Fold Prediction

Fold prediction

1. Fold recognition (threading)
2. *ab initio* fold prediction
3. Protein folding (MD with explicit solvent)

Threading

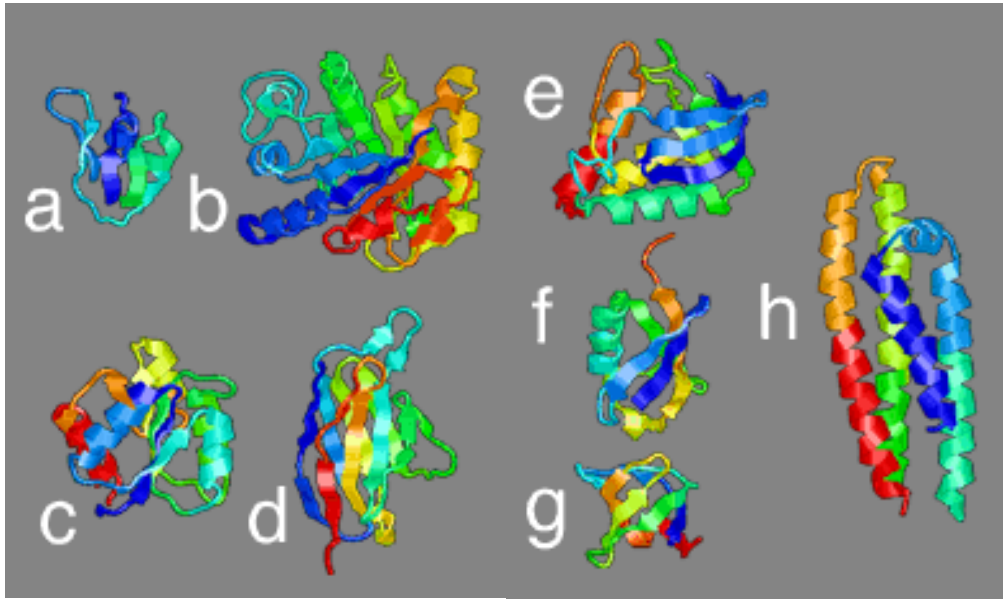
Idea: Find the optimal structure for a new (target) sequence in the set of known 3D-structures (templates) by threading the target sequence.

Fold recognition / Threading

Principle: Find a compatible fold for a given sequence ....

```
>Protein XY
MSTLYEKLGGTTAVDLA
VDKFYERVLQDDRIKH
FFADVDMAKQRAHQK
AFLTYAFGGTDKYDGR
YMREAHKELVENHGLN
GEHFDAVAEDLLATLKE
MGVPEDLIAEVAAVAG
APAHKRDVLNQ
```

≈?



Using ...

- 1D – 3D profile matching,
- mean force potentials,
- secondary structure predictions,
- position specific scoring matrices (PSSM),
- keyword statistics,
- ....

1. Fold recognition (threading)
    1. Knowledge-base potentials
        1. Distance dependent potentials
            - Atom-centered
            - Sequence distance
            - Reference state
        2. Solvation
        3. Z-scores and energy profiles
        4. Methods: Prosa, Anolea, DOPE
    2. Distance homology matrices (PSSM)
        1. Function association
        2. Methods: FUGUE, PHYRE, ModLink
    3. Secondary structure alignment
        1. Secondary structure prediction
            - Machine learning theory
            - Neural Networks
        2. Methods: TOPITS

1. Knowledge-base potentials

1. Distance dependent potentials

According to Boltzmann law

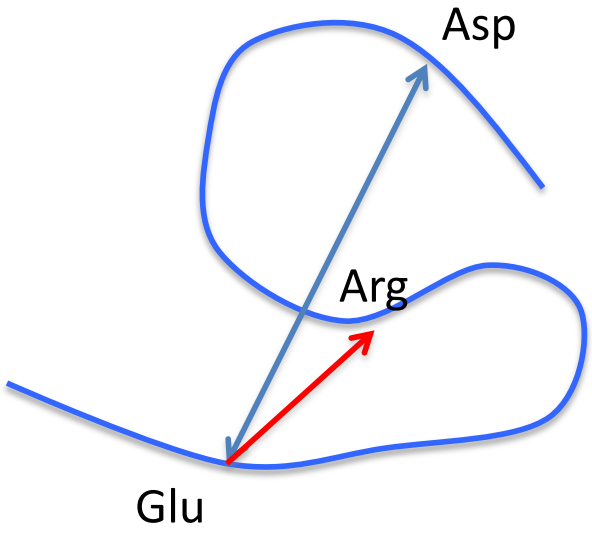
$$P(x) = \frac{1}{Z} e^{-E(x)/k_B T}$$

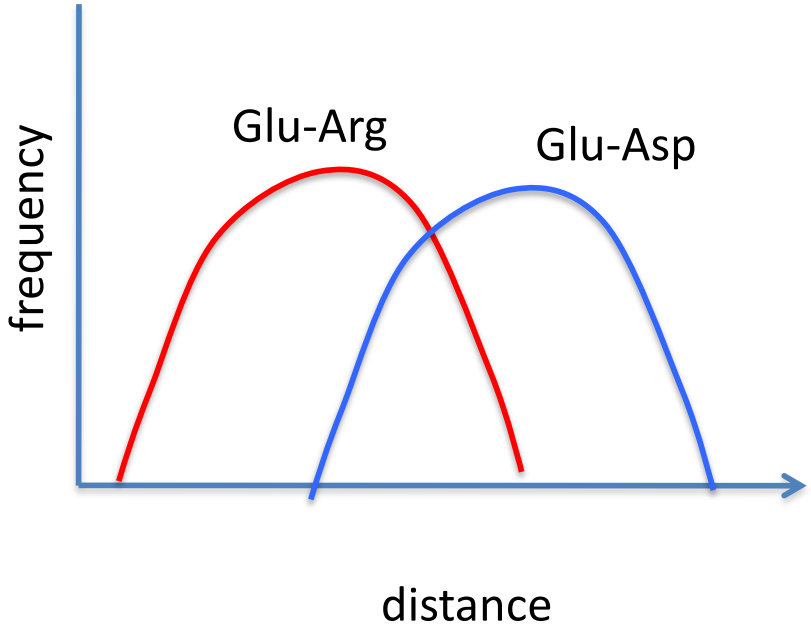
Therefore, energy is related with probability

$$P(Asp, Asp, d = 10A) \Rightarrow E(Asp, Asp, d = 10A)$$

1. Knowledge-base potentials

1. Distance dependent potentials

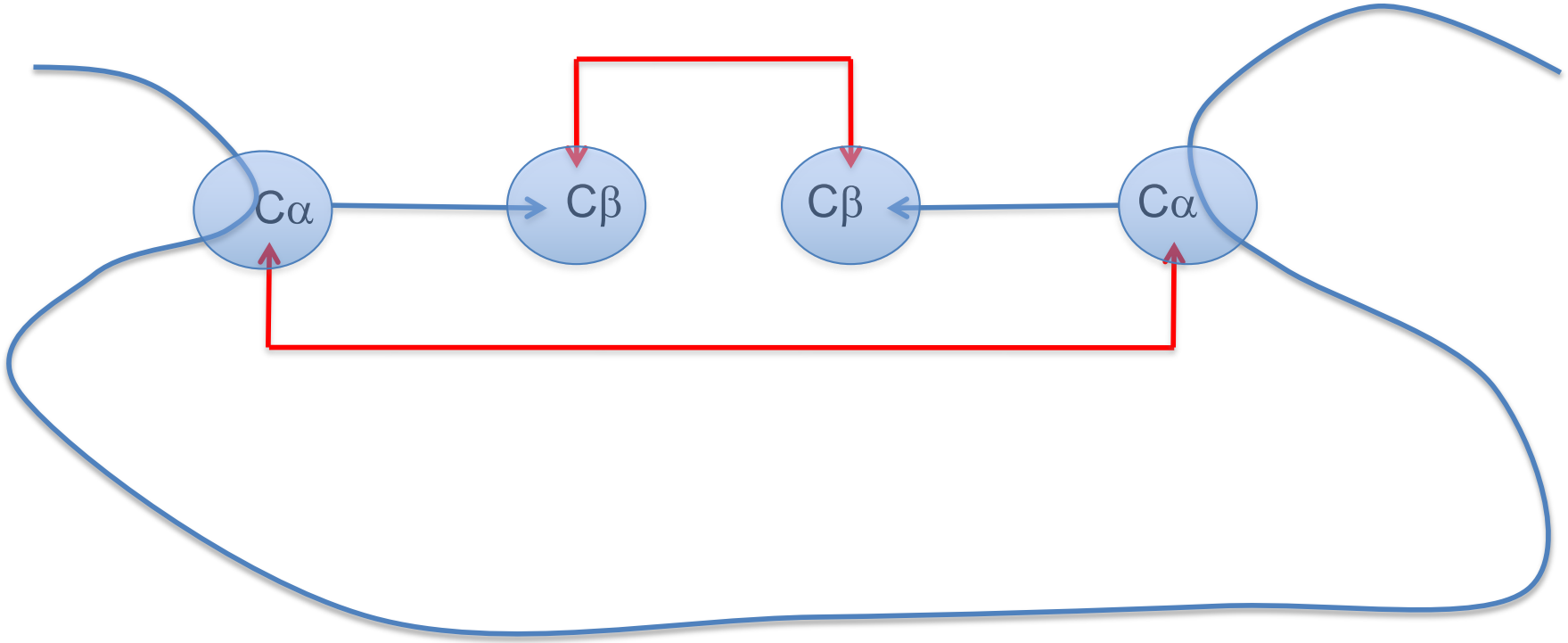




1. Knowledge-base potentials

1. Distance dependent potentials

1. Distances are calculated between atoms: We have to select what atom are we going to use
   •The best choice is Cβ because it indicates the direction of the side-chain



1. Knowledge-base potentials

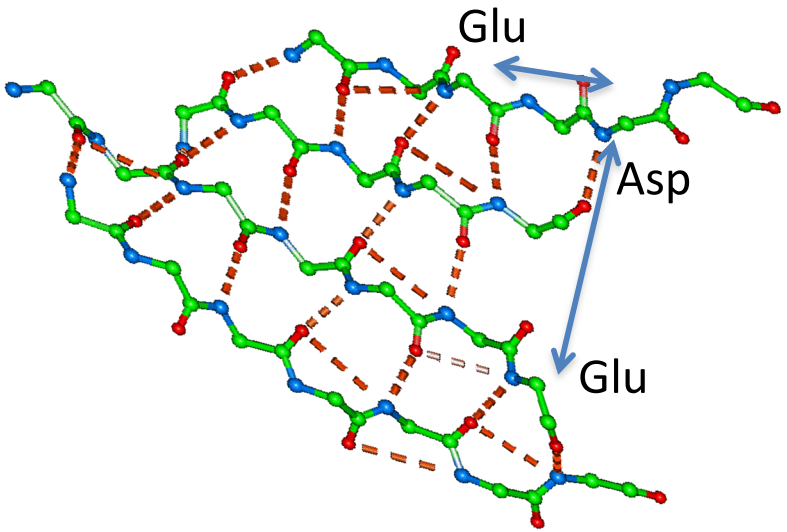
1. Distance dependent potentials

2. The database of structures to extract distances has to avoid redundant structures (between homologs and members of the same family/superfamily)
   - If we use all the structures of the same or similar protein there will be a bias. Thus, we use a set with less than 40% of sequence similarities

1. Knowledge-base potentials

1. Distance dependent potentials

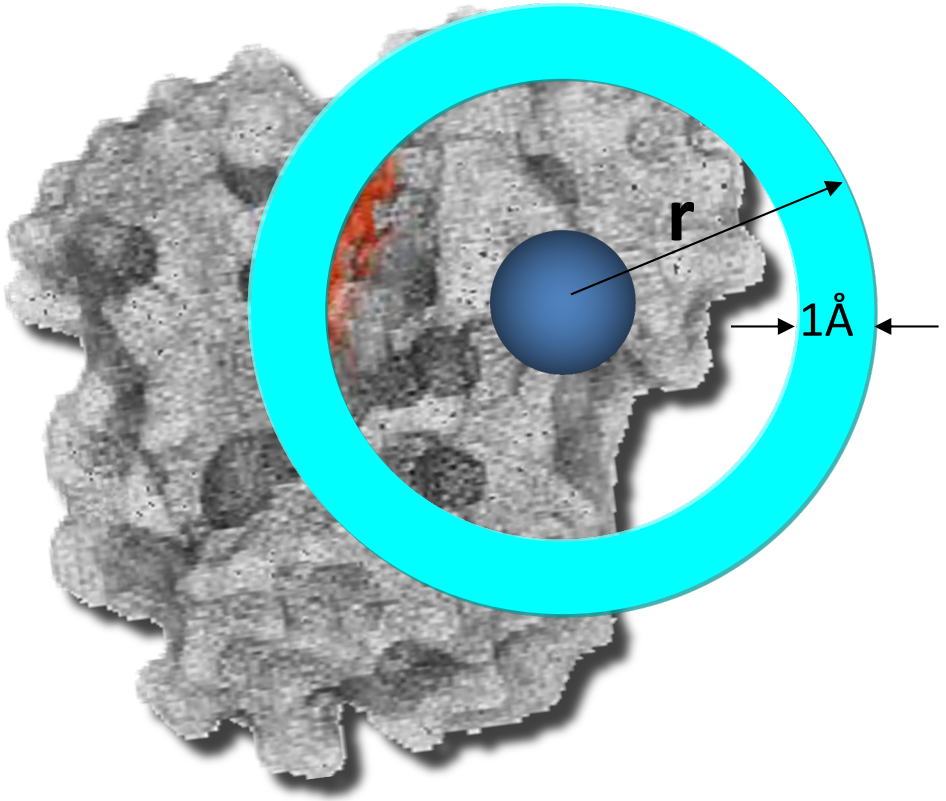
3. The frequency of a pair of residues at distance “r” is different if the residues are close or distant along the sequence
   - We split the calculation of frequencies depending on the sequence distance between residues



1. Knowledge-base potentials

1. Distance dependent potentials

4. Reference state: The density of residues around one residue is not a continuous model, it depends on the size and shape of the protein.
   - We need to normalize by the density ($4\pi r^2 \varepsilon(r)$) and thus defining a reference state



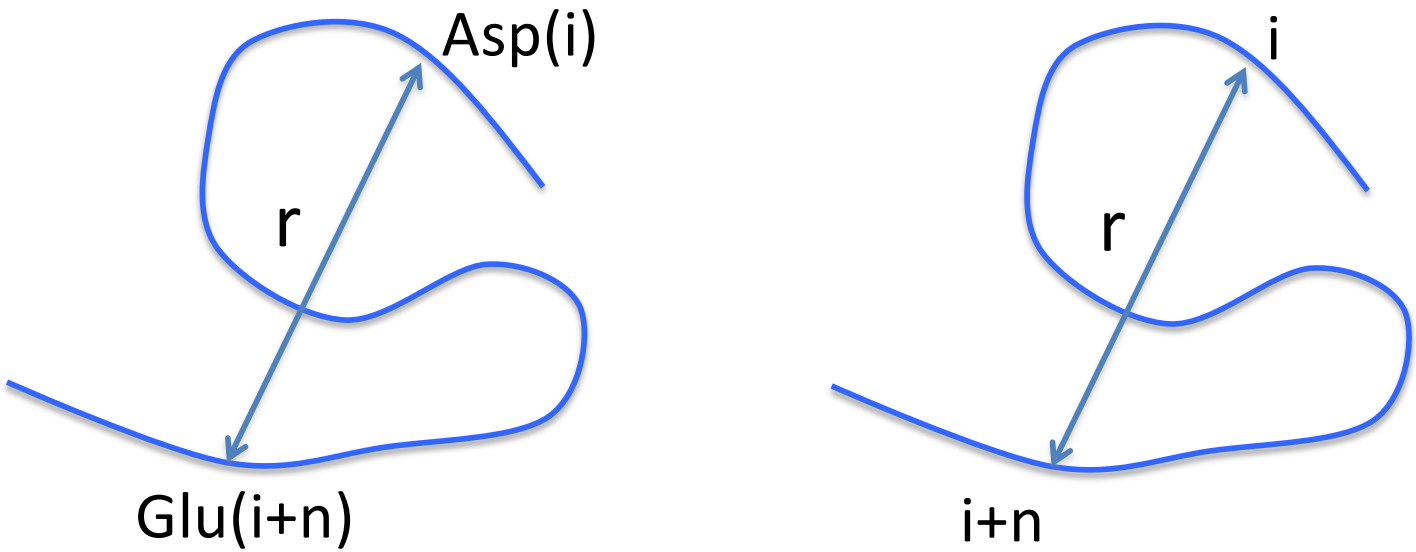
1. Knowledge-base potentials

1. Distance dependent potentials

4. Reference state: the simplest definition of the reference state is to use the whole data set of residue pairs, thus instead of using energies we use incremental energies.
   - Let be a pair of residues Asp and Glu at distance n in sequence. Let be N(r/ED,n) the number of pairs ED like this at distance r between their Cβ atoms, and N(r/n) the total of pairs of residues at distance n in sequence and r between their Cβ atoms

1. Knowledge-base potentials

1. Distance dependent potentials

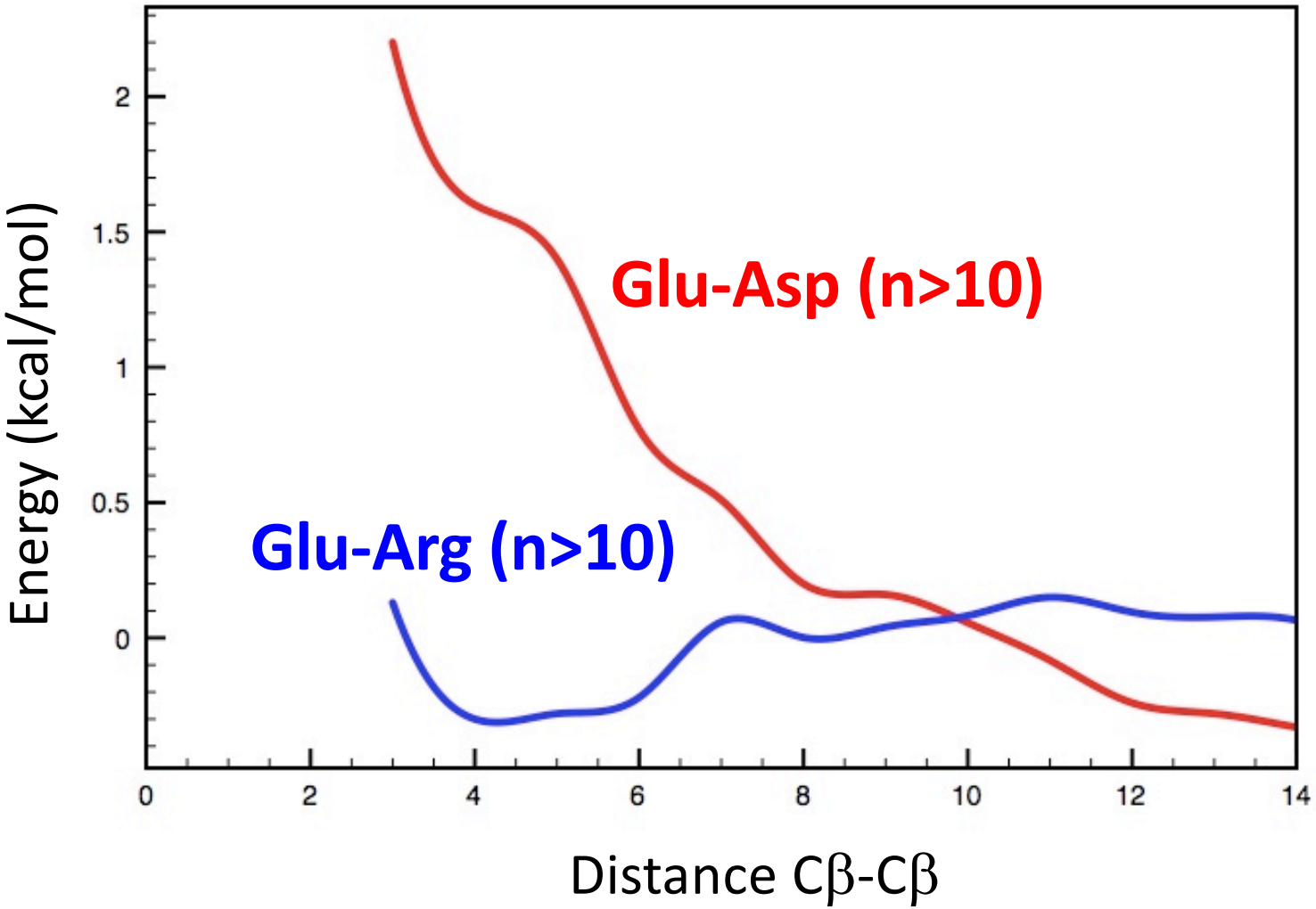


$$\Delta E(r/(Glu,Asp,C\beta,C\beta,n)) = -kT\ln\left(\frac{N(r/ED,n)}{N(r/n)}\right)$$

