JBC REVIEWS ASBMB AWARD ARTICLE

Toward the solution of the protein structure prediction problem

Received for publication, April 20, 2021, and in revised form, June 7, 2021 Published, Papers in Press, June 11, 2021,
https://doi.org/10.1016/j.jbc.2021.100870

Robin Pearce[1] **and Yang Zhang**[1,2,*]

From the [1]*Department of Computational Medicine and Bioinformatics,* [2]*Department of Biological Chemistry, University of Michigan, Ann Arbor, Michigan, USA*

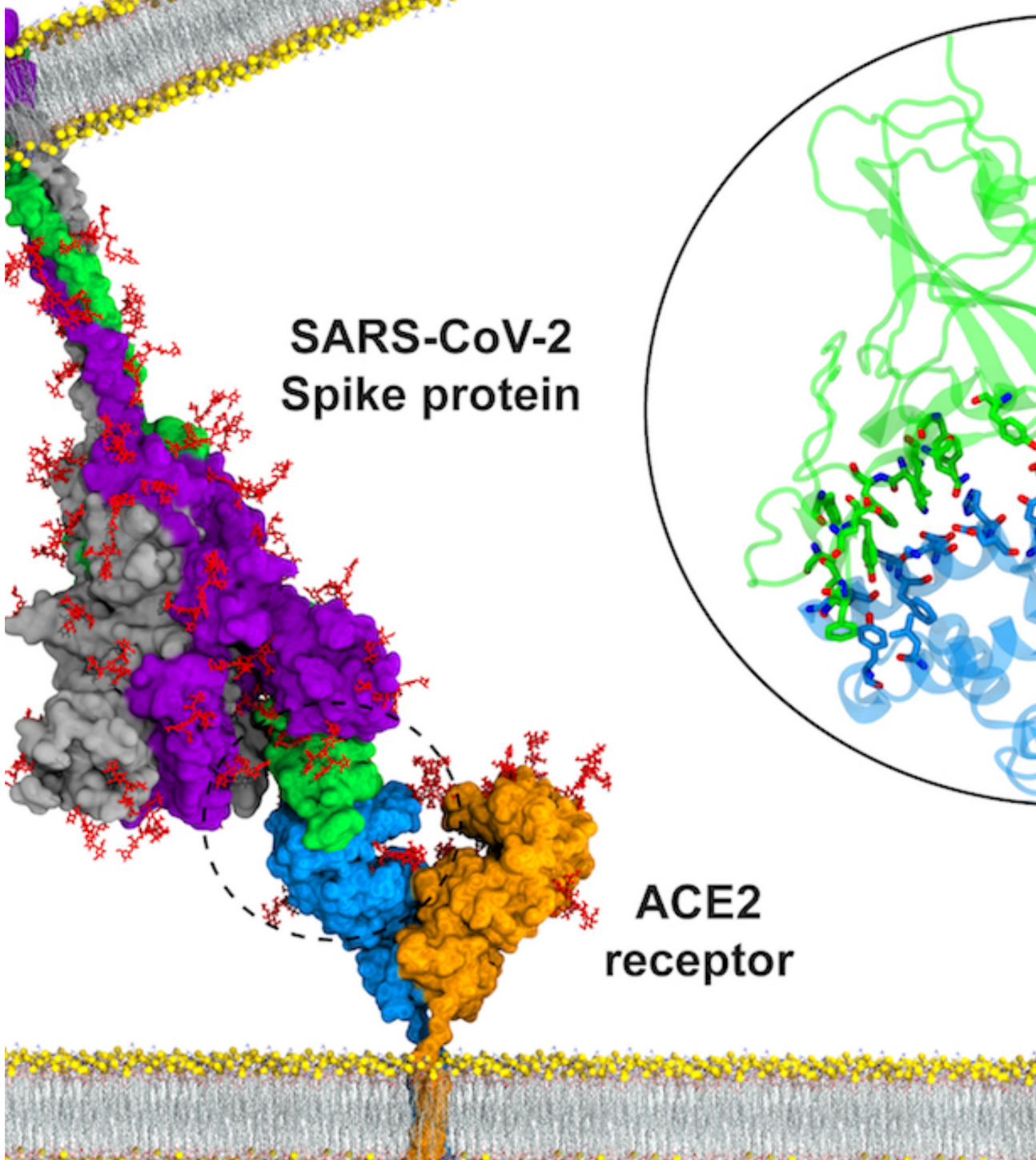
Edited by Wolfgang Peti

Structures predictions and molecular docking

Structural Bioinformatics

Outline

- **Structure prediction: concepts**
- Structure prediction: the origins
- Structure prediction: key advances
- State of the art and AI approaches
- Protein complexes and molecular docking
- AI approaches to protein complexes and molecular docking

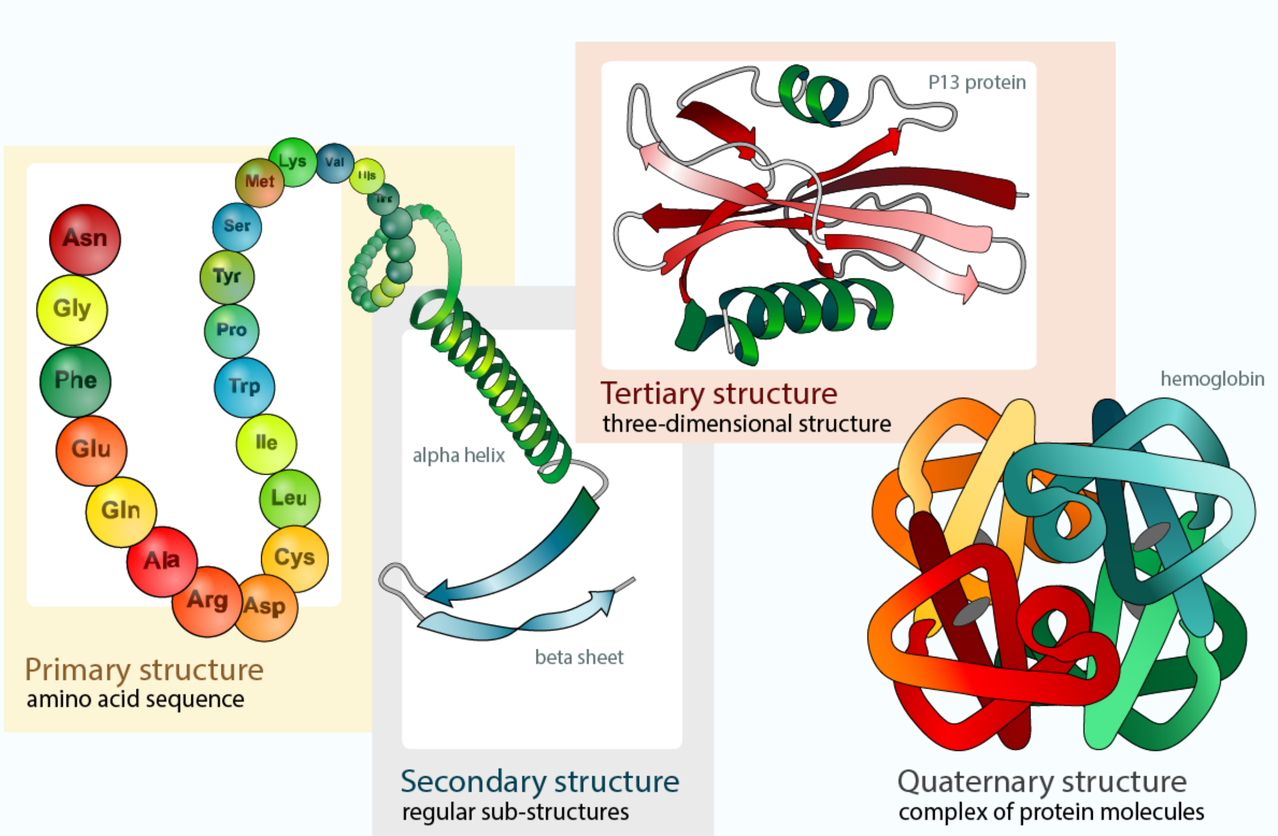


Motivations

- **High-throughput sequencing technology have greatly exacerbated the gap between the number of known sequences (~10^8) and the number of experimentally determined protein structures (~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)**
- **…**

Protein Structure Prediction and Design

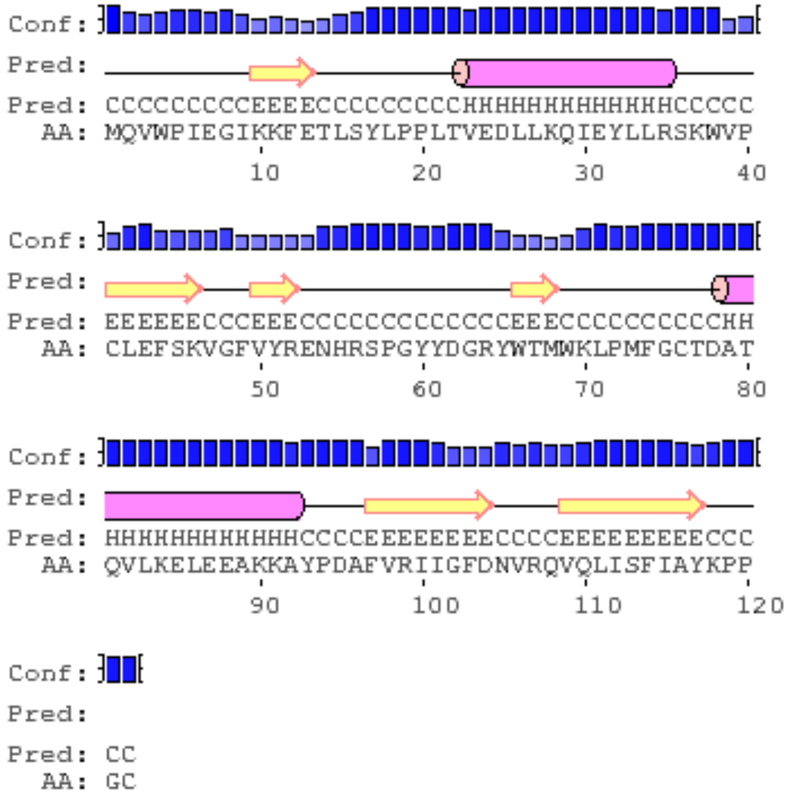
- From Sequence to Structure: remember the hierarchy of protein structural classification (primary, secondary, tertiary and quaternary)
- From Structure/Function to Sequence



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)



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

Useful as:

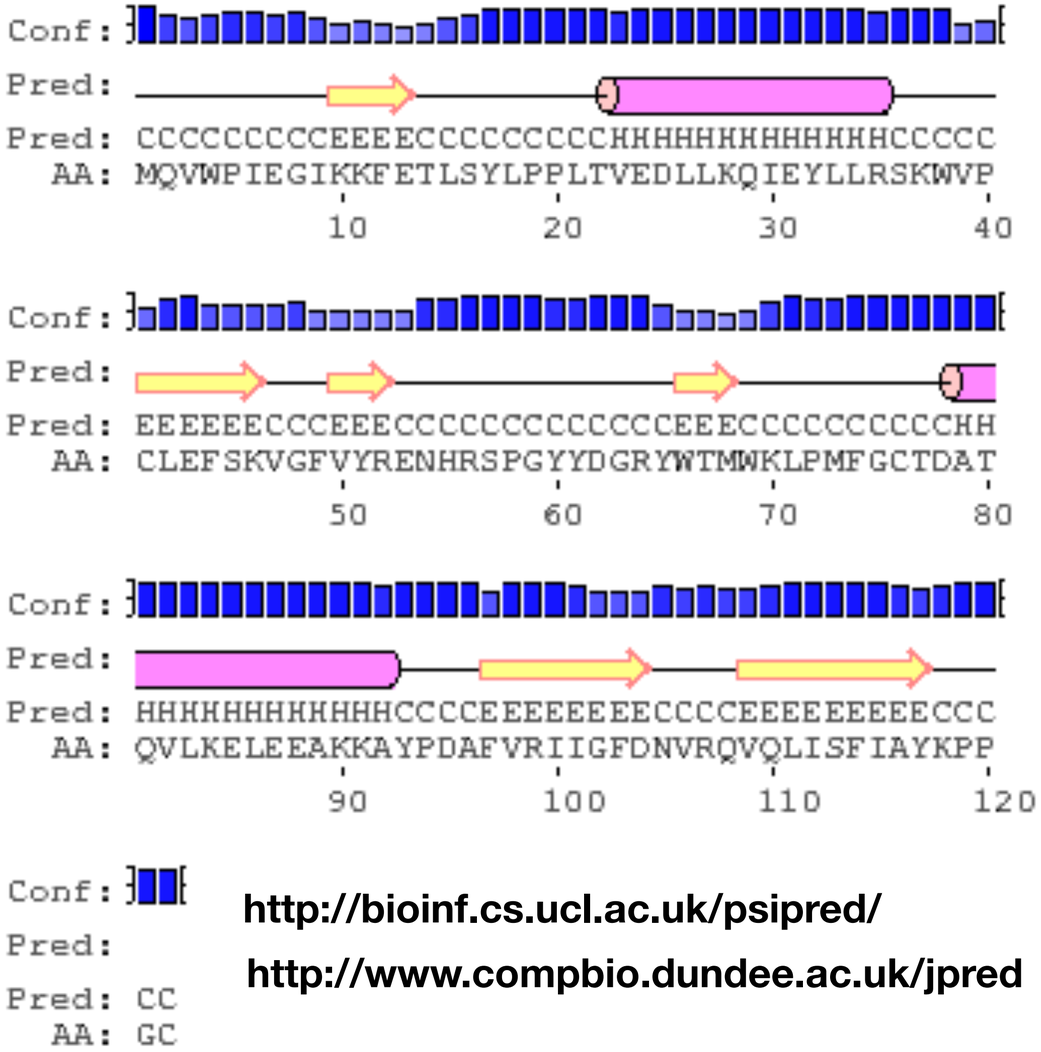
- **a step towards tertiary structure prediction**
- **Sequence alignment**
- **Structure determination at intermediate resolution**
- **Prediction of other properties (aggregation, trans-membrane, disorder)**

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%



http://bioinf.cs.ucl.ac.uk/psipred/

http://www.compbio.dundee.ac.uk/jpred

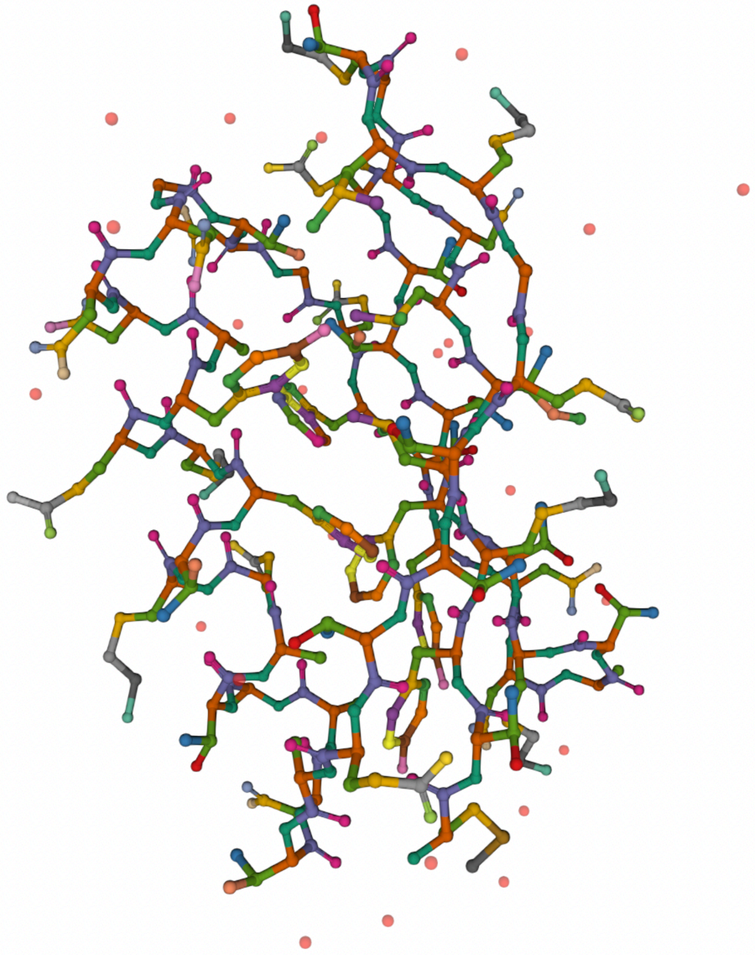
Let’s visualise the structure prediction problems

