all possibilities

Can we find what is the probability distribution of finding each amino acid in a given position in the sequence?

$$P(A_1, \dots, A_L)$$

This should be such that the probability of a single amino acid in a single position and the probabilities of the couples are those observed from the alignment

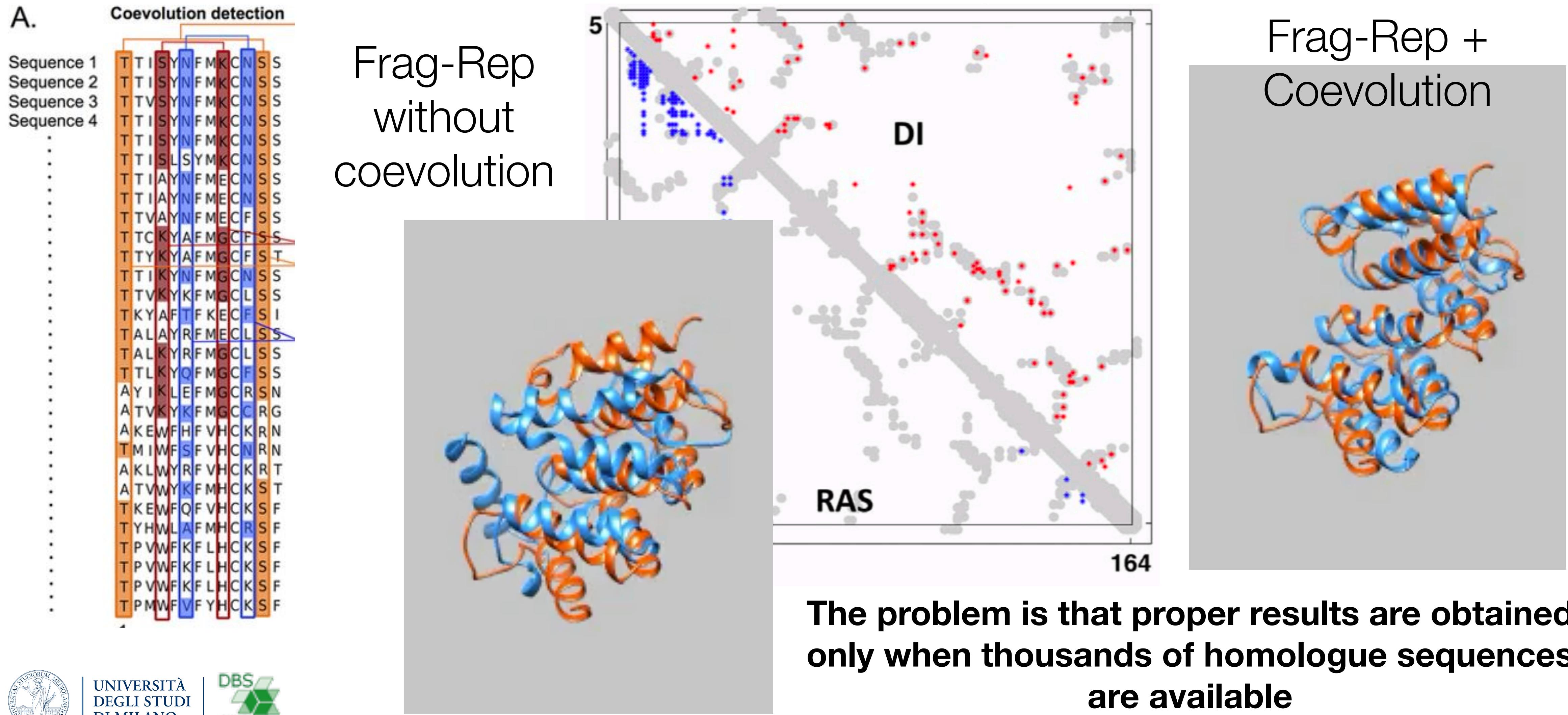
$$P_i(A_i) = \sum_{\{A_k | k \neq i\}} P(A_1, \dots, A_L) = f_i(A_i) \quad (6)$$

$$P_{ij}(A_i, A_j) = \sum_{\{A_k | k \neq i, j\}} P(A_1, \dots, A_L) = f_{ij}(A_i, A_j) .$$

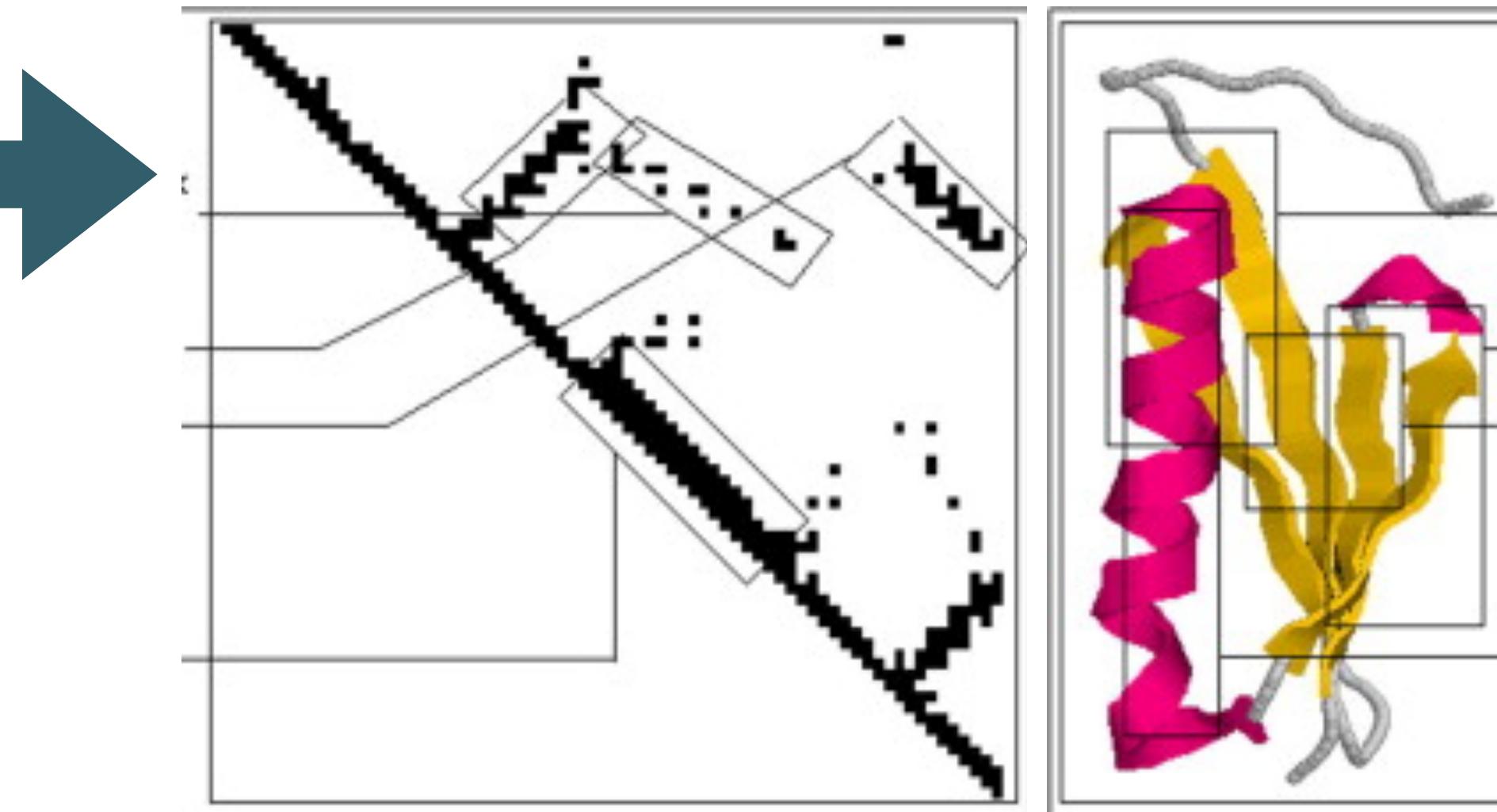
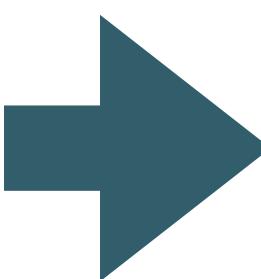
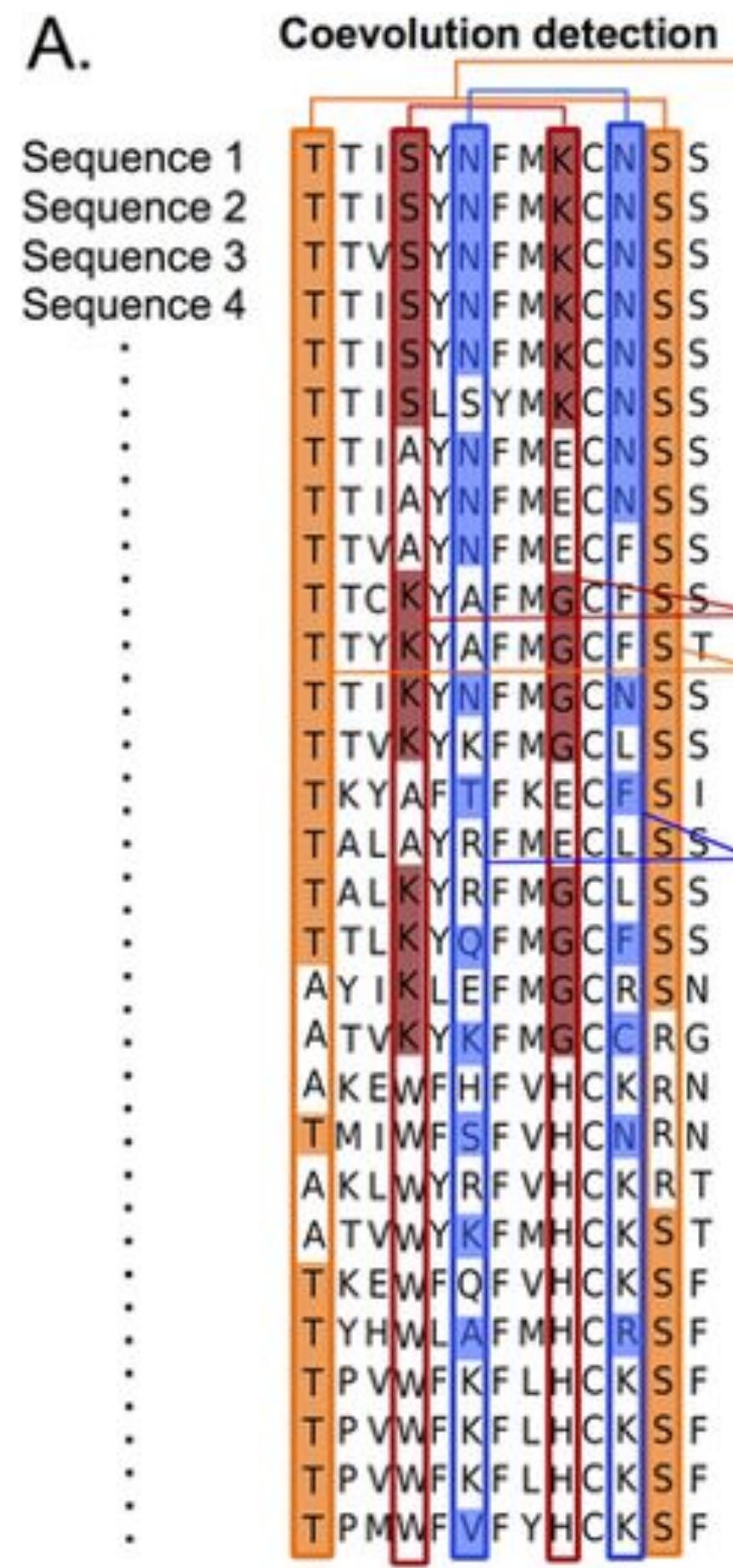
This Direct Information (DI) or Direct Coupling Analysis (DCA) can be obtained by mean field approaches, pseudolikelihood or Boltzmann Learning



2012: From Mutual to Direct Information



From Sequence->Structure to MSA->Structure

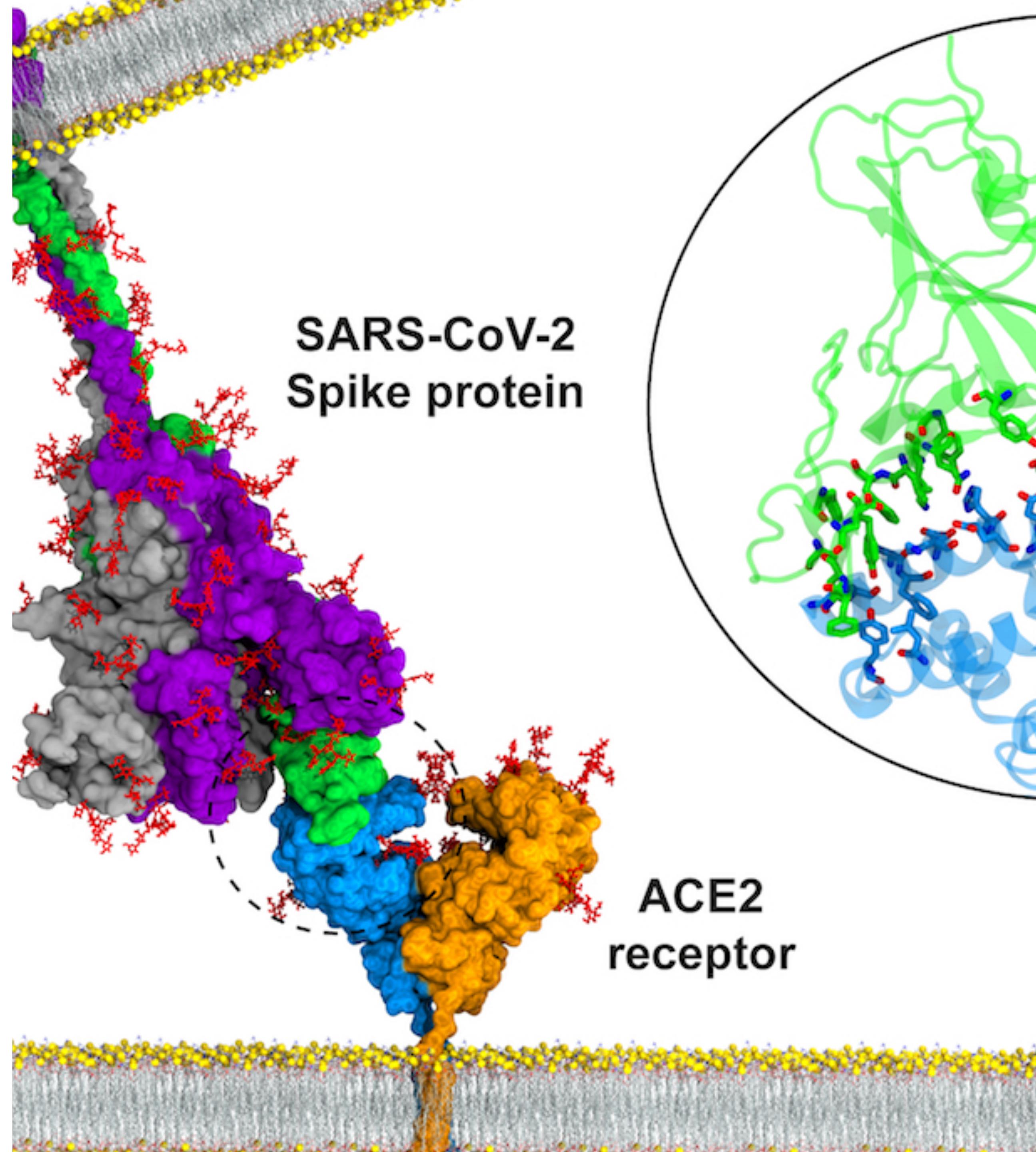


From 2012 it has been clear that when thousands of homologues sequences are available the problem of finding a representative structure associated to them can be solved, but in general they are not available for difficult cases.