1. Knowledge-base potentials

1. Distance dependent potentials

Example of distance dependent knowledge-based potentials



1. Knowledge-base potentials
2. Solvation

1. Solvation of a residue is calculated as proportional to accessible surface area (ASA)
   - The factor of proportion depends on the tendency of the residue (i.e. Asp in position "i" of the sequence) to be solvated (hydrophobicity calculated with water-octanol partition coefficient)

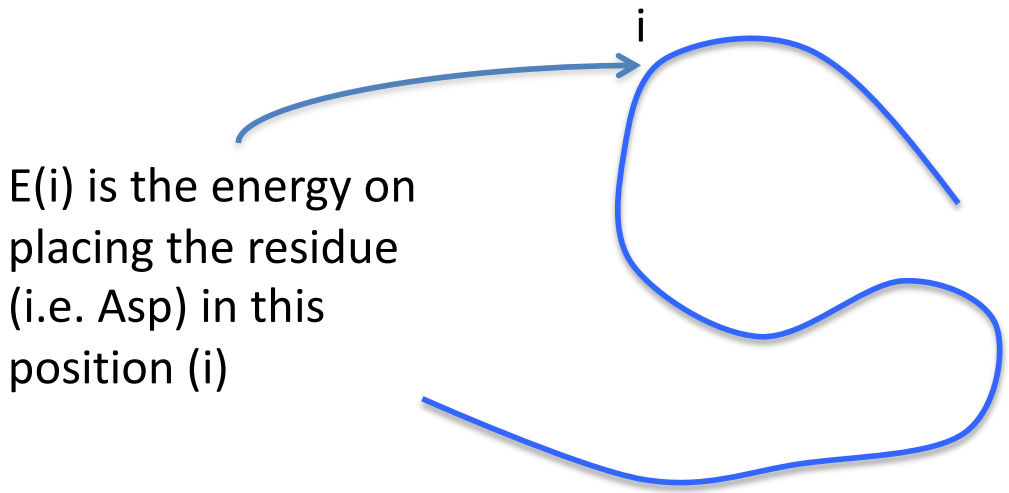
$$E_{sol}(i) = \sigma_{Asp} ASA(i)$$

2. Solvation can also be calculated using the frequency of the residue to be exposed on the surface

1. Knowledge-base potentials

3. Z-scores and energy profiles

Once we have a set of energies for pairs of residues (force field) we can calculate the energy of each residue along the sequence in a specific conformation



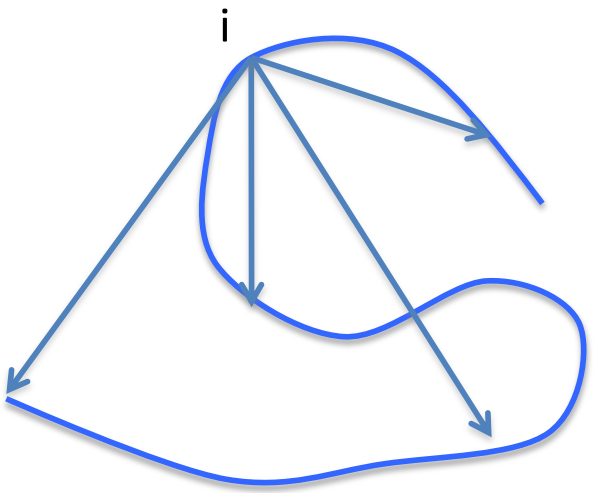
1. Knowledge-base potentials

3. Z-scores and energy profiles

$$E_i = \sum_{j \neq i} E_{ij}\left(r, n = |j - i|\right)$$

$$E_{ij}\left(r, n = |j - i|\right) = \Delta E(r/(Glu(j), Asp(i), C\beta, C\beta, n))$$

$$E_{sol}(i) = \sigma_{Asp} ASA(i)$$



Note: we have assumed that in position I we have placed Asp and Glu in position j=i+n

1. Knowledge-base potentials
3. Z-scores and energy profiles

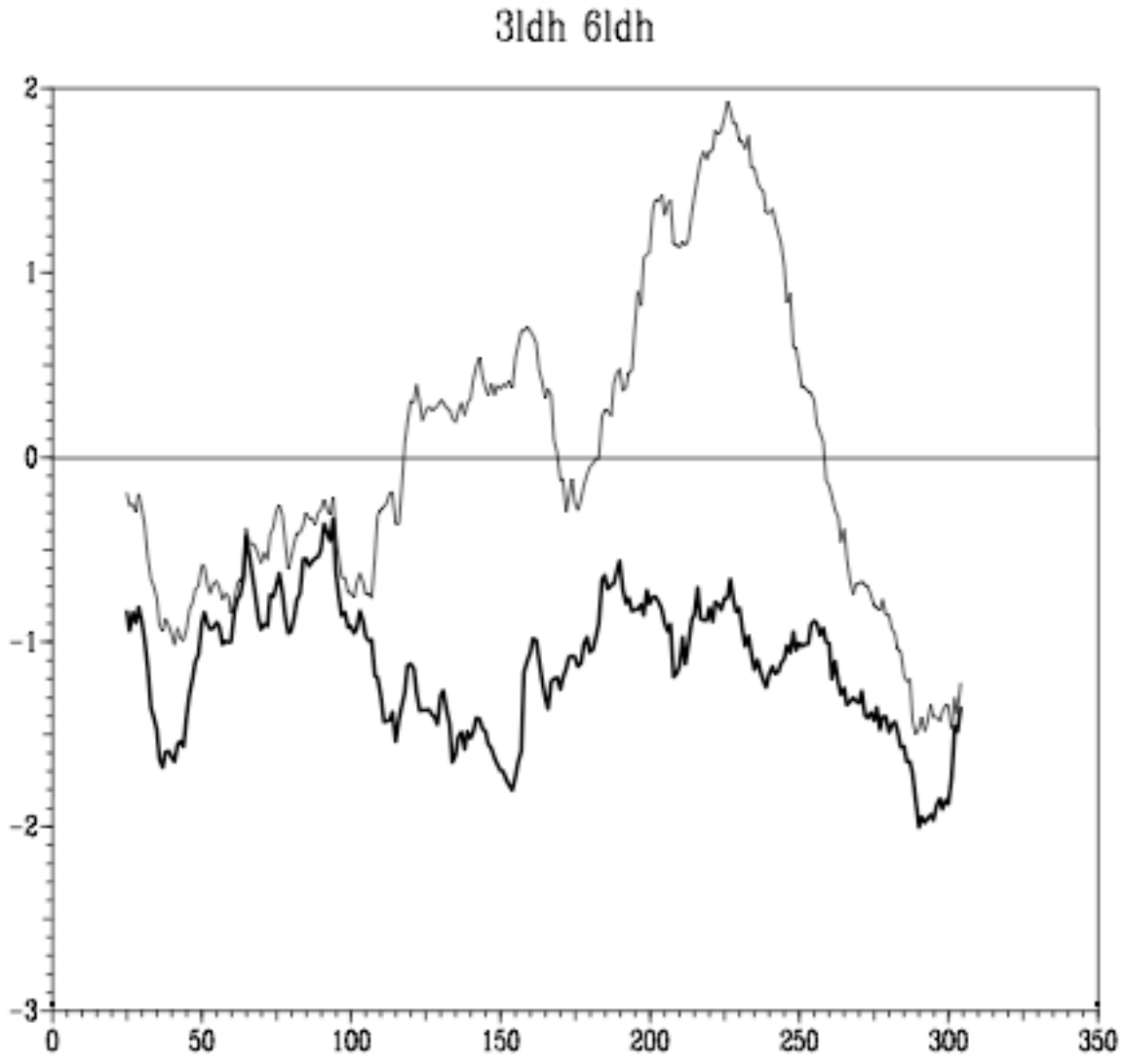
The total energy of a protein is obtained by the sum of the pair-energies and the energy from its surface (solvation)

$$E = \sum_i E_i + \beta \sum_i E_{sol}(i)$$

The profile energy is obtained by the curves of the pair-energies, surface energy and combined energy of both with respect to the residue position

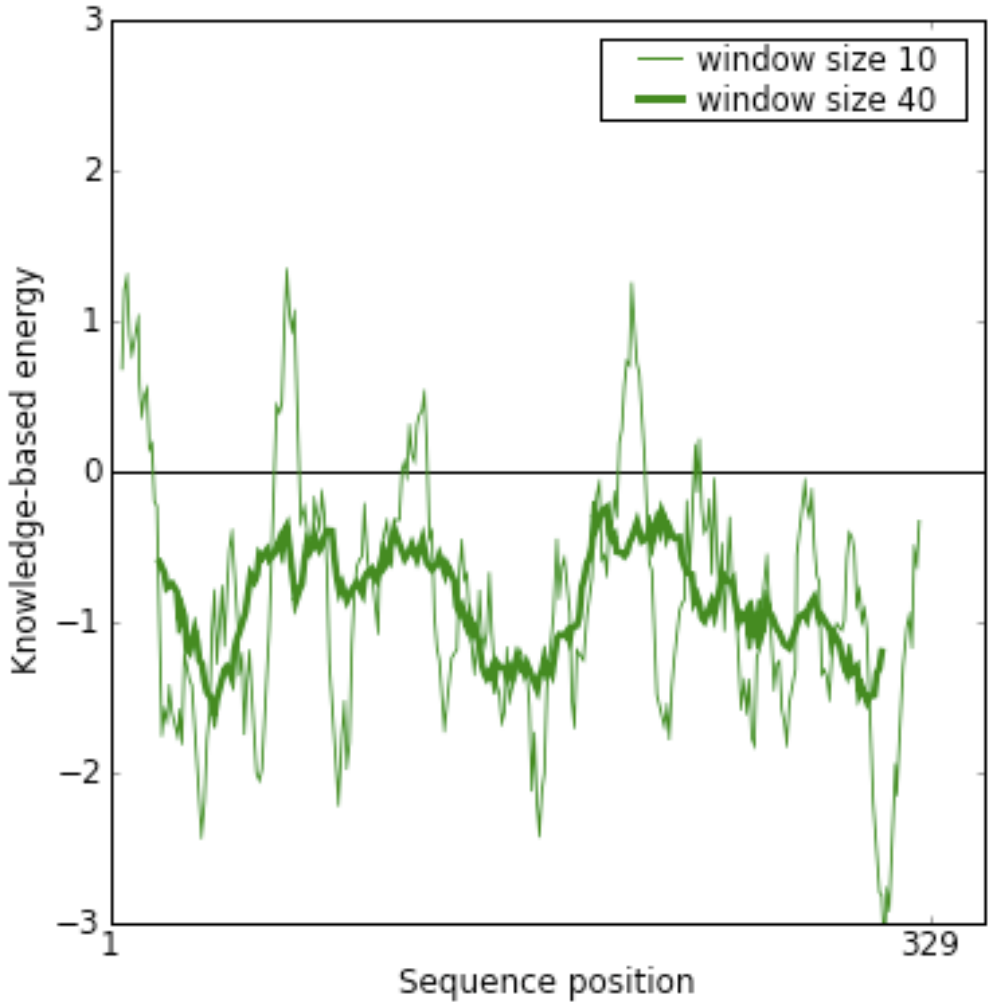
1. Knowledge-base potentials
3. Z-scores and energy profiles

Example of profile energy from PROSA



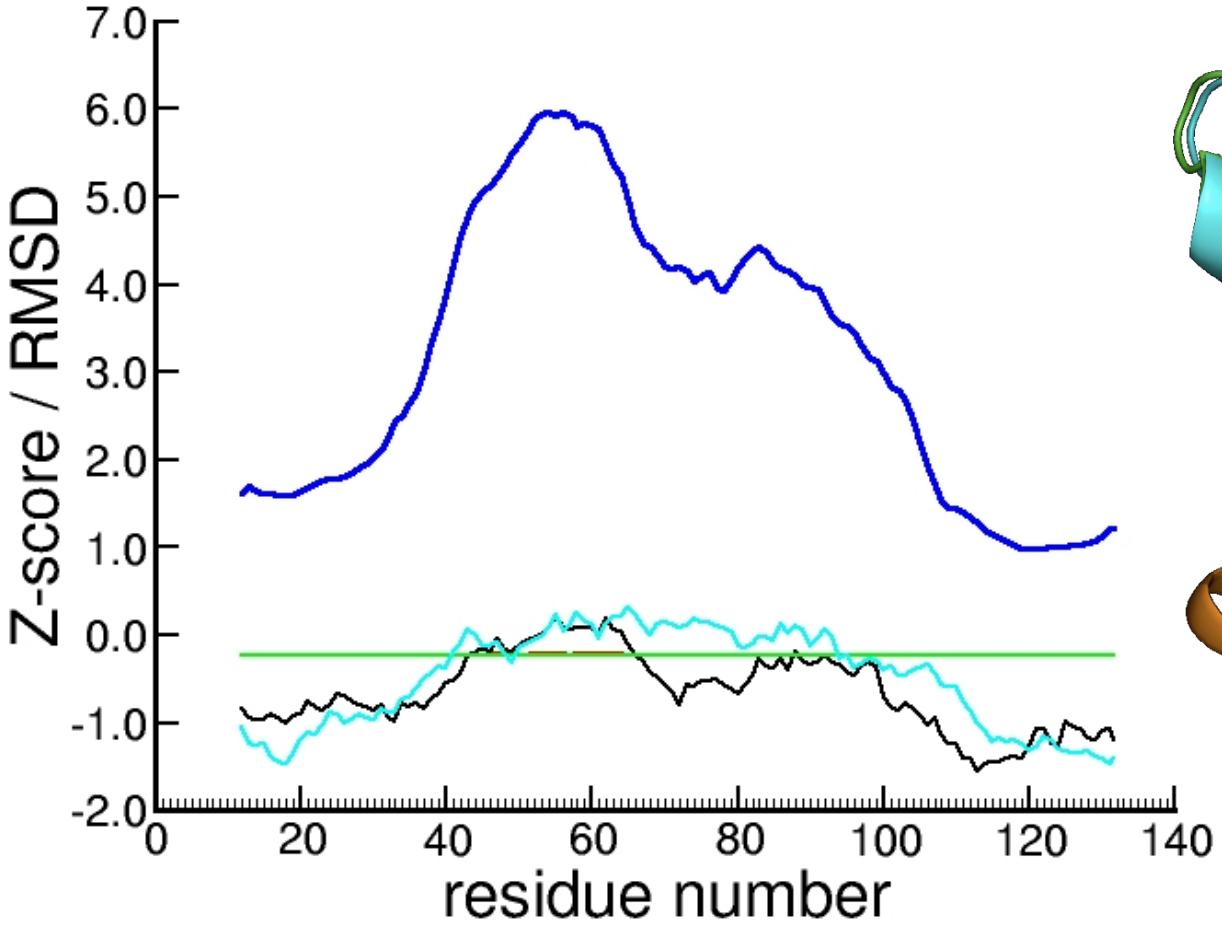
1. Knowledge-base potentials
3. Z-scores and energy profiles

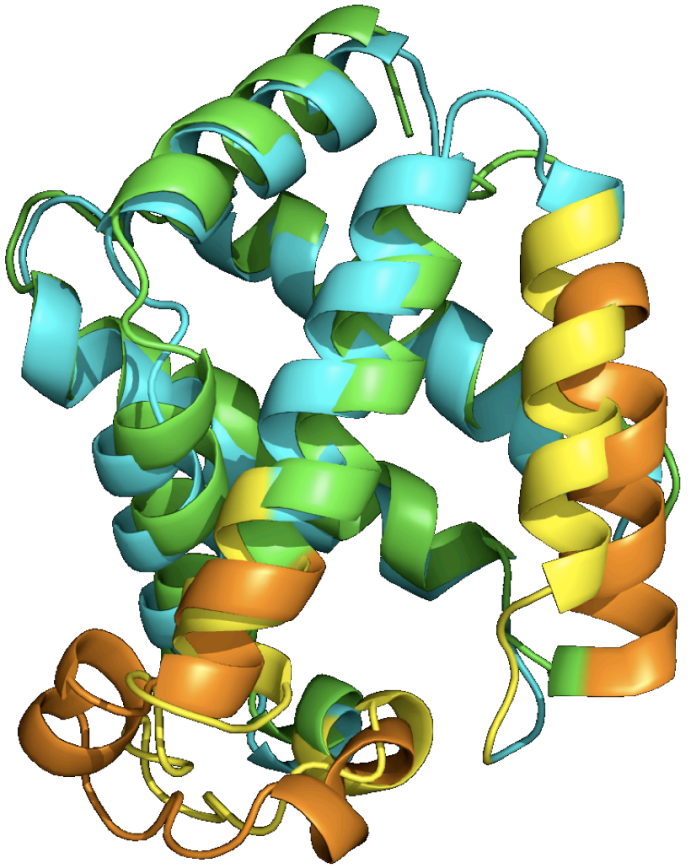
Often the curve is smoothed by windowing the curve: the value on each point is defined by the average of a window of W residues and the window moves along the X axis.