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 x N_{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
ATOM      5  CB  MET A   1      14.733  17.216  29.882  1.00 20.72
ATOM      6  CG  MET A   1      15.742  16.983  28.738  1.00 23.81
ATOM      7  SD  MET A   1      17.378  17.025  29.359  1.00 28.11
ATOM      8  CE  MET A   1      17.166  16.055  30.819  1.00 27.51
ATOM      9  N   THR A   2      13.719  19.413  27.573  1.00 12.63
ATOM     10  CA  THR A   2      13.088  19.661  26.283  1.00 12.68
ATOM     11  C   THR A   2      13.561  18.631  25.300  1.00 12.02
ATOM     12  O   THR A   2      14.763  18.432  25.121  1.00 13.07
ATOM     13  CB  THR A   2      13.527  20.980  25.667  1.00 14.62
ATOM     14  OG1 THR A   2      13.307  22.020  26.627  1.00 15.31
ATOM     15  CG2 THR A   2      12.704  21.284  24.409  1.00 14.47
ATOM     16  N   TYR A   3      12.574  18.048  24.642  1.00 11.17
ATOM     17  CA  TYR A   3      12.726  17.033  23.612  1.00 10.11
ATOM     18  C   TYR A   3      12.109  17.637  22.316  1.00 10.52
ATOM     19  O   TYR A   3      11.165  18.449  22.364  1.00  9.38
ATOM     20  CB  TYR A   3      11.907  15.809  24.042  1.00 10.96
ATOM     21  CG  TYR A   3      12.497  15.093  25.196  1.00 10.60
ATOM     22  CD1 TYR A   3      13.560  14.276  25.012  1.00 12.20
ATOM     23  CD2 TYR A   3      12.045  15.324  26.492  1.00 11.77
ATOM     24  CE1 TYR A   3      14.205  13.693  26.058  1.00 13.25
ATOM     25  CE2 TYR A   3      12.663  14.737  27.567  1.00 12.45
ATOM     26  CZ  TYR A   3      13.772  13.910  27.323  1.00 11.39
ATOM     27  OH  TYR A   3      14.476  13.300  28.344  1.00 14.48
ATOM     28  N   LYS A   4      12.633  17.222  21.175  1.00  9.63
ATOM     29  CA  LYS A   4      12.179  17.659  19.887  1.00  9.41
ATOM     30  C   LYS A   4      11.677  16.470  19.087  1.00  9.49
ATOM     31  O   LYS A   4      12.151  15.336  19.237  1.00  8.55
ATOM     32  CB  LYS A   4      13.376  18.247  19.100  1.00 12.36
ATOM     33  CG  LYS A   4      12.954  19.035  17.857  1.00 17.46
ATOM     34  CD  LYS A   4      14.119  19.494  16.982  1.00 20.77
ATOM     35  CE  LYS A   4      14.184  21.056  16.755  1.00 24.12
ATOM     36  NZ  LYS A   4      12.929  21.820  16.303  1.00 25.14
ATOM     37  N   LEU A   5      10.771  16.761  18.157  1.00  8.76
ATOM     38  CA  LEU A   5      10.253  15.790  17.221  1.00  7.91
ATOM     39  C   LEU A   5      10.360  16.415  15.781  1.00  9.03
ATOM     40  O   LEU A   5       9.916  17.539  15.553  1.00  6.35
ATOM     41  CB  LEU A   5       8.765  15.468  17.506  1.00  8.63
ATOM     42  CG  LEU A   5       8.058  14.607  16.411  1.00  8.98
ATOM     43  CD1 LEU A   5       8.626  13.160  16.373  1.00  8.38
ATOM     44  CD2 LEU A   5       6.577  14.522  16.660  1.00  9.13
ATOM     45  N   ILE A   6      10.995  15.689  14.856  1.00  7.05
ATOM     46  CA  ILE A   6      11.082  16.103  13.475  1.00  9.67
ATOM     47  C   ILE A   6      10.046  15.228  12.753  1.00  8.83