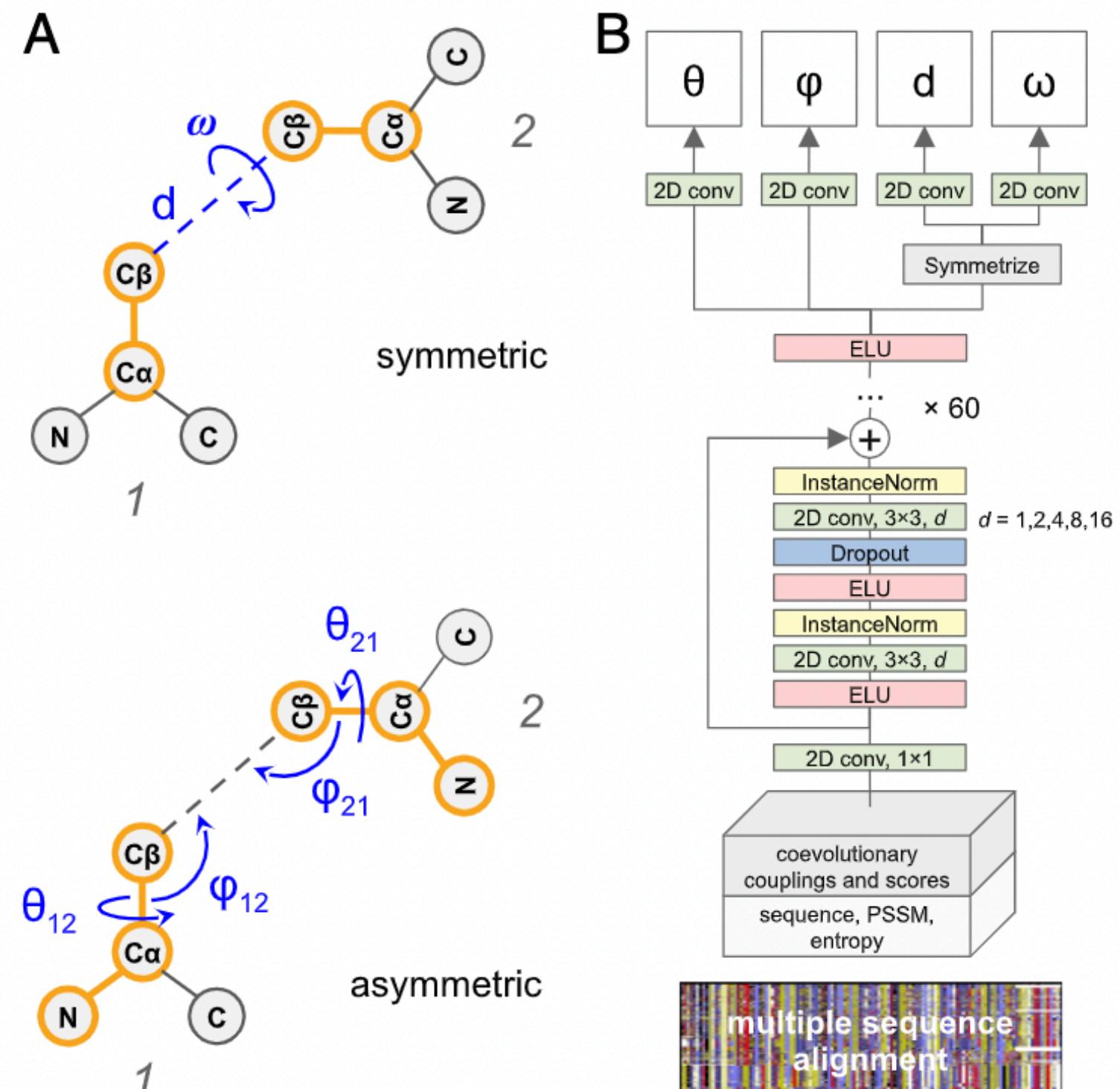
How to improve the quality of the prediction when only few hundreds or less homologue sequences are available?

Outline

- Structure prediction: the origins
- Structure prediction: key advances
- **State of the art and AI approaches**
- Protein Design
- Challenges: conformational changes, effect of mutations, binding, quaternary structure



CASP13 - AlphaFold, RaptorX and TrRosetta: deep-learning distances and orientations from coevolution features



~15,000 proteins structures PDB
Histograms for the distances among all couples of amino acids (distance distributions):

ALA-ALA: distances from all couples in all selected PDBs

ALA-CYS:
ALA-ASP:....

All sequences are also associated to their relative MSA derived covariance matrix

A deep neural network is trained to go from the covariance matrix to the distance distribution

NOTE: we move from contact, binary, maps to distance distribution maps, this allow doing direct minimisation of the structure using the propagation property of the network



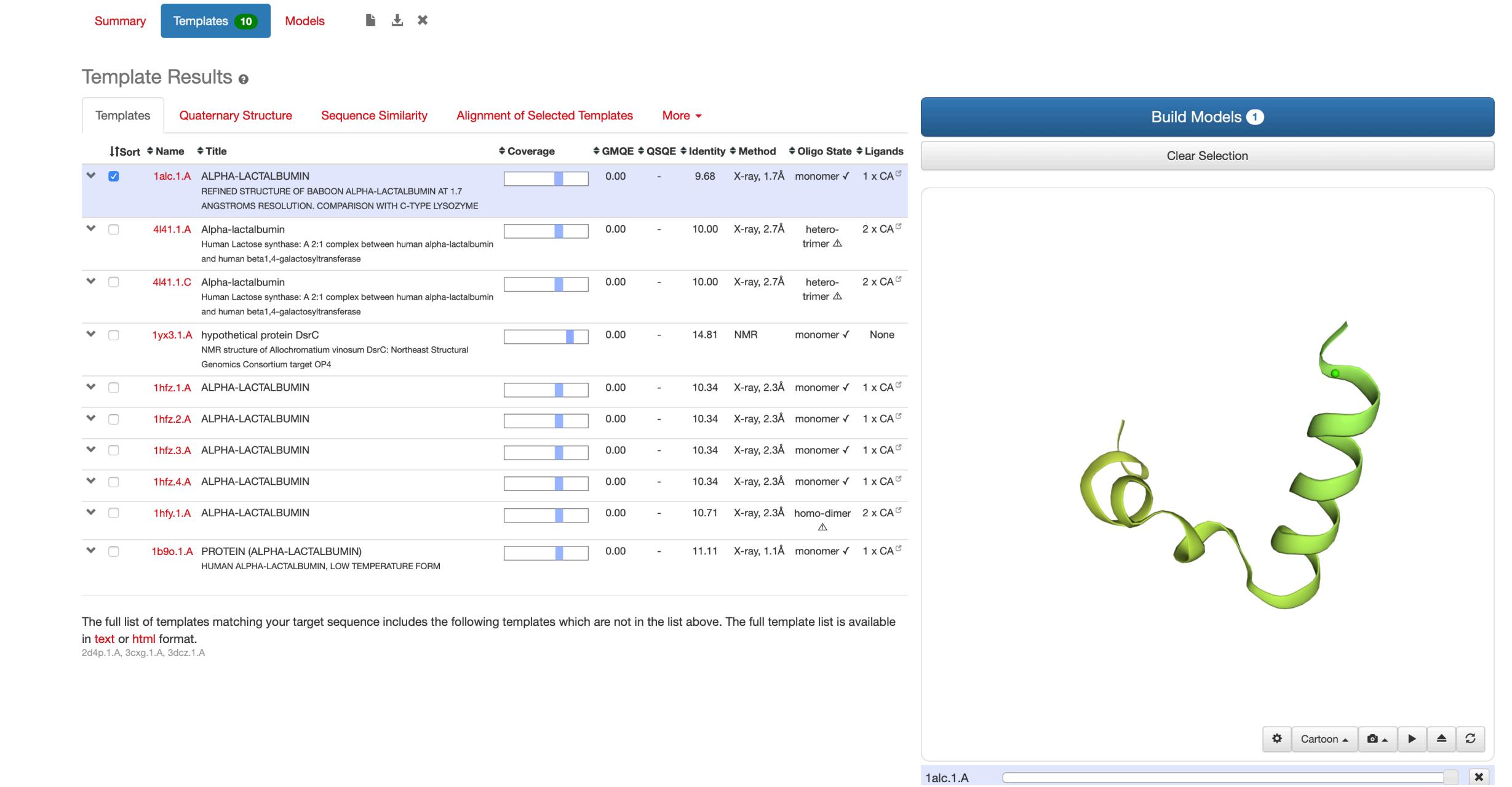
MIZ1: an example from my own experience

MIZ1 is a protein that is studied here in the department for its functional role in plants were it plays a key role in root development. Its structure is unknown and we wanted to determine it, or at least to help in design a construct for further experiments, because the full length protein gets into the inclusion bodies in E. Coli.

MVPYQELTLQRSFSYNSRKINPVTSPARSSHVRSPSSALIPSIPEHEFLVPCRRCSYV
PLSSSSSASHNIGKFHLKFSLLRSFINIINIPACKMLSLPSPPPSSSSVSNQLISLVTG
GSSSLGRRVTGTYGHKRKGHTFSVQYNQRSDPVLLLLAMSTATLVKEMSSGLVRIALE
CEKRHRSGTKLFQEPKWTMYCNGRKCGYAVSRGGACTTDWRLNTSRVTVGAGVIPTP
KTIDDVSGVGSGTELGELLYMRGKFERVVGSRDSEAFYMMNPDKNGGPELSIFLLRI

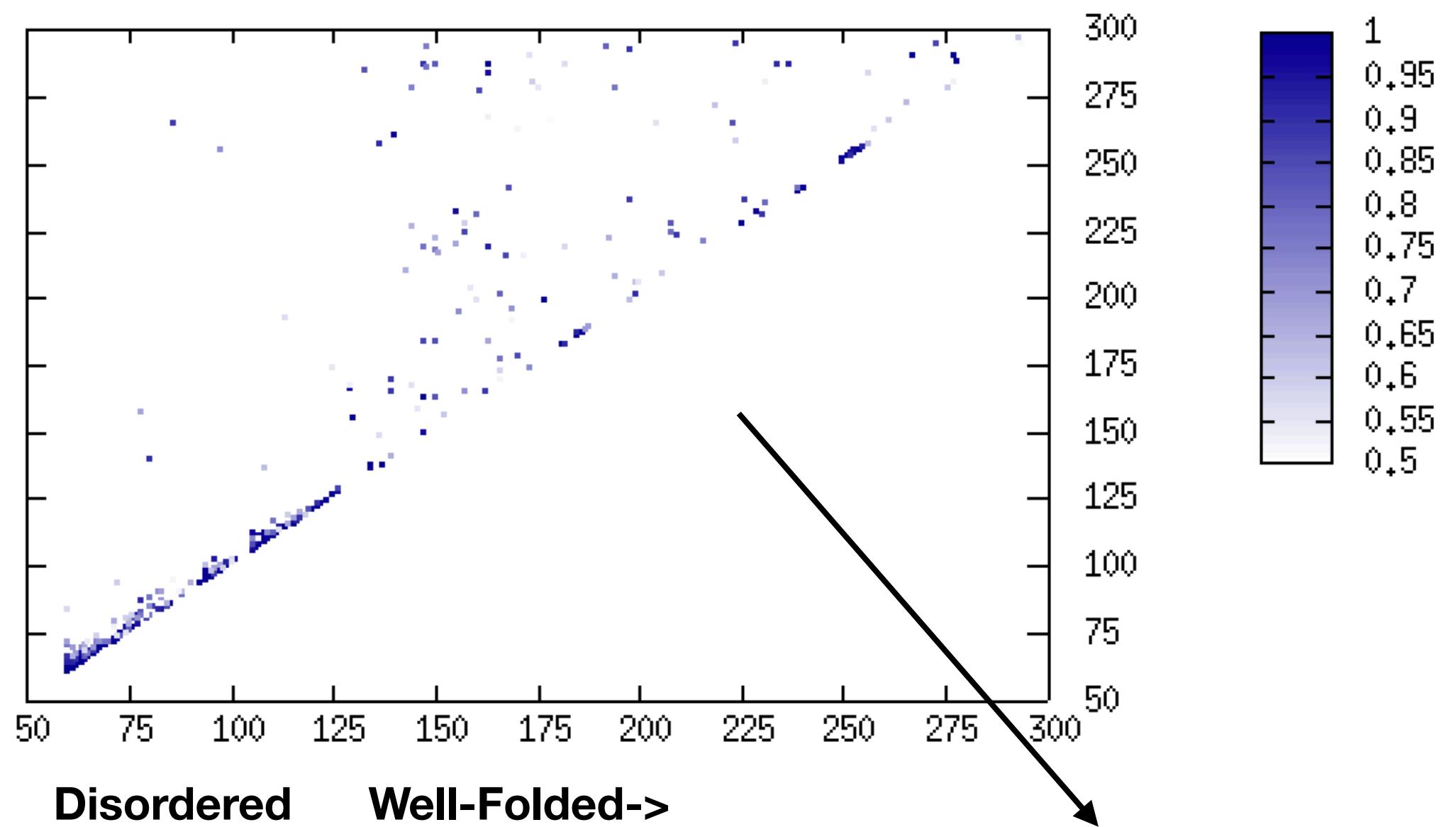
First test - homology modelling - but there are no homologues with known structure:

NOPE!



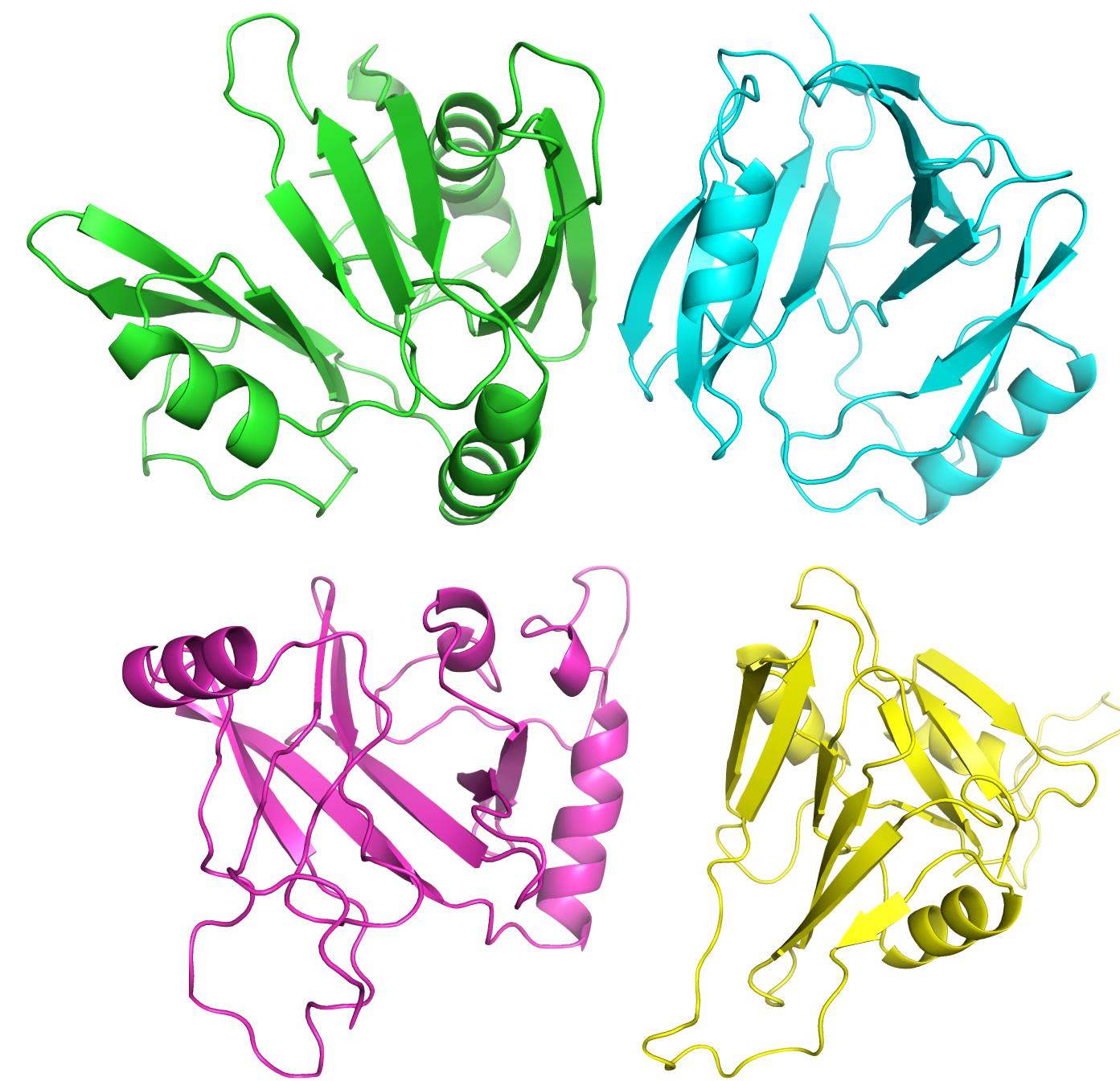
MIZ1: an example from my own experience

MSA and coevolution analysis, we found ~700 homologues sequences, that is a small number for pre-AF methods.