1. Knowledge-base potentials
3. Z-scores and energy profiles

Energy profiles can be used to detect errors in modeling

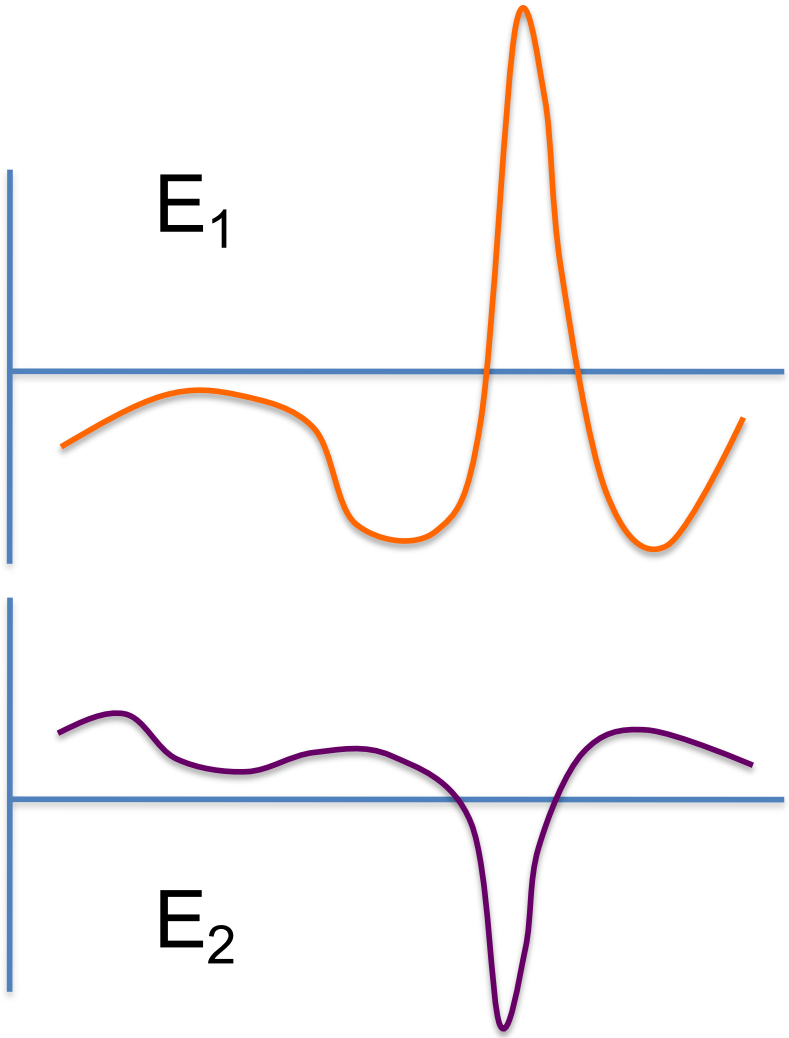




1. Knowledge-base potentials
3. Z-scores and energy profiles

Question:

Can we use the total energy to discriminate correct folds among wrong conformations (decoys)?



$$E_1 > E_2$$

Wrong solution

1. Knowledge-base potentials
3. Z-scores and energy profiles

Question:
Can we use the total energy to discriminate correct folds among wrong conformations (decoys)?

$$0 > E_1 > E_2 > E_3 \mathrm{L} > E_n$$

Many solutions is a wrong solution

Solution:
Define a new function statistically meaningful, the Z-score

1. Knowledge-base potentials
3. Z-scores and energy profiles

Threading Z-score is defined by comparing the energy on one fold (j) with the average of the real folds from the database (i.e. transforms the function “energy” into a Gaussian distribution centered at zero)

$$Zscore_j = \frac{E_j - \langle E \rangle}{\sigma}$$

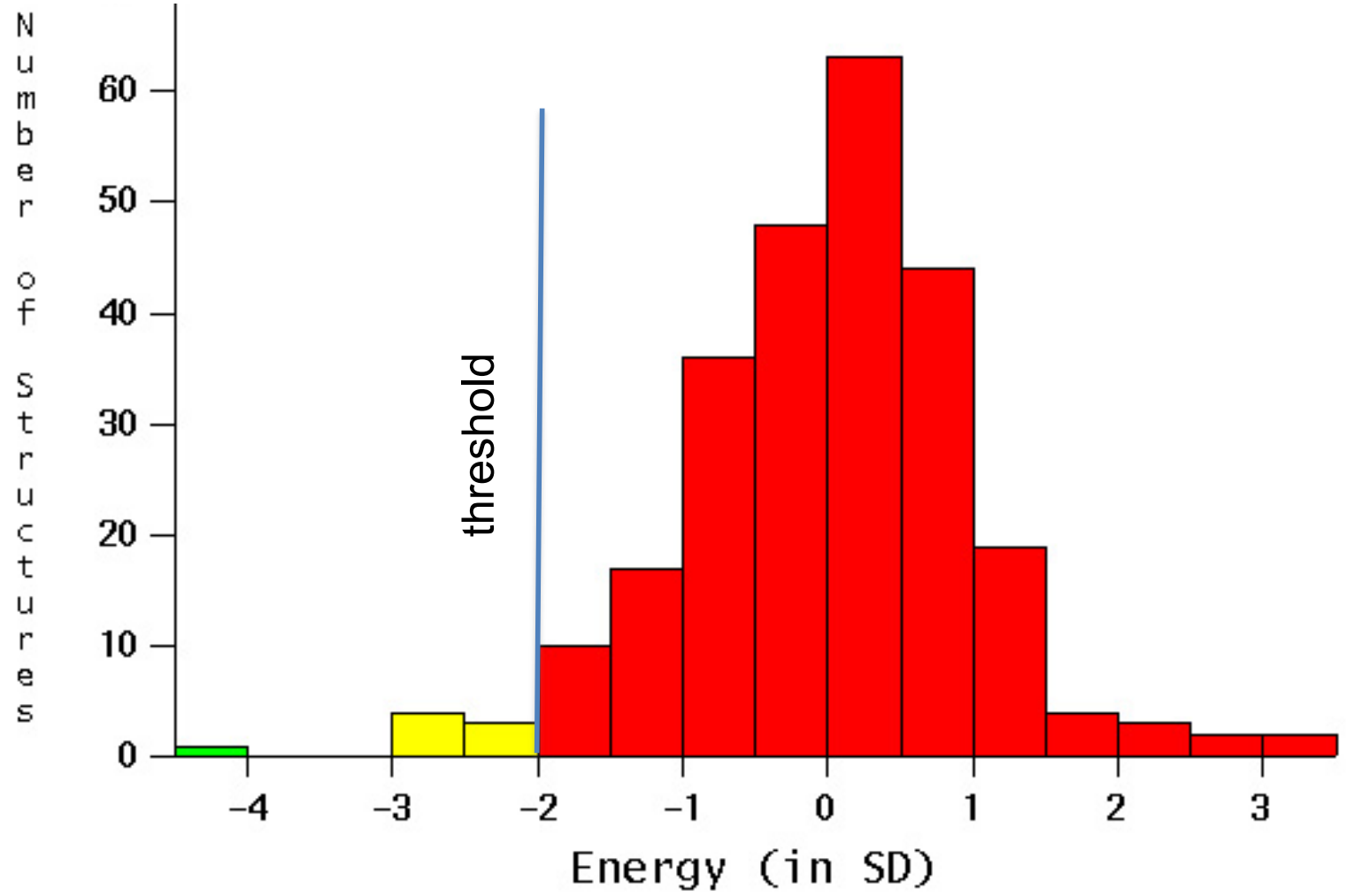
$$\langle E \rangle = \frac{\sum_{i=1}^{N_{folds}} E_i^{real}}{N_{folds}}$$

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N_{folds}} \left(E_i - \langle E \rangle\right)^2}{N_{folds} - 1}}$$

<u>This is the same problem as the following</u>:
Consider the final marks in the class after the exam. We can calculate the 10 best alumni according to their marks.
Are these the best alumni of SBI in the world?
We have to weight their marks with the best students of the world, assuming the exam was the same.
To do that, we use the set of marks of the total of SBI teachers in the world, and we assume they are the best set.
Then, we compare our 10 alumni with them.
If their marks are similar (close to the average of teachers), they are indeed the best.

1. Knowledge-base potentials

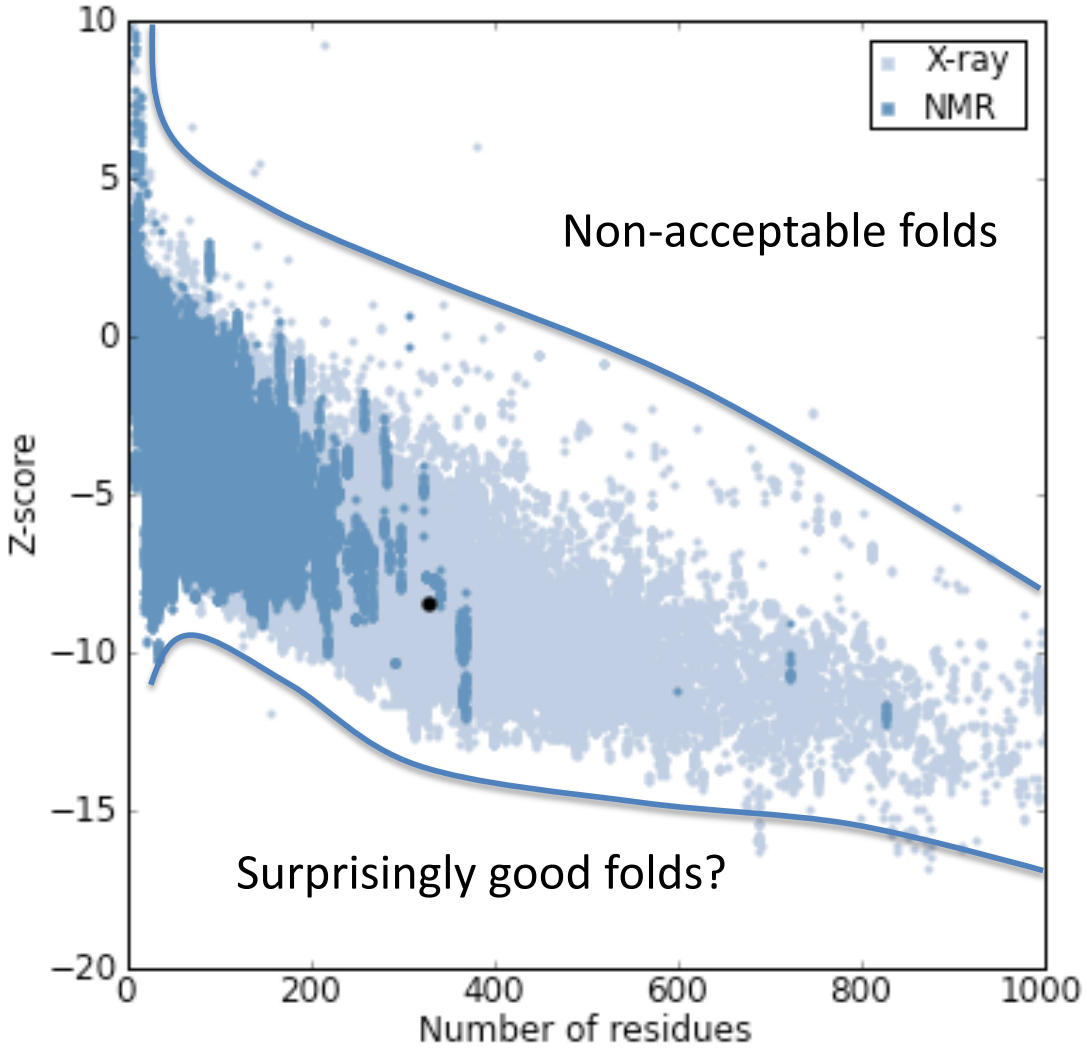
3. Z-scores and energy profiles



	Fingerprint	Protein	Frozen	Thawed
*	1bbk.A	METHYLAMINE DEHYDROGENASE	-3.195	-4.211
+	1apb	L-*ARABINOSE-BINDING PROTEIN	-1.978	-2.742
+	2fbj.L	IG*A FV FRAGMENT (H)	-0.843	-2.636

1. Knowledge-base potentials
3. Z-scores and energy profiles

Zscores can also be presented as a function of the length of the protein sequence



2. Remote homologs (PSSM)

We can use sequence alignments with position specific substitution matrices (PSSM) (see theory in practices)

1. Alignment between one sequence and a Hidden Markov Model profile (hmmpfam, hmmscan)
2. Alignment between two Hidden Markov Model profiles (HHSearch, PRC)
3. Alignment between sequences using PSSMs (BLAST, fugue)

2. Remote homologs (PSSM)

1. Function association

PHYRE / 3D-PSSM

Remotely homologous structures that can't be found by conventional methods are detected by using profiles (or PSSMs) generated by PSI-Blast for both target sequence and the sequences of the known structures. Phyre performs a profile-profile matching algorithm together with predicted secondary structure matching.

The functional keywords are found by gathering homologues of the target sequence from Swissprot, taking the keywords associated with the Swissprot homologues and weighting them according to their background frequency across the whole Swissprot database using SAWTED

1. Knowledge-base potentials
3. Z-scores and energy profiles

SAWTED

What is SAWTED?

SAWTED stands for **S**tructure **A**ssignment **W**ith **Te**xt **D**escription. It is a method to improve the coverage of the detection of remote homologues of known structure by sequence searches (e.g. PSI-BLAST) and fold recognition programs.

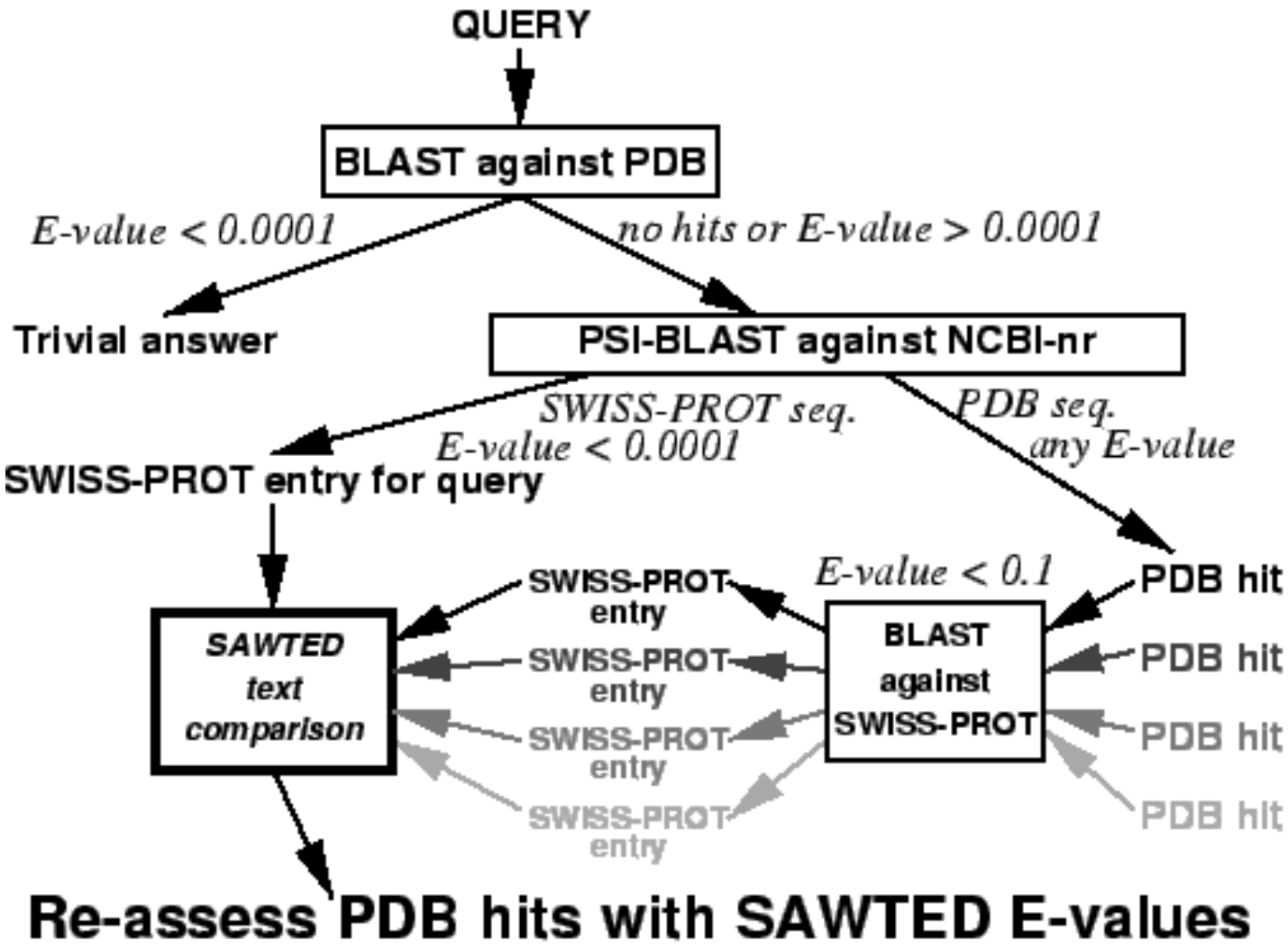
What does it do?

When sequence database searches return only hits with scores worse than an accepted threshold for reliability the user will often compare what is known about the function of the query sequence with that known about the poor scoring hits. Some hits may appear more sensible than others and deserve closer inspection. In SAWTED this comparison is made automatically using an algorithm to compare the text of SWISS-PROT annotations related to the query and to the poor scoring hits. A single E-value is given for the user to assess the similarity of function.

SAWTED is currently implemented to enhance PSI-BLAST searches against the PDB, and as part of our 3D-PSSM fold recognition server

1. Knowledge-base potentials
3. Z-scores and energy profiles

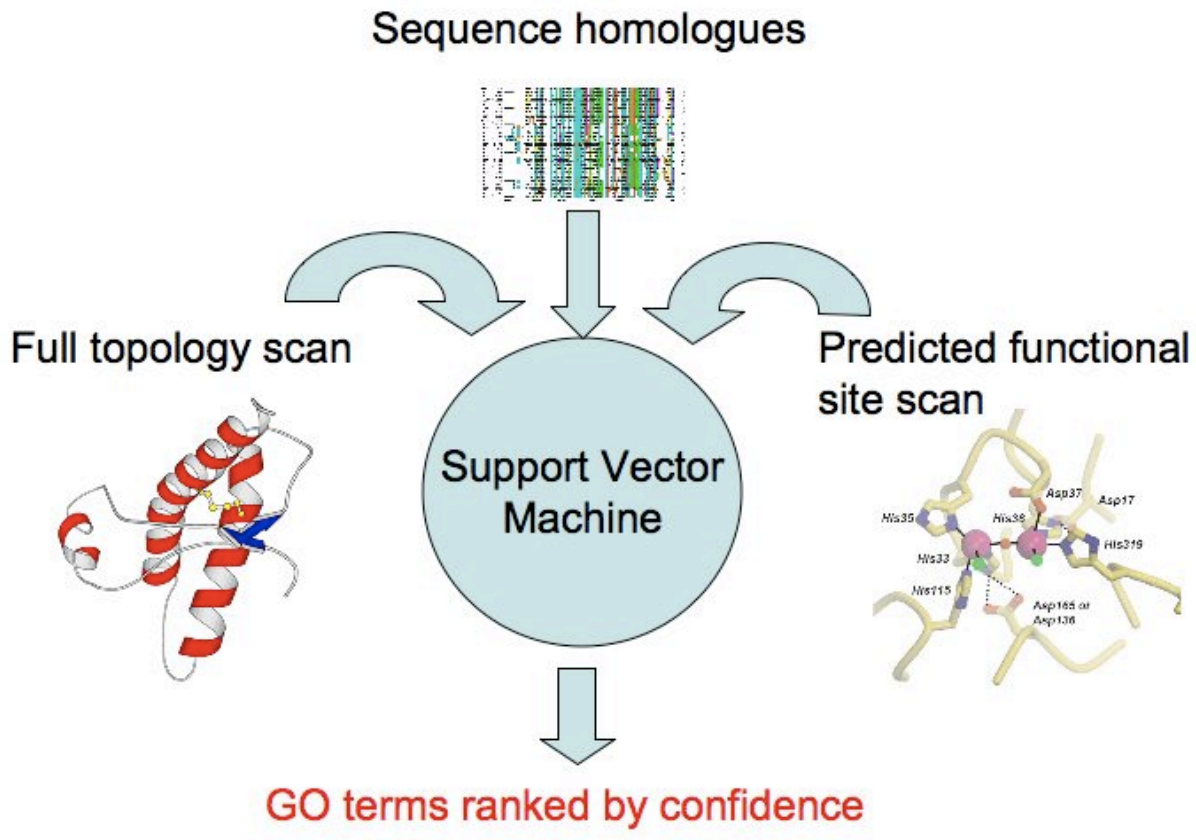
SAWTED in PHYRE & 3D-PSSM



2. Remote homologs (PSSM)
1. Function association

3D2GO

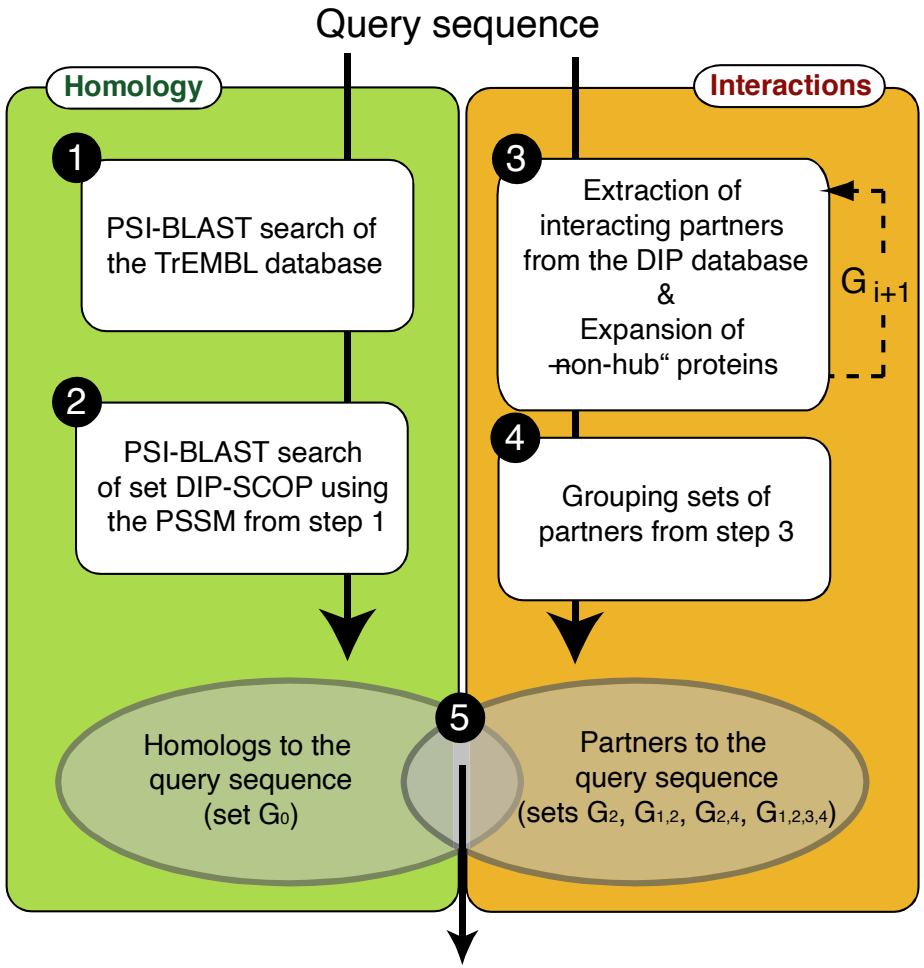
Requires the use of Machine learning methods (SVM) to select the best associated terms