ATOM     48  O   ILE A   6      10.068  14.016  12.892  1.00  7.29
ATOM     49  CB  ILE A   6      12.484  15.880  12.922  1.00  9.90
ATOM     50  CG1 ILE A   6      13.453  16.788  13.678  1.00 13.88
ATOM     51  CG2 ILE A   6      12.520  16.275  11.428  1.00  9.38
ATOM     52  CD1 ILE A   6      14.844  16.502  13.276  1.00 15.90
ATOM     53  N   LEU A   7       9.085  15.850  12.087  1.00  9.18
ATOM     54  CA  LEU A   7       8.009  15.163  11.389  1.00  7.98
ATOM     55  C   LEU A   7       8.312  15.181   9.902  1.00  9.81
ATOM     56  O   LEU A   7       8.580  16.234   9.291  1.00  7.52
ATOM     57  CB  LEU A   7       6.690  15.903  11.602  1.00 11.14
ATOM     58  CG  LEU A   7       6.147  16.007  13.030  1.00 13.03
ATOM     59  CD1 LEU A   7       5.647  17.410  13.283  1.00 14.31
ATOM     60  CD2 LEU A   7       5.037  14.952  13.247  1.00 12.67
ATOM     61  N   ASN A   8       8.383  14.018   9.323  1.00 10.71
ATOM     62  CA  ASN A   8       8.628  13.975   7.913  1.00 13.03
ATOM     63  C   ASN A   8       7.575  13.039   7.320  1.00 12.38
ATOM     64  O   ASN A   8       7.885  11.921   6.936  1.00 11.56
ATOM     65  CB  ASN A   8      10.025  13.485   7.676  1.00 15.38
ATOM     66  CG  ASN A   8      10.270  13.214   6.226  1.00 19.45
ATOM     67  OD1 ASN A   8      10.134  14.101   5.386  1.00 19.46
ATOM     68  ND2 ASN A   8      10.497  11.955   5.900  1.00 22.31
ATOM     69  N   GLY A   9       6.309  13.415   7.484  1.00 10.44
ATOM     70  CA  GLY A   9       5.213  12.642   6.966  1.00 12.52
ATOM     71  C   GLY A   9       4.914  12.990   5.506  1.00 12.67
ATOM     72  O   GLY A   9       5.571  13.842   4.929  1.00 14.71
ATOM     73  N   LYS A  10       3.922  12.342   4.909  1.00 14.25
ATOM     74  CA  LYS A  10       3.589  12.601   3.497  1.00 15.77
ATOM     75  C   LYS A  10       2.910  13.955   3.387  1.00 16.15
ATOM     76  O   LYS A  10       3.170  14.710   2.434  1.00 15.37
ATOM     77  CB  LYS A  10       2.611  11.571   2.910  1.00 18.26           C
ATOM     78  CG  LYS A  10       2.910  10.139   3.208  1.00 21.67           C
ATOM     79  CD  LYS A  10       1.958   9.176   2.456  1.00 23.96           C
```



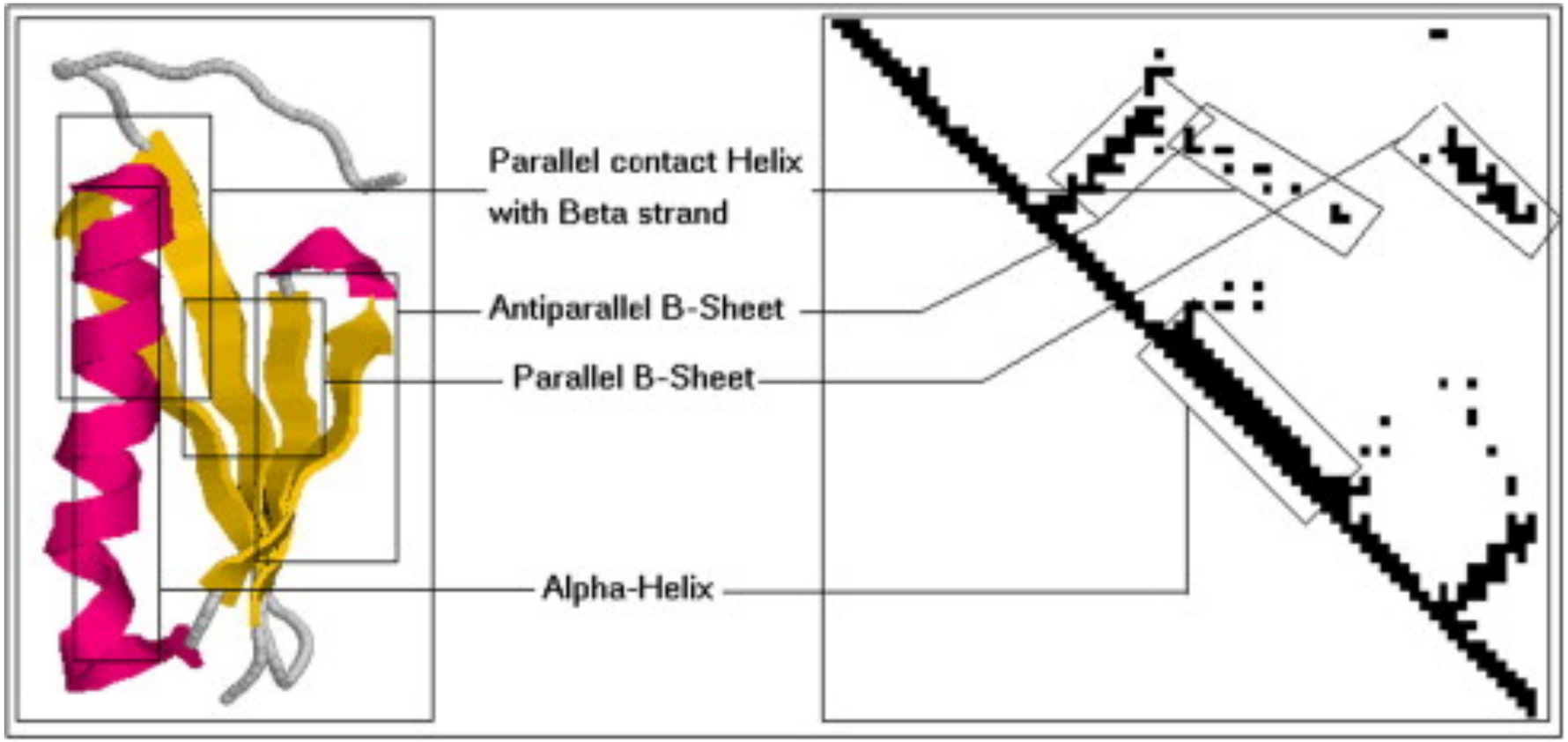
UNIVERSITÀ DEGLI STUDI DI MILANO

DBS

An intermediate step: contact map predictions

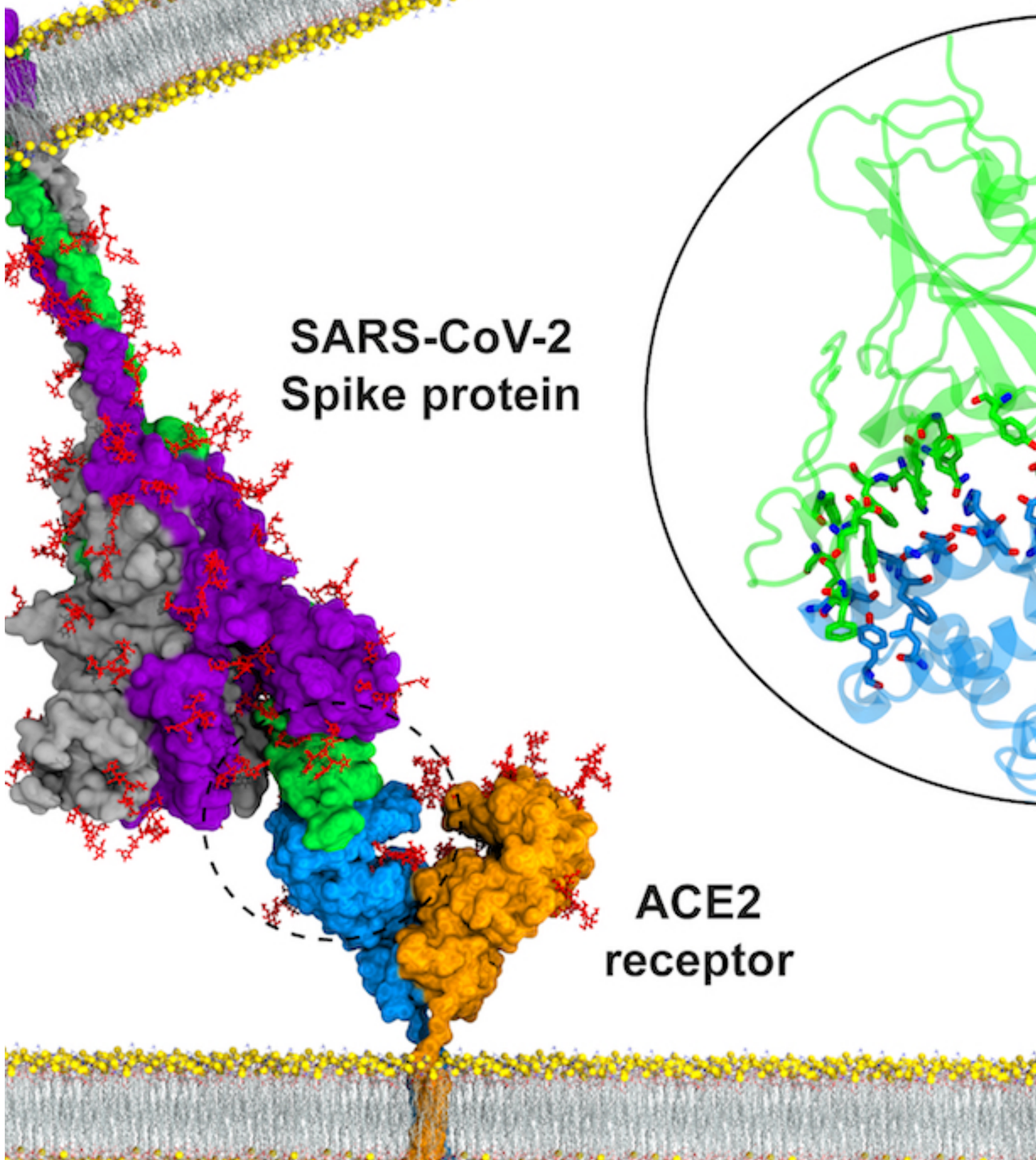
Contact prediction:

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



Outline

- Structure prediction: concepts
- **Structure prediction: the origins**
- Structure prediction: key advances
- State of the art and AI approaches
- Protein complexes and molecular docking
- AI approaches to protein complexes and molecular docking



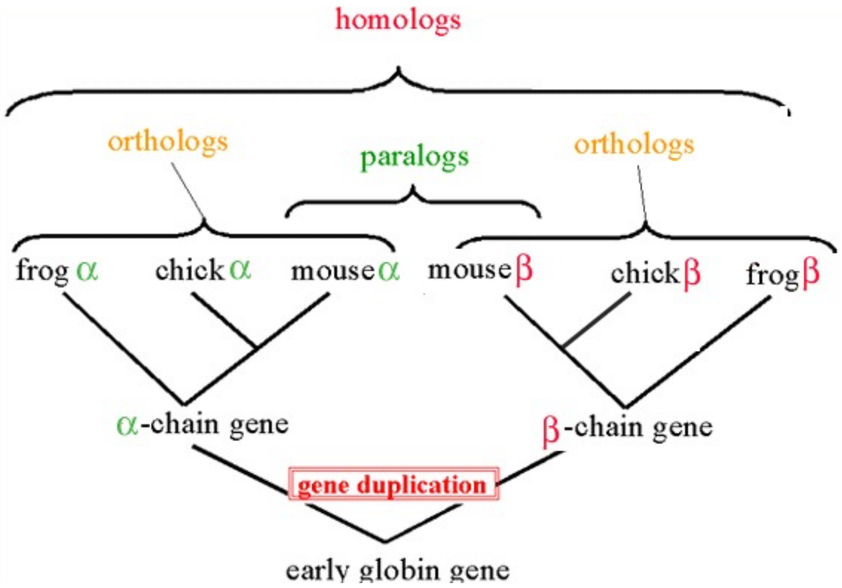
Homologs: The Sequence - Structure - Function Paradigm

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:

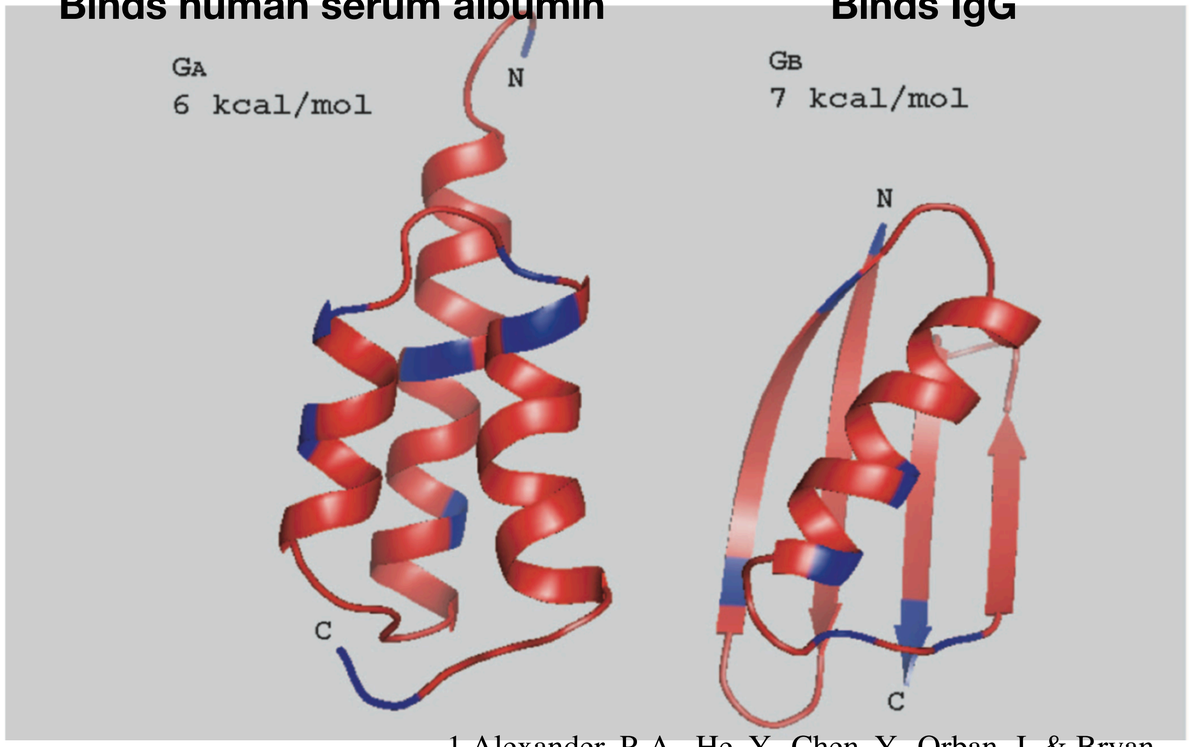
A) <u>Either</u> predicting a protein fold should be not too difficult because the details of the sequence don't matter