Yet, we used the map to design a construct to produce the protein that actually worked well

Robetta + ev-couplings - ‘old’ ab-initio structure prediction:



Four structures with nothing in common... nope.

MIZ1: an example from my own experience

RaptorX

(first time of ev-couplings+deep learning):

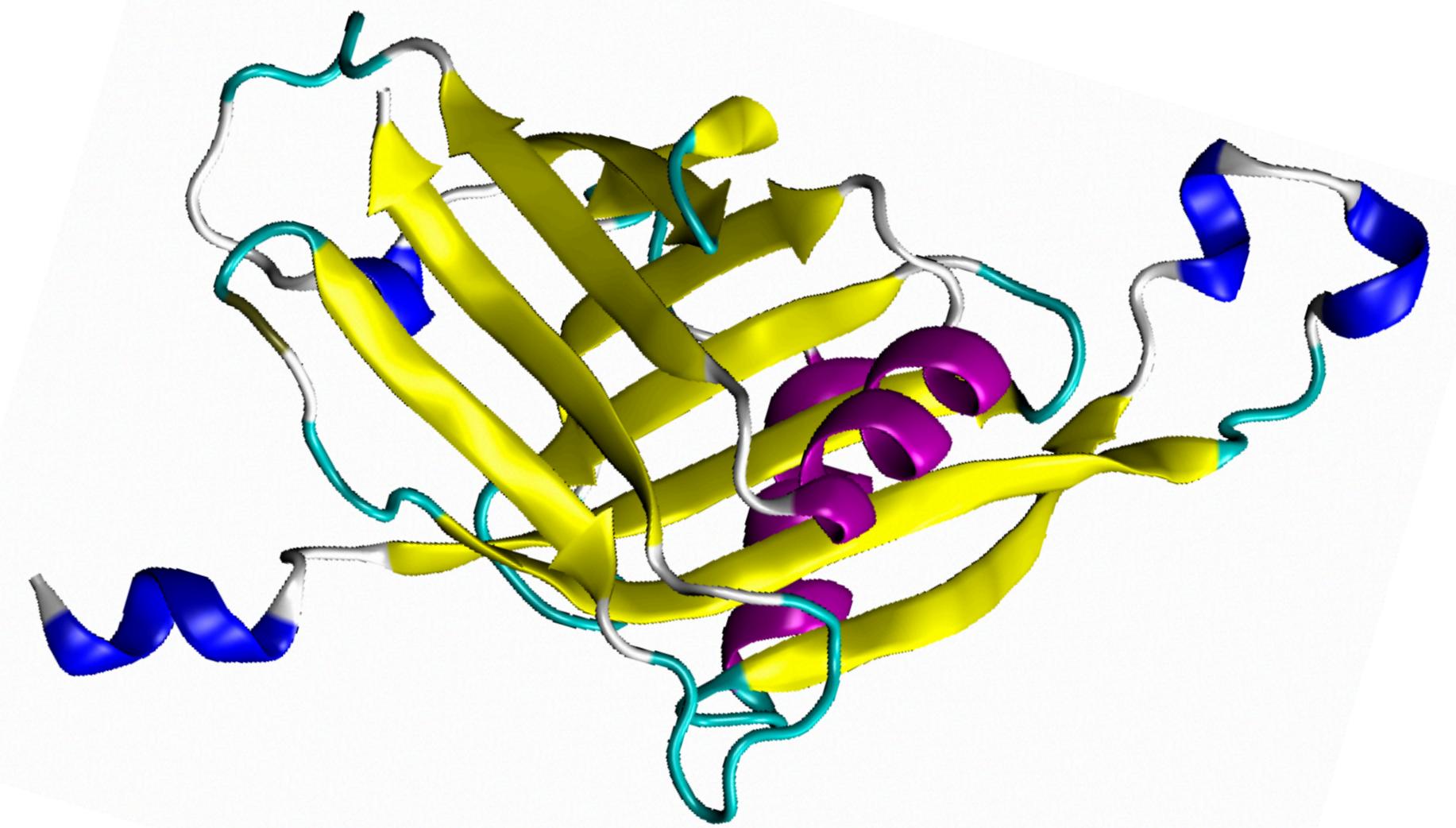


**Five structures with something in common,
but when looked in the details they are
unphysical**

NOTE 1: this is a case with ~700 hundred homologue sequences, this is still a relatively large number

trRosetta

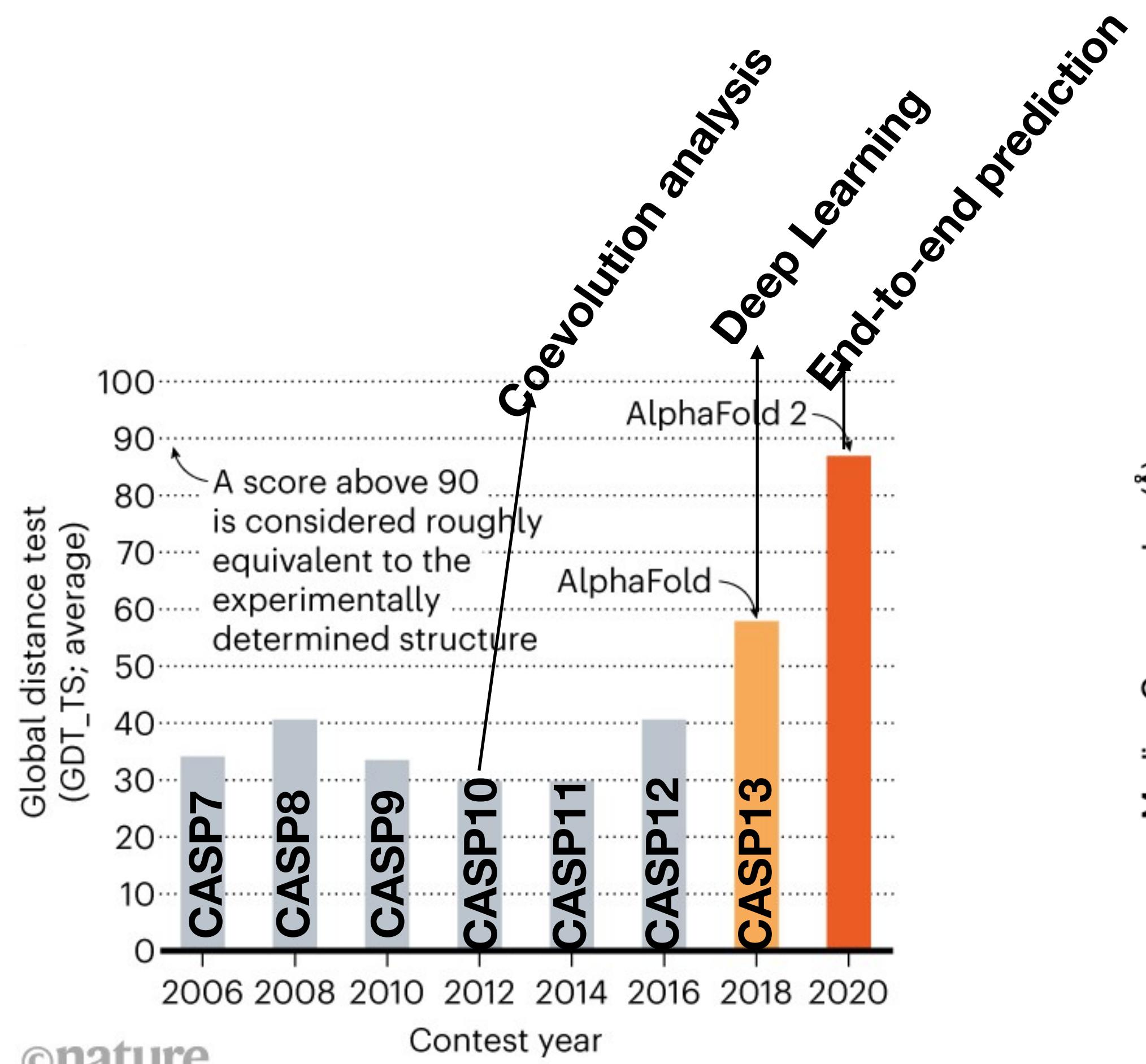
(post AF ev-couplings+deep learning):



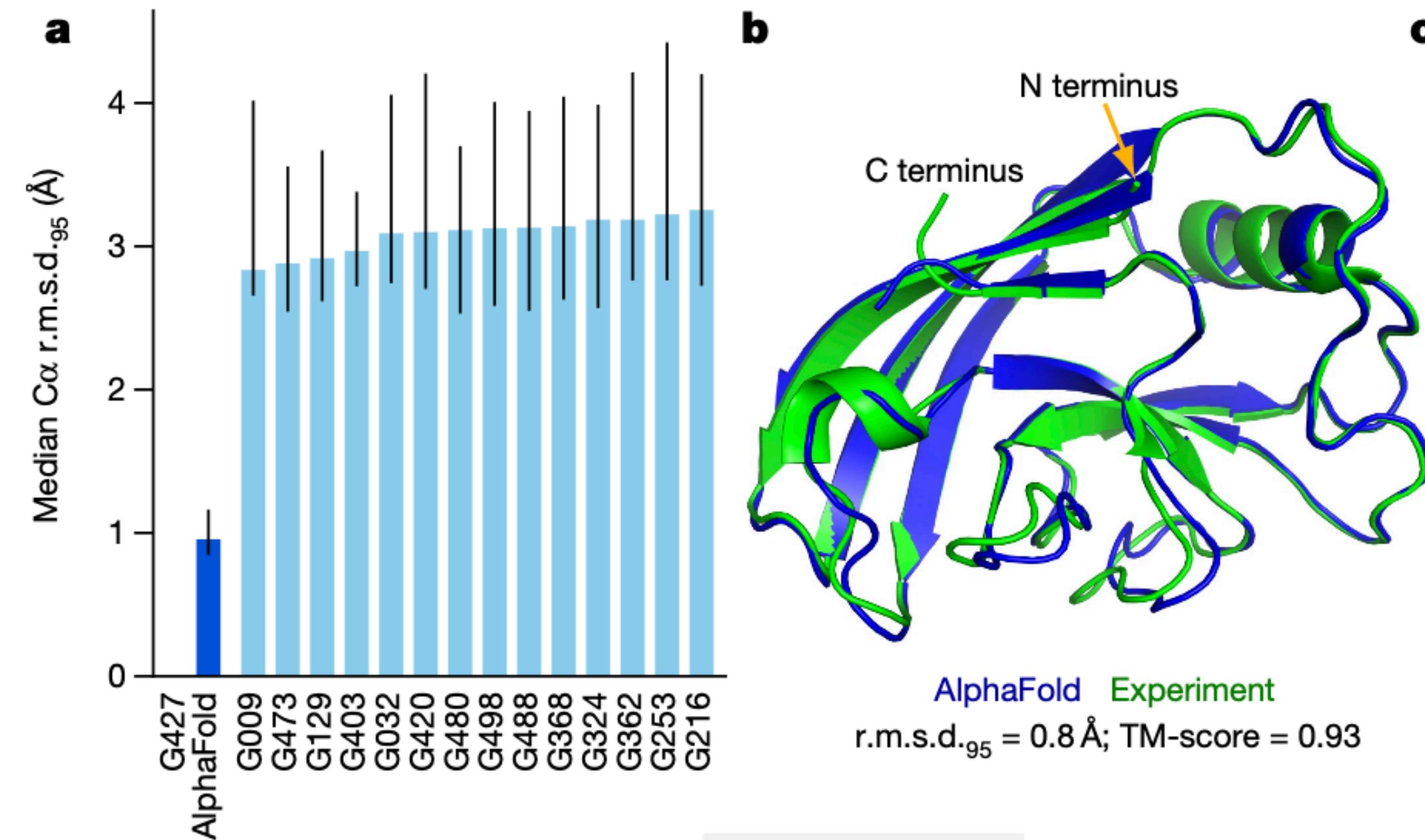
Well converged, well-defined unique structural model with perfectly reasonable features.

NOTE 2: I am talking about trRosetta and not AF, because the first AlphaFold as never been made publicly available

CASP14 - AlphaFold2: not just a technical improvement but a completely ad hoc newly designed network architecture



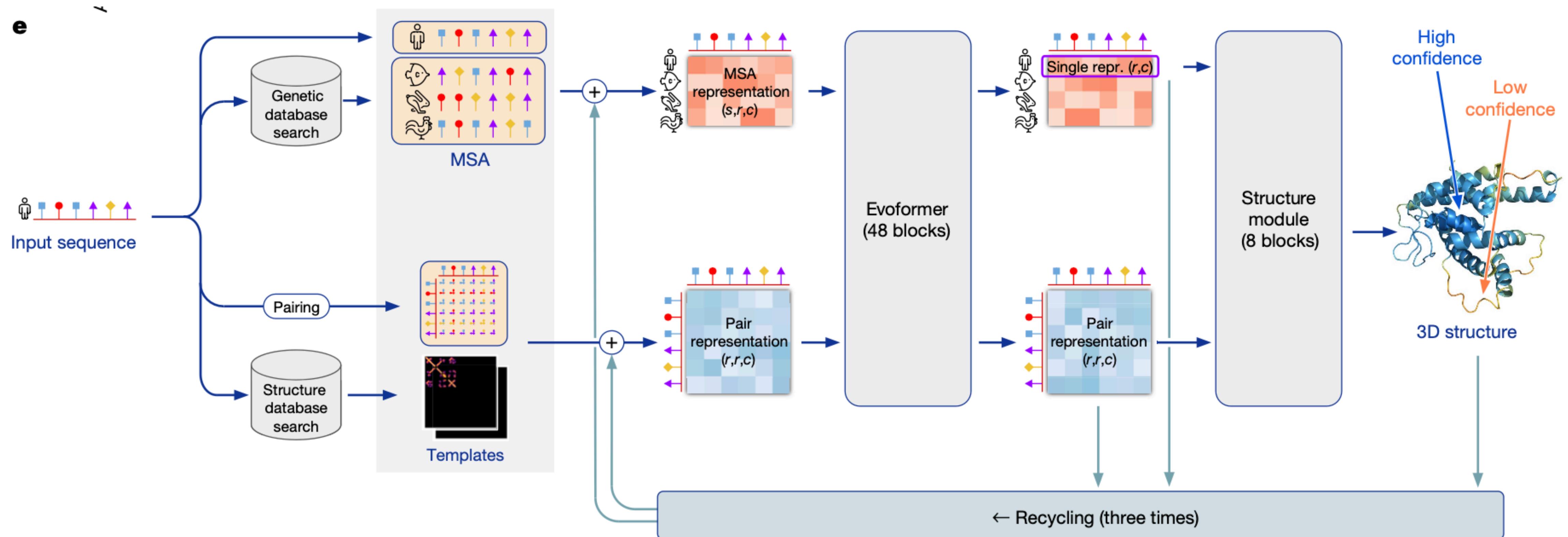
How to improve the quality of the prediction when less than one hundred homologue sequences are available?