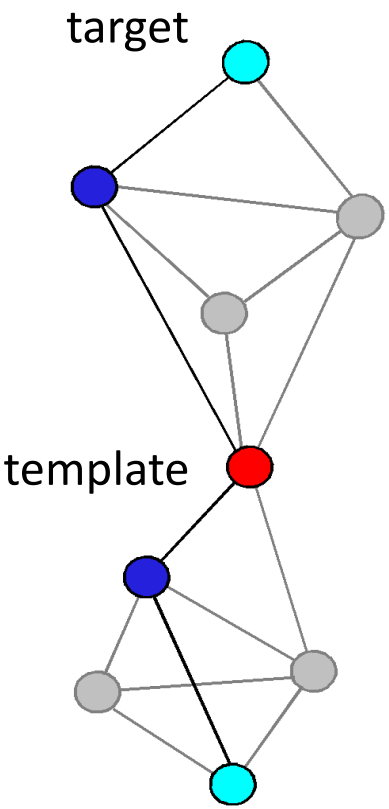


2. Remote homologs (PSSM)
1. Function association

ModLink

Uses the knowledge of protein-protein interactions to select the best candidates (according to sequence-based alignments) among the homologs with known structure.





2. Remote homologs (PSSM)

1. Function association

Example

PSI-BLAST search of the C-terminal domain of yeast Elongation Factor 1γ (**Ferredoxin like fold**)

Hits in SwissProt	E-value	Shares Fold	Appears in G_2
SYEC_YEAST	0.027	no	**?**
EF1B_YEAST	0.036	yes	**?**
SC14_YEAST	0.83	no	**?**

2. Remote homologs (PSSM)

1. Function association

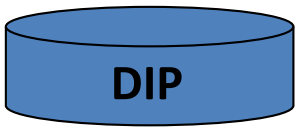
Example

PSI-BLAST search of the C-terminal domain of yeast Elongation Factor 1γ (**Ferredoxin like fold**)

Hits in SwissProt	E-value	Shares Fold	Appears in G_2
SYEC_YEAST	0.027	no	**?**
EF1B_YEAST	0.036	yes	**?**
SC14_YEAST	0.83	no	**?**

DIP entry 17026E

EF1G_YEAST —— EF1A_YEAST



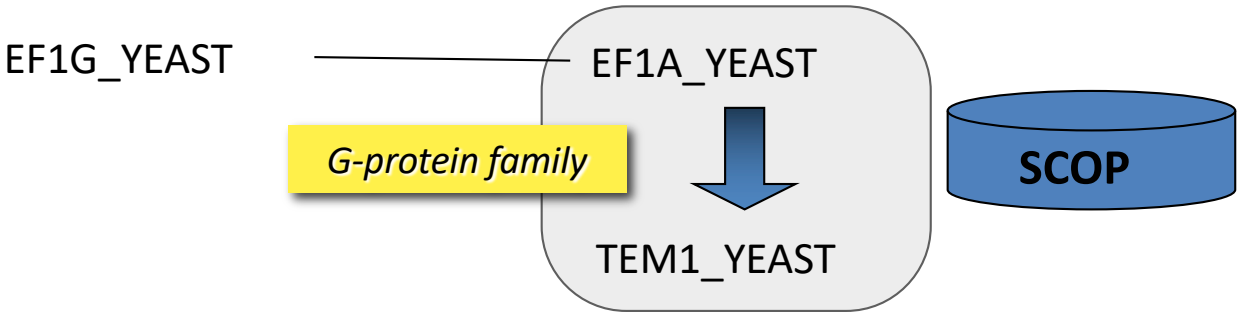
2. Remote homologs (PSSM)

1. Function association

Example

PSI-BLAST search of the C-terminal domain of yeast Elongation Factor 1γ (**Ferredoxin like fold**)

Hits in SwissProt	E-value	Shares Fold	Appears in G_2
SYEC_YEAST	0.027	no	**?**
EF1B_YEAST	0.036	yes	**?**
SC14_YEAST	0.83	no	**?**



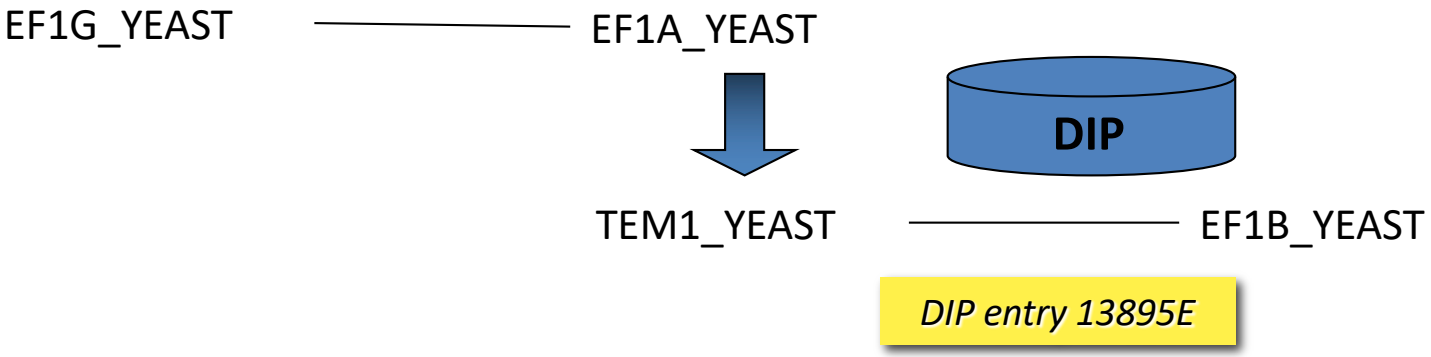
2. Remote homologs (PSSM)

1. Function association

Example

PSI-BLAST search of the C-terminal domain of yeast Elongation Factor 1γ (**Ferredoxin like fold**)

Hits in SwissProt	E-value	Shares Fold	Appears in G_2
SYEC_YEAST	0.027	no	**?**
EF1B_YEAST	0.036	yes	**?**
SC14_YEAST	0.83	no	**?**



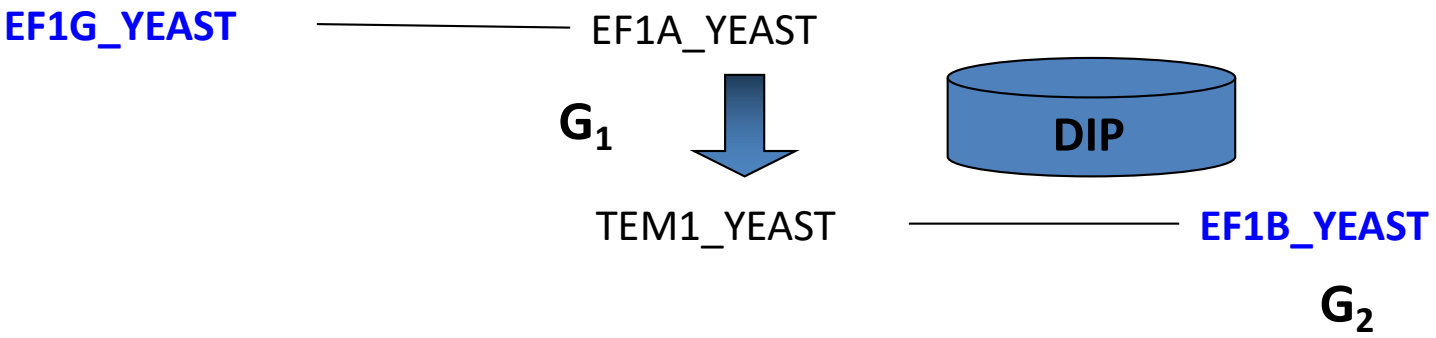
2. Remote homologs (PSSM)

1. Function association

Example

PSI-BLAST search of the C-terminal domain of yeast Elongation Factor 1γ (**Ferredoxin like fold**)

Hits in SwissProt	E-value	Shares Fold	Appears in G_2
SYEC_YEAST	0.027	no	**no**
EF1B_YEAST	**0.036**	**yes**	**yes**
SC14_YEAST	0.83	no	**no**



3. Secondary structure alignment

1. secondary structure prediction (machine learning)

M = { set of data obtained with a predictive model}

D = { set of data known}

Bayes Theorem

$$P(D/M) = \frac{P(D \cap M)}{P(M)}$$

$$P(M/D) = \frac{P(D \cap M)}{P(D)}$$

$$P(M/D) = P(D/M)\frac{P(M)}{P(D)}$$

3. Secondary structure alignment

1. secondary structure prediction (machine learning)

M = { set of data obtained with a predictive model}

D = { set of data known}

Optimizing Function Φ (minimum Φ)

$$\Phi = -\log\left(P(M/D)\right)$$

$$\Phi = -\log\left(P(D/M)\right) - \log\left(P(M)\right) + \log\left(P(D)\right)$$

$$Min\left(\Phi\right) = Min\left(-\log\left(P(D/M)\right) - \log\left(P(M)\right)\right)$$

$$Min\left(\Phi\right) \approx Min\left(-\log\left(P(D/M)\right)\right)$$

Maximum a priori

Maximum likelihood

3. Secondary structure alignment

1. secondary structure prediction (machine learning)

Training set

Set of data without redundancies (i.e. a set of non-homologous sequences). This is used to optimize the parameters describing the model

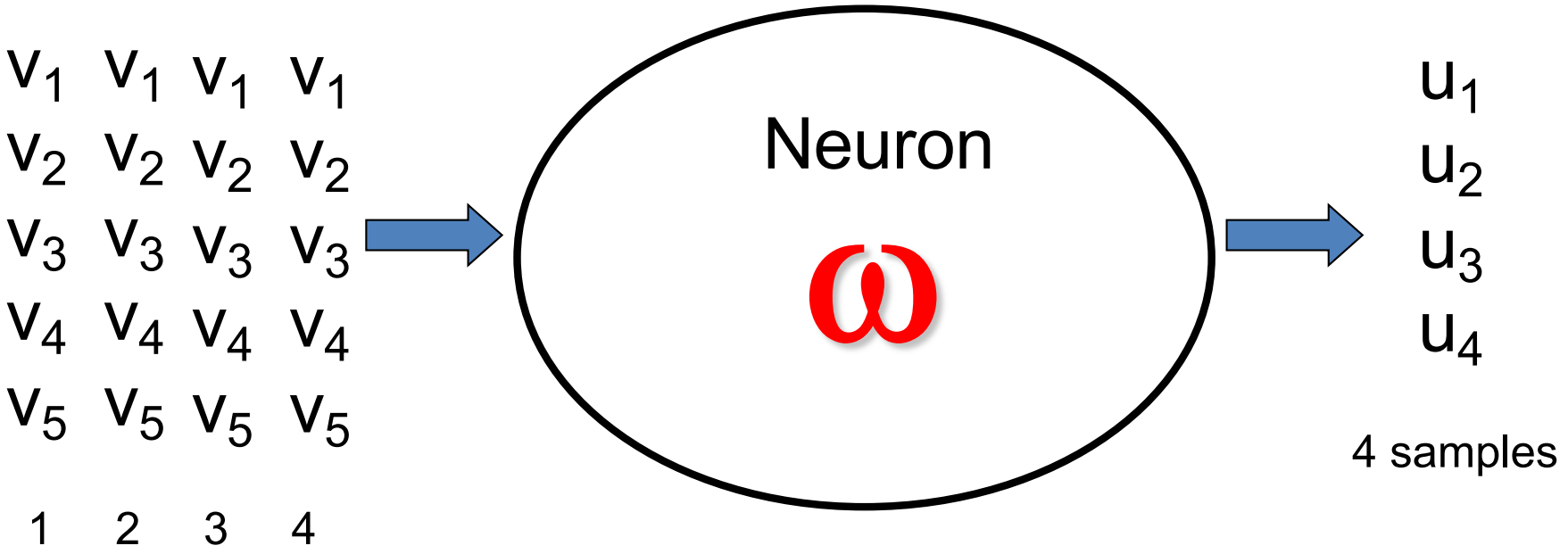
Test set

Set of data without any element used on the training set or similar to some element of the training set (i.e. a set of sequences non-homologous between them and non-homologous to any of the elements of the training set). This set is used to test the approach and validate the statistical accuracy of the method.

3. Secondary structure alignment

1. secondary structure prediction (Neural Network)

$$\text{input} = \{v_i / v_i \; i=1,n\}_m$$
$$\text{output} = \{u_i / u_i \; j=1,m\}$$



3. Secondary structure alignment

1. secondary structure prediction (Neural Network)

Parameters for the model: ω

$$x_j = \sum_{k=1}^{5} w_k V_k^j + w_0$$

$$y_j = f(x_j) = \frac{1}{1 + e^{-x_j}}$$

We need to optimize the parameters in order to get y_j as close as possible to u_j

3. Secondary structure alignment

1. secondary structure prediction (Neural Network)

Working hypothesis:

The error between the expected output values (u) and the output obtained with this “neuron” approach follows a multiple gaussian distribution. Therefore, the probability to obtain the output data, given the parameters of the neuron (ω and function f), is:

$$P(D \mid M) = P(u \mid \omega, f) = \prod_{j=1}^{m} \frac{1}{\sigma\sqrt{2\pi}} \times e^{\frac{-(u_j - y_j)^2}{2\sigma^2}}$$

$$\sigma = \sqrt{\frac{\sum_{j=1}^{m} (u_j - y_j)^2}{m-1}}$$

Where m is the number of samples

Maximum Likelihood solution:

This implies we can solve the optimization by means of the maximum likelihood approach. It also can be further simplified by assuming a constant standard deviation.

$$\Phi(w) = -\log(P) = -\frac{\log(2\pi)}{2} - \log(\sigma) + \sum_j \left(u_j - y_j\right)^2 / 2\sigma^2$$

$$\frac{\partial \Phi}{\partial w_k} = -\sum_j \frac{\left(u_j - y_j\right)}{\sigma^2} \frac{e^{-x_j}}{1 - e^{-x_j}} V_k^j$$

3. Secondary structure alignment

1. secondary structure prediction (Neural Network)

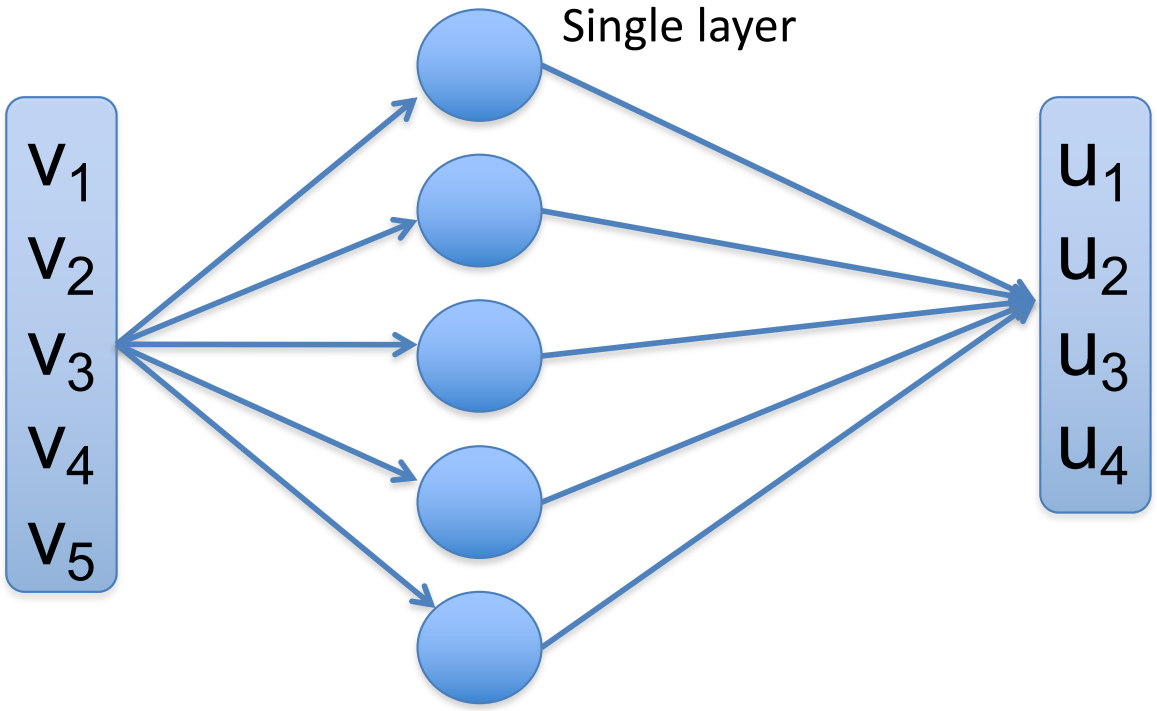
Neural Network

The protein sequence can be transformed into a set of vectors on the space of residues (dimension 20)

Inputs can check by windows of 15 Aa along the sequence

We can use more than one neuron, forming a layer of neurons.

We can add multiple layers formed by neurons.