B) <u>Or</u> most proteins have common ancestors and there wasn't enough time to decorrelate

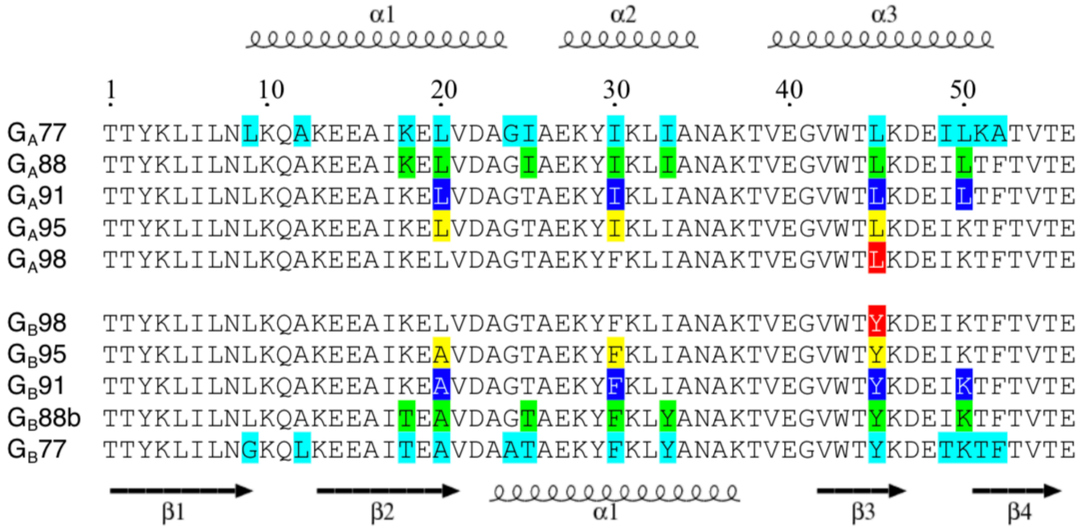
<u>PARALOGS</u>: Homologous biological components (genes, proteins, structures) within a single species that arose by gene duplication.

<u>ORTHOLOGS</u>: 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:





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.

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).

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):

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

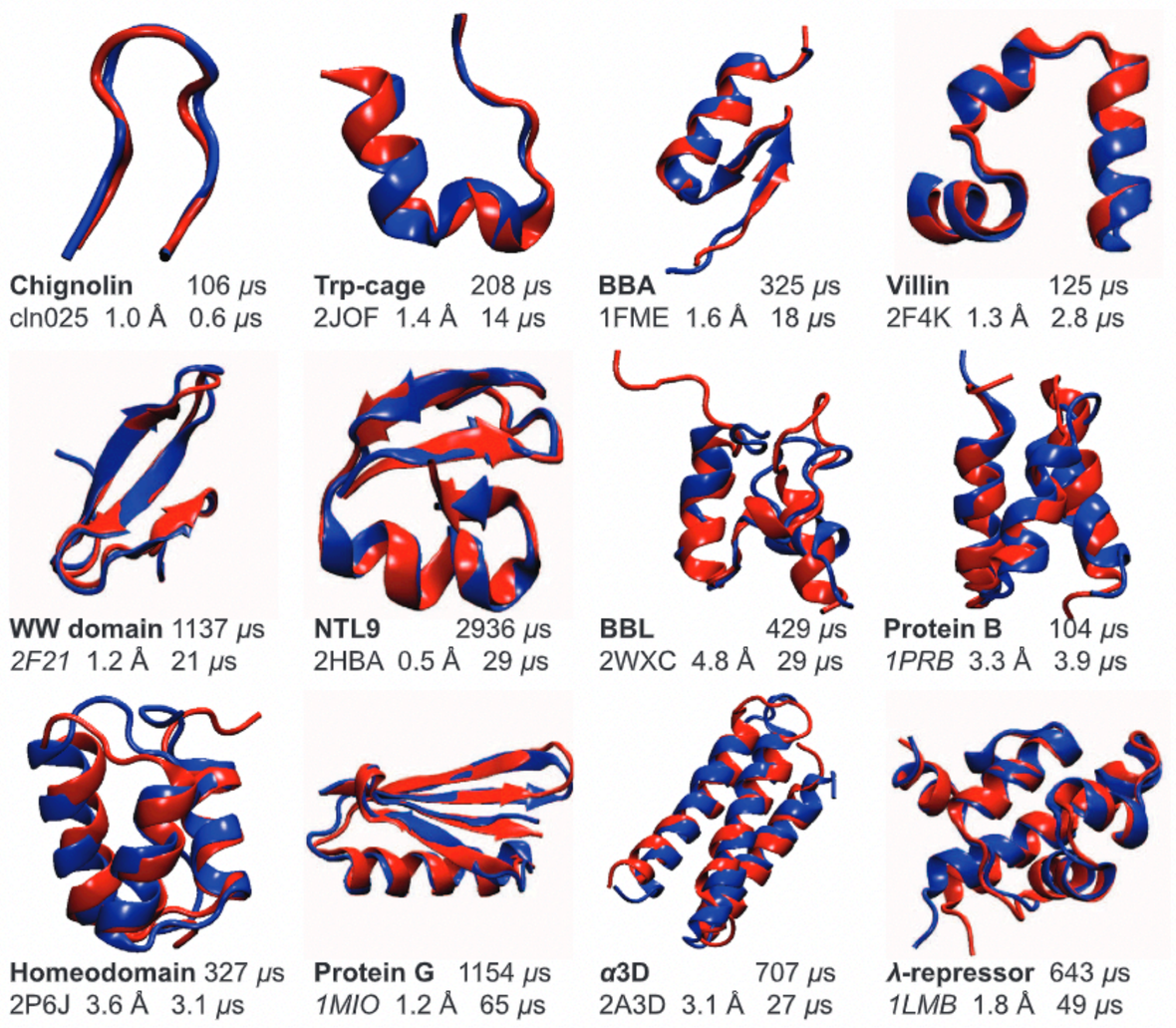
geometrical consideration pi-bonds, etc

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

point charge Coulomb

excluded volume

Dispersions (interactions of neutral molecules)



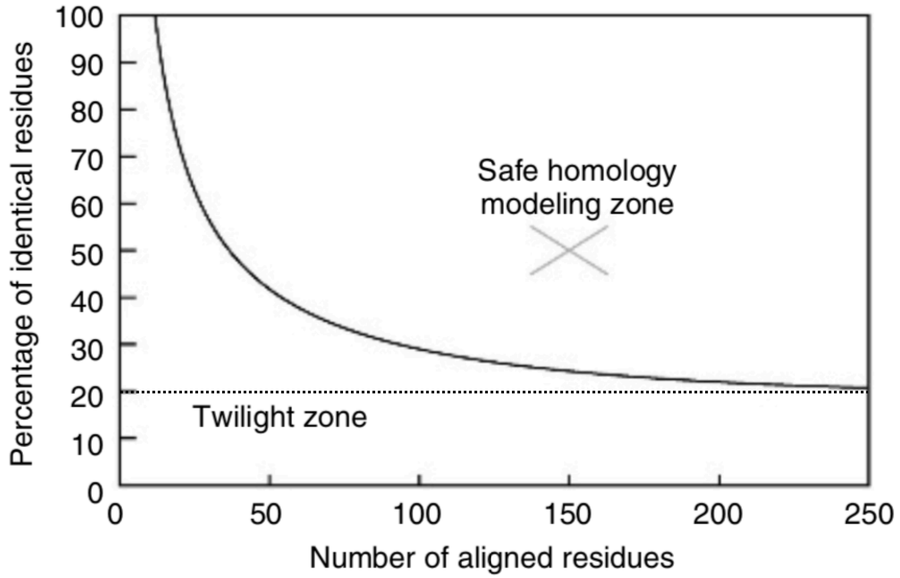
How fast-folding proteins fold. *Sci New York N Y* **334**, 517–20 (2011).

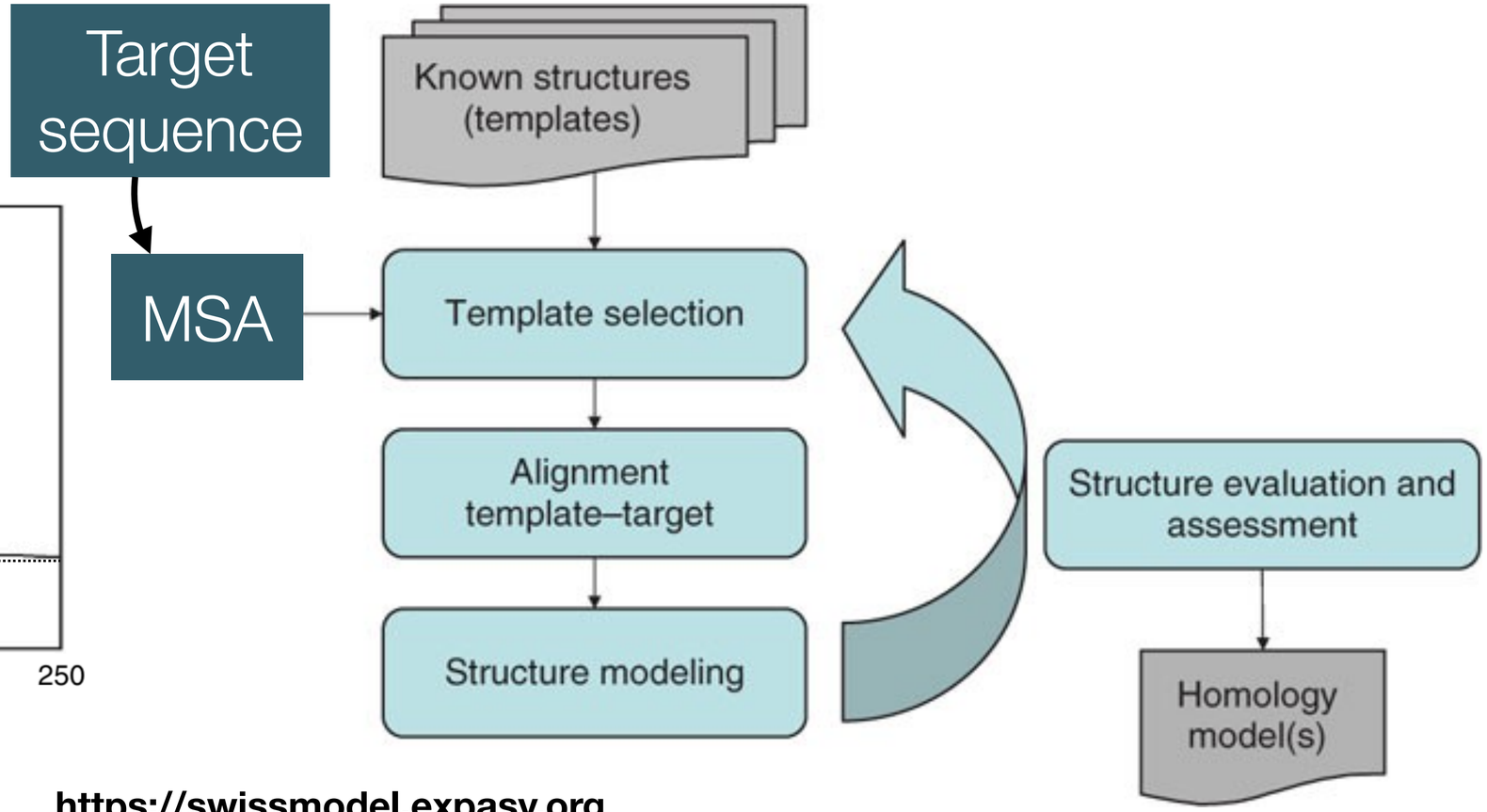
UNIVERSITÀ DEGLI STUDI DI MILANO

DBS

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.





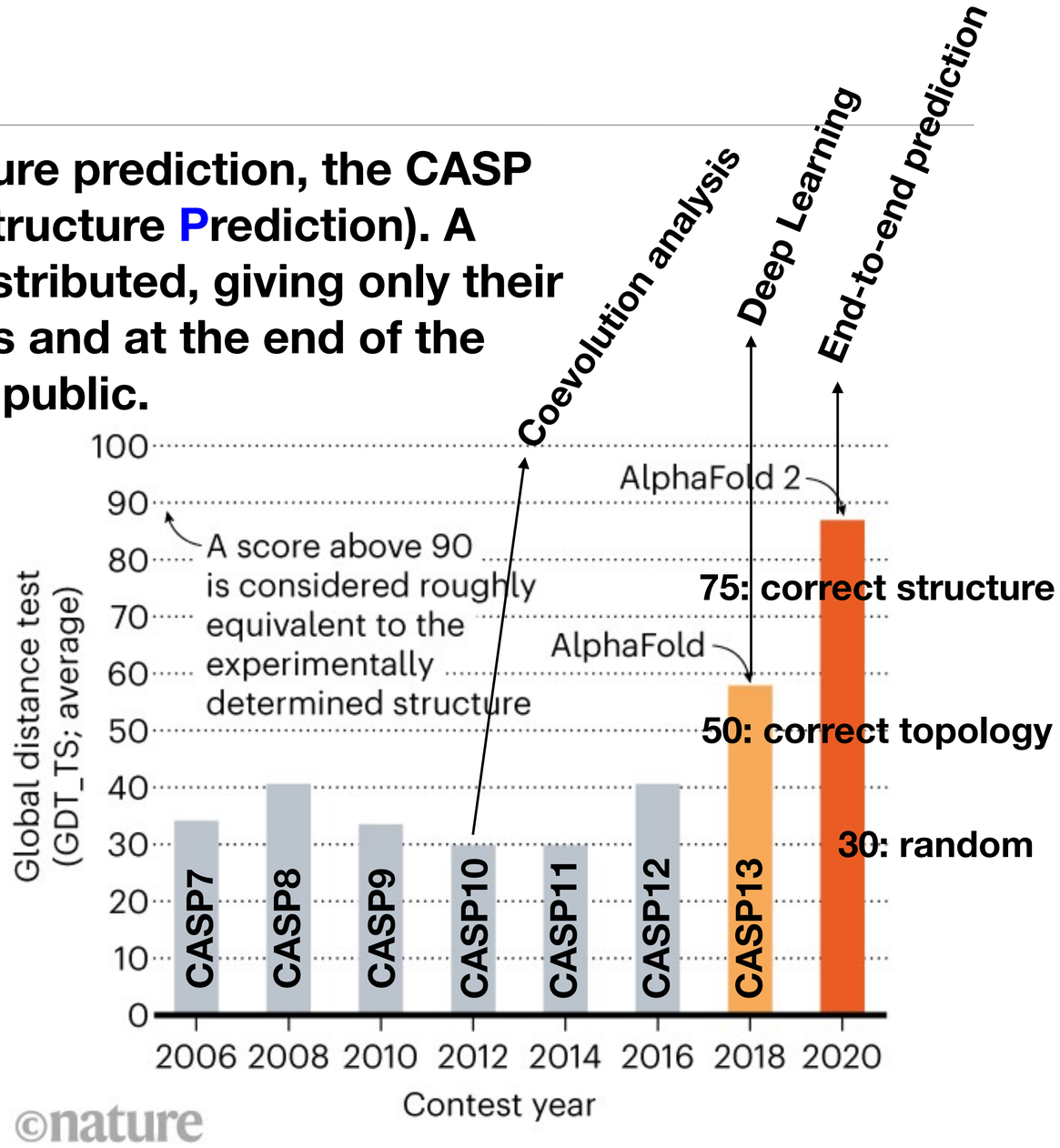
https://swissmodel.expasy.org

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"**



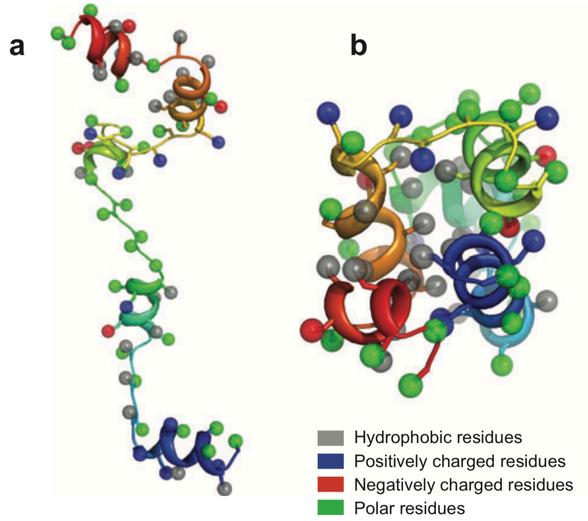
UNIVERSITÀ DEGLI STUDI DI MILANO

DBS

https://predictioncenter.org

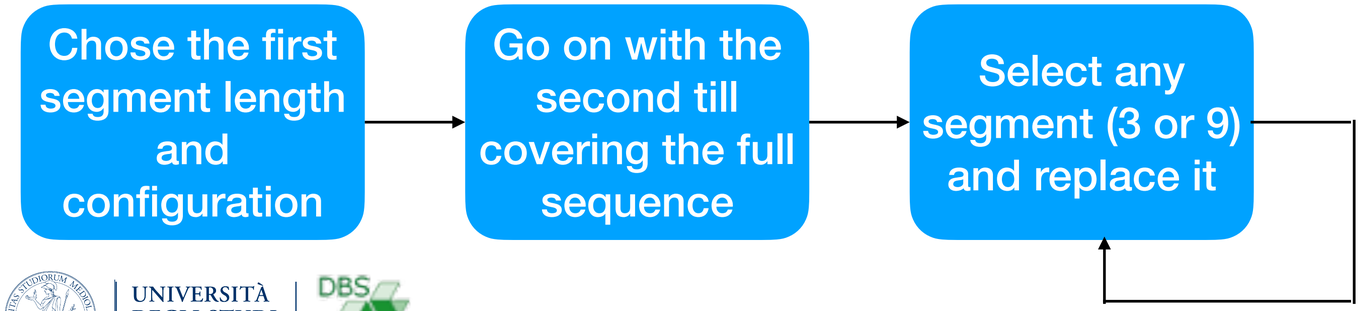
Before 2012: Fragment Replacement

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



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:

Fragment replacement:



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.

Metropolis Monte-Carlo (but often refereed simply as Monte Carlo) is an algorithm to evaluate a probability distribution.

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

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

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

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

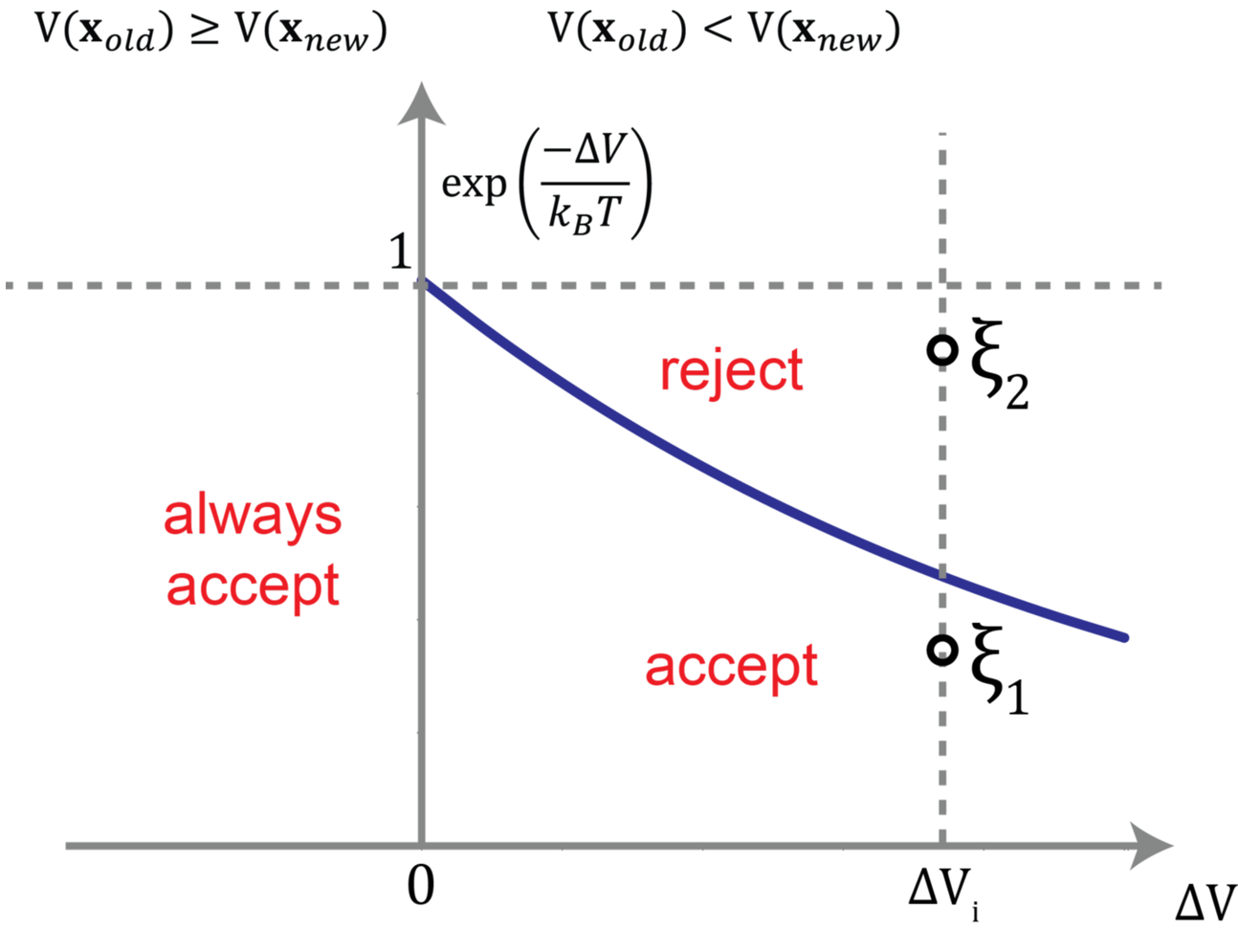
Starting from a configuration we can calculate its energy or score, then we can generate a second configuration and get its score and we need to decide if to keep it or not. How? Their relative probability is:

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

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.

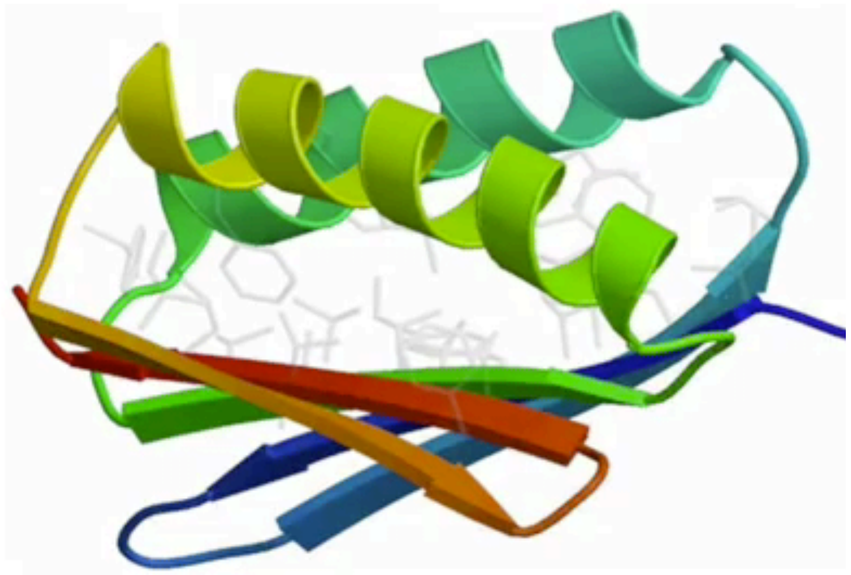
Monte-Carlo sampling

1. Given a starting configuration x_0, suggest a new configuration x_1
2. Calculate score(x_0) and score(x_1)
3. Calculate a=exp[(score(x_0)-score(x_1))/optim_rate]
4. Generate a random number u between 0..1
5. If u<=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

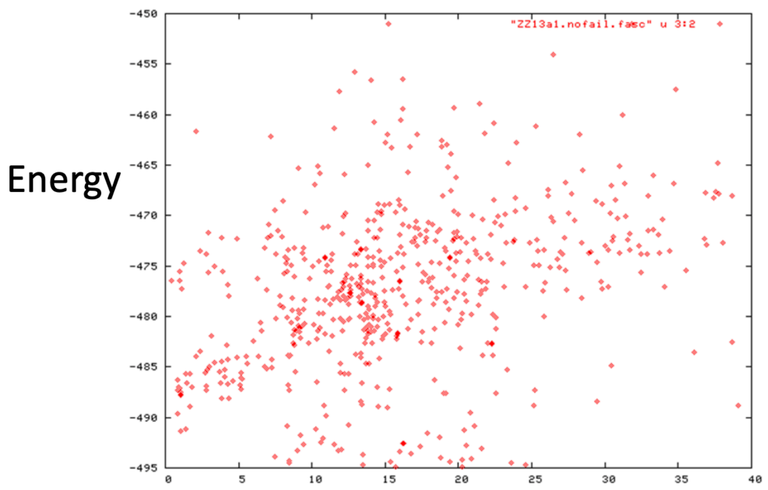


Convergence of a prediction is checked by scores distributions

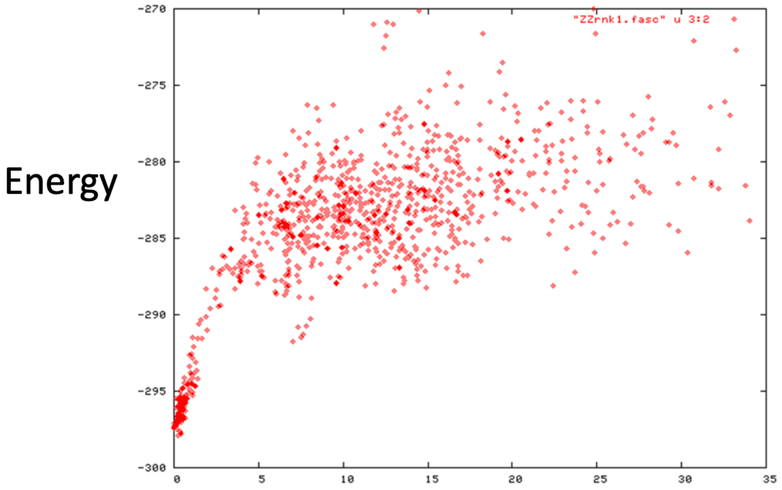
Then the best possible folds are further sampled at full atomistic resolution (without water)



Tens of thousands of structures are then scored:



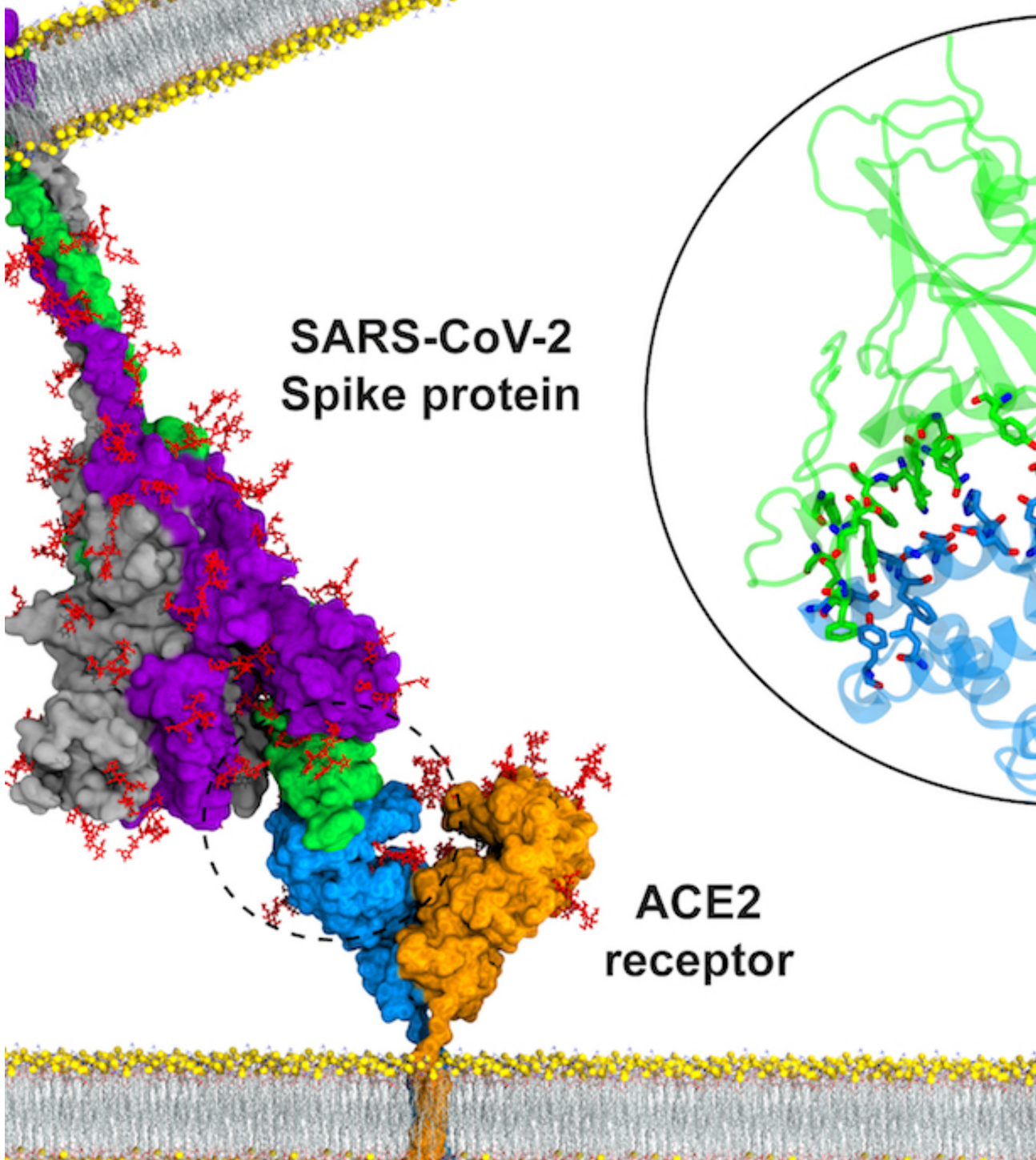
Bad



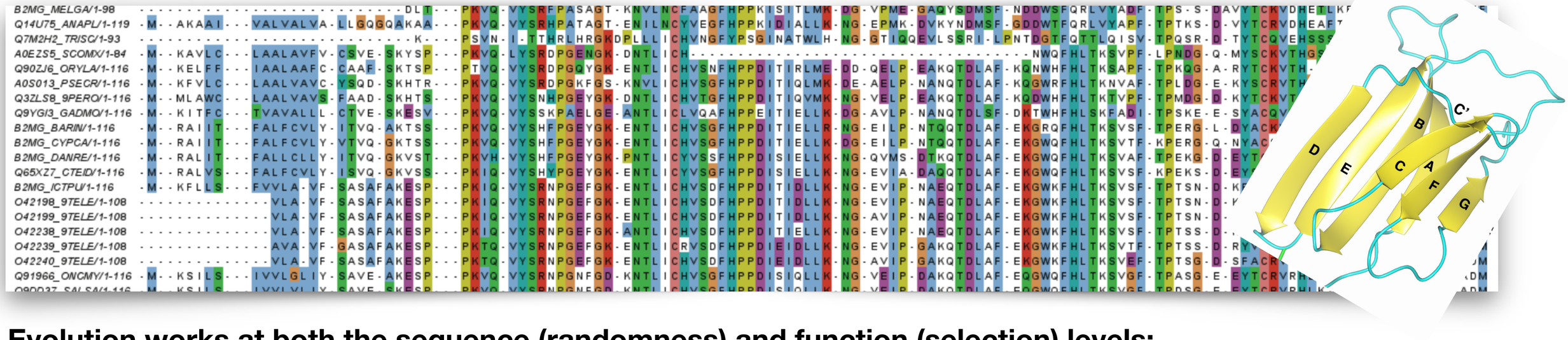
Good

Outline

- Structure prediction: concepts