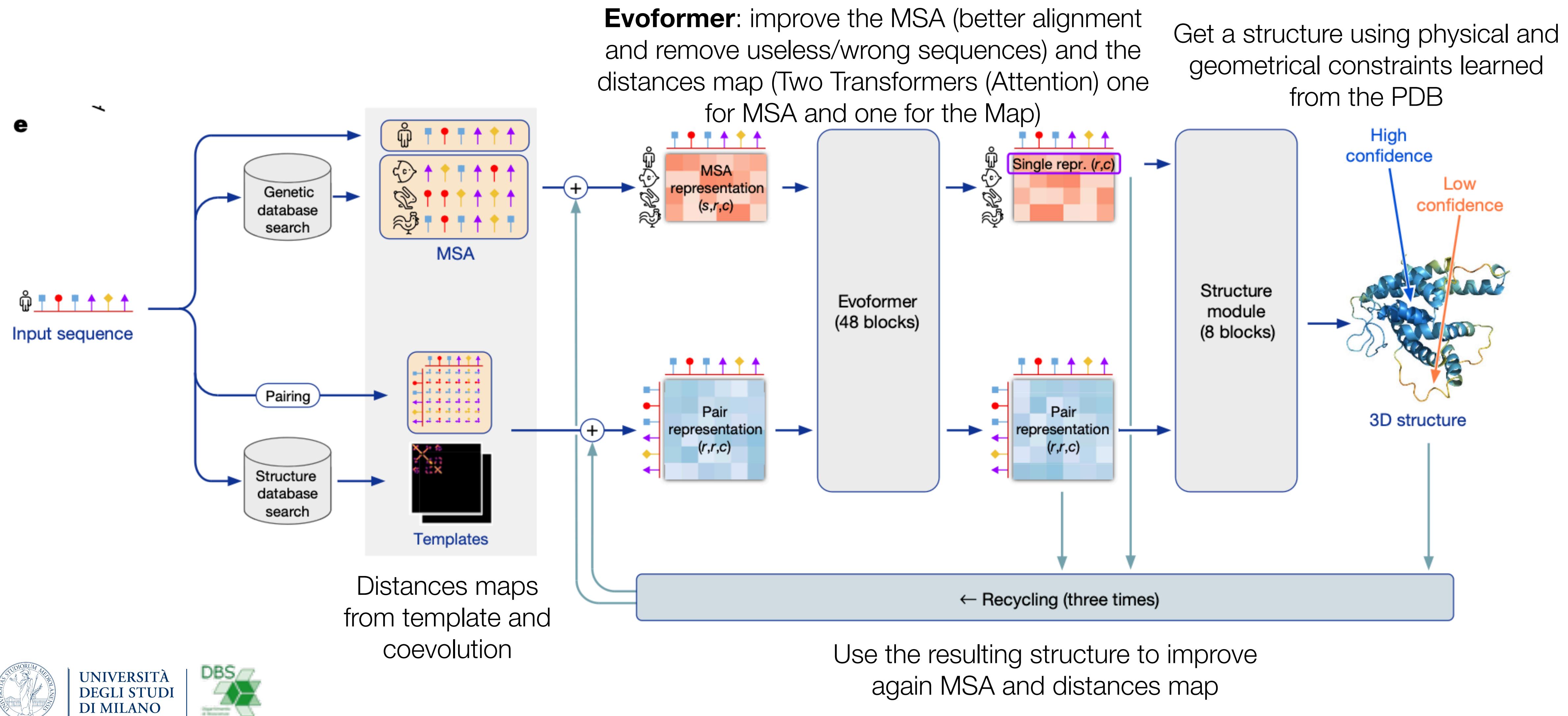
AlphaFold2: some ideas on the architecture

In pre AF2 approaches the network allowed to increase the amount of information extracted from the MSA.

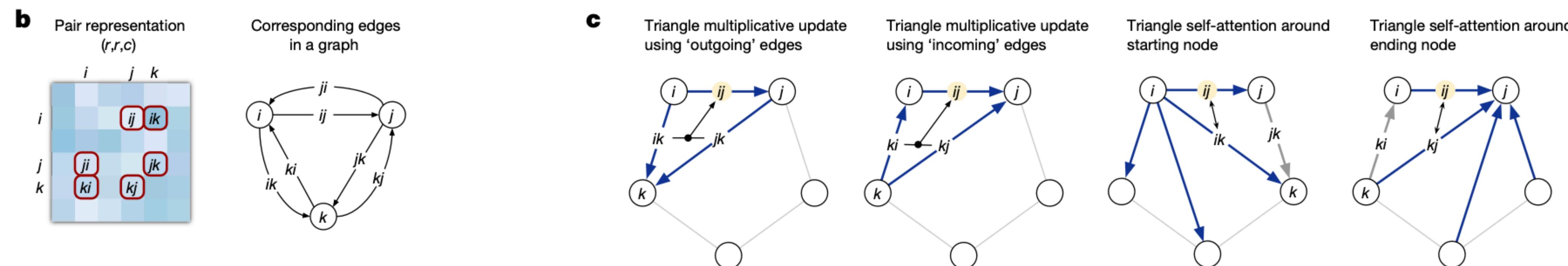
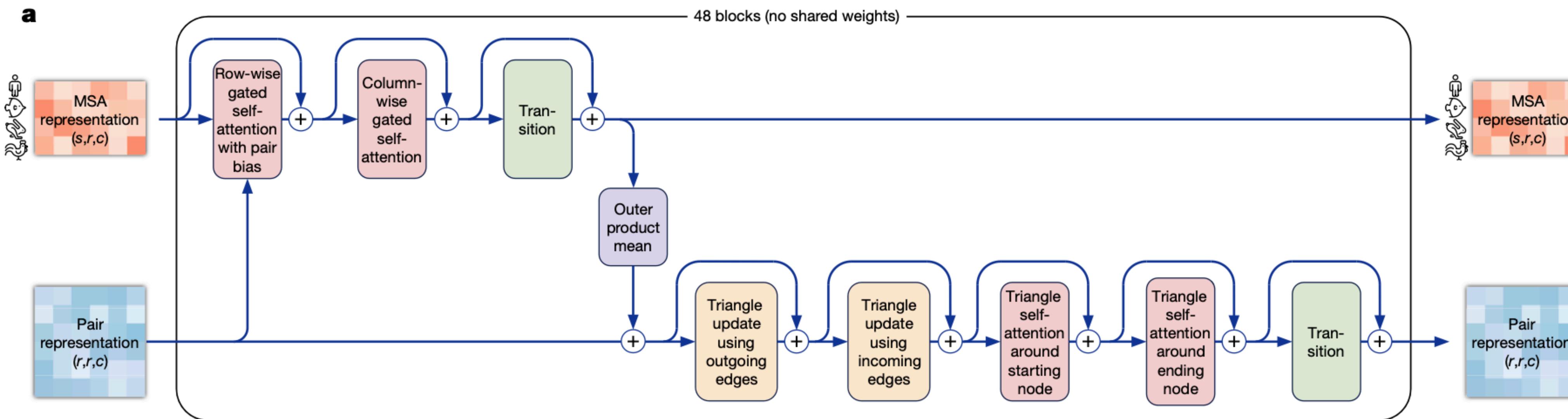
AF2 first addition is to improve the quality of the MSA itself. The second addition is to have an end-to-end network instead than using the network to generate the distance distribution map and then applying it to a conventional fragment replacement method.



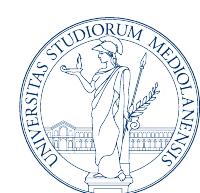
AlphaFold2: some ideas on the architecture



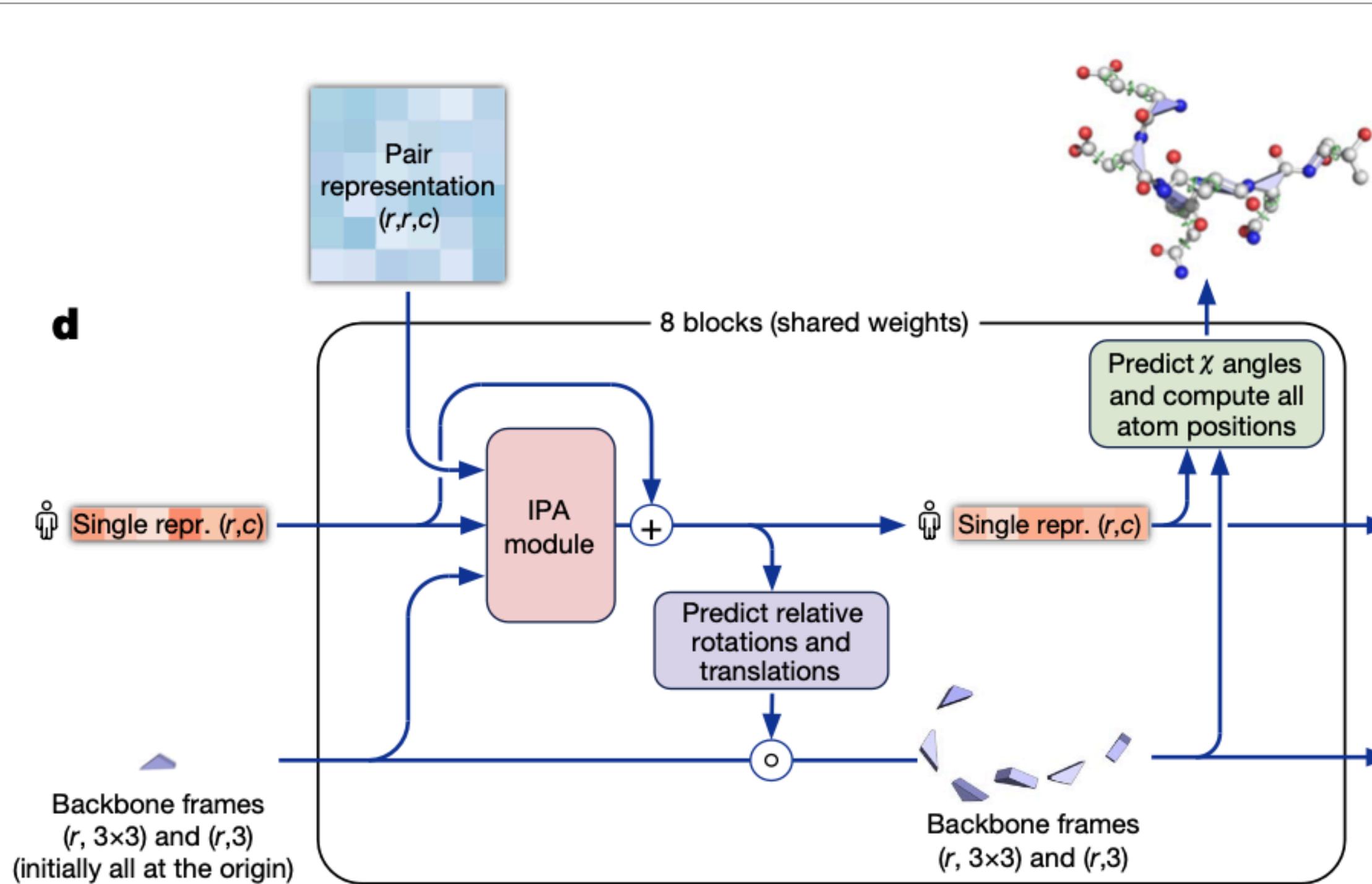
AlphaFold2: Evoformer



The Distances Map (Pair representation) has to obey the triangle inequality.

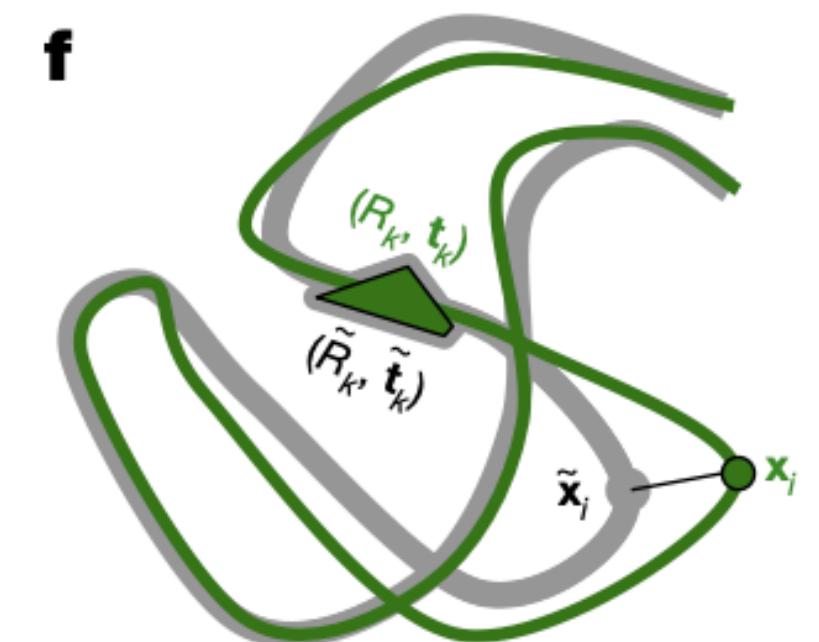
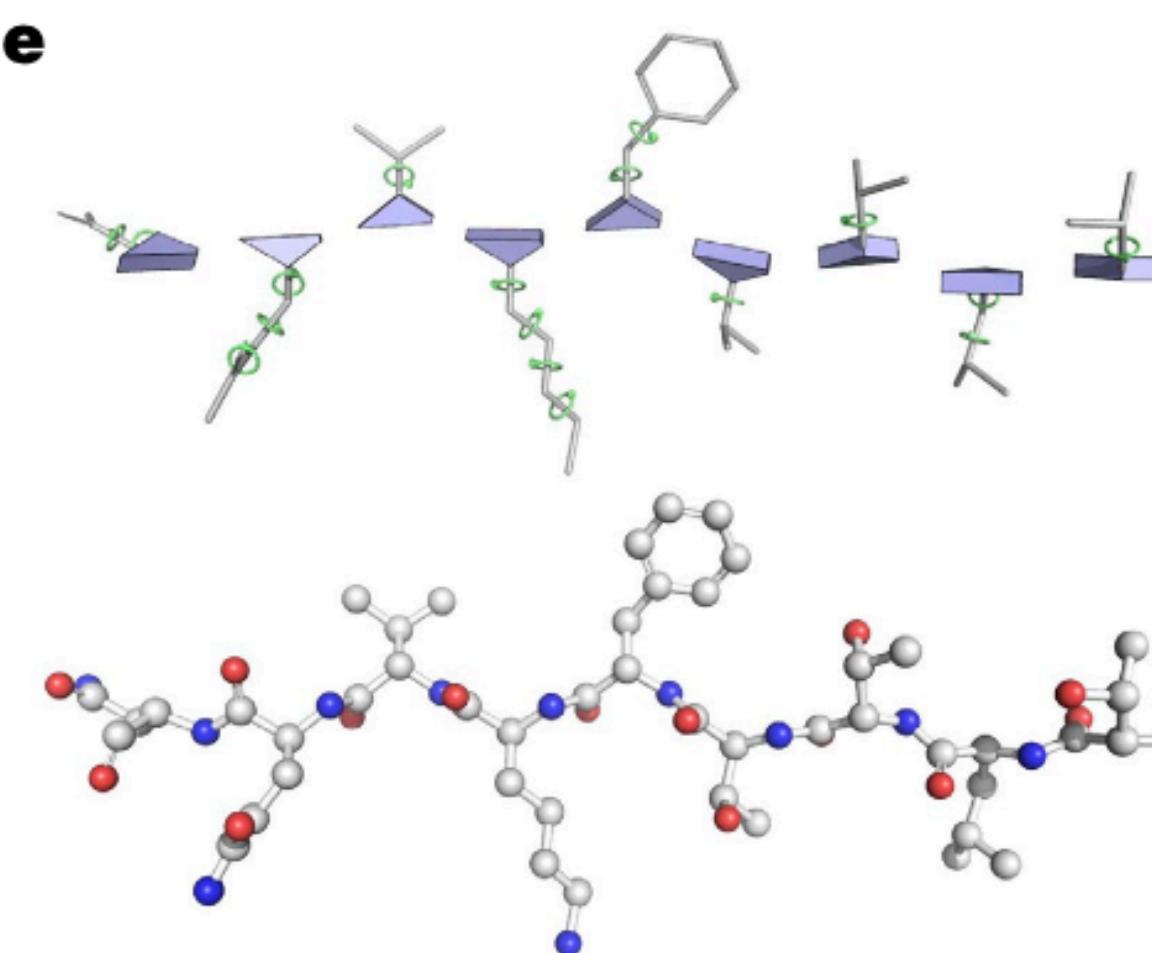