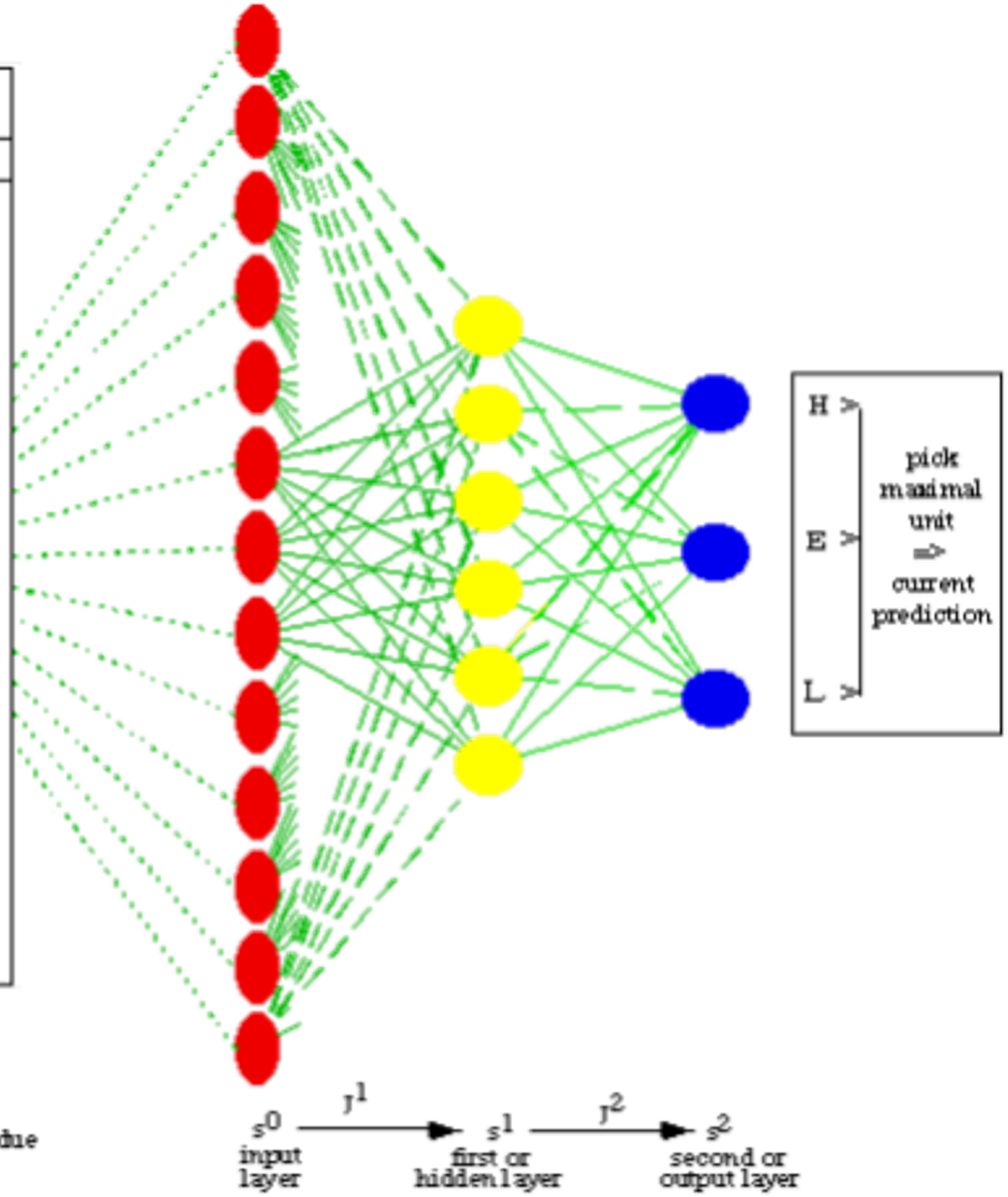


3. Secondary structure alignment

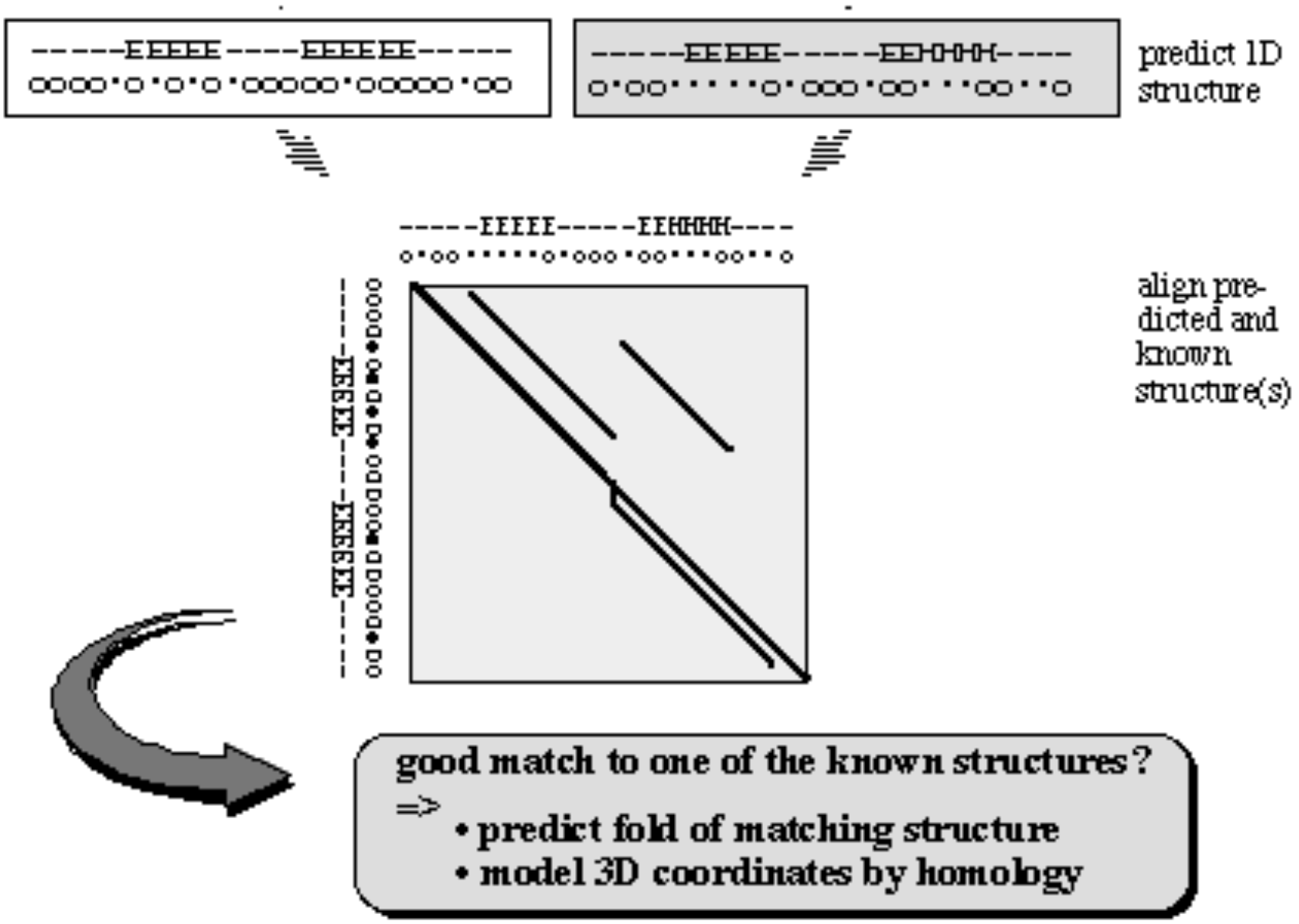
1. secondary structure prediction (Neural Network)

Neural Network (PHD)

Protein	Alignments	profile table
		GSAPD NTEKQ CVHIR LMYFW
G	G G G G	5.... ....
Y	Y Y Y Y	....5..
I	I I E E	....2.. ...3. ....
Y	Y Y Y Y	....5..
D	D D D D	....5 ....
P	P P P P	...5. ....
E	A E A A	..3.. ..2.. ....
D	V V E E	....1 ..2.. .2... ....
G	G G G G	5.... ....
D	D D D D	....5 ....
P	P P P P	...5. ....
D	D T D D	....4 .1... ....
D	N Q N N	....1 3...1 ....
G	G N G G	4.... 1.... ....
V	V I V V	.... .4.1. ....
N	E P K K	...1. 1.12. ....
P	P P P P	...5. ....
G	G G G G	5.... ....
T	T T T T	..... .5... ....
D	E K S A	.11.1 ..11. ....
F	F F F F	....5.



3. Secondary structure alignment
2. Method of fold recognition TOPITS and THREADER



Fold prediction

2. *ab initio* fold prediction (Rosetta)
    1. Revisiting the knowledge-based potential
    2. New potential based on conditional probabilities
    3. 9-Fragment database of structures
    4. Simulated Annealing construction
    5. Mutual Information
    6. Examples

1. Revisiting the knowledge-based potential

Given the radius of gyration of a protein structure (RG), we approximate the probability that this is the structure for a given sequence, where the sequence is defined as the

$$P(structure \mid sequence) = P(structure) \times \frac{P(sequence \mid structure)}{P(sequence)}$$

$$P(sequence \mid structure) = \prod_{i<j} P(aa_i, aa_j) \times \frac{P(r_{ij} \mid aa_i, aa_j)}{P(r_{ij})}$$

$$P(structure \mid sequence) \cong e^{-RG^2} \times \prod_{i<j} \frac{P(r_{ij} \mid aa_i, aa_j)}{P(r_{ij})} \qquad \text{(Equation 1)}$$

Where the term on the right contains the distance dependent knowledge-based potential: $P(r_{ij} \mid aa_i, aa_j) / P(r_{ij})$

2. New potential based on conditional probabilities

By applying Bayes theorem on a sequence (set of elements amino-acids), we can approach the conditional probability with respect to the structure in which the sequence is folded with the first two terms of the expansion:

$$P(x_1,x_2,x_3,\ldots,x_n) \cong \prod_i P(x_i) \times \prod_{i<j} \frac{P(x_i,x_j)}{P(x_i)P(x_j)} \cdots$$

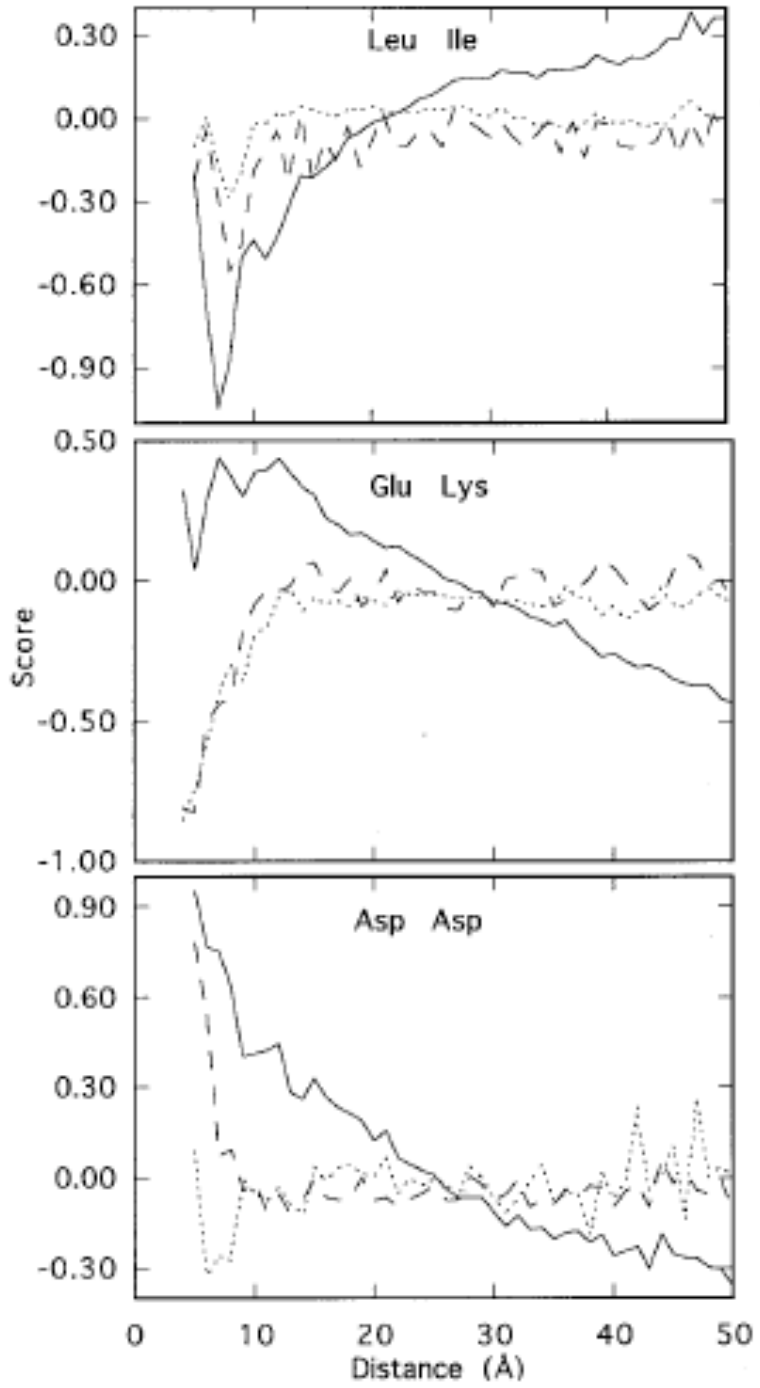
$$P(sequence \mid structure) = P(aa_1,aa_2,\ldots aa_n \mid structure)$$

$$P(aa_1,aa_2,\ldots aa_n \mid structure) \cong \prod_i P(aa_i \mid E_i) \times \prod_{i<j} \frac{P(aa_i,aa_j \mid r_{ij},E_i,E_j)}{P(aa_i \mid r_{ij},E_i,E_j)P(aa_j \mid r_{ij},E_i,E_j)}$$

$$P(structure \mid sequence) \cong e^{-RG^2} \times P(aa_1,aa_2,\ldots aa_n \mid structure) \qquad \text{(Equation 2)}$$

Where E_i is the environment (secondary structure, accessibility, etc.) of residue aa_i

2. New potential based on conditional probabilities



Example of differences between otentials calclated with equation 1 and equation 2.

Equation 1 is in continuous line

Equation 2 for two buried residues is in dotted line

Equation 2 for two exposed residues is in dashed line

3. 9-Fragment database of structures

Rosetta splits the sequence in fragments of 9 residues, using a window-like method

Rosetta contains a database of 9-residue fragments extracted from the total set of protein structures

Rosetta assigns the first 25 most probable 9-fragment segments to a 9-residue fragment of the target sequence by selecting those with smallest score:

$$score = \sum_{i=1}^{9} \sum_{aa=1}^{20} \left| S(aa,i) - X(aa,i) \right|$$

Where S(aa,i) is the frequency of residue aa in position i of the target sequence and its homologs in the same 9-residues fragment. Similarly, X(aa,i) is the frequency of amino-acid aa in position i for all similar 9-residue fragments (with the same structure)

4. Simulated annealing construction

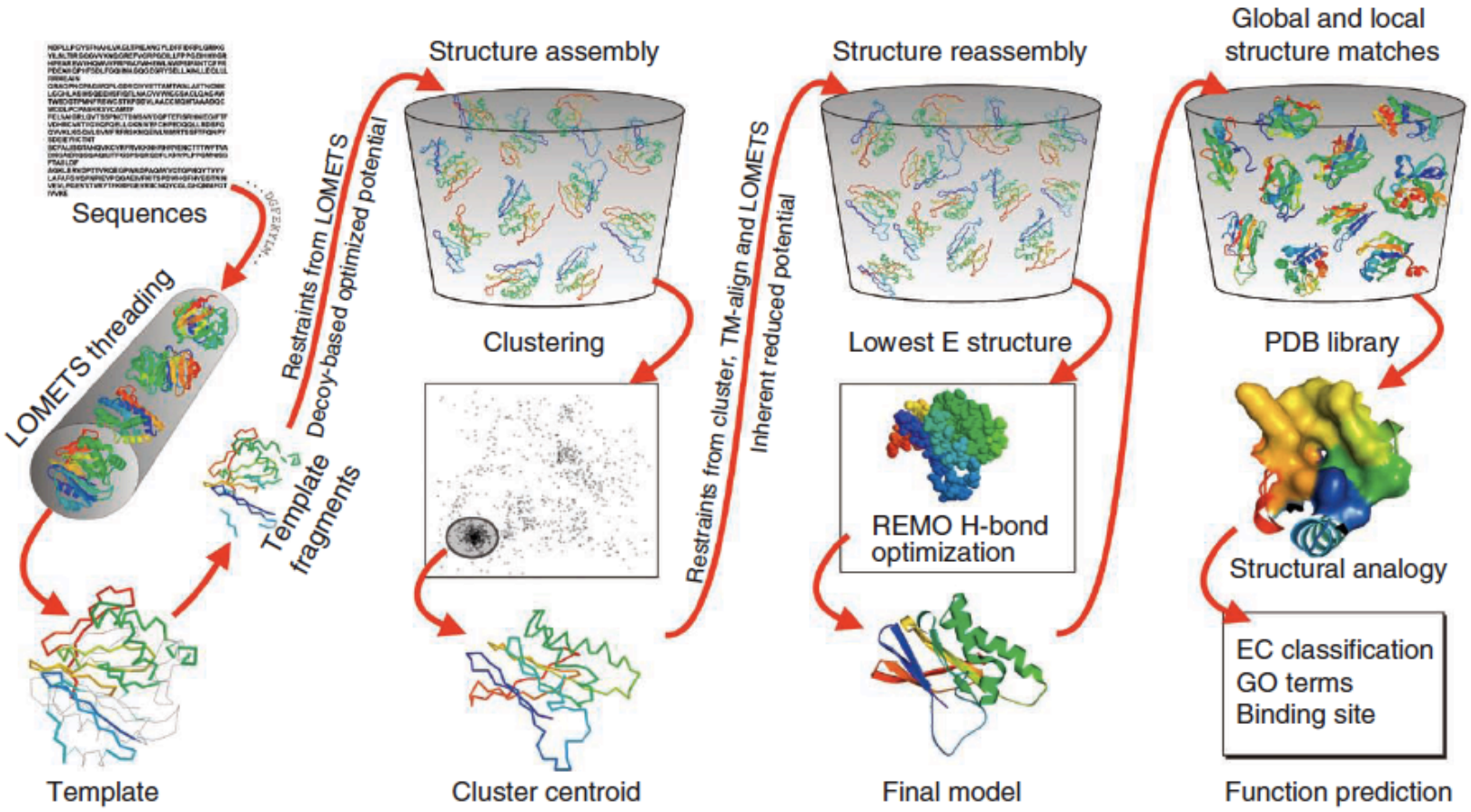
Rosetta applies small changes in torsional angles for each fragment considered in order to join the 9-residue fragmented structures assigned to the 9-residue segment of the target

A conformation is selected according to the most probable structure-score: P(structure|sequence). A Metropolis-Montecarlo simulation is applied using a simulated annealing

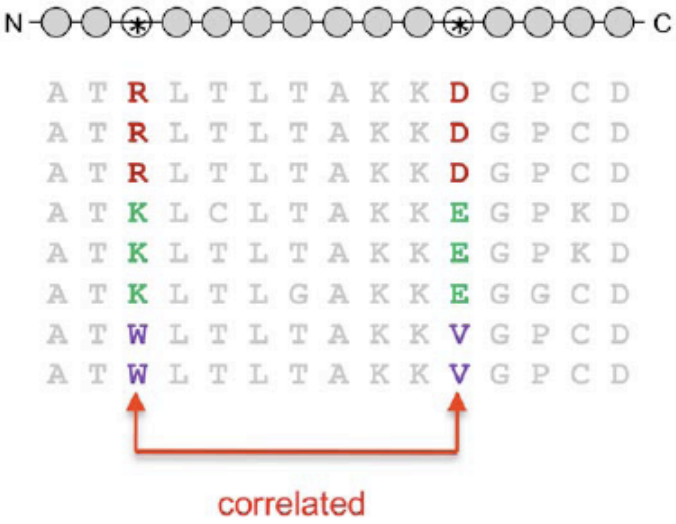
The structure-score is first calculated with equation 1, and when the simulation obtains a closer and more definite structure equation 2 (with more detailed potential) is applied.