- Structure prediction: the origins
- **Structure prediction: key advances**
- State of the art and AI approaches
- Protein complexes and molecular docking
- AI approaches to protein complexes and molecular docking



Proteins structures are more conserved than sequences…

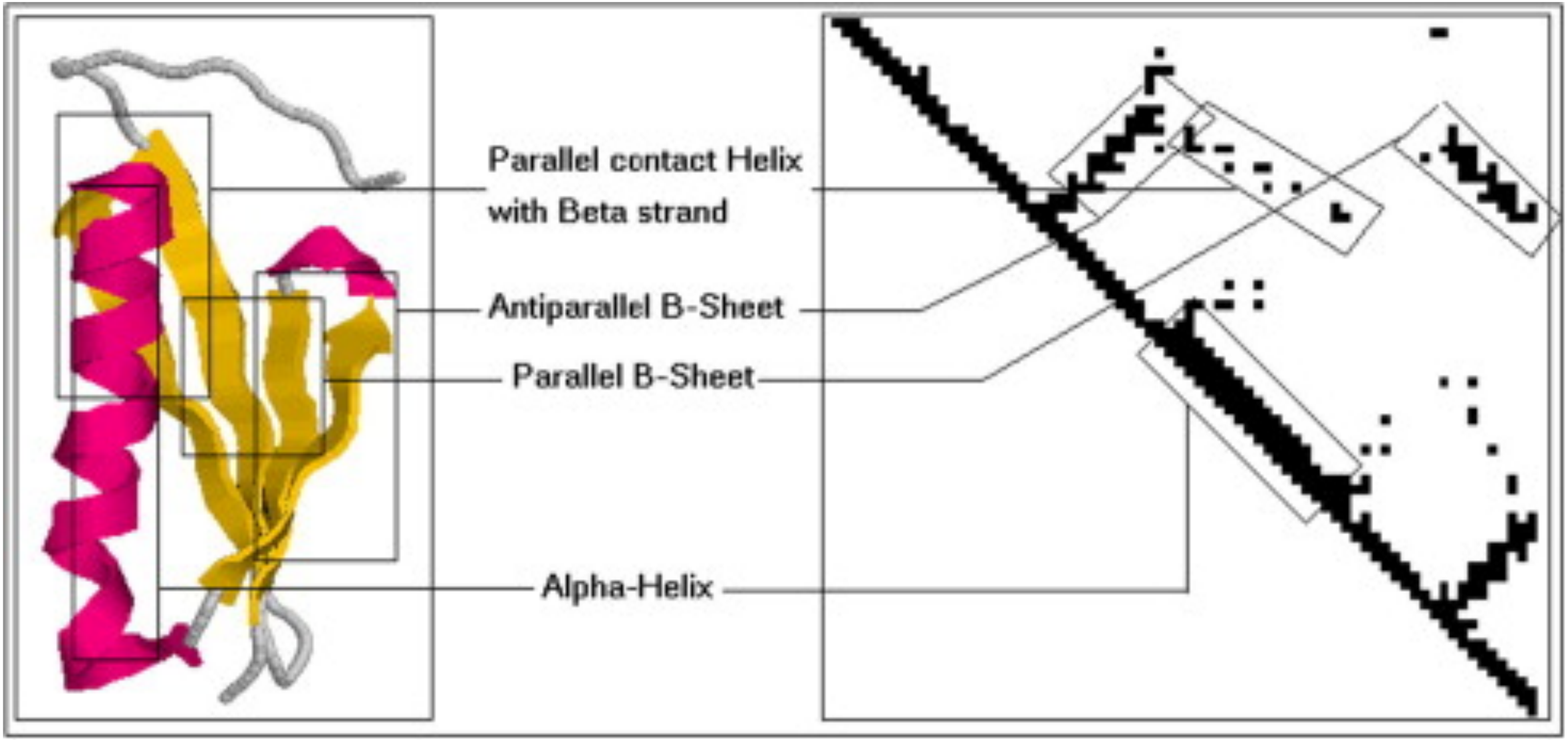


Evolution works at both the sequence (randomness) and function (selection) levels:

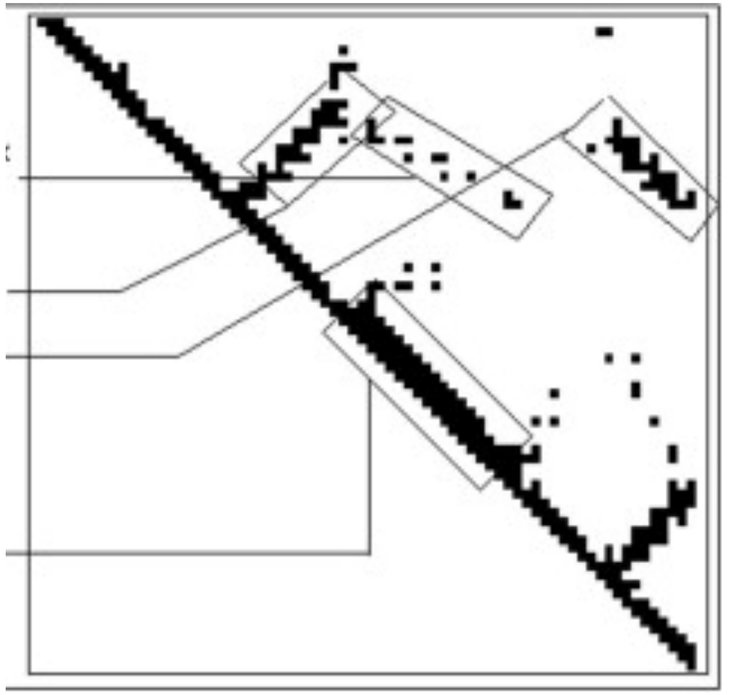
Random mutations (or insertions or deletions) of the sequence are selected by their effect on the function. This determines the exploration of the sequence space under the restraint of conserving/improving the function, that often corresponds to conserving the structure (the idea is that the cell is 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.

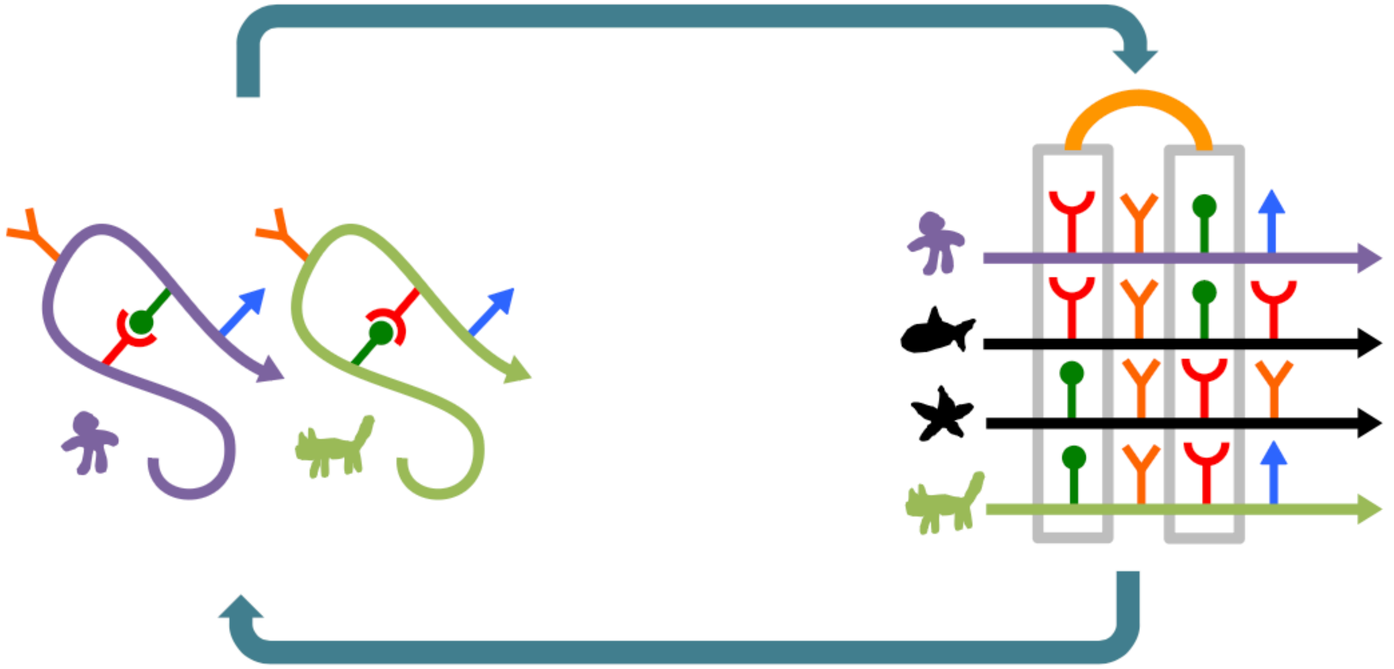
From a sequence of N letters (protein sequence) to a matrix NxN with 1 and 0s:



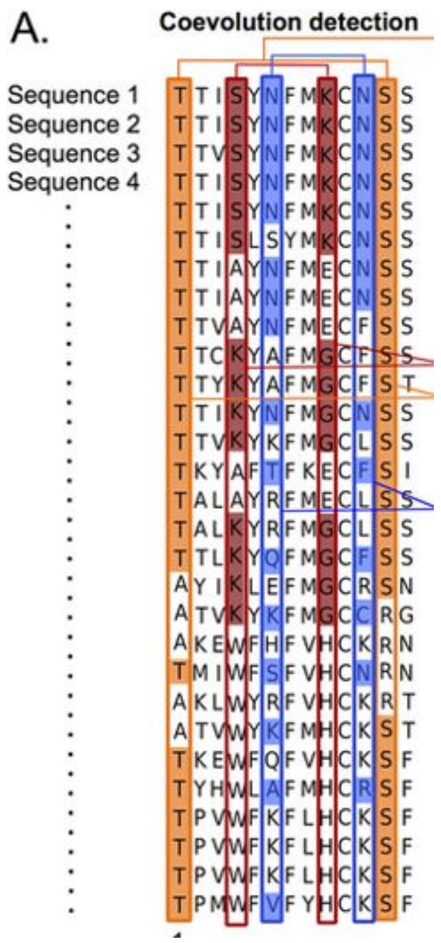
From a sequence of N letters (protein sequence) to a matrix NxN with 1 and 0s:

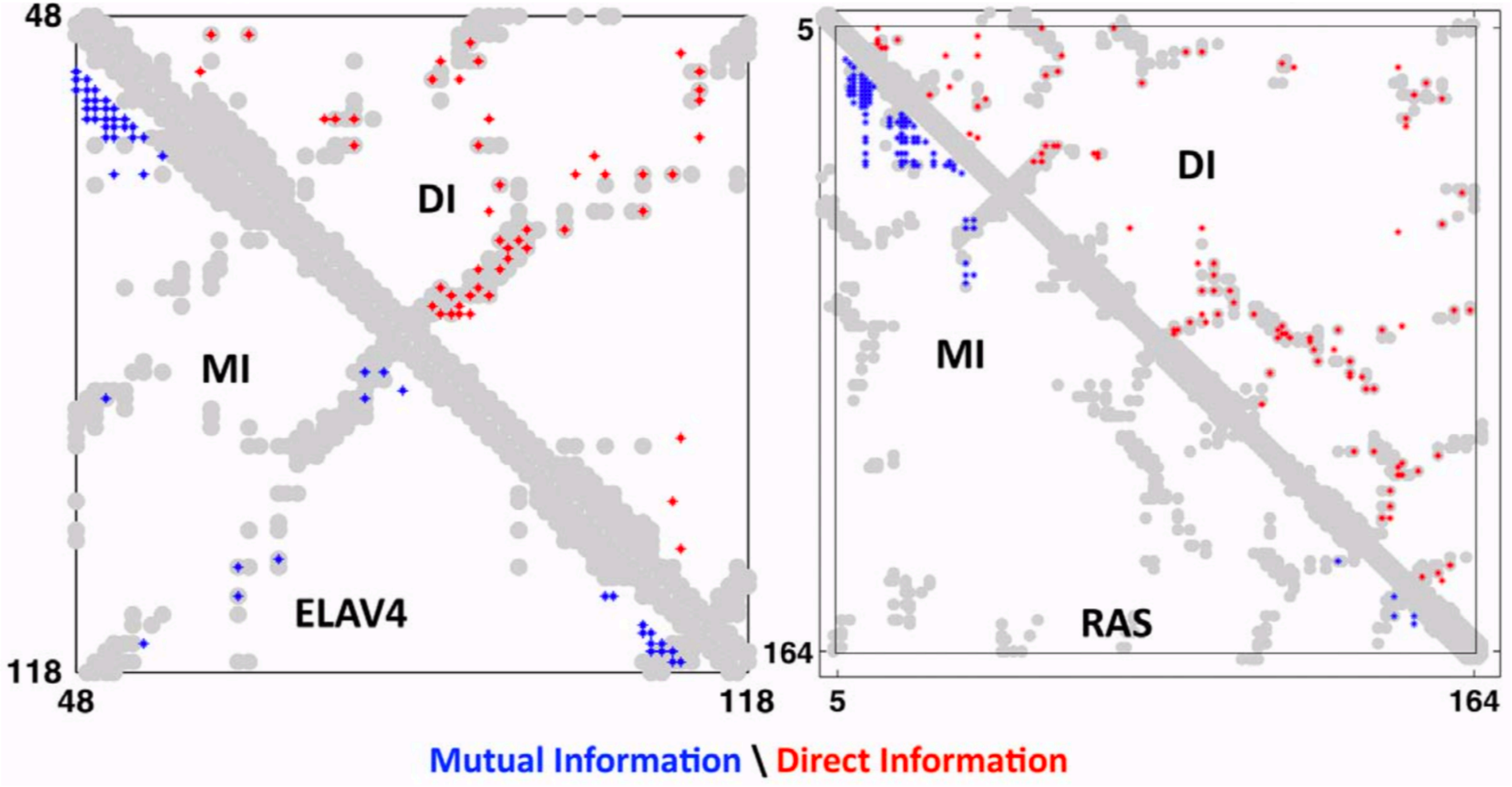


Contacts in proteins are evolutionarily conserved and encoded in a **MSA** (Multiple Sequence Alignment) due to **coevolution**



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

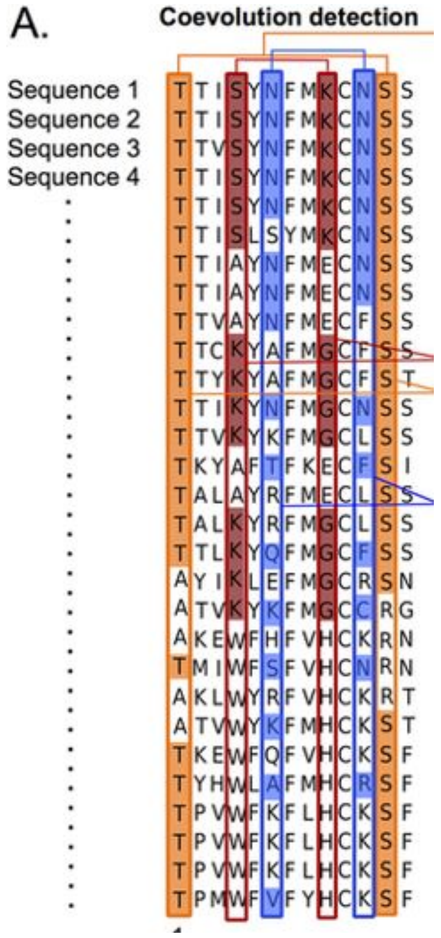




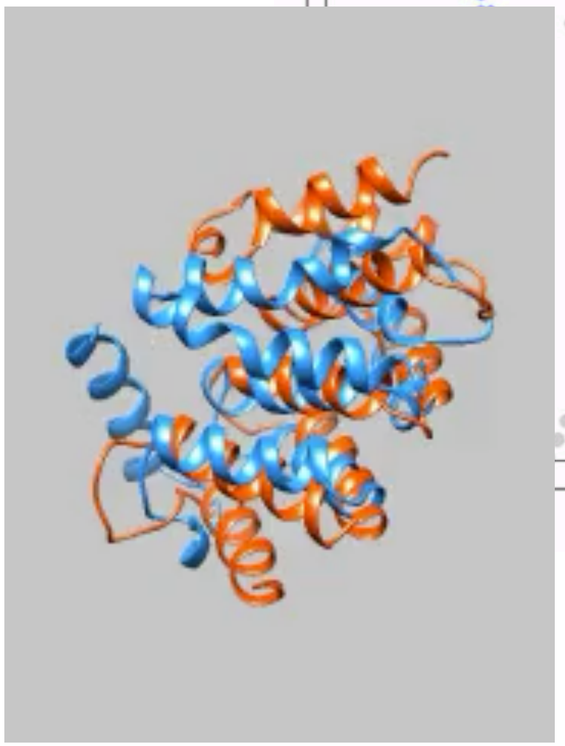
$$MI_{ij} = \sum_{A,B} f_{ij}(A,B) \ln \frac{f_{ij}(A,B)}{f_i(A) f_j(B)} \cdot$$

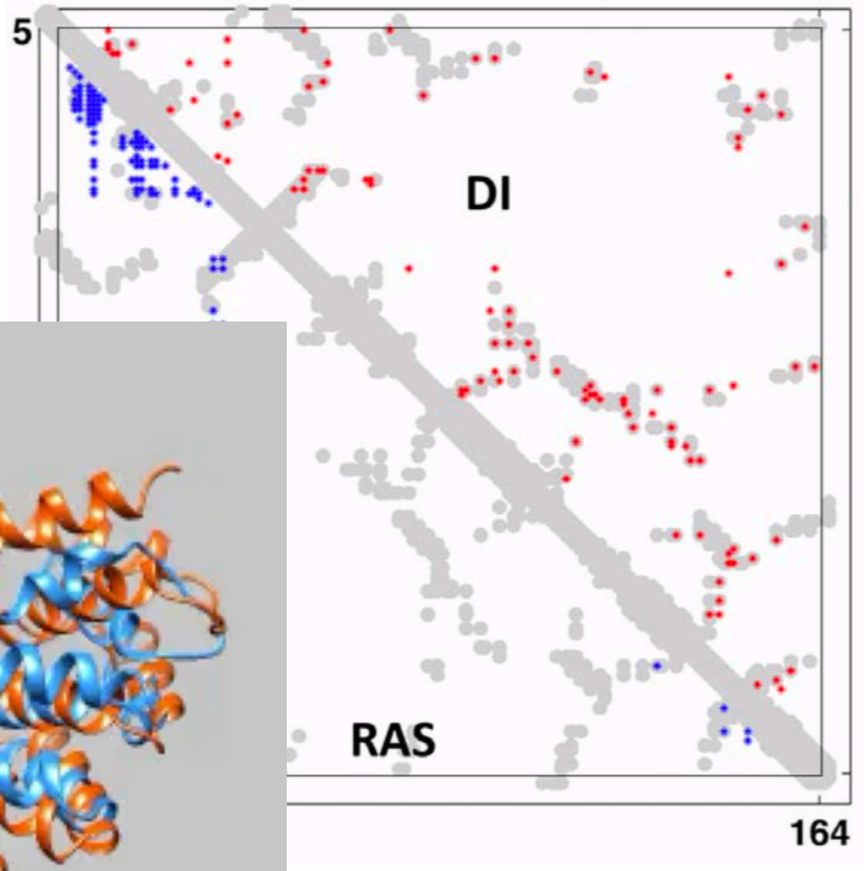
The problem is that MSA do not include all possibilities, so we need to correct it for all the not observed combinations.

2012: From Mutual to Direct Information

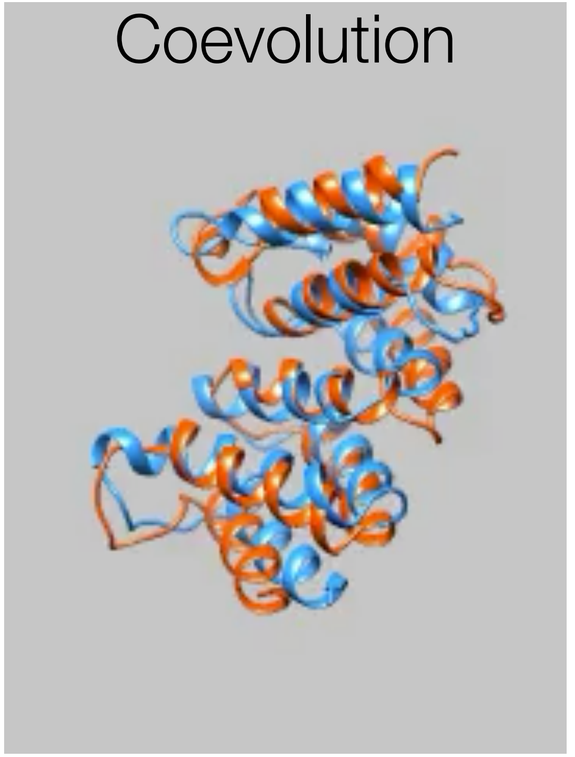


Frag-Rep without coevolution



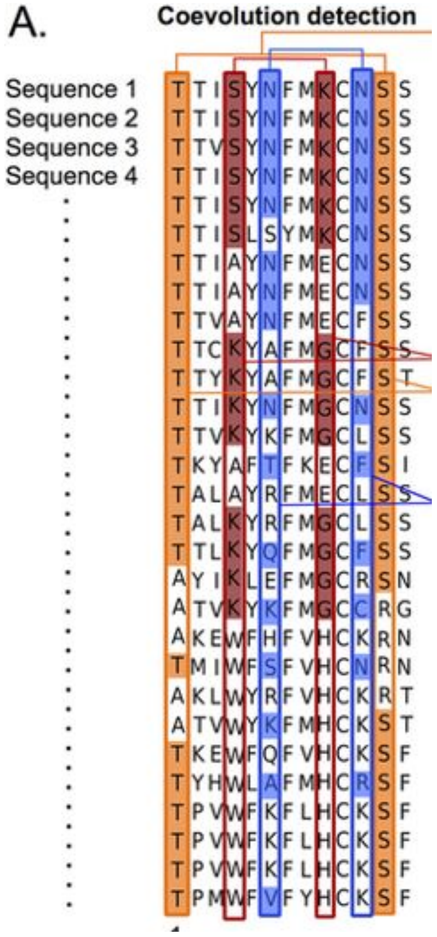


Frag-Rep + Coevolution

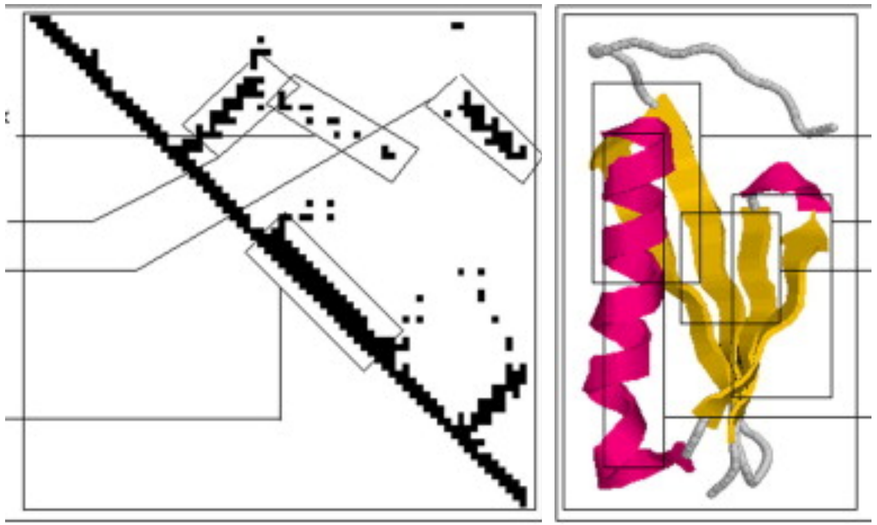


The problem is that proper results are obtained only when thousands of homologue sequences are available

From Sequence->Structure to MSA->Structure



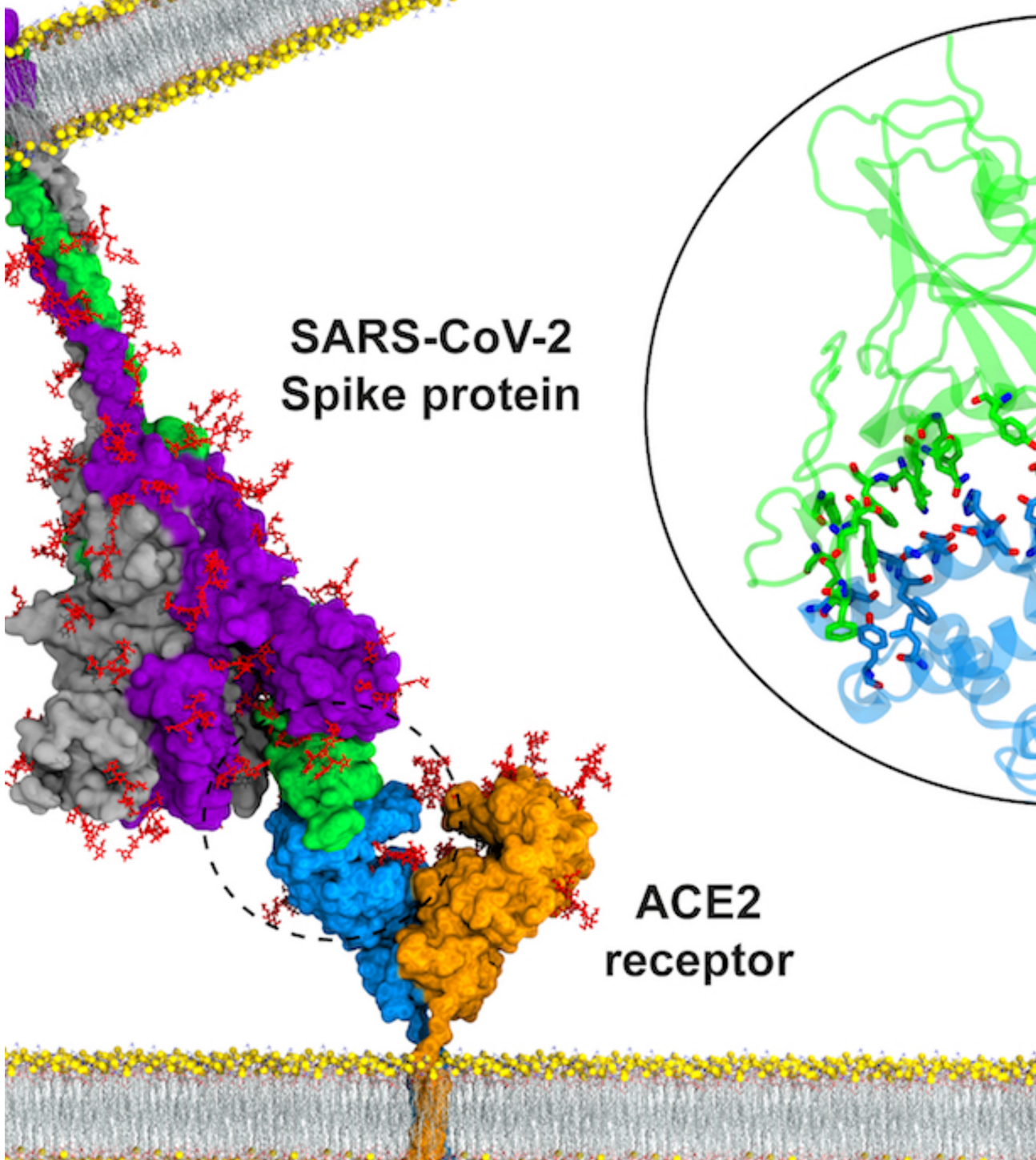
From 2012 it has been clear that when <u>thousands of homologues sequences</u> are available the problem of <u>finding a representative structure</u> associated to them can be solved, but in general they are not available for difficult cases.



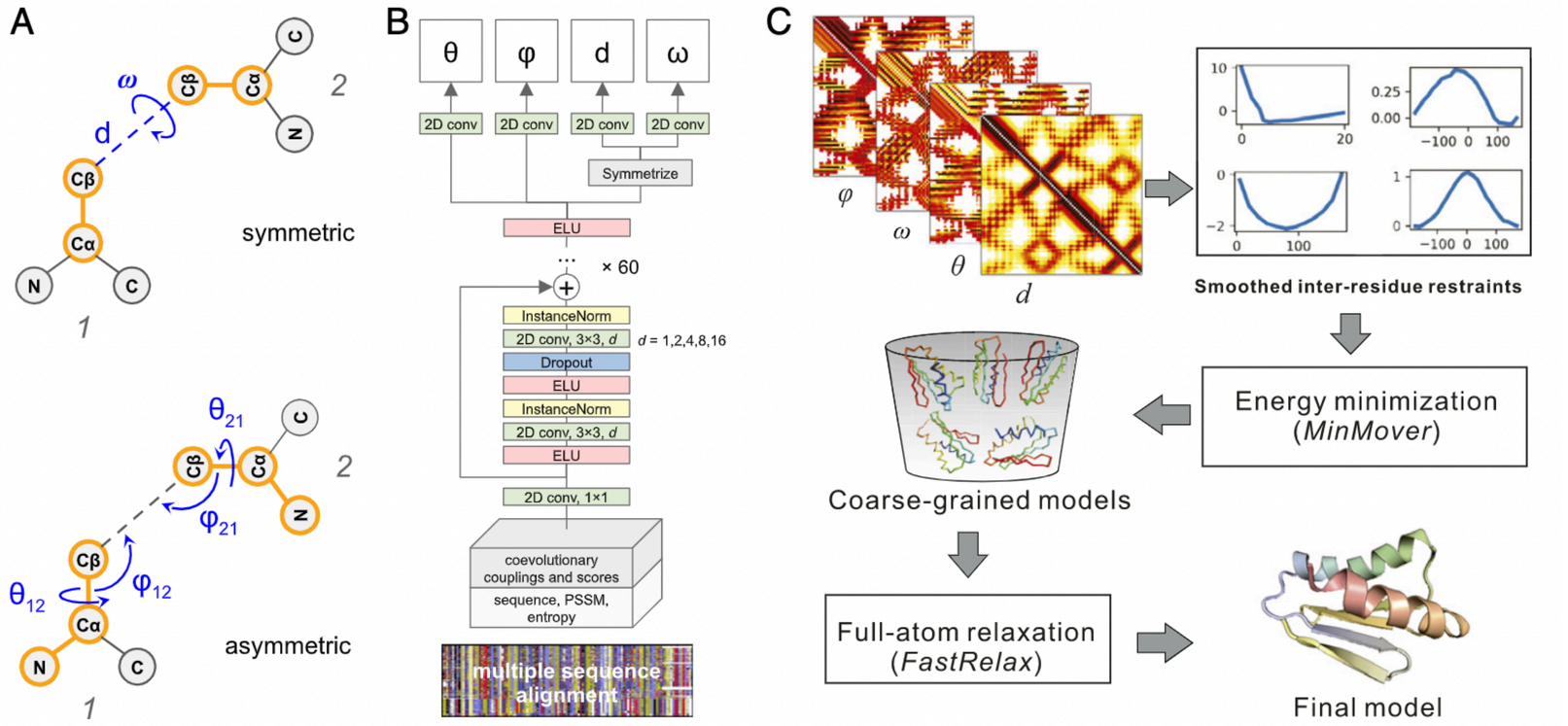
How to improve the quality of the prediction when <u>only few hundreds or less homologue sequences</u> are available?

Outline

- Structure prediction: concepts
- Structure prediction: the origins
- Structure prediction: key advances
- **State of the art and AI approaches**
- Protein complexes and molecular docking
- AI approaches to protein complexes and molecular docking



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

1.Yang, J. *et al*. Improved protein structure prediction using predicted interresidue orientations. *Proc National Acad Sci* **117**, 1496–1503 (2020).

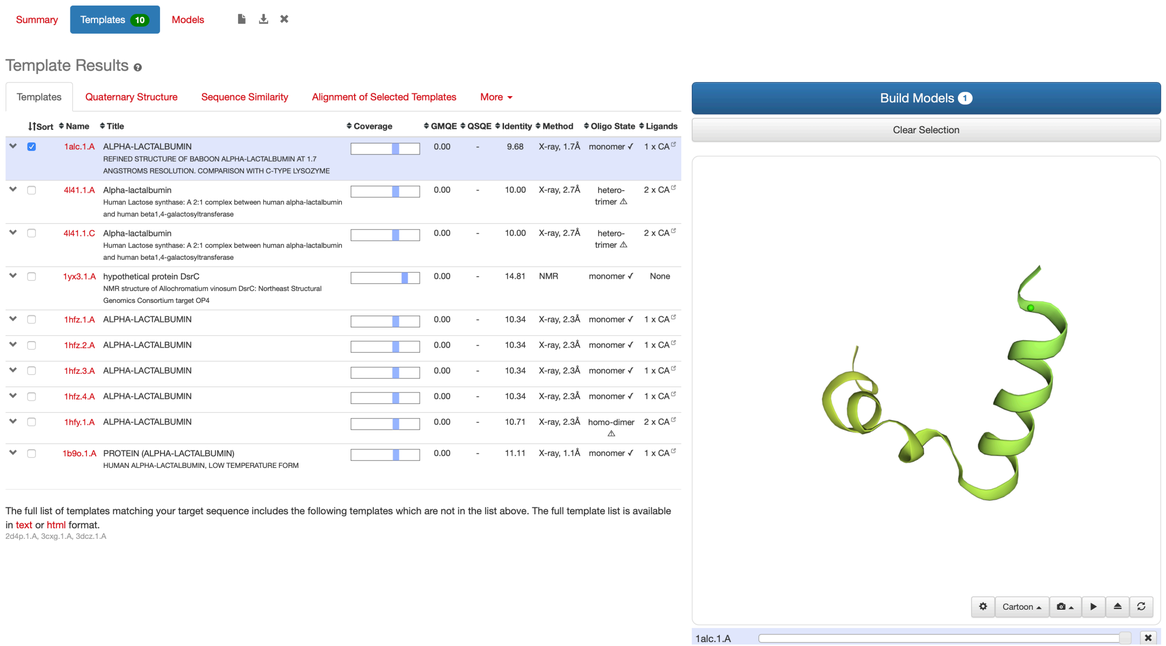
https://github.com/gjoni/trRosetta

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.