AlphaFold2: Structure Module



The single repr are Nres matrices representing the rotation and translation of the residue.

$$\mathbf{M} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$



The structure module considers the protein as a “residue gas”. Every amino acid is modelled as a triangle, representing the three atoms of the backbone. These triangles float around in space, and are moved by the network to form the structure. The structure module considers the protein as a “residue gas”. Every amino acid is modelled as a triangle, representing the three atoms of the backbone. These triangles float around in space, and are moved by the network to form the structure.

AlphaFold2: Loss Function

The network is trained end-to-end, with gradients coming from the main Frame Aligned Point Error (FAPE) loss $\mathcal{L}_{\text{FAPE}}$ and a number of auxiliary losses. The total per-example loss can be defined as follows

$$\mathcal{L} = \begin{cases} 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} & \text{training} \\ 0.5\mathcal{L}_{\text{FAPE}} + 0.5\mathcal{L}_{\text{aux}} + 0.3\mathcal{L}_{\text{dist}} + 2.0\mathcal{L}_{\text{msa}} + 0.01\mathcal{L}_{\text{conf}} + 0.01\mathcal{L}_{\text{exp resolved}} + 1.0\mathcal{L}_{\text{viol}} & \text{fine-tuning} \end{cases}, \quad (7)$$

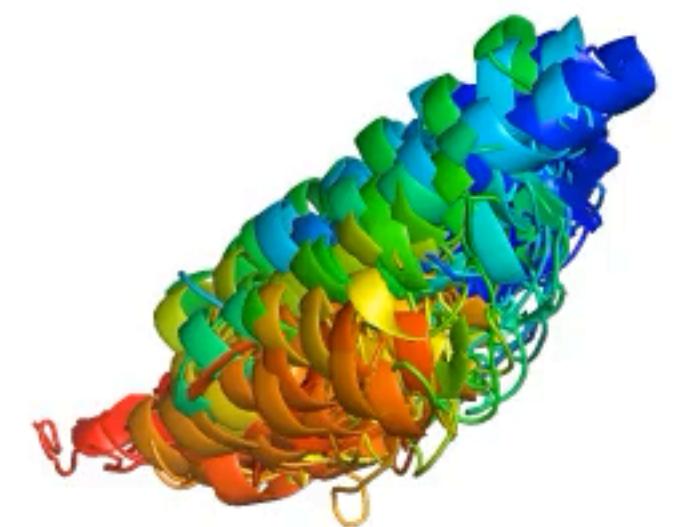
where \mathcal{L}_{aux} is the auxiliary loss from the Structure Module (averaged FAPE and torsion losses on the intermediate structures, defined in [Algorithm 20 line 23](#)), $\mathcal{L}_{\text{dist}}$ is an averaged cross-entropy loss for distogram prediction, \mathcal{L}_{msa} is an averaged cross-entropy loss for masked MSA prediction, $\mathcal{L}_{\text{conf}}$ is the model confidence loss defined in [subsubsection 1.9.6](#), $\mathcal{L}_{\text{exp resolved}}$ is the “experimentally resolved” loss defined in [subsubsection 1.9.10](#), and $\mathcal{L}_{\text{viol}}$ is the violation loss defined in [subsubsection 1.9.11](#). The last two losses are only used during fine-tuning.

To decrease the relative importance of short sequences, we multiply the final loss of each training example by the square root of the number of residues after cropping. This implies equal weighting for all proteins that are longer than the crop size, and a square-root penalty for the shorter ones.

$\mathcal{L}_{\text{FAPE}}$ is a clever deviation of the atomic position with respect to the experiment



Examples:



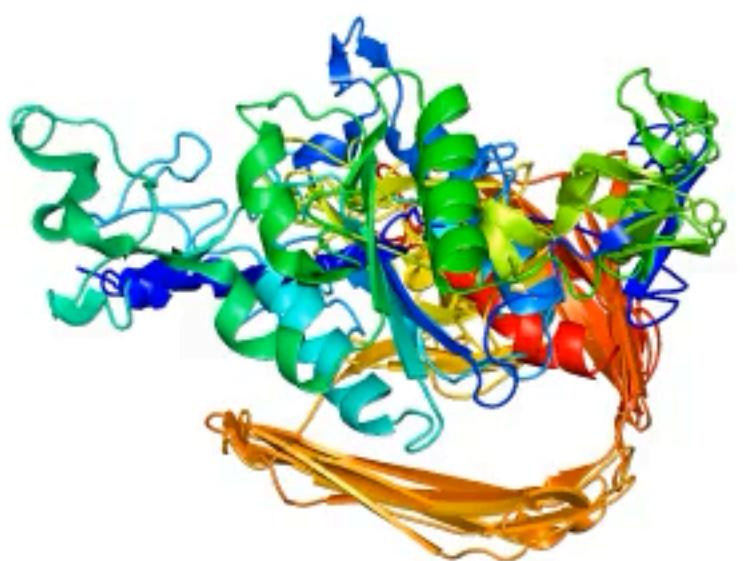
Recycling iteration 0, block 01
Secondary structure assigned from the final prediction



UNIVERSITÀ
DEGLI STUDI
DI MILANO



Examples:



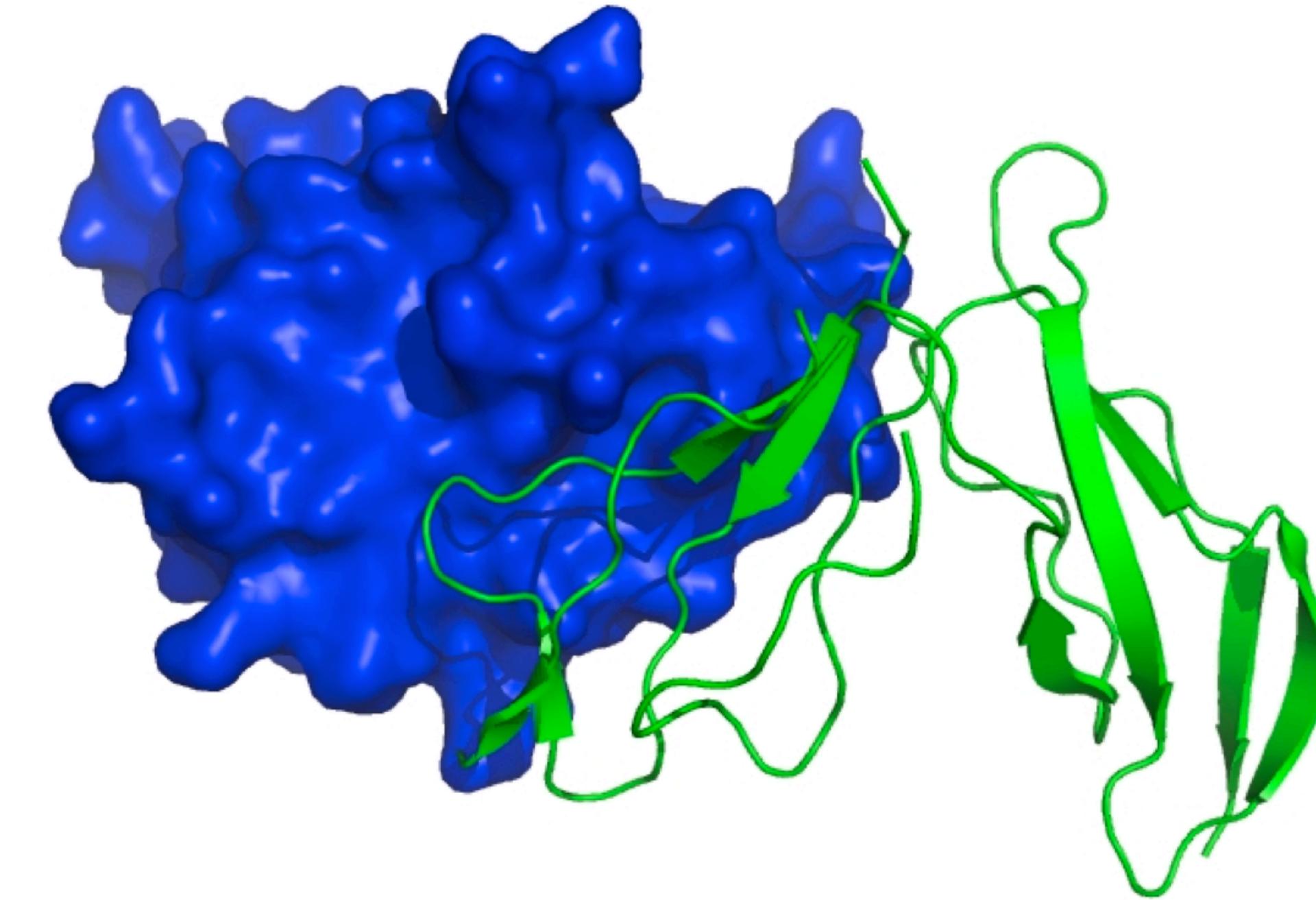
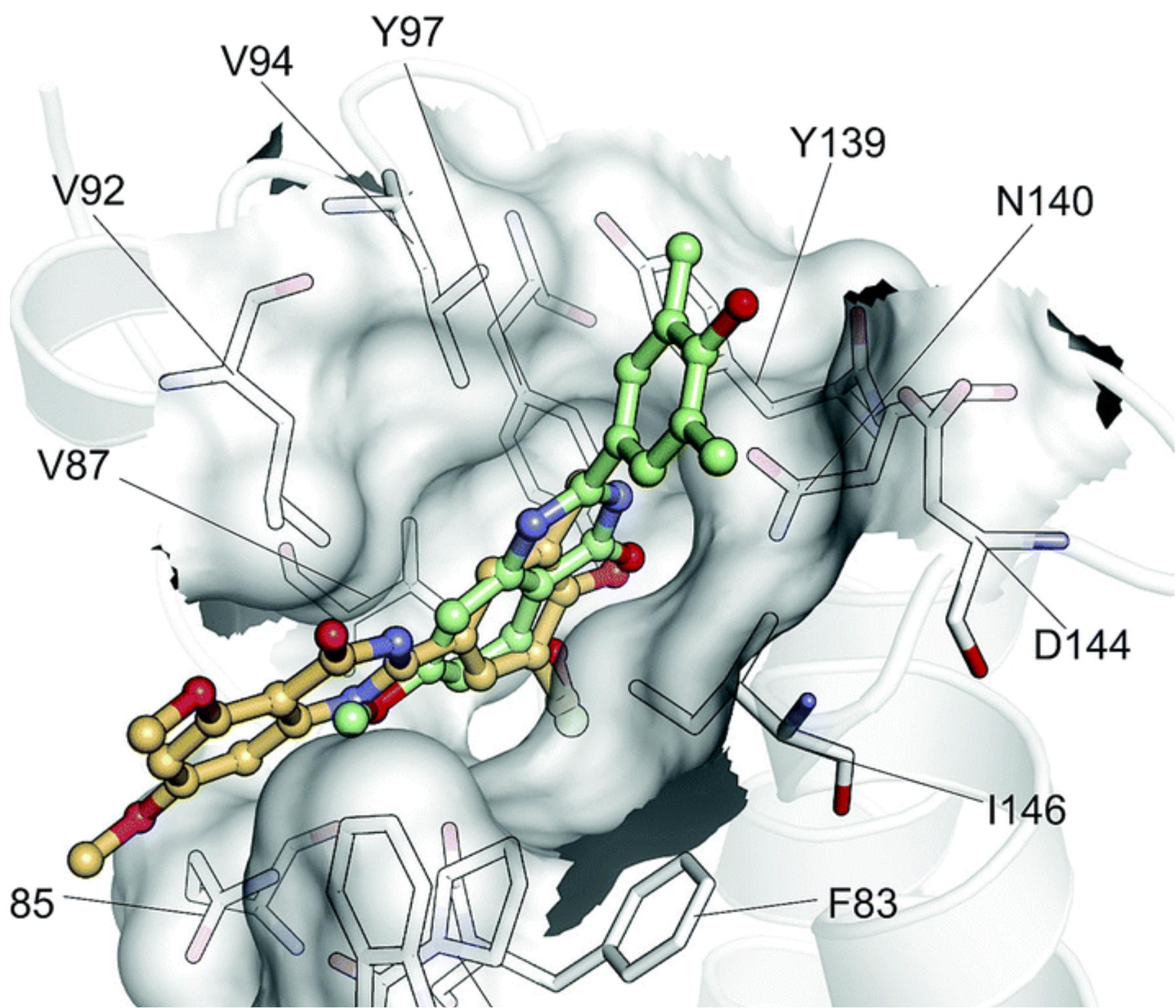
Recycling iteration 0, block 01
Secondary structure assigned from the final prediction