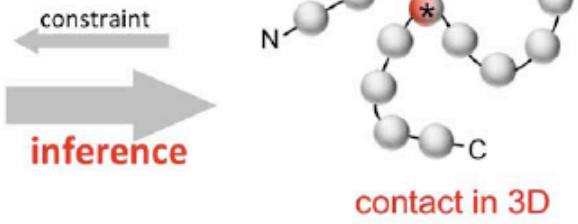
5. iTASSER

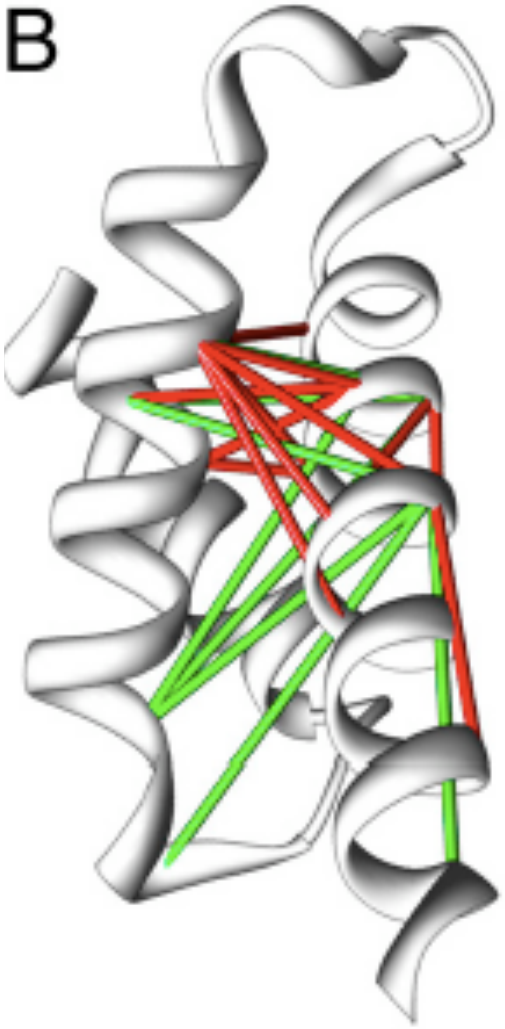
iTASSER uses LOMETS threading. LOMETS uses the results of several threading approaches based on remote homology (i.e. FUGUE, HHSEARCH, etc.) and selects the common fragment-templates to assemble the target structure. Then it follows a similar approach to Rosetta



6. Mutual Information







$$\mathrm{MI}_{ij} = \sum_{A,B} f_{ij}(A,B) \ \ln \frac{f_{ij}(A,B)}{f_i(A) f_j(B)},$$

6. Mutual Information

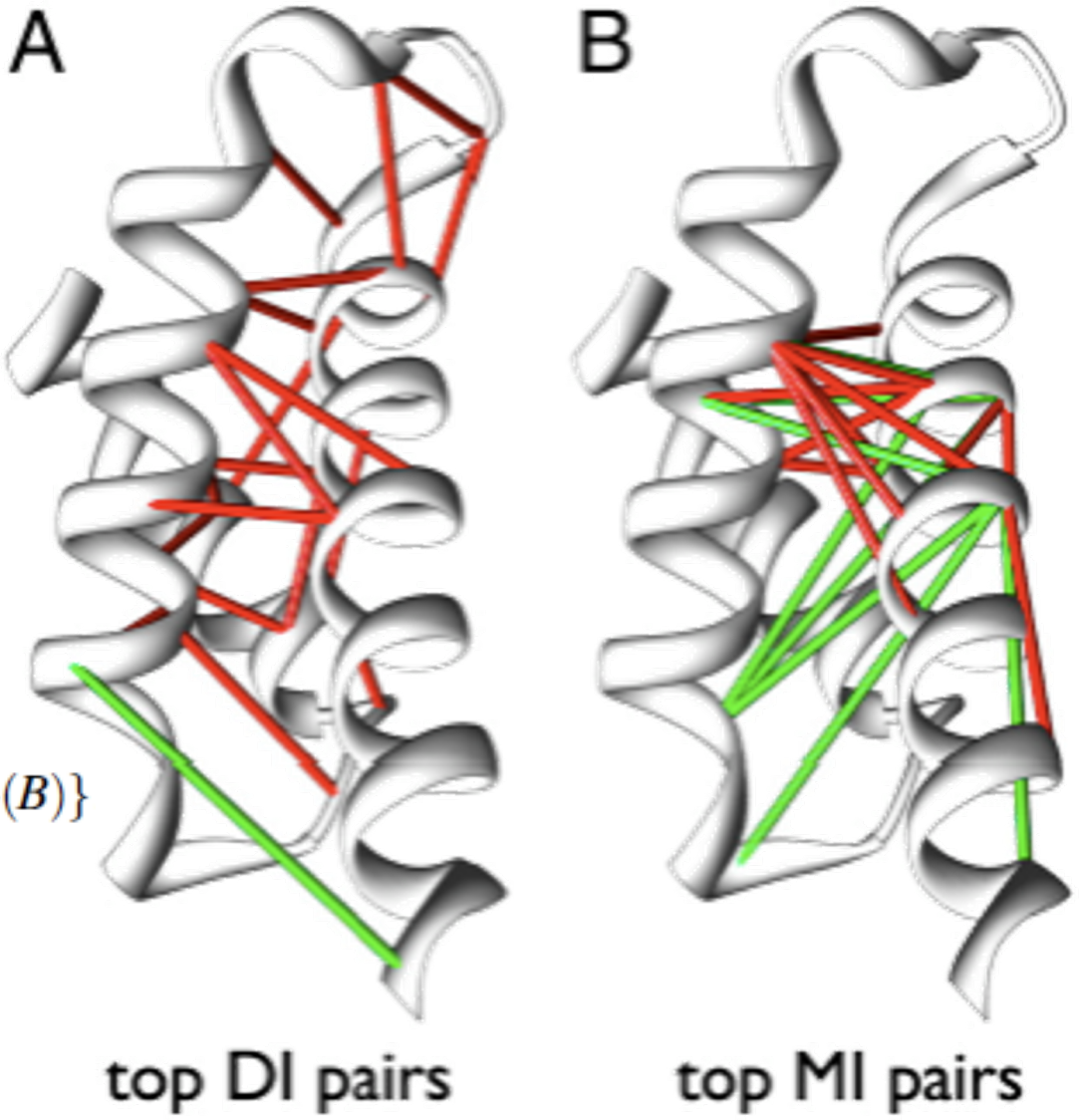
$$\mathrm{MI}_{ij} = \sum_{A,B} f_{ij}(A,B) \ \ln \frac{f_{ij}(A,B)}{f_i(A) f_j(B)},$$

$$\mathrm{DI}_{ij} = \sum_{AB} P_{ij}^{(\mathrm{dir})}(A,B) \ \ln \frac{P_{ij}^{(\mathrm{dir})}(A,B)}{f_i(A)\, f_j(B)}.$$

$$f_i(A) = \sum_{B} P_{ij}^{(\mathrm{dir})}(A,B),$$

$$f_j(B) = \sum_{A} P_{ij}^{(\mathrm{dir})}(A,B).$$

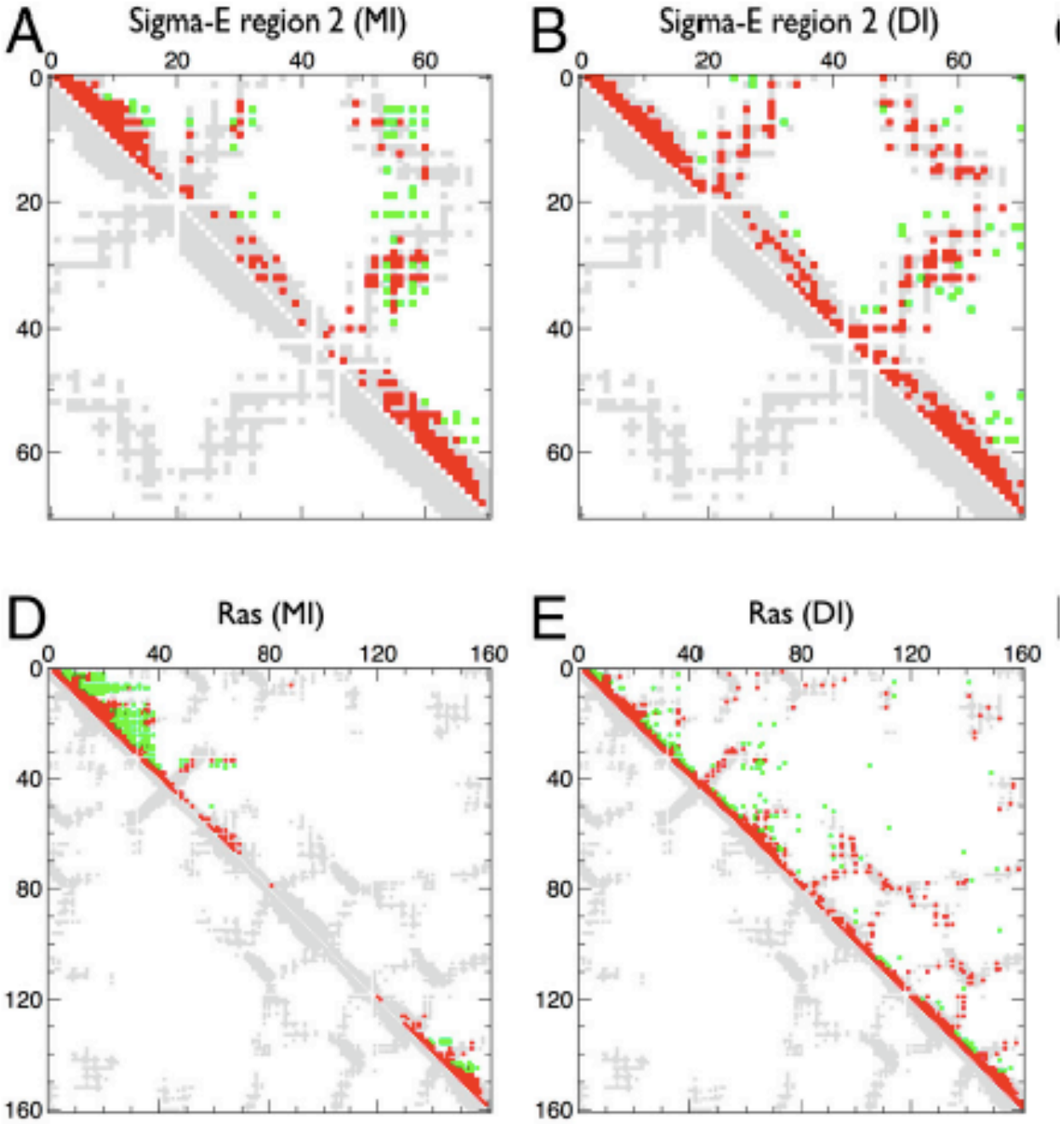
$$P_{ij}^{(\mathrm{dir})}(A,B) = \frac{1}{Z_{ij}} \exp\{e_{ij}(A,B) + \tilde{h}_i(A) + \tilde{h}_j(B)\}$$



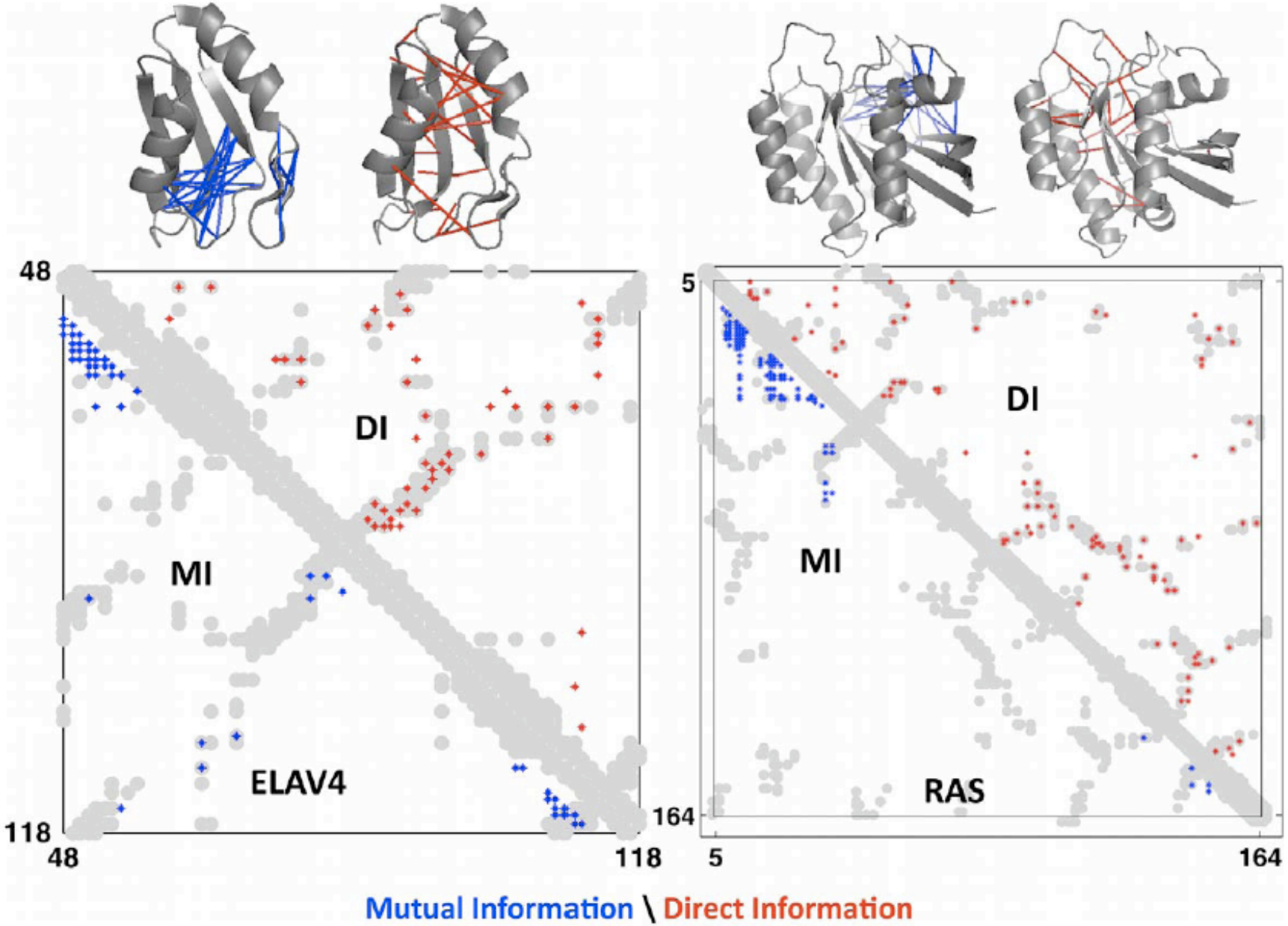
Marks DS et al.. PLoS One. 2011;6(12):e28766. Epub 2011

Morcos F, et al. . Proc Natl Acad Sci U S A. 2011 Dec 6;108(49):E1293-301.