```
MVPYQELTLQRSFSYNSRKINPVTSPARSSHVRSPSSSALIPSIPEHELFLVPCRRCSYV
PLSSSSSASHNIGKFHLKFSLLLRSFINIINIPACKMLSLPSPPSSSSSVSNQLISLVTG
GSSSLGRRVTGTLYGHKRGHVTFSVQYNQRSDPVLLLDLAMSTATLVKEMSSGLVRIALE
CEKRHRSGTKLFQEPKWTMYCNGRKCGYAVSRGGACTDTDWRVLNTVSRVTVGAGVIPTP
KTIDDVSGVGSGTELGELLYMRGKFERVVGSRDSEAFYMMNPDKNGGPELSIFLLRI
```

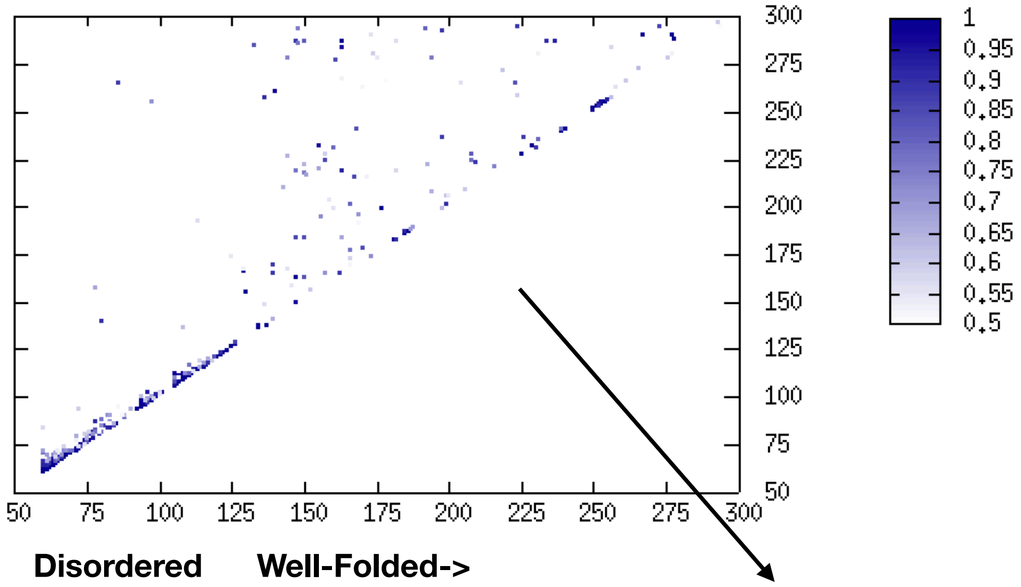
First test - homology modelling - but there are no homologues with known structure:



NOPE!

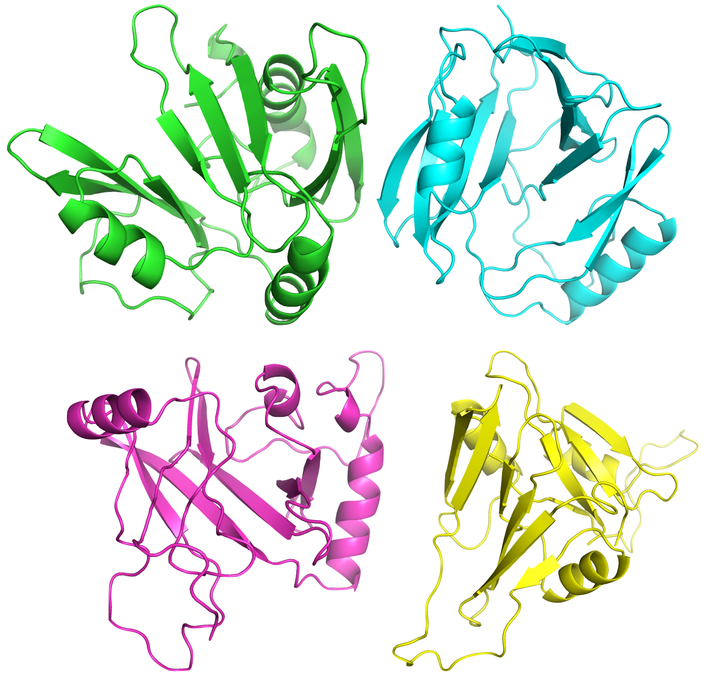
MIZ1: an example from my own experience

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:



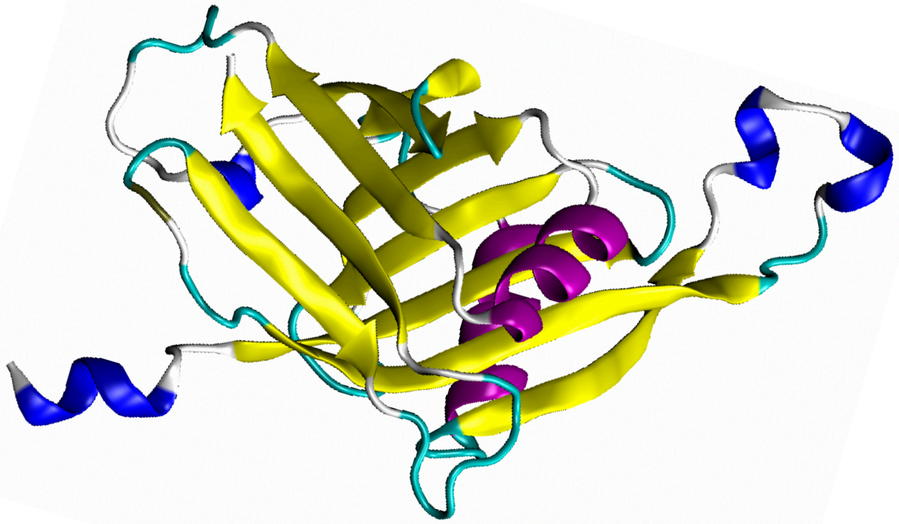
Four structures with nothing in common… nope.

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

trRosetta
(post AF ev-couplings+deep learning):

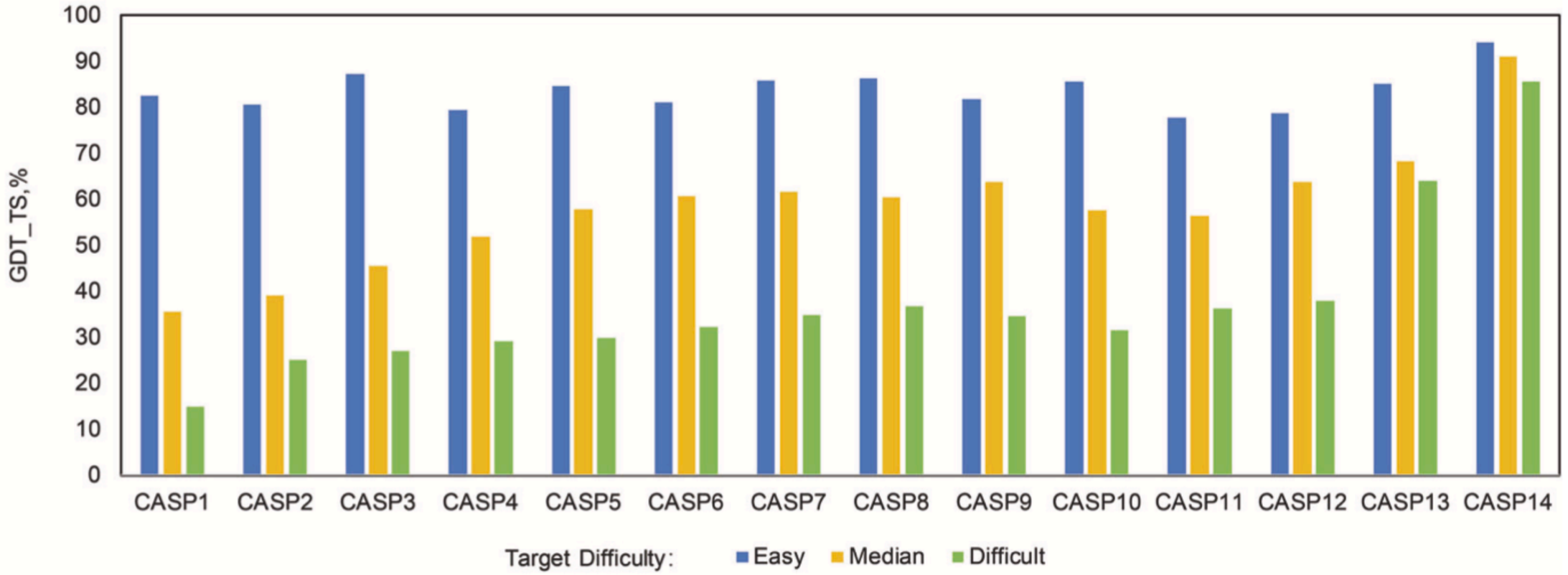


Well converged, well-defined unique structural model with perfectly reasonable features.

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

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 <u>less than one hundred homologue sequences</u> are available?

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

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

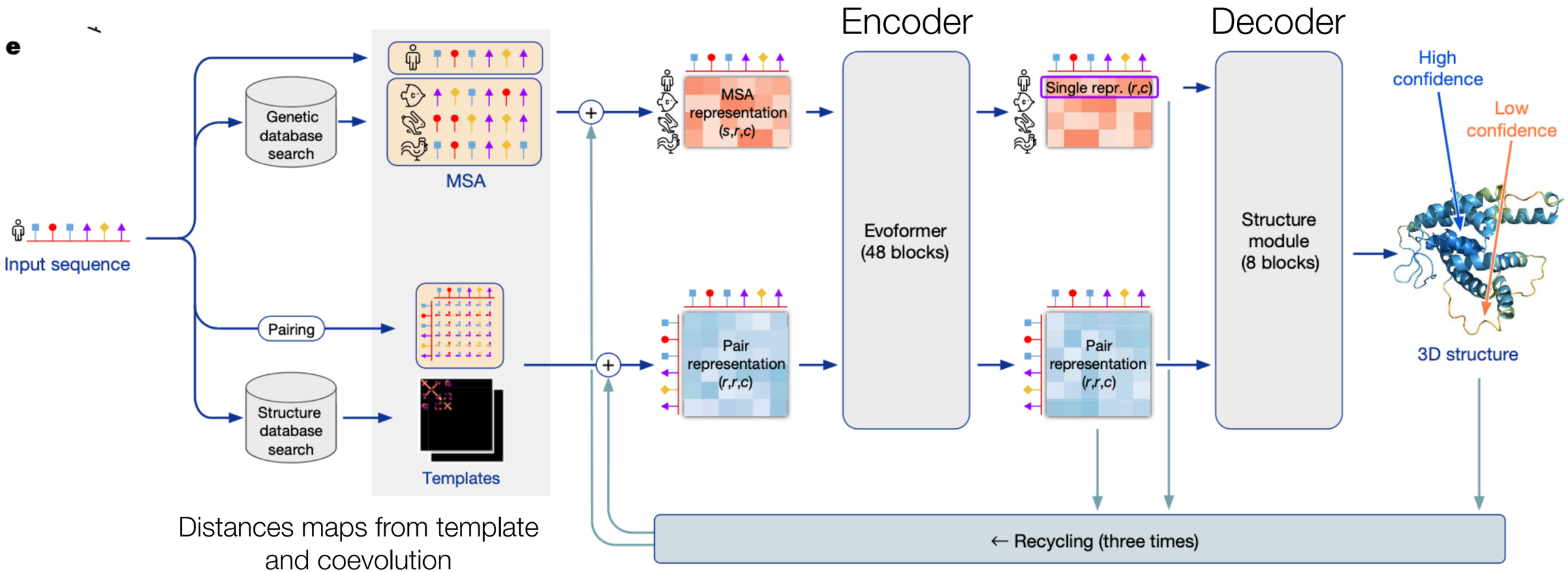
where $\mathcal{L}_{\text{aux}}$ is the auxiliary loss from the Structure Module (averaged FAPE and torsion losses on the intermediate structures, defined in Algorithm 20 line 23), $\mathcal{L}_{\text{dist}}$ is an averaged cross-entropy loss for distogram prediction, $\mathcal{L}_{\text{msa}}$ is an 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.

L_{FAPE} is a clever deviation of the atomic position with respect to the experiment

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.

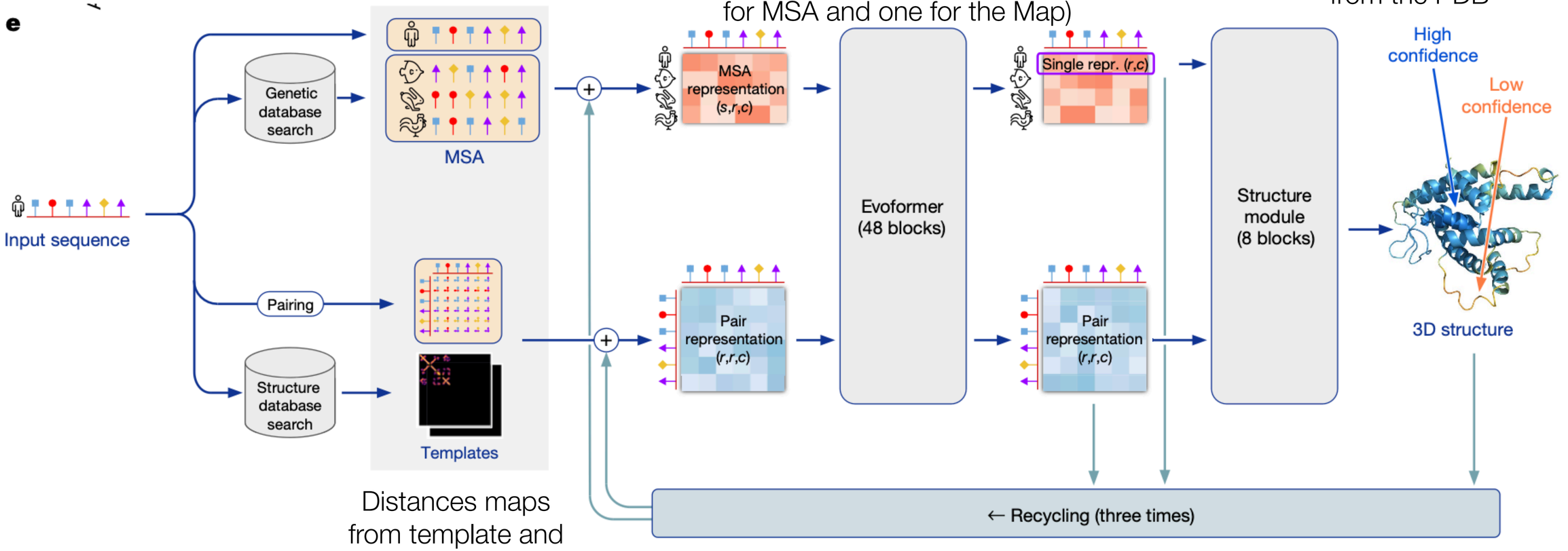


Distances maps from template and coevolution

AlphaFold2: some ideas on the architecture

Evoformer: improve the MSA (better alignment and remove useless/wrong sequences) and the distances map (Two Transformers (Attention) one for MSA and one for the Map)

Get a structure using physical and geometrical constraints learned from the PDB



Distances maps from template and coevolution

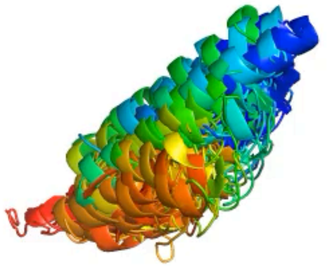
Use the resulting structure to improve again MSA and distances map

AlphaFold2: some ideas on the architecture

The Evoformer encoder tries to build an optimal representation of the MSA and pair distance matrix that at once gives an optimal aligment and optimal geometries (that is triplet of amino acids should be characterised by distances that are compatible with triangles).

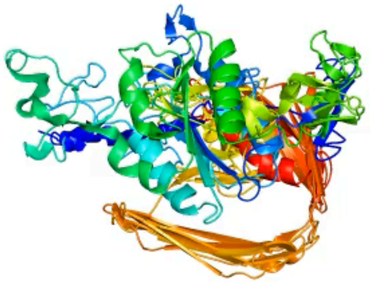
This information is passed to the structure module that considers the protein as a “residue gas”. Every amino acid is modelled as a triangle, representing the three atoms of the backbone. These triangles float around in space, and are moved by the network to form the structure.

Examples:



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

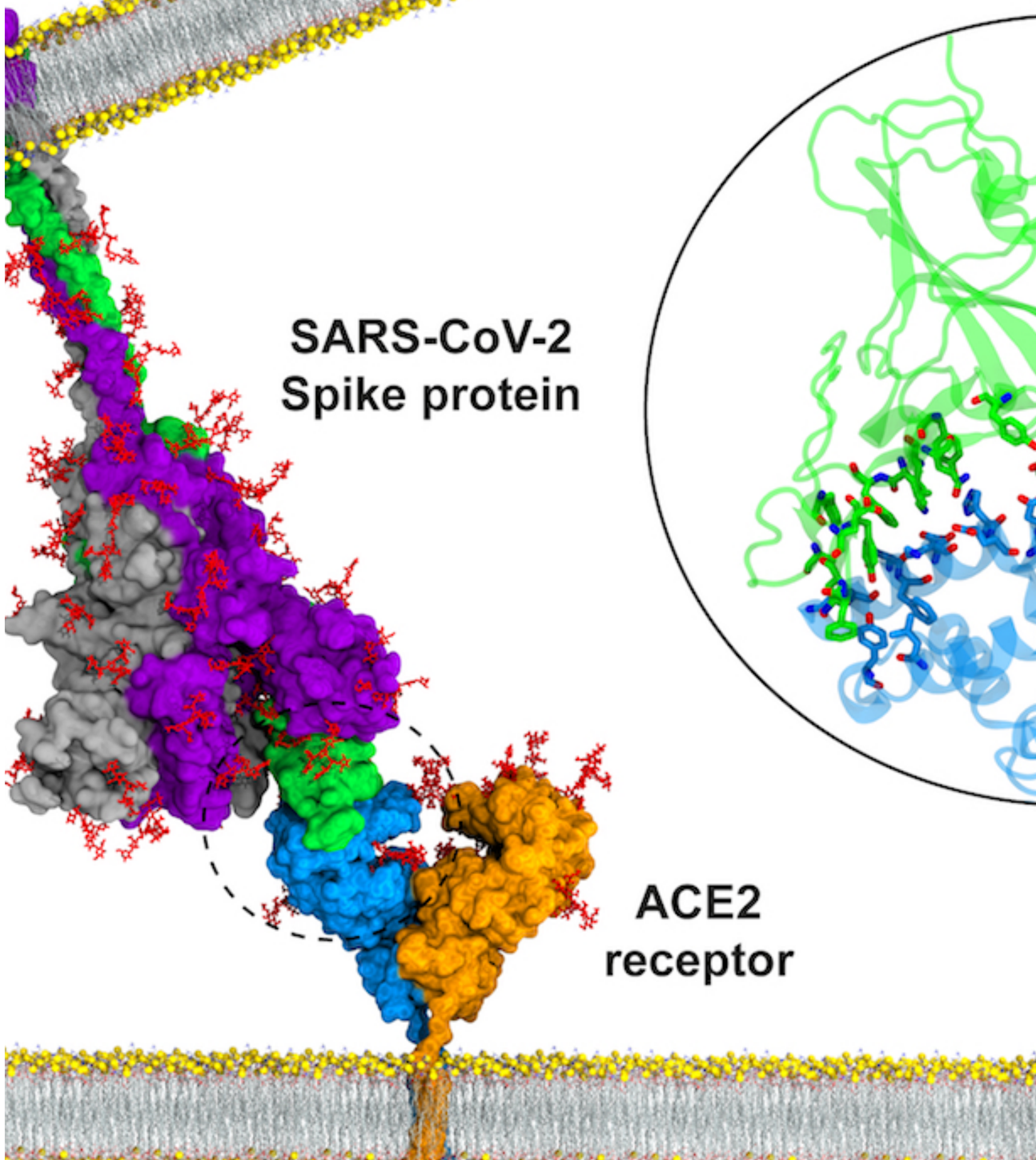
Examples:



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

Outline

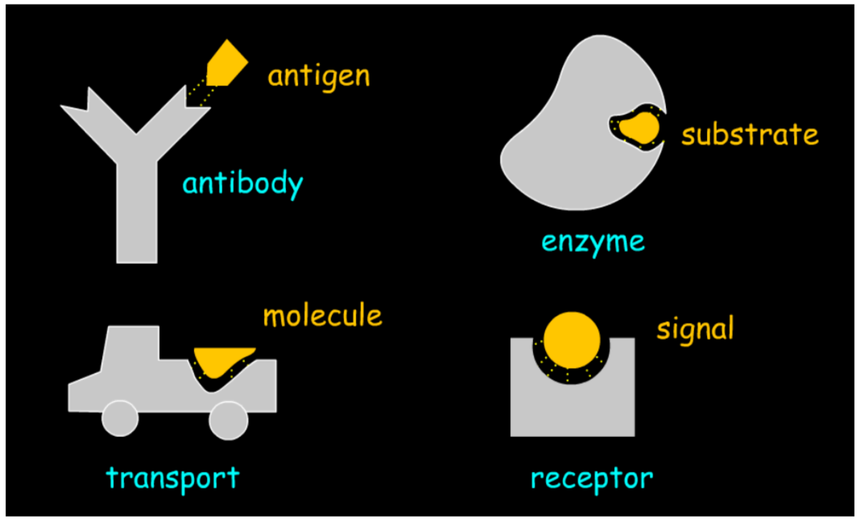
- Structure prediction: concepts
- Structure prediction: the origins
- Structure prediction: key advances
- State of the art and AI approaches
- **Protein complexes and molecular docking**
- AI approaches to protein complexes and molecular docking



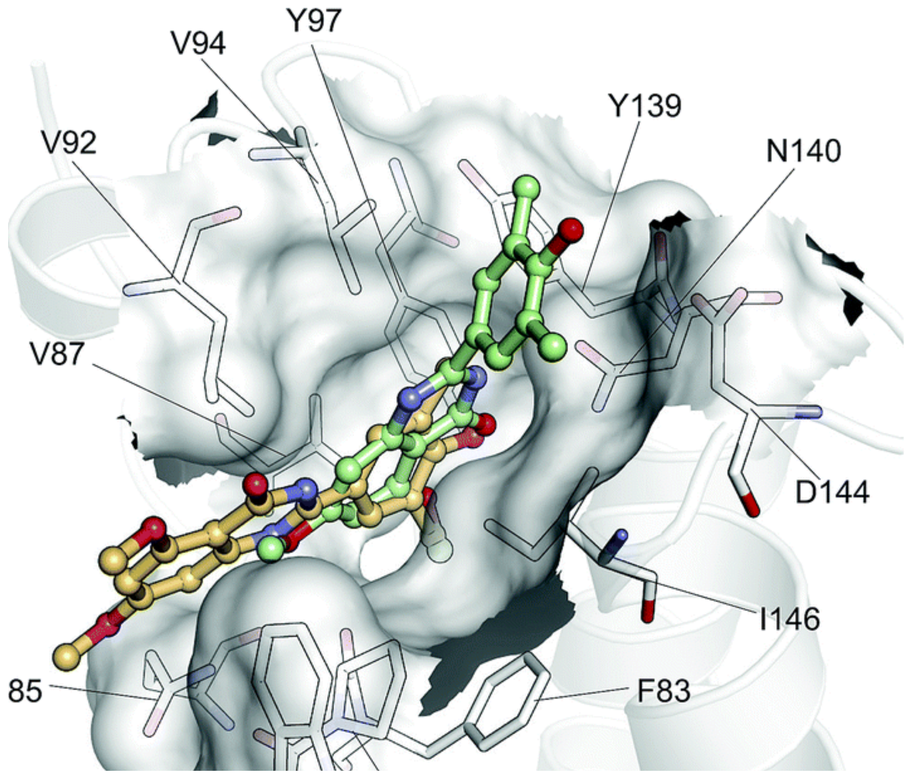
Molecular Recognition and Docking

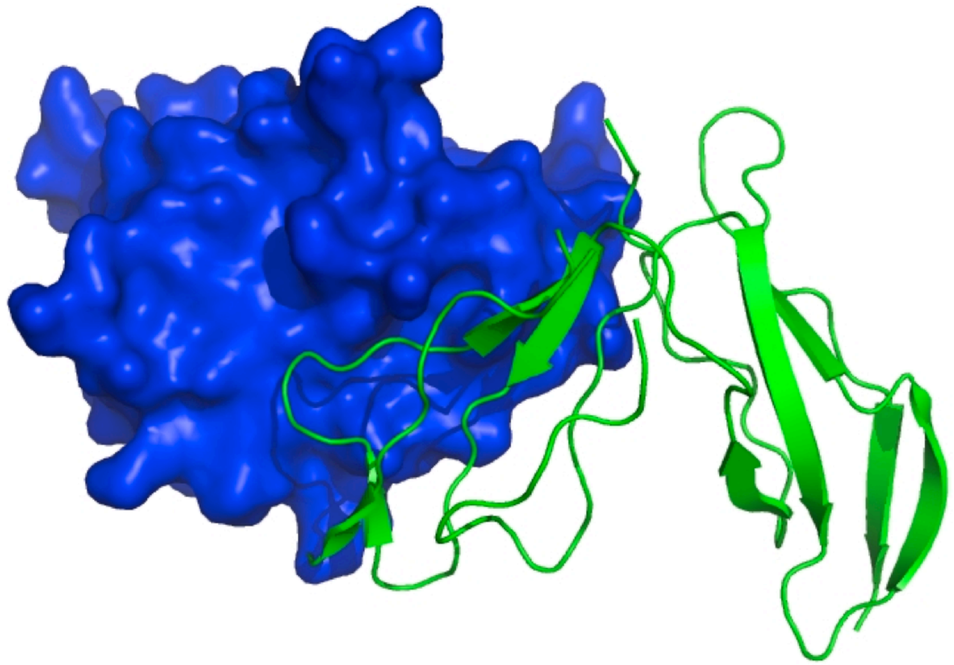
Molecular recognition is the ability of biomolecules to recognize other biomolecules and selectively interact with them in order to promote fundamental biological events such as transcription, translation, signal transduction, transport, regulation, enzymatic catalysis, viral and bacterial infection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