UNIVERSITÀ
DEGLI STUDI
DI MILANO



Molecular Docking: binding and quaternary structures

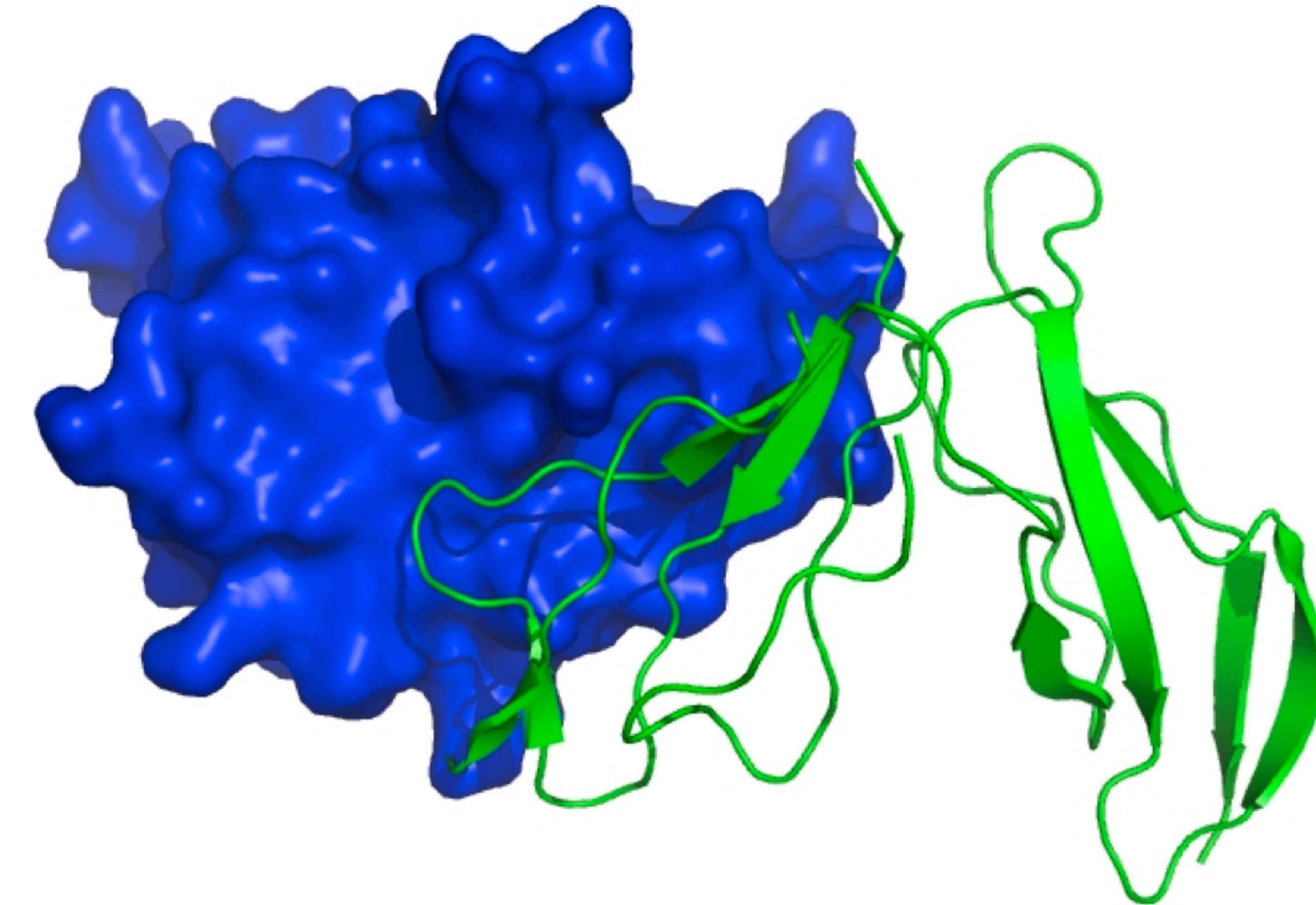
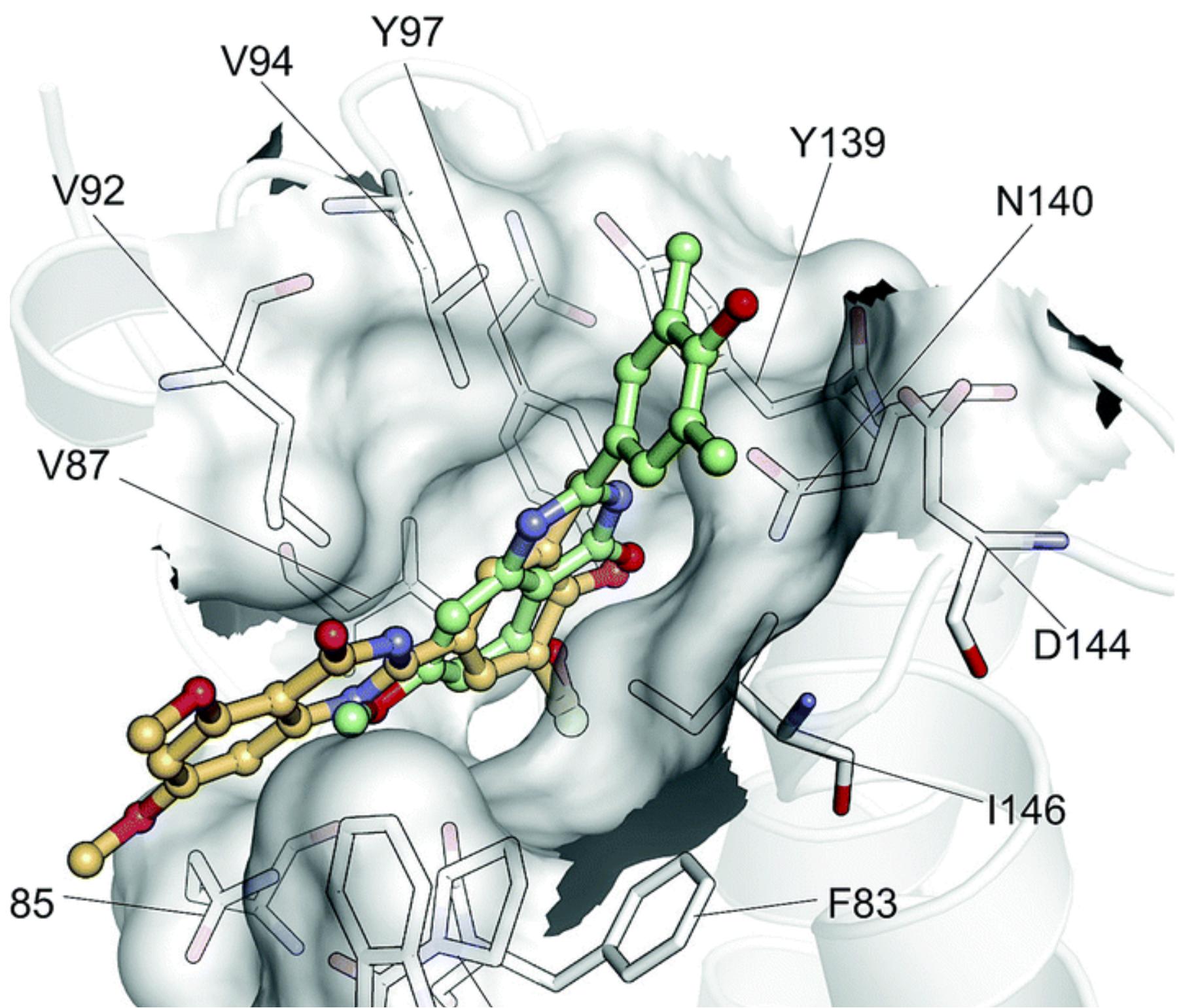


Prediction of the bound structure and binding strength for

- Small molecules (virtual screening, drug discovery, ..)**
- Macromolecular complexes (protein-protein, protein-nucleic, ...)**

Complexation can be associated with conformational changes

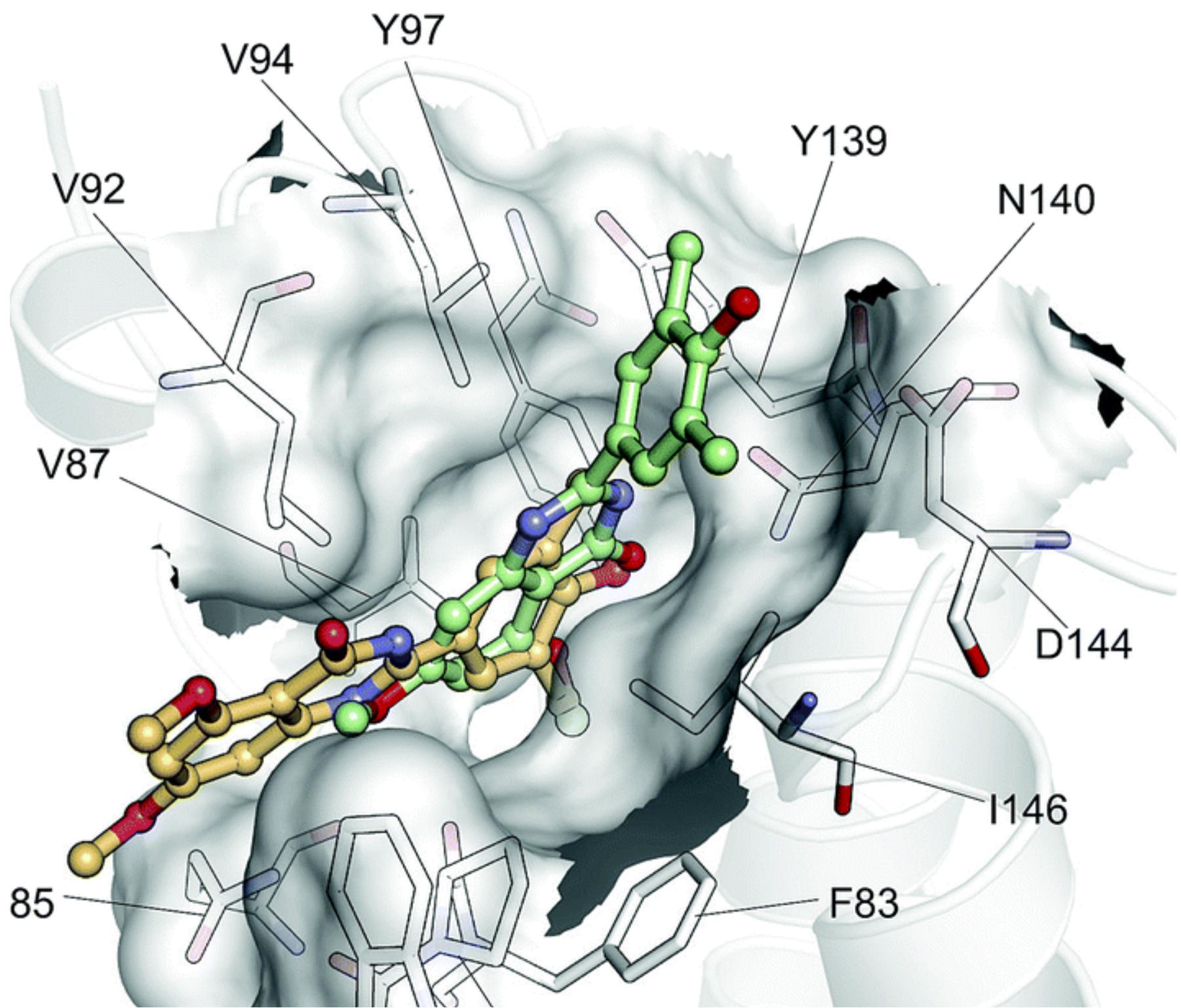
Molecular Docking



The general principles and aims are the same, but the practical implementations are generally very different so you will find different tools for the two cases

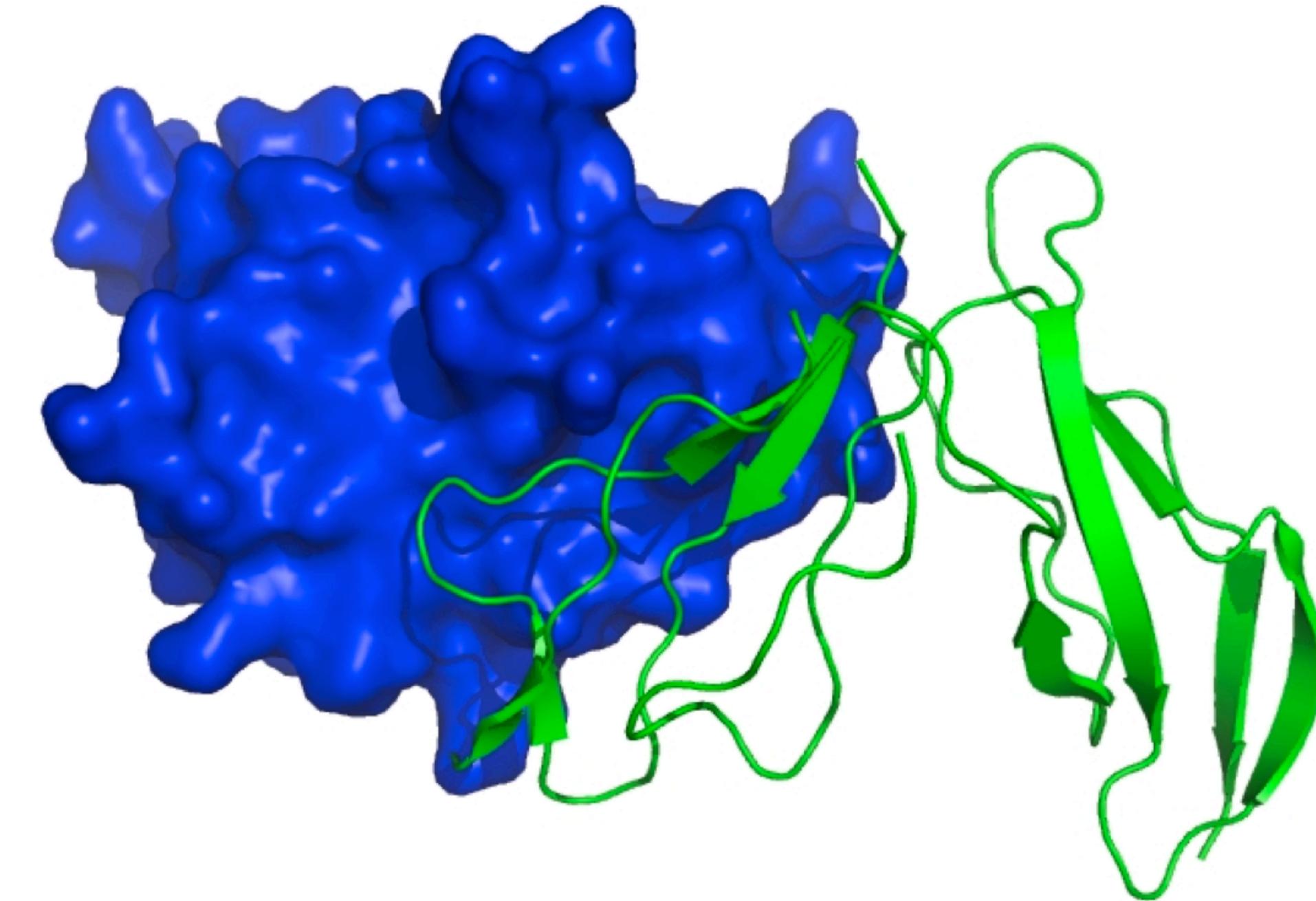


Molecular Docking



**Autodock vina
UCSF dock
Glide**

...