6. Mutual Information

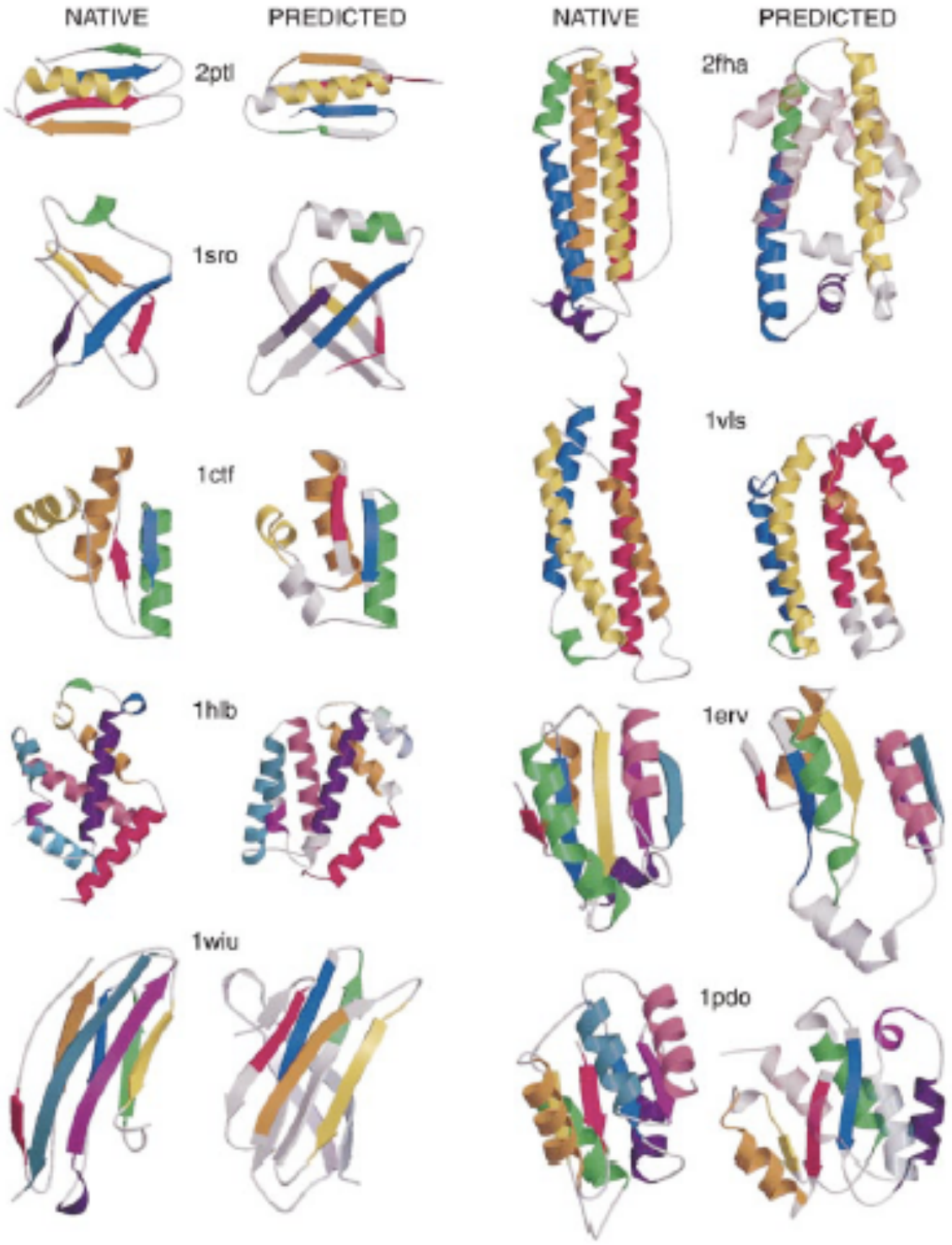


6. Mutual Information

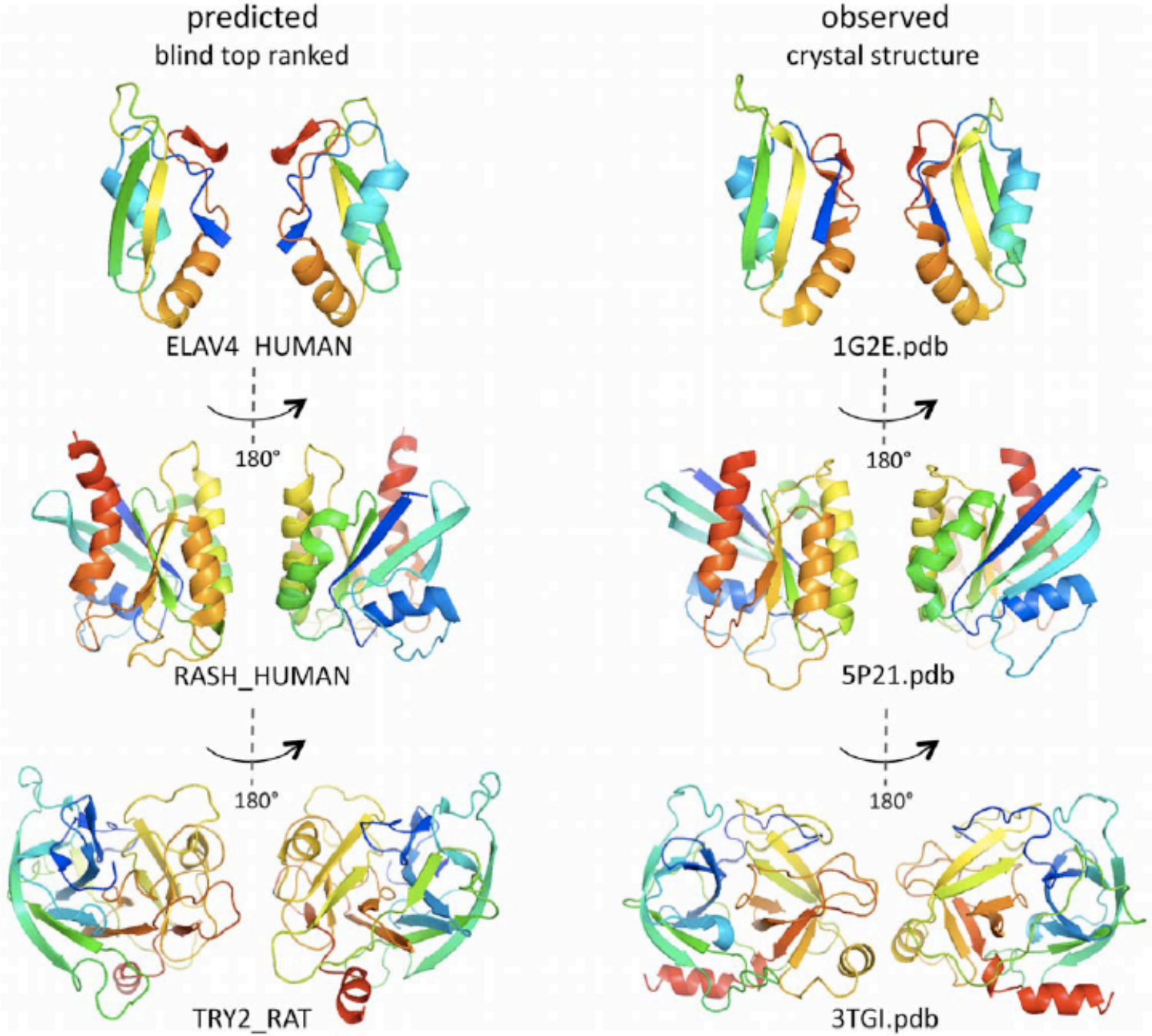


7. Examples

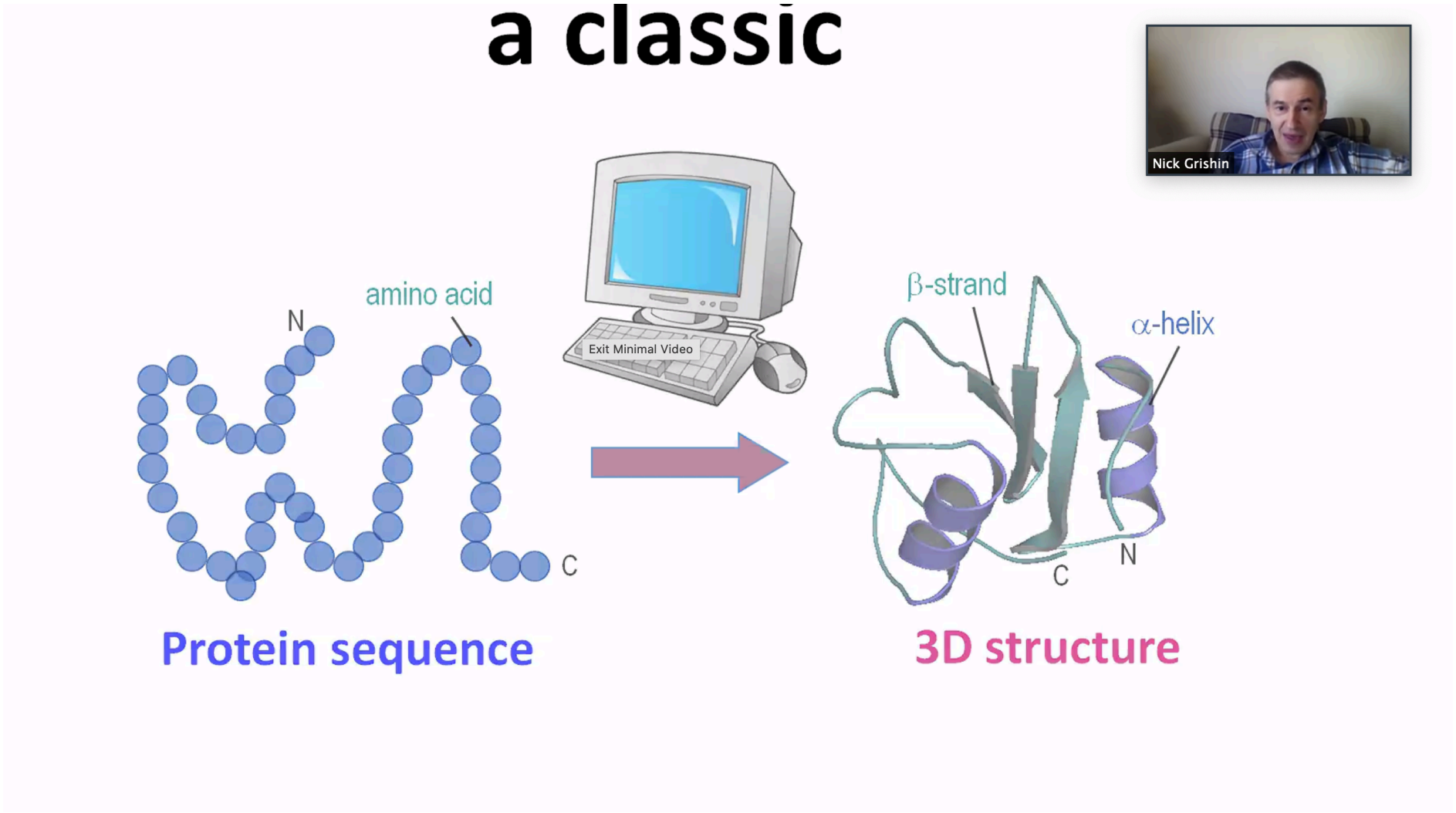
Rosetta



7. Examples Direct information



8. CASP (Critical Assessment of Structure Prediction)



8. CASP (Critical Assessment of Structure Prediction)

EVALUATION: GDT_TS

The GDT score is calculated as the largest set of amino acid residues' alpha carbon atoms in the model structure falling within a defined distance cutoff of their position in the experimental structure, after superimposing two structures.

By the original design the GDT algorithm calculates 20 GDT scores,i.e. for each of 20 consecutive distance cutoffs (0.5 Å, 1.0 Å, 1.5 Å, ... 10.0 Å).

For structure similarity assessment it is intended to use the GDT scores from several cutoff distances, and scores generally increase with increasing cutoff.
A plateau in this increase may indicate an extreme divergence between the experimental and predicted structures, such that no additional atoms are included in any cutoff of a reasonable distance.

The conventional GDT_TS total score in CASP is the average result of cutoffs at 1, 2, 4, and 8 Å

GDT_TS was implemented in the **Local-Global Alignment** (LGA) program.

8. CASP (Critical Assessment of Structure Prediction)

EVALUATION: TM_score

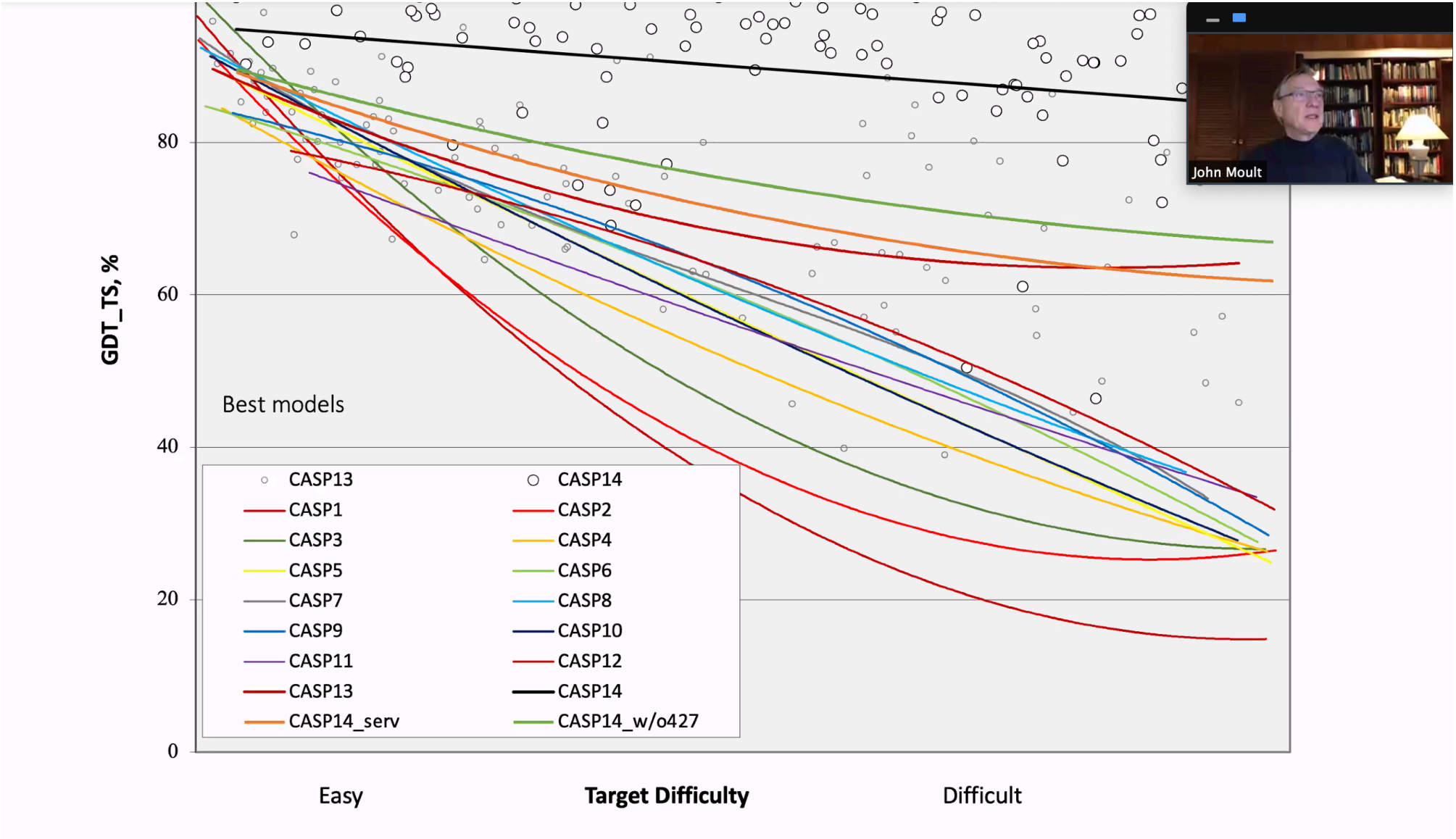
The template modeling score or TM-score is a measure of similarity between two protein structures. The TM-score indicates their difference: 1 indicates a perfect match between two structures (thus the higher the better). Generally, scores below 0.20 corresponds to randomly chosen unrelated proteins, whereas structures with a score higher than 0.5 assume roughly the same fold.

$$\text{TM-score} = \max_{\text{alignment}} \left[\frac{1}{L_{\text{target}}} \sum_{i}^{L_{\text{aligned}}} \frac{1}{1 + \left(\frac{d_i}{d_0(L_{\text{target}})} \right)^2} \right]$$

where

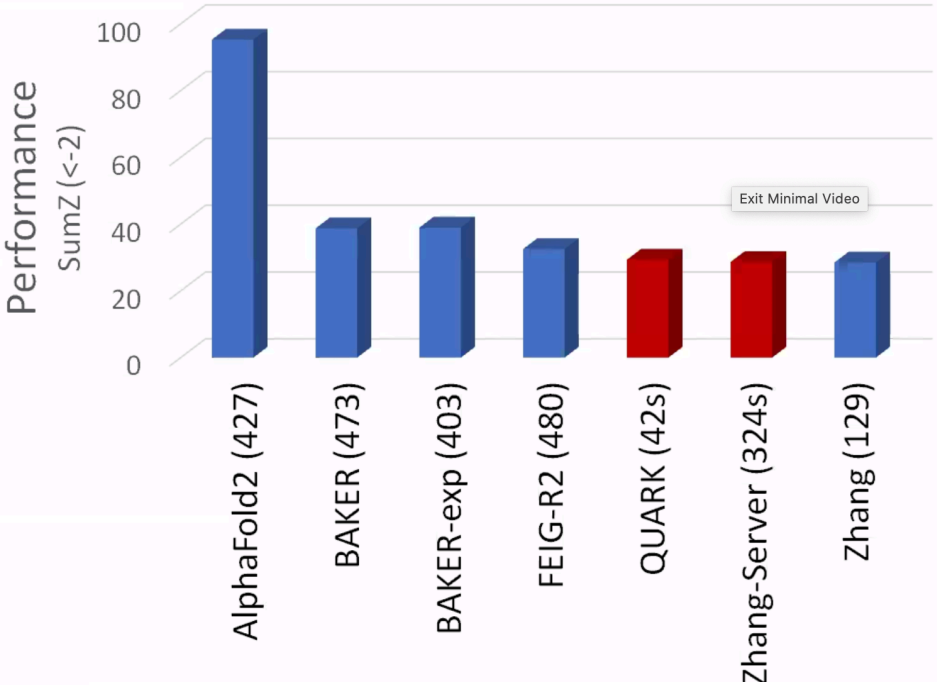
$$d_0(L_{\text{target}}) = 1.24 \sqrt[3]{L_{\text{target}} - 15} - 1.8$$

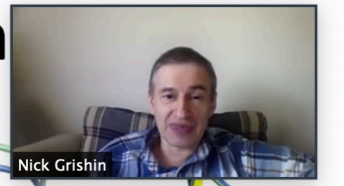
8. CASP (Critical Assessment of Structure Prediction)