Molecular Docking: binding and quaternary structures



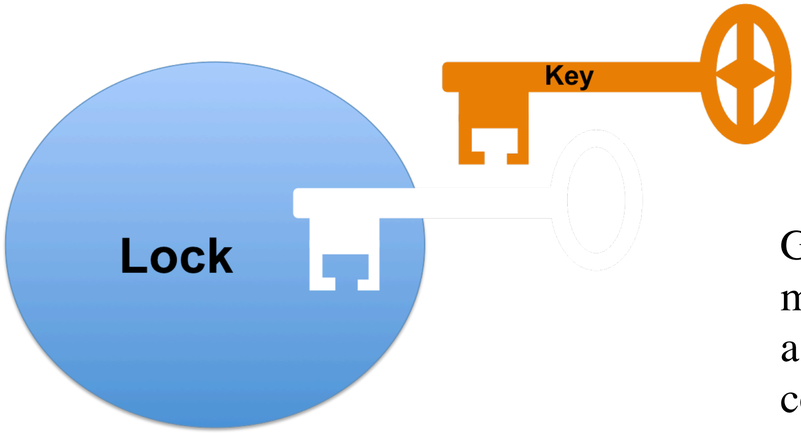


Prediction of the <u>bound structure and binding strength</u> for

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

Complexation can be associated with conformational changes

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)**

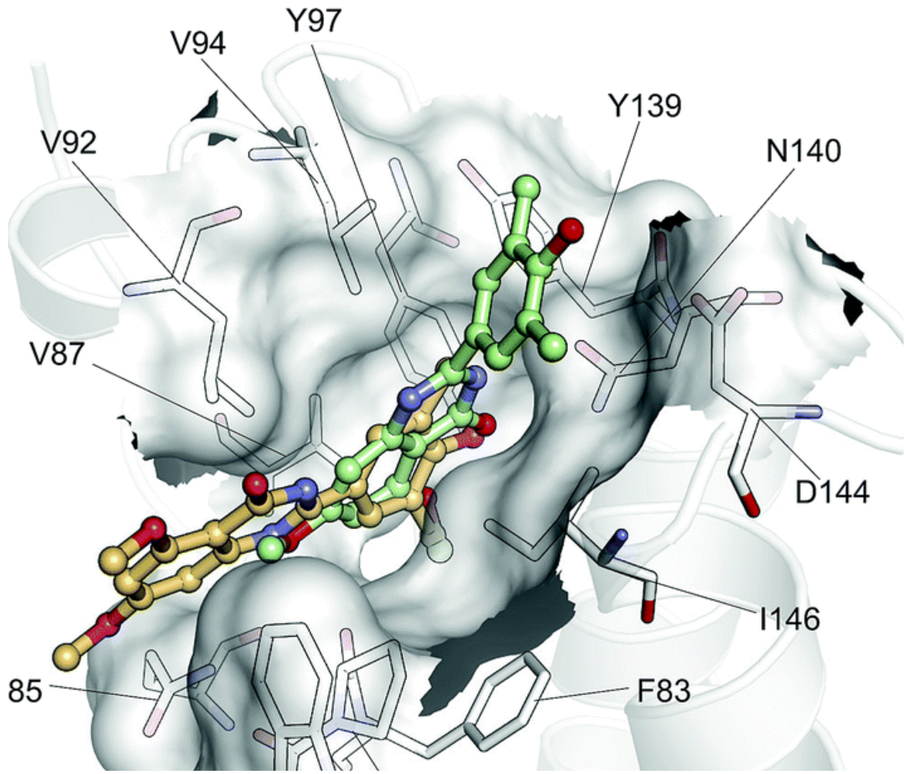
Docking mechanisms

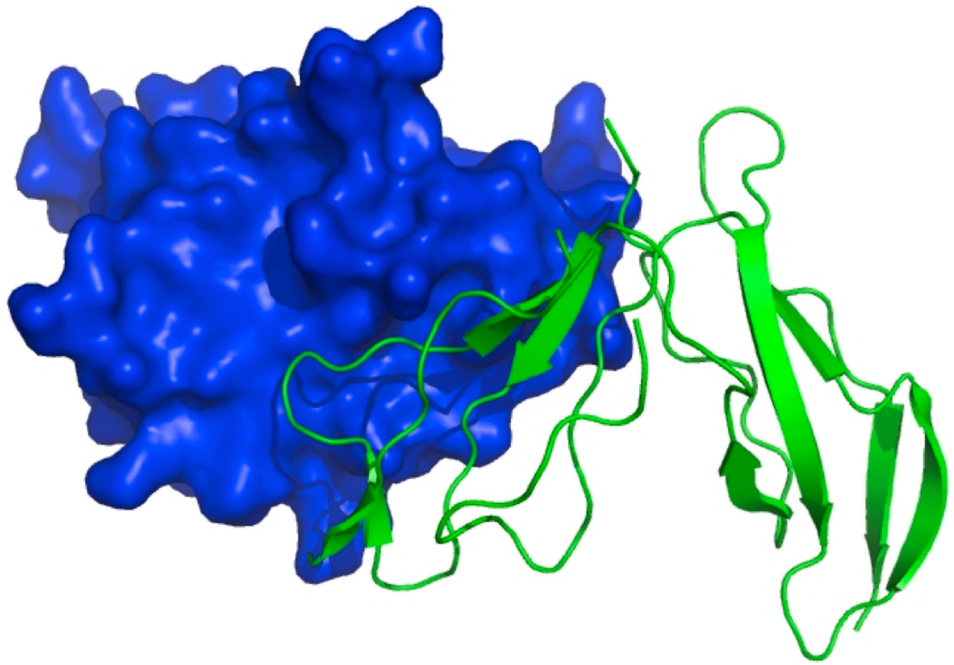
Lock-and-Key: this is simple, the assumption is that both the receptor and the ligand are rigid and their structure is ideal for binding. The next step is to accept that the ligand (this is more specific for the binding of small molecules) can be flexible, while the receptor is rigid.

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 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.

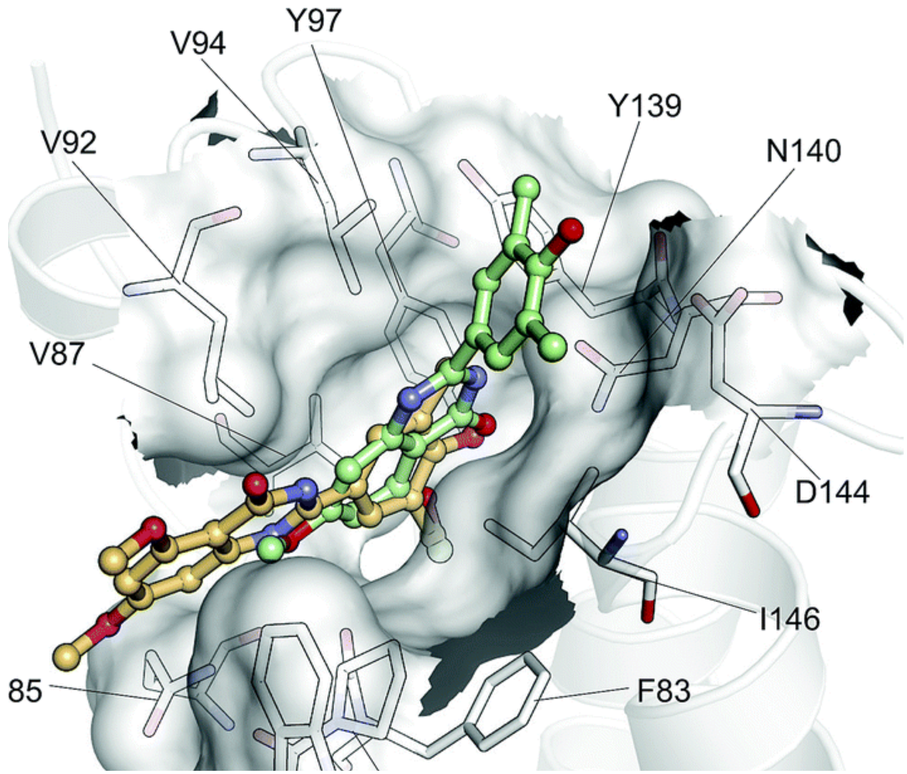
Molecular Docking: binding and quaternary structures





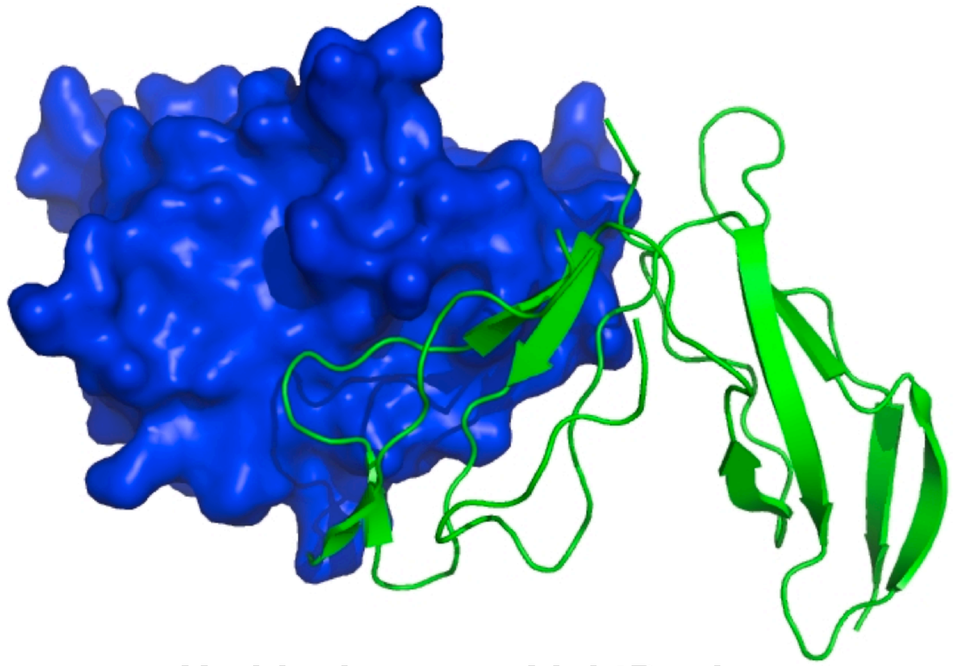
The general principles and aims are the same, but the practical implementations are generally very different so you will find different tools for the two cases

Molecular Docking: softwares



Autodock vina
UCSF dock
Glide
...

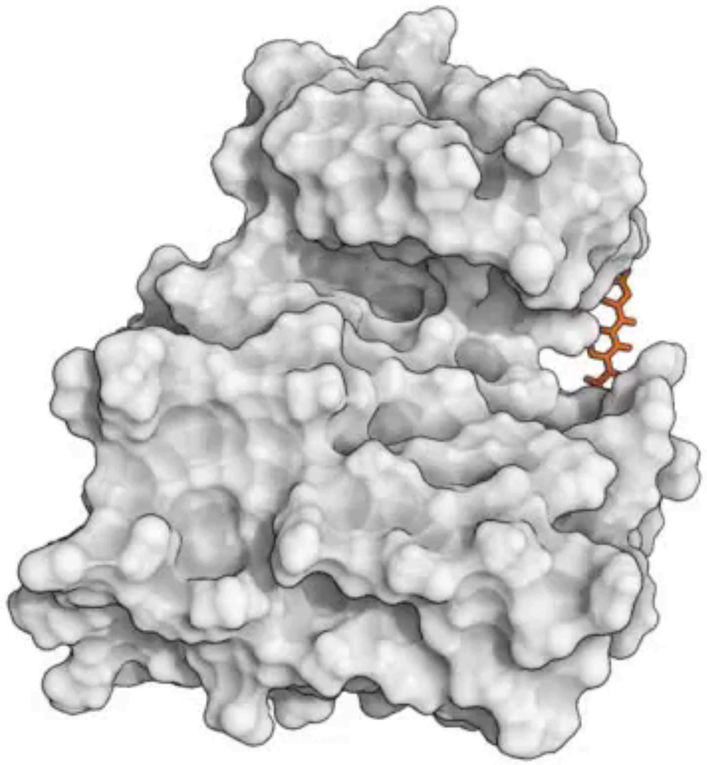
DiffDock
...



Haddock
Rosetta Dock
ClusPro
...

LightDock
AFMultime
RosettaFold2
...

MD simulations



In principle MD simulations can be used to search for ligand binding, but the are problems:

1) Parametrize force-fields for all the possible chemicals is hard (this is more relevant for ligand binding than for protein complexes)
2) The binding time scale can be slow in particular when associated with conformational changes (the problem of sampling). This is even more relevant for protein complexes where the systems become large and so the simulations become slow.

1.Shan, Y., Kim, E. T., Eastwood, M. P., Dror, R. O., Seeliger, M. A. & Shaw, D. E. How does a drug molecule find its target binding site? *J Am Chem Soc* **133,** 9181–3 (2011).

Rigid or semi-flexible docking: a simplified simulation approach

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?)**

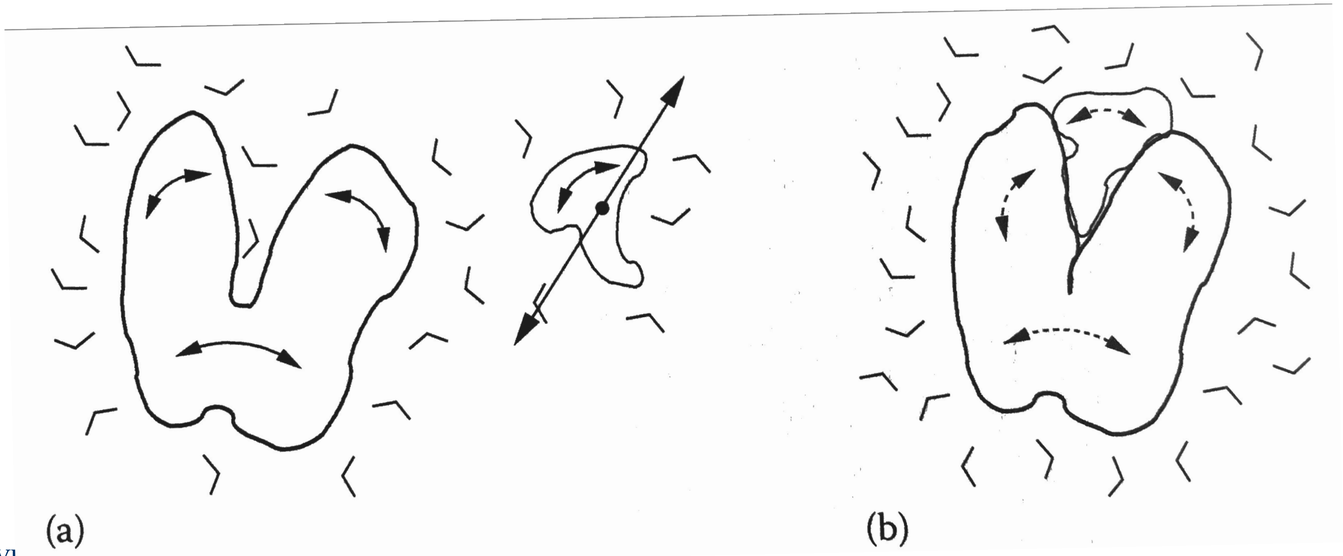
A number of poses can be generated at random by rotation and translation of a molecule around the other (using Monte Carlo or genetic algorithms) and then they can be scored for example using a physicochemical inspired scoring function (a simplified force-field). Protein Flexibility can be introduced using multiple structures, while ligand flexibility can be introduced allowing the molecule to rotate around “rotatable” bonds.

Rigid or semi-flexible docking: a simplified simulation approach

From calorimetry we can measure binding free energies ($\Delta G_{binding}$). 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.

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

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

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

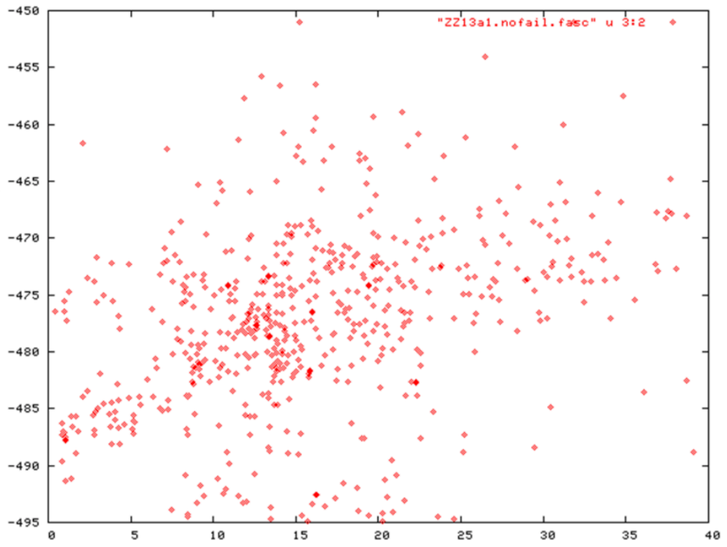
These are all related to intermolecular interactions. The hydrophobic term can be seen as the one accounting for solvation.

This is then only source of internal energy

http://www.pdbbind.org.cn/index.asp

Scoring function and convergence

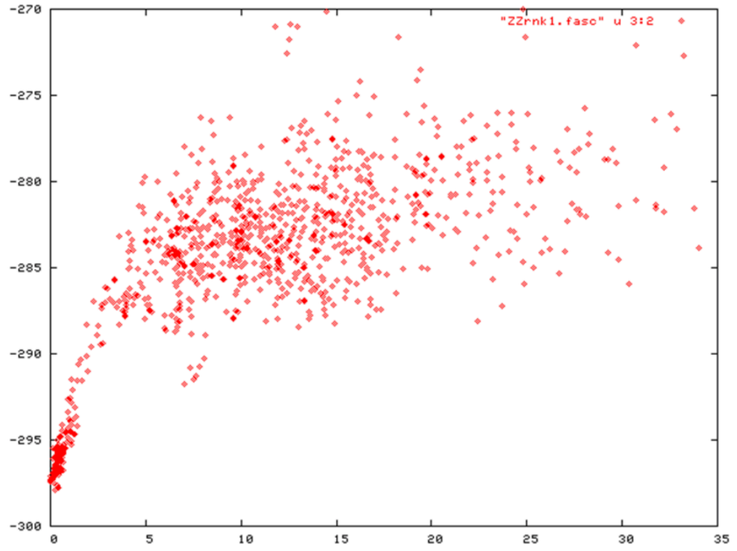
Energy



RMSD

Bad

Energy

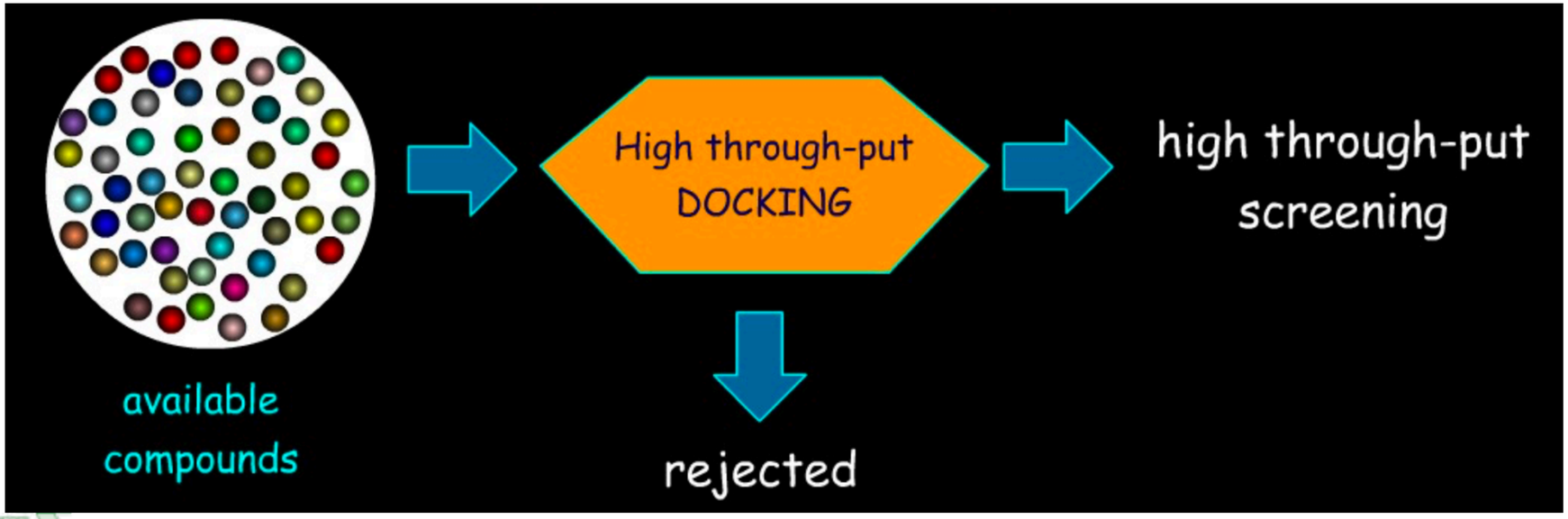


RMSD

Good

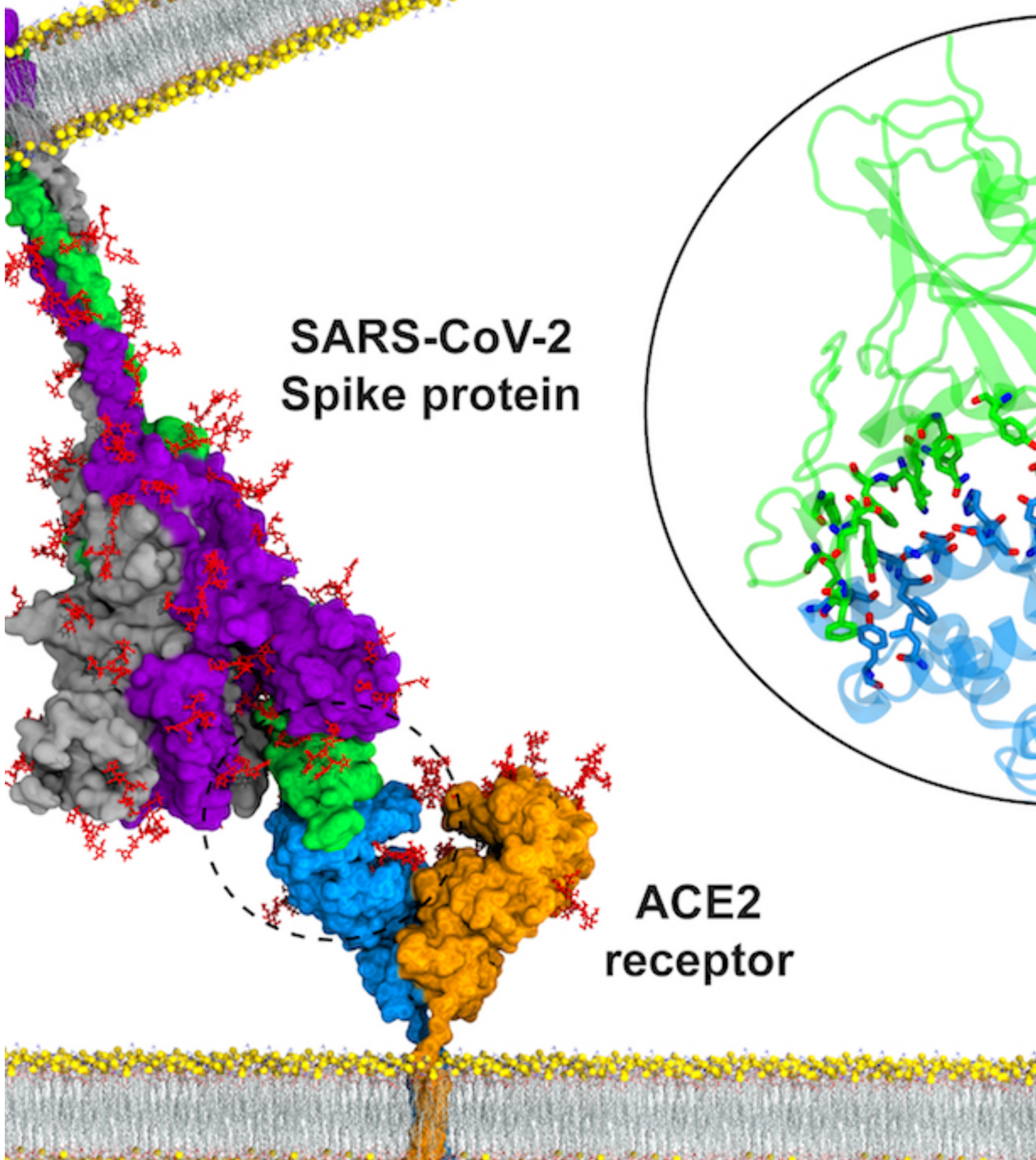
Virtual Screening

- 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



Outline