**Haddock
Rosetta Dock
ClusPro**

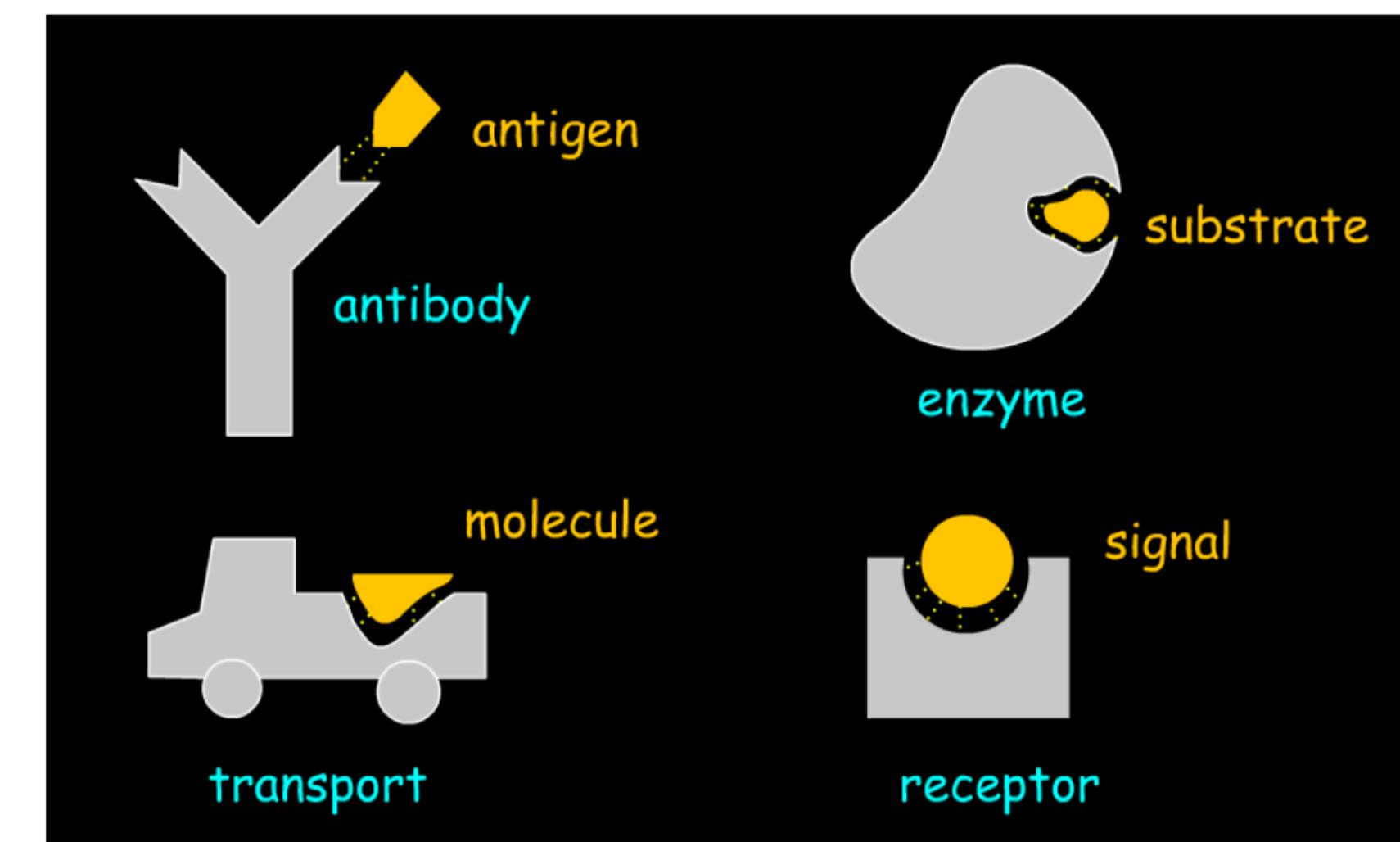
...



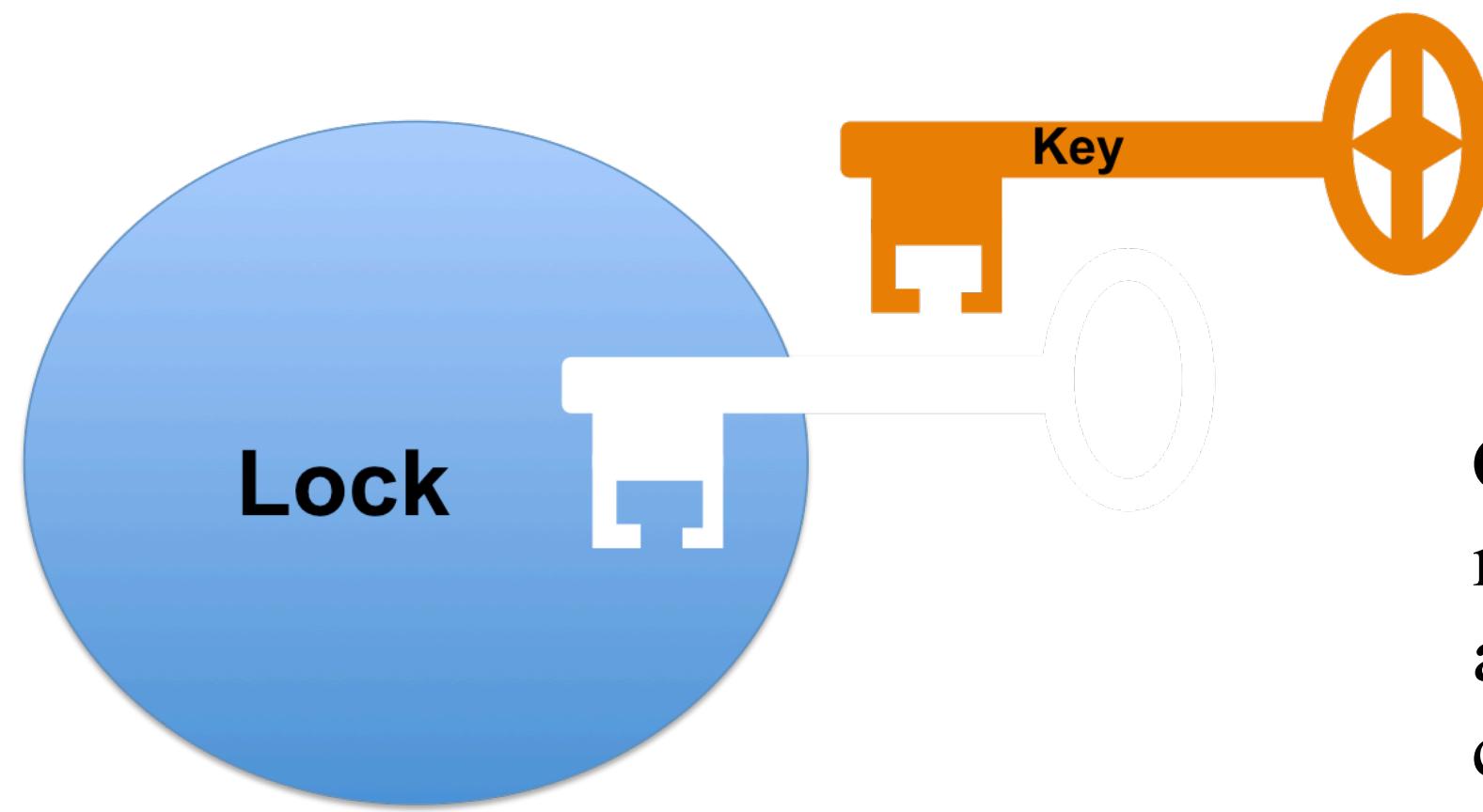
Molecular Recognition and Docking

Molecular recognition is the ability of biomolecules to recognize other biomolecules and selectively interact with them in order to promote fundamental biological events such as transcription, translation, signal transduction, transport, regulation, enzymatic catalysis, viral and bacterial infection and immune response.

Molecular docking is the process that involves placing molecules in appropriate configurations to interact with a receptor. Molecular docking is a natural process which occurs within seconds in a cell.



Lock and Key



Emil Fischer (1894)

Specificity in enzyme-substrate recognition

Generally speaking the idea is that specificity in molecular recognition is the result of rigid surface that as a consequence can bind together only when exactly complementary (**Lock and Key model**)

Docking algorithm: (0 level)

- analyse surface of the two molecules
- find possible regions of binding (surface compatibility -> shape and electrostatic interactions)
- minimise the energy.

Actually in many cases recognition mechanisms are more complex, with in principle flexibility on both the receptor and the ligand
(Flexible Docking)



UNIVERSITÀ
DEGLI STUDI
DI MILANO



Docking mechanisms

Lock-and-Key: this is simple, the assumption is that both the receptor and the ligand are rigid and their structure is ideal for binding. The next step is to accept that the ligand (this is more specific for the binding of small molecules) can be flexible, while the receptor is rigid.

Induced Fit: In 1958 Daniel Koshland introduced the "induced-fit theory". The basic idea is that in the recognition process, both ligand and target mutually adapt to each other through small conformational changes, until an optimal fit is achieved. This is traduced in what is usually called the flexible-docking where both active site region and the ligand are allowed a certain degree of flexibility.

Conformational Selection: this recognises that both the ligand and the receptor can be in many different configurations, and that binding can only occur when relatively optimal configurations are populated. This is traduced in the so called ensemble docking, where multiple configurations of the receptor are used at the same time each can also be considered locally flexible to account for additional induced fit.



Molecular Docking

There are two problems:

- 1. How to generate quickly as many ‘reasonable’ poses as possible (including or not flexibility)**
- 2. How to distinguish (‘score’) a good pose from a bad one (will the true complex score better than all other possible complexes?)**



Ligand-receptor interactions

From calorimetry we can measure binding free energies ($\Delta G_{\text{binding}}$). A free energy is a measure of the effective energy it is possible to extract from a reaction. Furthermore a free energy is always the result of two contributions, enthalpy and entropy.

$$\Delta G_{\text{binding}} = \Delta H_{\text{binding}} - T\Delta S_{\text{binding}}$$

Can we think of what are these contributions?



Enthalpy:

- Internal energy (R, L, S)
- Interaction with the solvent (RS, LS)
- Interaction RL

Entropy:

- Internal entropy of R, L and S



Ligand-receptor interactions



$$\Delta G_0^{\text{bind}} = \Delta H_{\text{RL}} + \Delta H_{\text{RS}} + \Delta H_{\text{LS}} + \Delta H_{\text{R}} + \Delta H_{\text{L}} + \Delta H_{\text{S}} - T(\Delta S_{\text{R}} + \Delta S_{\text{L}} + \Delta S_{\text{S}}).$$

Ligand-Receptor enthalpy is always gained upon binding.

RS and LS enthalpy should be lost due to water leaving the ligand and the receptor.

The ligand often loses entropy upon binding (rotation and translation but also internal motion)

Solvent instead gains entropy (more free molecules)

Receptor could lose or gain entropy (usually lose)



VINA scoring function

This is what is usually called a “scoring function” that is a mathematical object that given some numbers return a single number, e.g. all the x, y, z coordinates of a given configuration put into the scoring function return a single number that is the score for that configuration.

$$\Delta G_{binding} = \Delta G_{gauss} + \Delta G_{repulsion} + \Delta G_{hbond} + \Delta G_{hydrophobic} + \Delta G_{tors}$$

ΔG_{gauss}

Attractive term for dispersion, two gaussian functions

$\Delta G_{repulsion}$

Square of the distance if closer than a threshold value

ΔG_{hbond}

Ramp function - also used for interactions with metal ions

$\Delta G_{hydrophobic}$

Ramp function

ΔG_{tors}

Proportional to the number of rotatable bonds

These are all related to intermolecular interactions. The hydrophobic term can be seen as the one accounting for solvation.



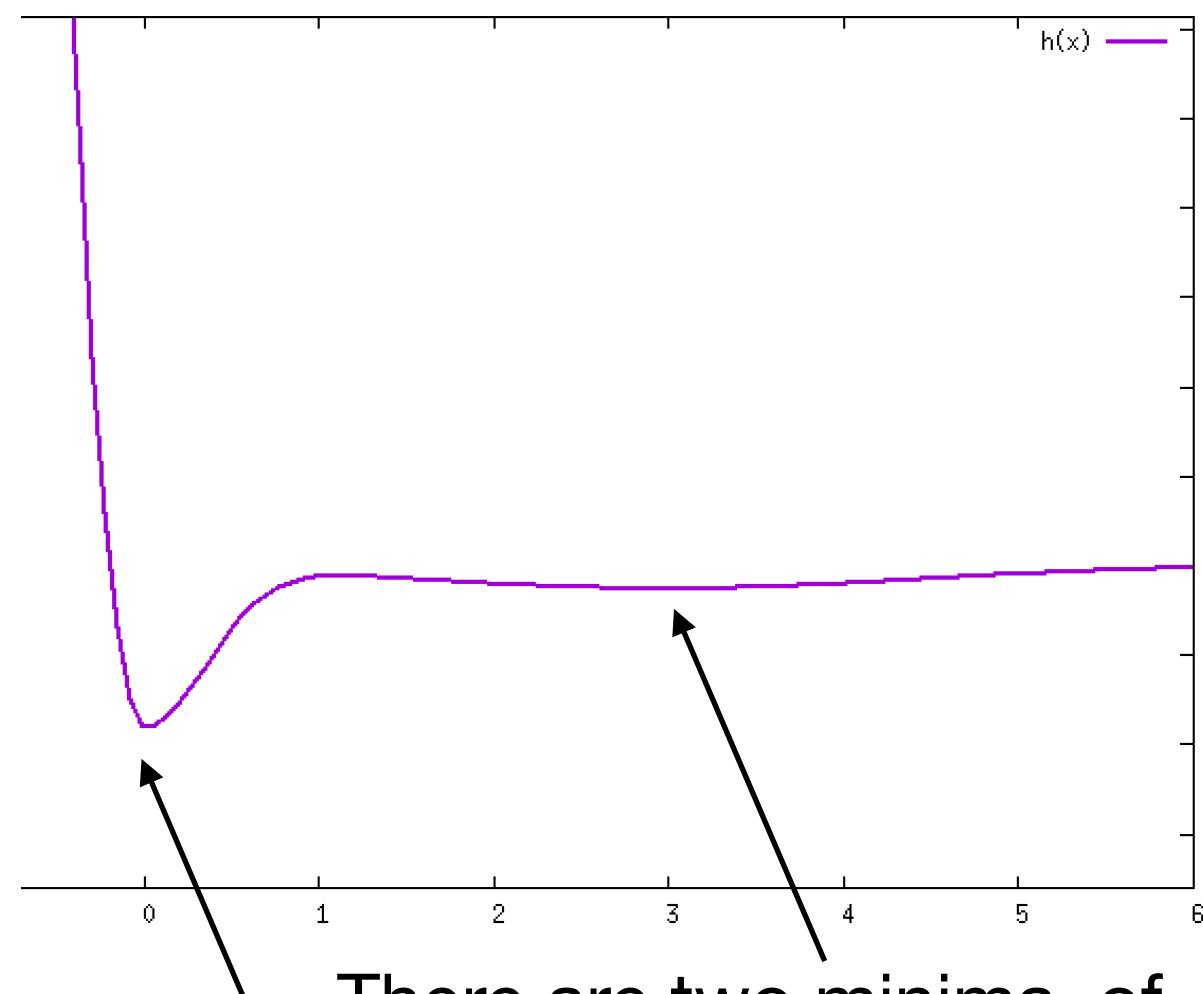
This is then only source
of internal energy



Searching for the best score!

Given a ligand and a receptor what a docking program does is try to find the global minimum of the scoring function.

If you remember your first year class finding the minimum of a 1D function is easy, you calculate the zeros of the first derivative and then calculate the function for those values, the lowest value is the global minimum.



There are two minima, of which the first is global minimum

The problem is that in our case the scoring function is not a 1D function, because the number of variables is $3 \times \text{number of atoms}$. In this case it is not possible to approach the problem analytically but we need to approach it numerically (although it is possible to calculate analytically all the derivatives of the scoring function).