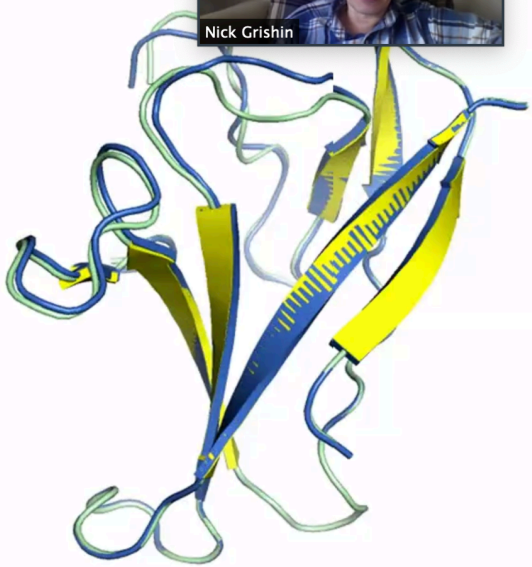


8. CASP (Critical Assessment of Structure Prediction)

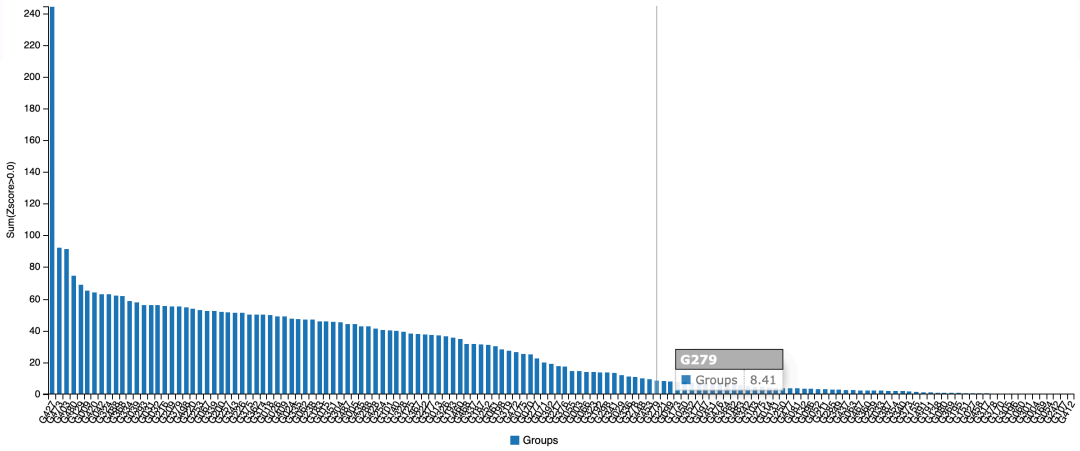
Group 427 (AlphaFold2) Outperforms th





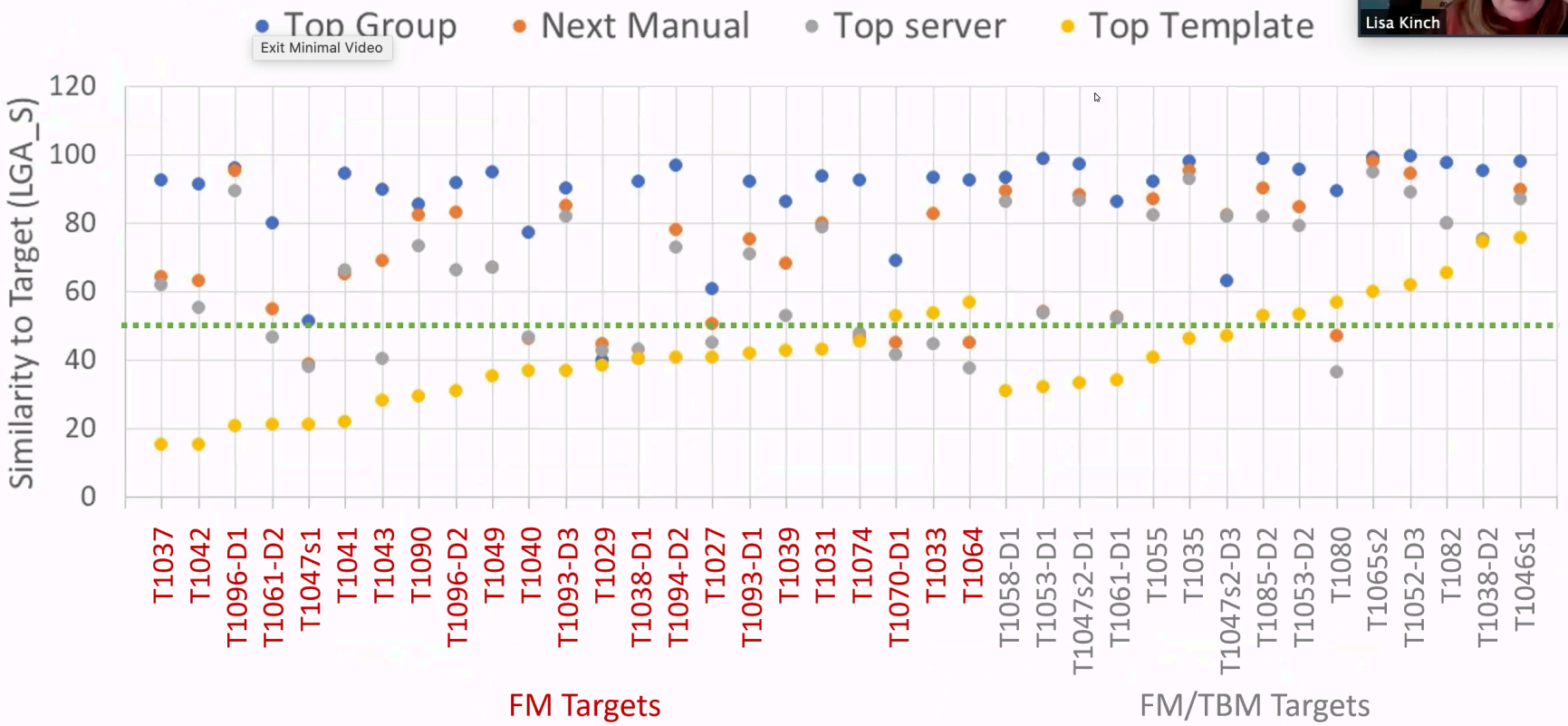


T1064 SARS CoV2 ORF8
AlphaFold2 Model 1
GDT_TS 86.96

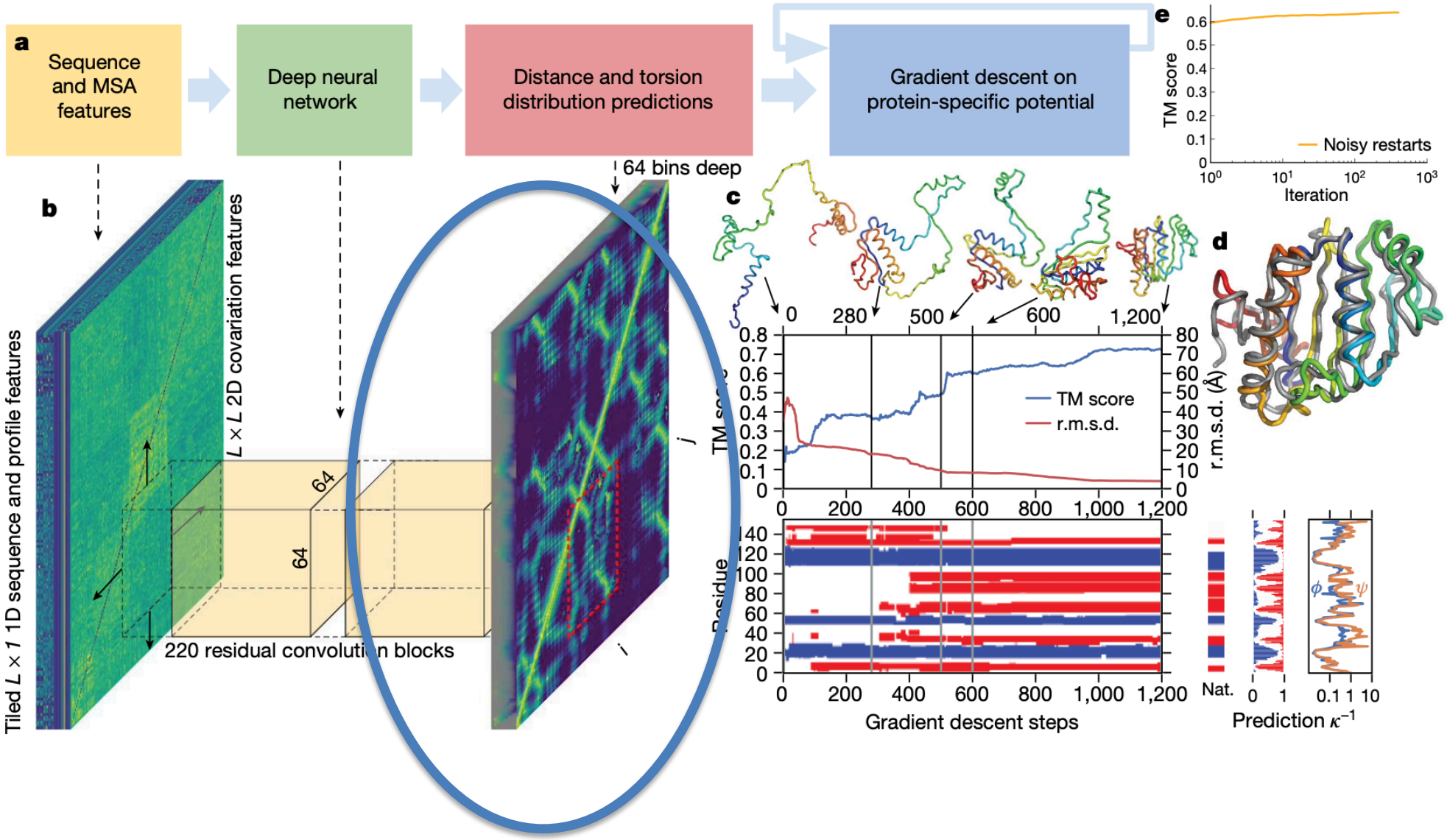


8. CASP (Critical Assessment of Structure Prediction)

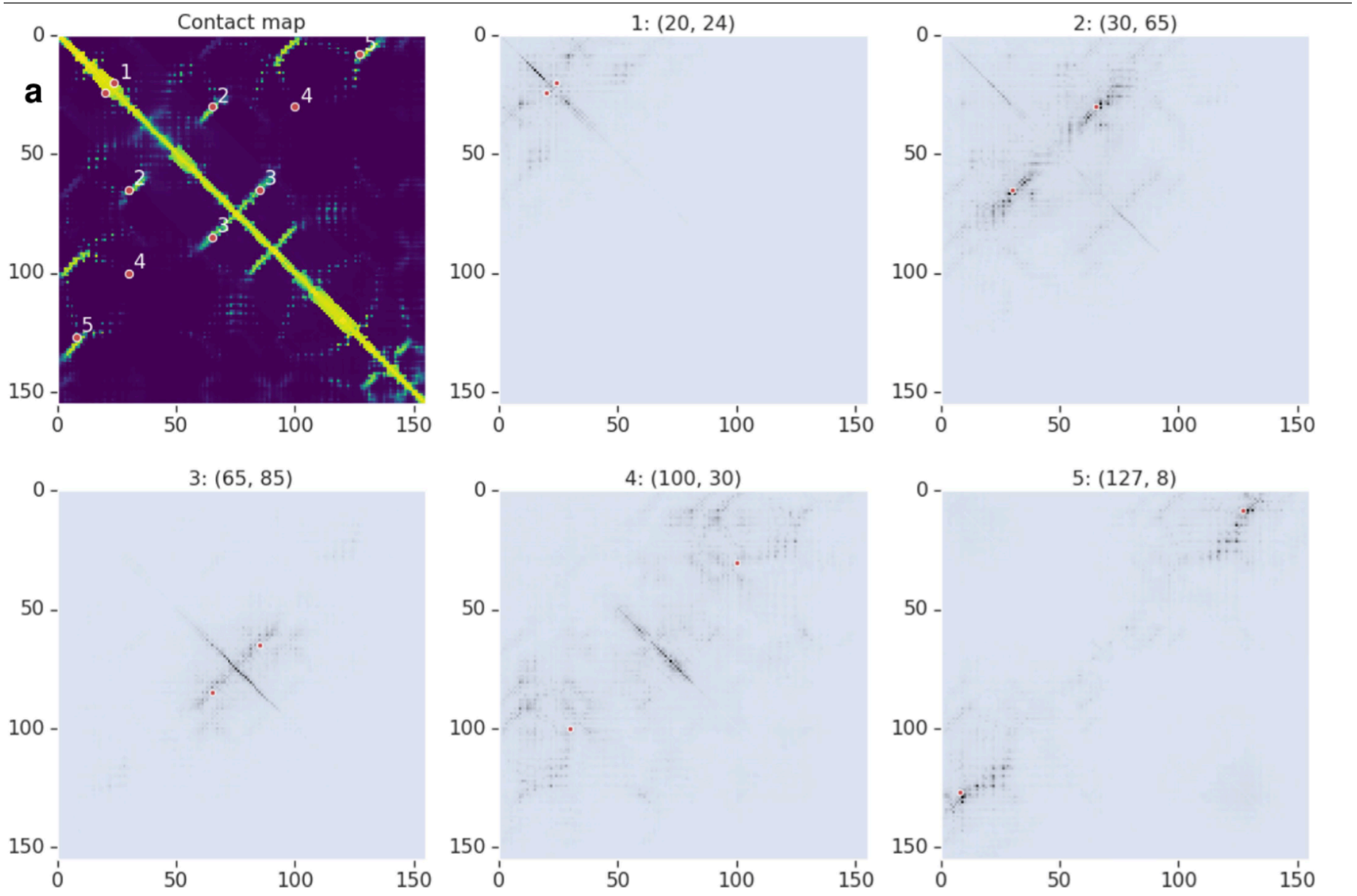
Top Prediction Models Beat Top Templates



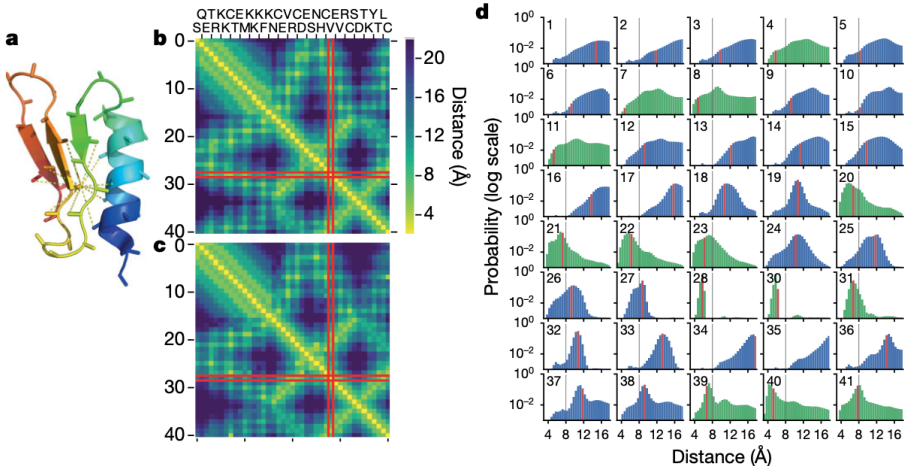
9. AlphaFold 1

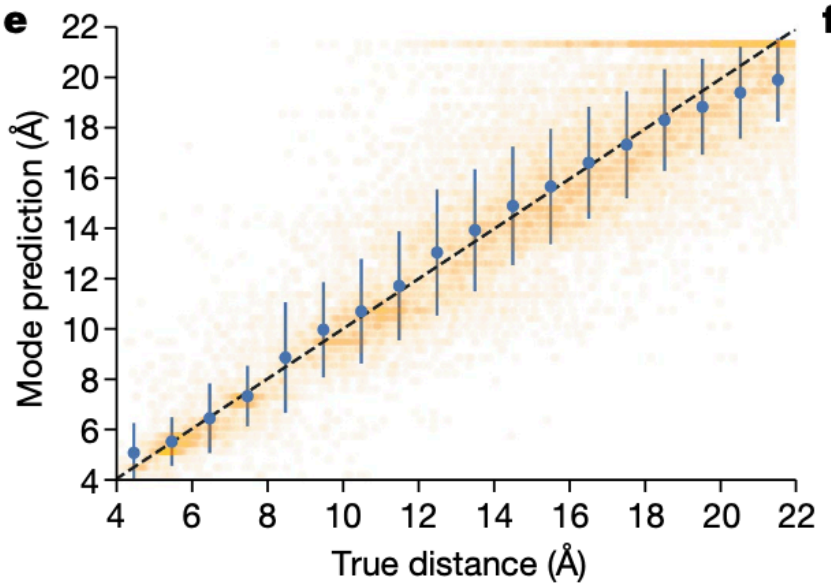


9. AlphaFold 1 (Distogram)



9. AlphaFold 1 (Potentials)





Distance potentials The basic distance potential is computed as a sum over all residue pairs of the likelihood of the inter-residue distances:

$$V_{\text{distance}}(\mathbf{x}) = -\sum_{i,j,\ i\neq j} \log P(d_{ij} \mid \mathcal{S}, \text{MSA}(\mathcal{S})). \tag{1}$$

The distance potential with a reference state becomes:

$$V_{\text{distance}}(\mathbf{x}) = -\sum_{i,j,\ i\neq j} \log P(d_{ij} \mid \mathcal{S}, \text{MSA}(\mathcal{S})) - \log P(d_{ij} \mid \text{length}, \delta_{\alpha\beta}). \tag{2}$$

The torsions are modelled with a von Mises distribution for each residue:

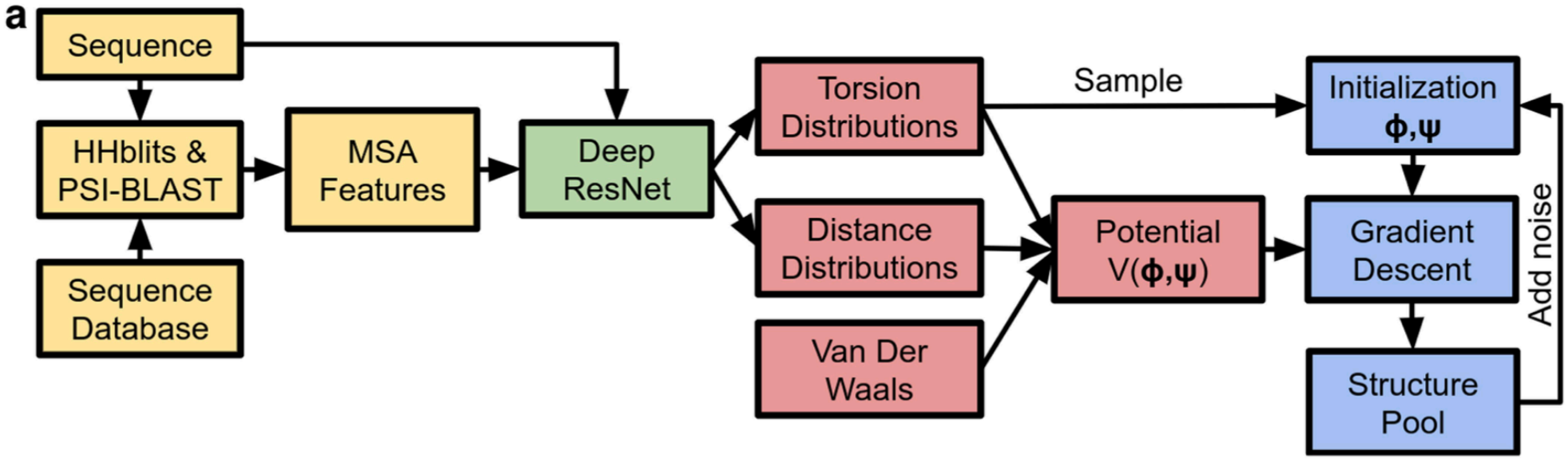
$$V_{\text{torsion}}(\boldsymbol{\phi}, \boldsymbol{\psi}) = -\sum_{i} \log p_{\text{vonMises}}(\phi_i, \psi_i \mid \mathcal{S}, \text{MSA}(\mathcal{S})). \tag{3}$$

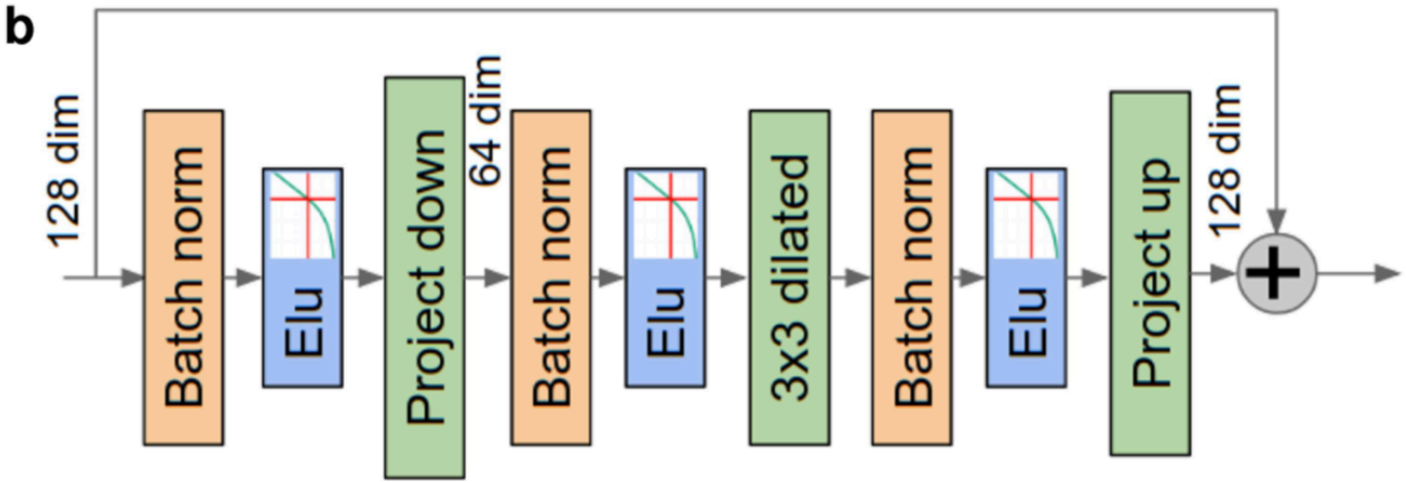
The total potential that we optimise is thus:

$$V_{\text{total}}(\boldsymbol{\phi}, \boldsymbol{\psi}) = V_{\text{distance}}(G(\boldsymbol{\phi}, \boldsymbol{\psi})) + V_{\text{torsion}}(\boldsymbol{\phi}, \boldsymbol{\psi}) + V_{\texttt{score2\_smooth}}(G(\boldsymbol{\phi}, \boldsymbol{\psi})). \tag{4}$$

The terms are weighted equally as determined by cross-validation.

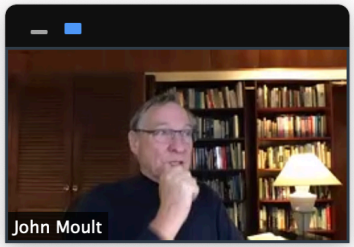
9. AlphaFold 1



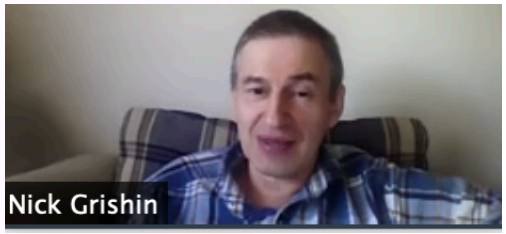


10. …where the future goes

Not an end, but a Beginning: The door is open to?:



- Protein complexes
- Accuracy estimation
- Protein design
- Protein dynamics
- Protein conformational change
- Preferred conformations of disordered proteins
- Mutation interpretation
- Ligand docking



We are approaching the times when

Exit Minimal Video

computational biology

will be used to **VALIDATE**

experimental structures