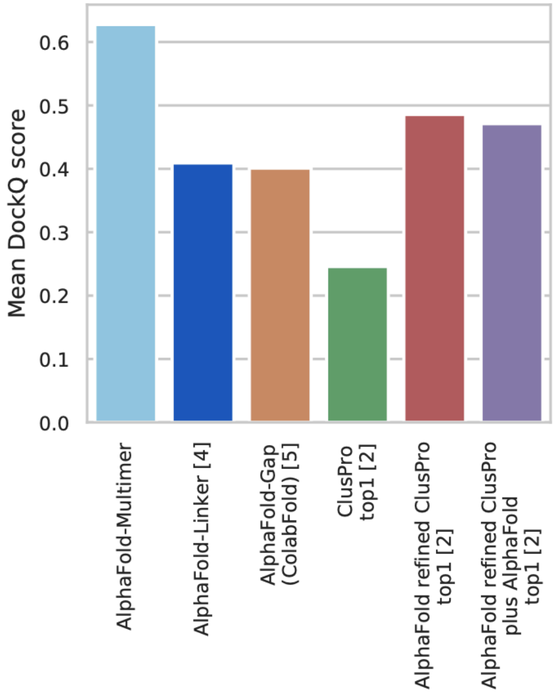
- Structure prediction: concepts
- Structure prediction: the origins
- Structure prediction: key advances
- State of the art and AI approaches
- Protein complexes and molecular docking
- **AI approaches to protein complexes and molecular docking**



AF2 Multimer: quaternary structure and protein complexes

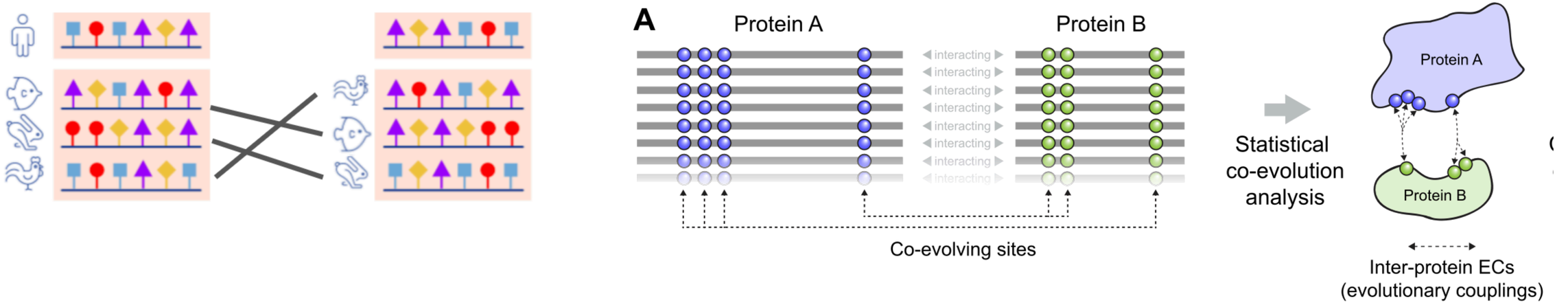
AF2 multimer is a retraining of AF2 specialised for complexes. The network architecture is the same of AF2 but the scoring function is modified to work better for complexes:

- The FAPE scoring that scores the local geometry is applied with 3 nm instead than 1 nm
- An additional loss is applied to keep the center of mass of the chains sufficiently farther apart
- The number of atomic clashes is less relevant to increase the chances of getting interfaces (at the expense of structure quality
- The network is trained on complexes



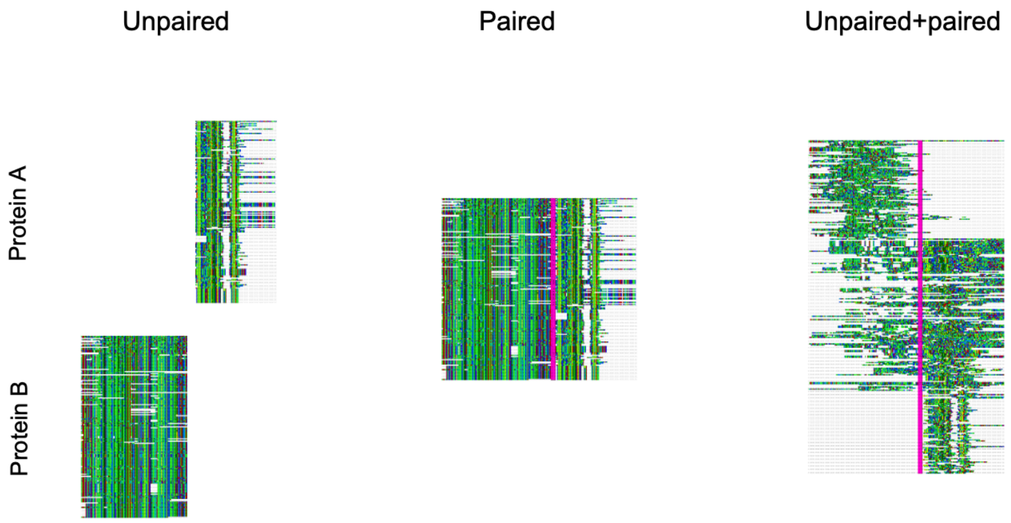
(a)

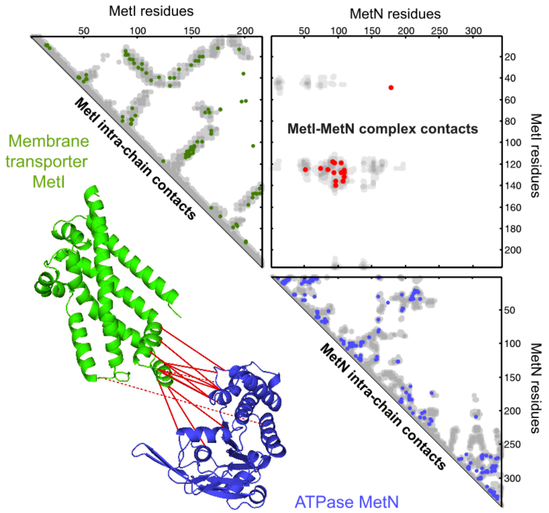
The key is still the co-evolution analysis from MSA

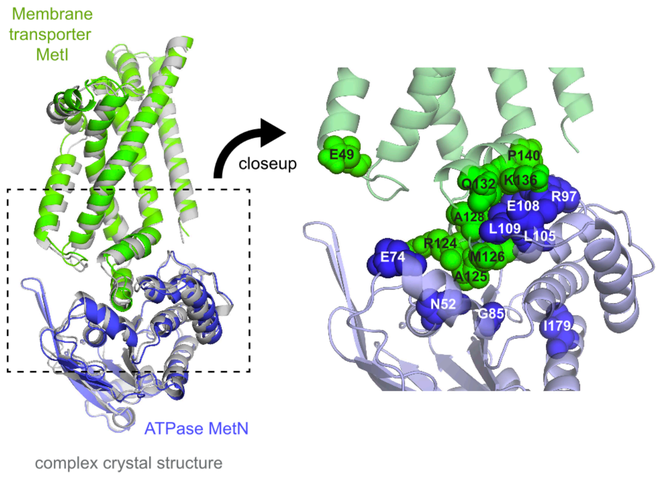


The classical approach is to use this information for rigid docking, that is to solve the jigsaw of combining two 3D structure together. One common issue is that the docking is generally not rigid. This can be due to local changes or to global ones. Furthermore PAIRED MSA are more difficult to obtain because there is less statistics

The key is still the co-evolution analysis from MSA







AF2 multimer partially solve the problem of the flexibility because it solve the structure of the complex all the once. Nonetheless is still strongly limited by the lack of MSA statistics and by the fact that the conformational space to search is much larger.

RosettaFold2: a combination of AF2 and RF

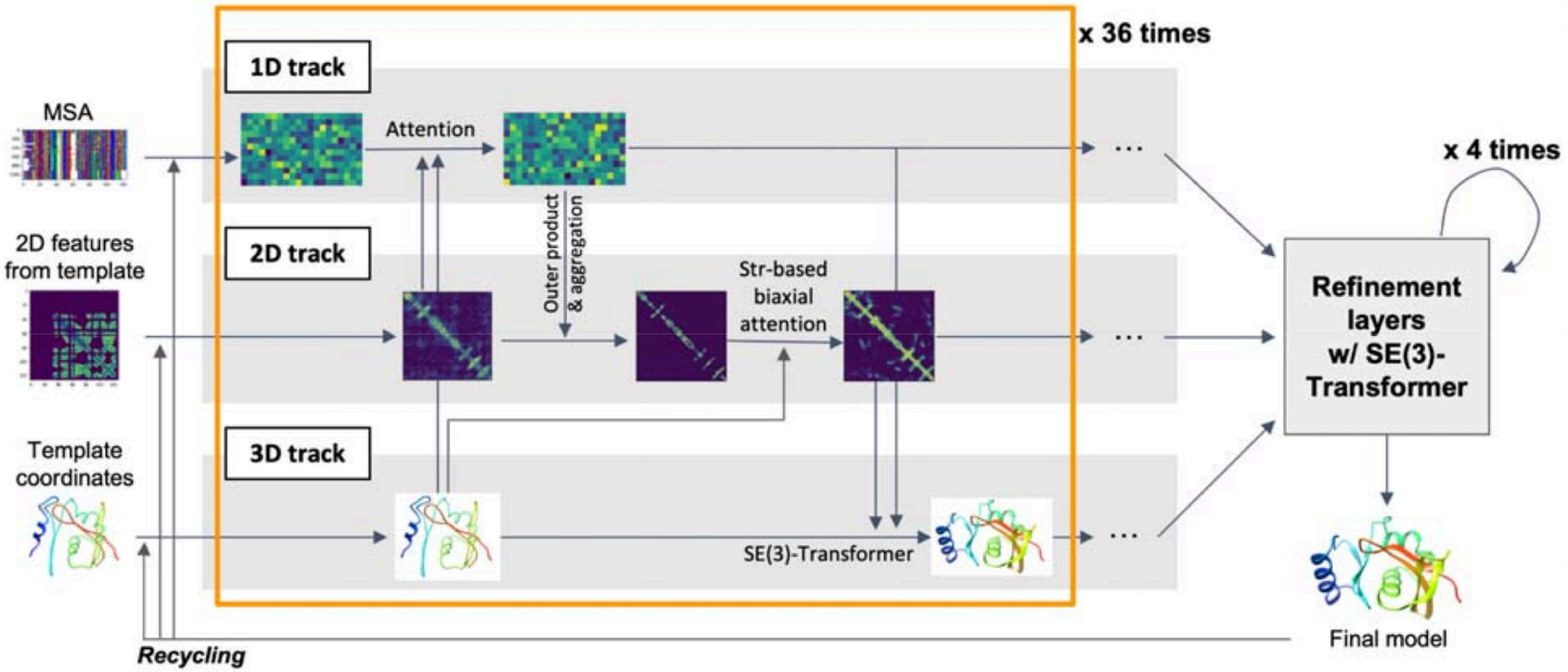


Figure 1. An overview of the RF2 3-track architecture. Three parallel tracks synchronize and update a representation of the sequence, residue-pair, and 3D structure, producing a fine 3-dimensional structure.

A key difference of RF is that instead getting a structure from the MSA plus distance matrix, it brings forward at the same time the MSA, the distance matrix and the 3D structure. This has implication for protein design because a 3D structure can provide input information.

RosettaFold2: a combination of AF2 and RF

The model works for both tertiary and quaternary (complex) structures. Its prediction performances are comparable to AF2 but with better timings.

The key result is that of showing that AF2 is not the only solution to the structure prediction problem and that different architectures can actually work equally well.

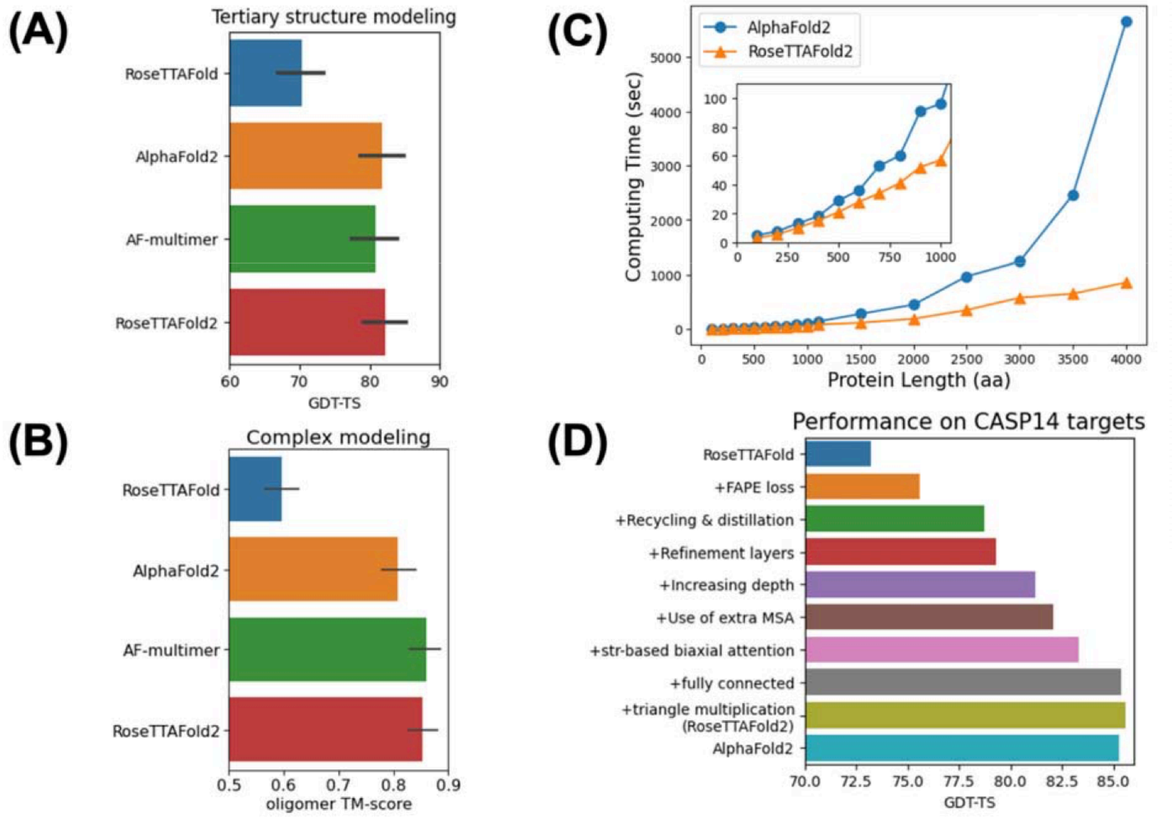
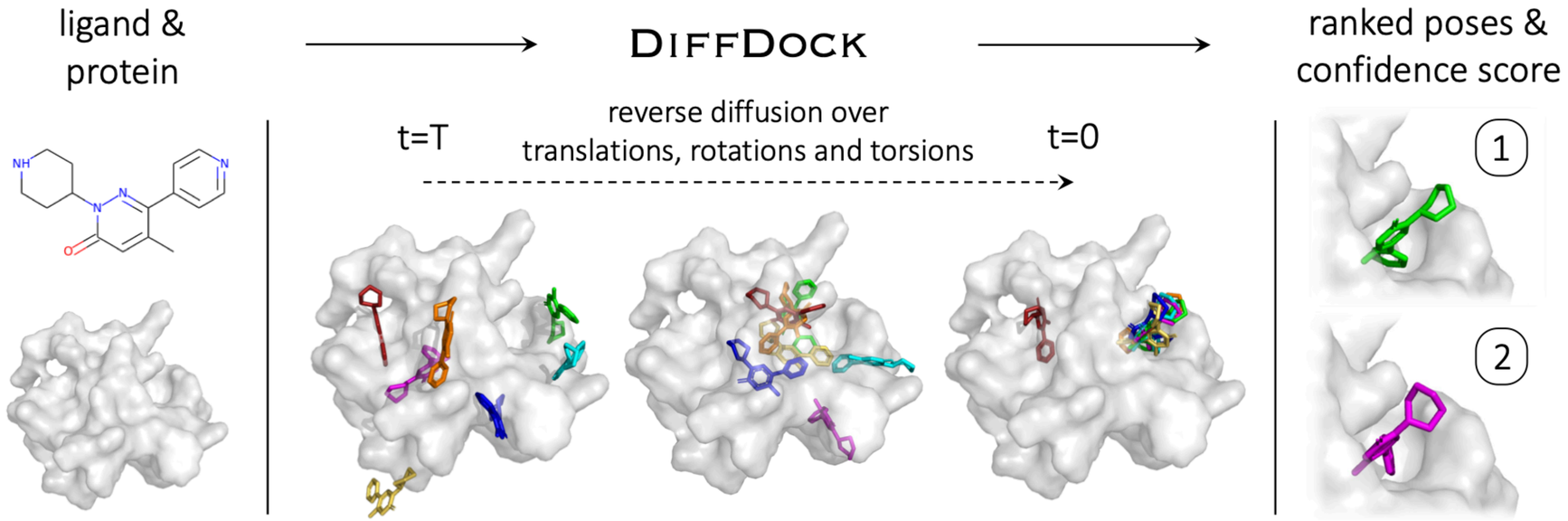


Figure 2. Comparison of the results of RF2 versus AF2. On a set of recently-solved structures (released after training of AF and RF2), we compare the average accuracy of structure prediction over A) 113 protein monomers and B) 140 protein hetero-dimer complexes. C) A comparison of AF2 and RF2 in terms of runtime scaling. D) CASP14 performance as individual features were added to the model.

Diffdock



pirically, DIFFDOCK obtains a 38% top-1 success rate (RMSD<2Å) on PDB-Bind, significantly outperforming the previous state-of-the-art of traditional docking (23%) and deep learning (20%) methods. Moreover, while previous methods are not able to dock on computationally folded structures (maximum accuracy 10.4%), DIFFDOCK maintains significantly higher precision (21.7%). Finally, DIFFDOCK has fast inference times and provides confidence estimates with high selective accuracy.

https://doi.org/10.48550/arXiv.2210.01776