The numerical solution is easier if we simplify the search

- **Rigid docking**
- **Flexible ligand**
- **Both flexible**
- **Ensemble**



Minimization Algorithms: steepest (or gradient) descent

- 1. Starting configuration**
- 2. Calculate the scoring function**
- 3. Calculate all the first order partial derivatives (the gradient)**

$$\nabla f = \frac{\partial f}{\partial x} \mathbf{i} + \frac{\partial f}{\partial y} \mathbf{j} + \frac{\partial f}{\partial z} \mathbf{k},$$

$$f(x, y, z) = 2x + 3y^2 - \sin(z)$$

$$\nabla f = 2\mathbf{i} + 6y\mathbf{j} - \cos(z)\mathbf{k}.$$

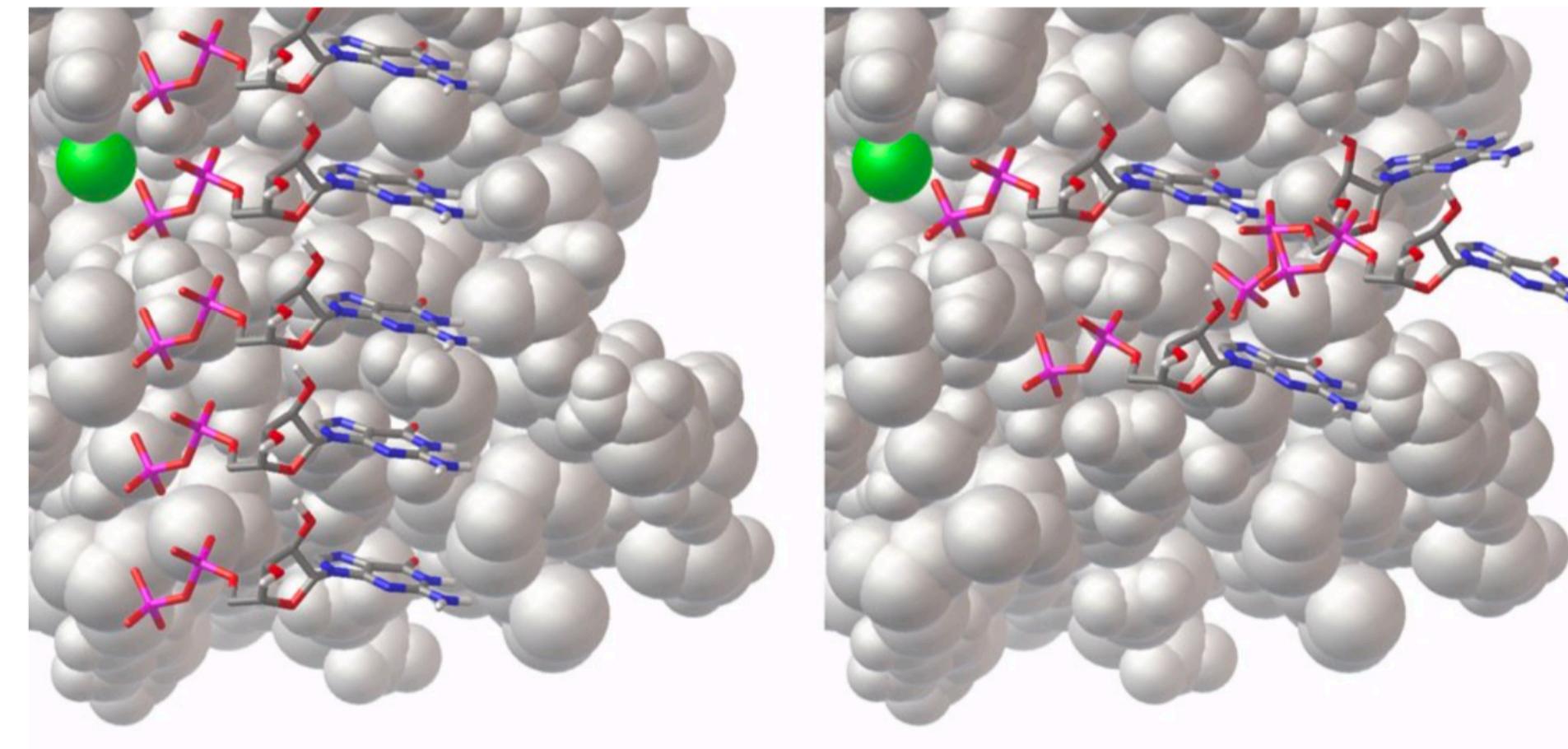
The gradient is the direction of maximum growth of the curve. (if you think in 1D, positive would mean to the right, negative to the left)

- 4. Move following the negative of the gradient** $\mathbf{x}_{n+1} = \mathbf{x}_n - \gamma_n \nabla F(\mathbf{x}_n),$

By definition if γ is small it will stop when the gradient is zero. But also by definition this will be just the closest minimum to the starting configuration.



Searching for the best score!



Systematic

Exhaustive, deterministic
Outcome is dependent on granularity of sampling
Feasible only for low-dimensional problems

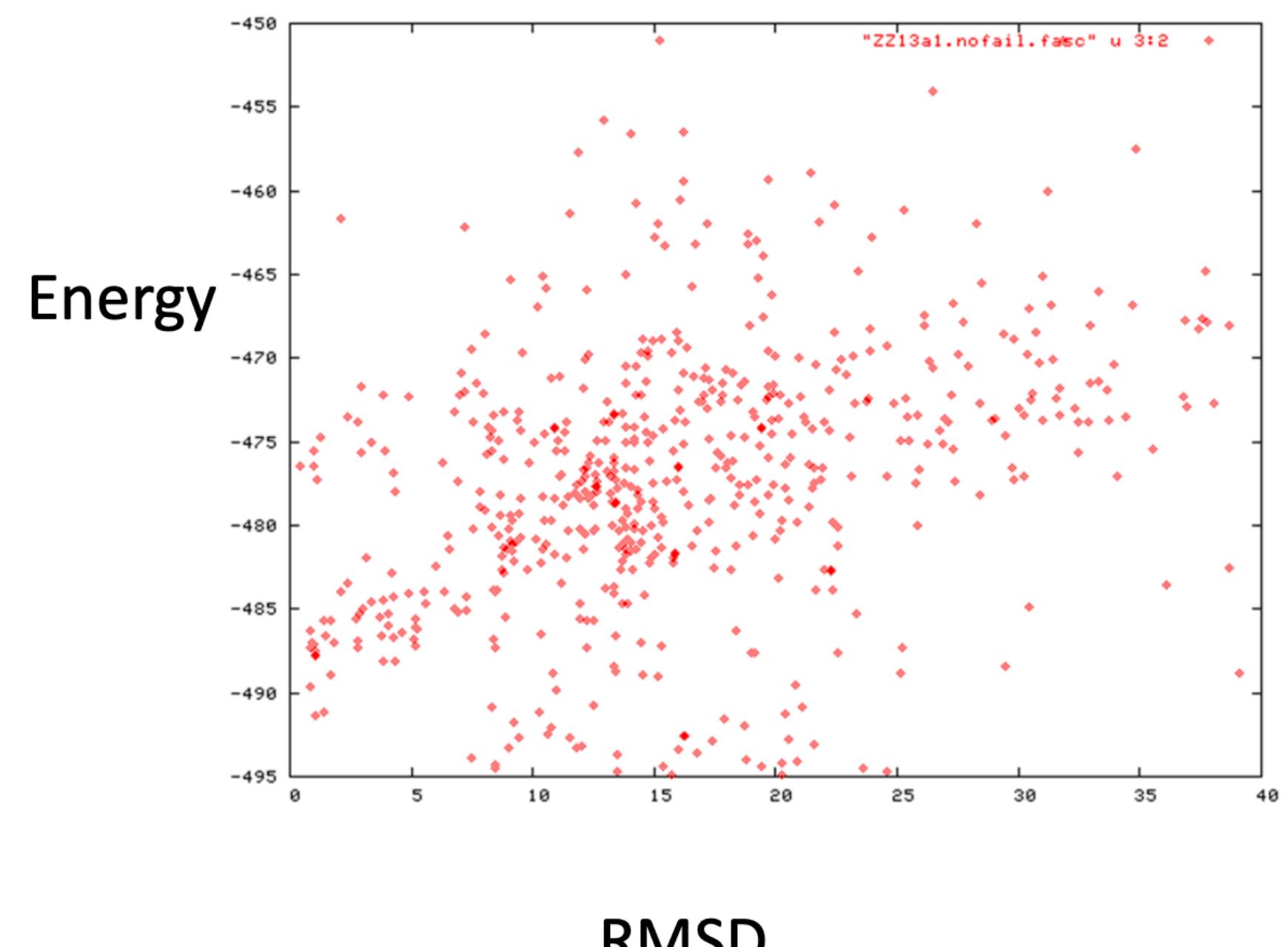
Stochastic

Random, outcome varies
Must repeat the search or perform more steps to improve chances of success
Feasible for larger problems

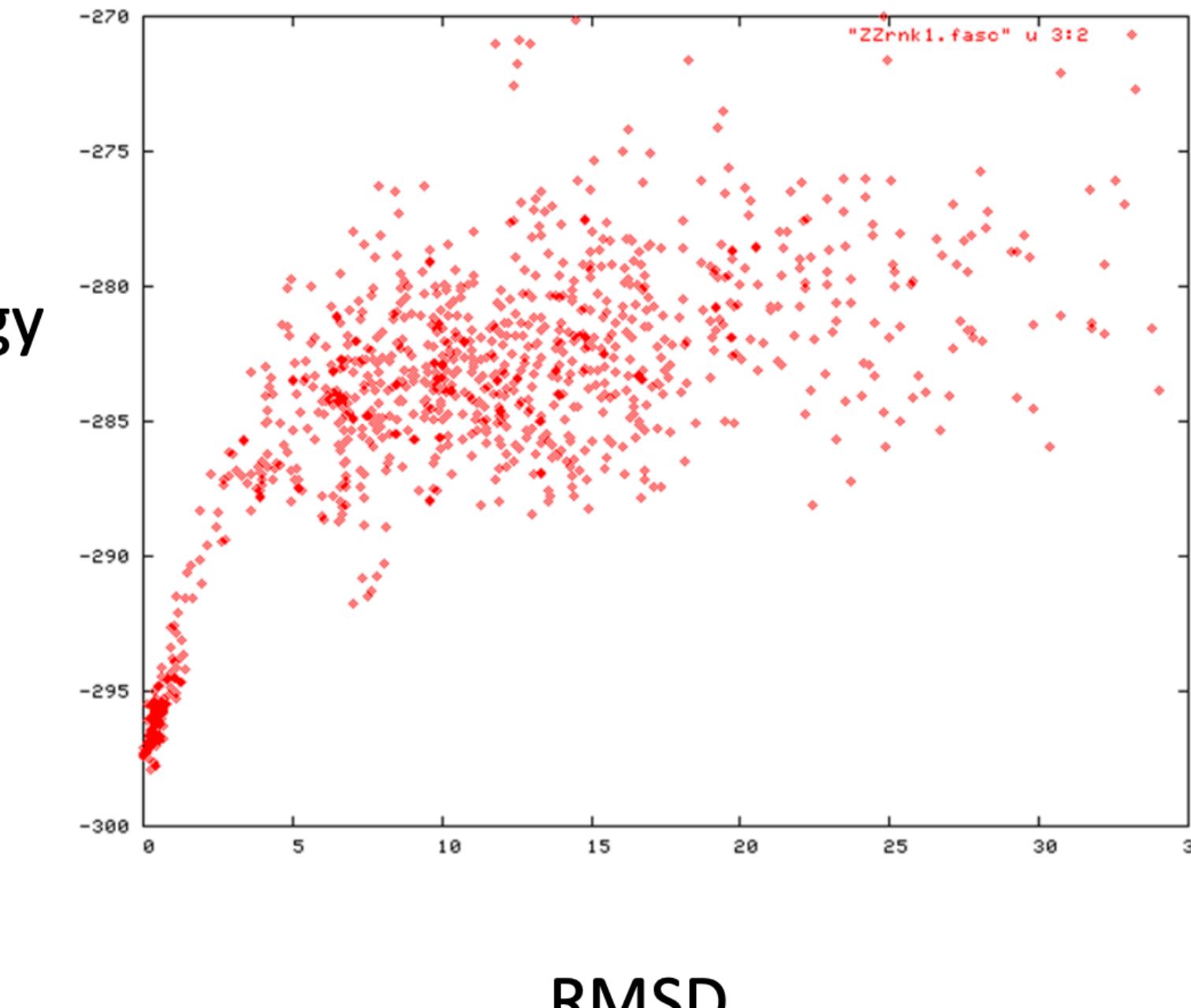
Monte Carlo (standard or + simulated annealing)
Genetic algorithms
Both can be coupled with local minimisation (Gradient Descent, conjugate gradient, L-BFGS)



Scoring function and convergence



Bad

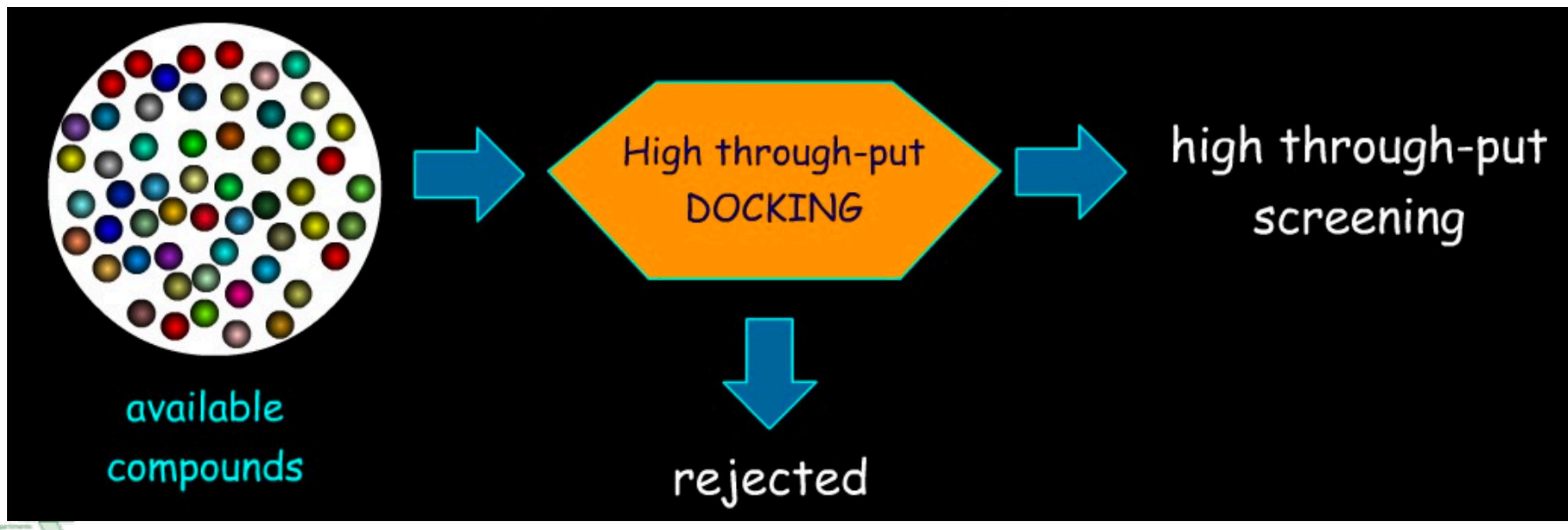


Good



Virtual Screening

- When the goal of docking is to dock all the compounds of a library (the molecules being available or not yet synthesized), the process is **called virtual screening** or **high throughput docking**
- Virtual screening identifies active compounds in a large database and ranks them by their affinity to the receptor
- The method is not used to recognize active molecules but to **eliminate those that are likely to be inactive**



Limitations

Rigid docking with flexible ligand is currently extremely efficient

if the active site is well identified some flexibility of the active site can be also considered successfully

Additional flexibility can only be incorporated by increasing the number of receptor structures (ensemble docking)

Is it possible to perform fully flexible - first principles - docking?

Yes but is generally computational demanding (so unsuited for drug developments) and a far more detailed description of the interactions is needed.

