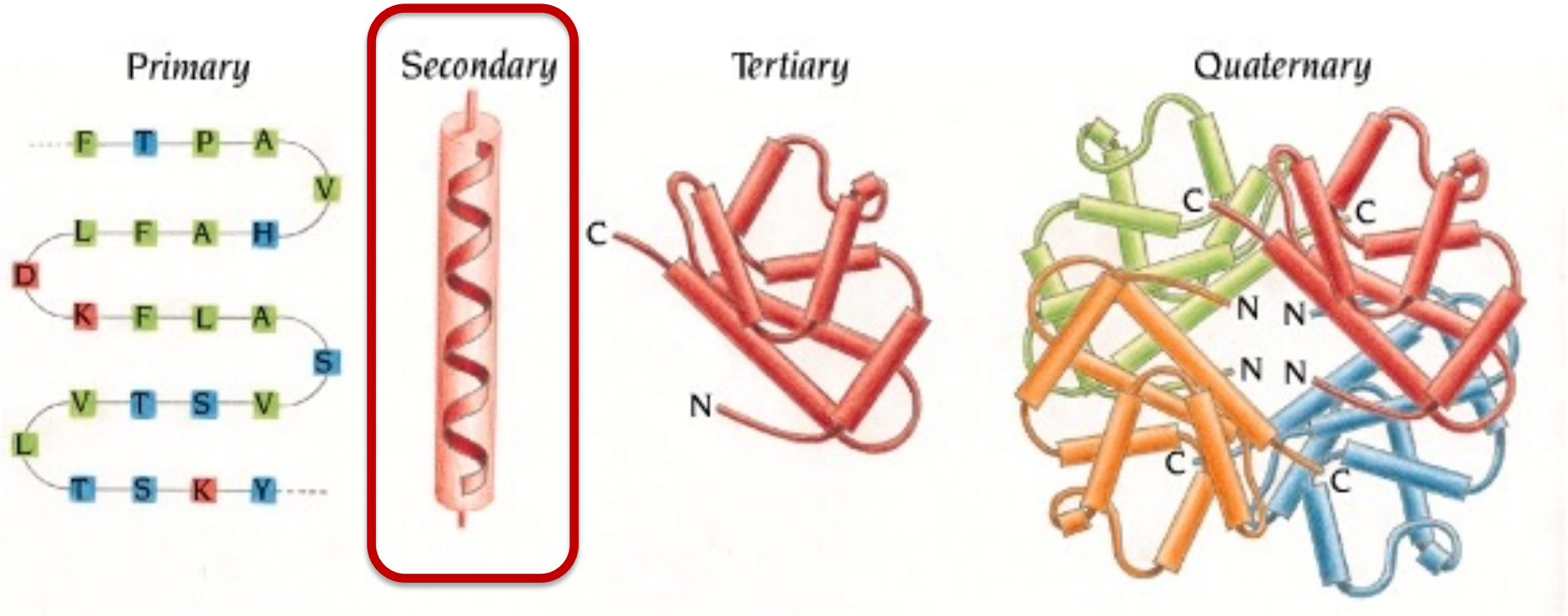
BIOENG-320
Synthetic Biology

Protein design Lecture 2
March 3, 2025

Patrick Barth
EPFL

What are the common protein structure building blocks that we can design ?

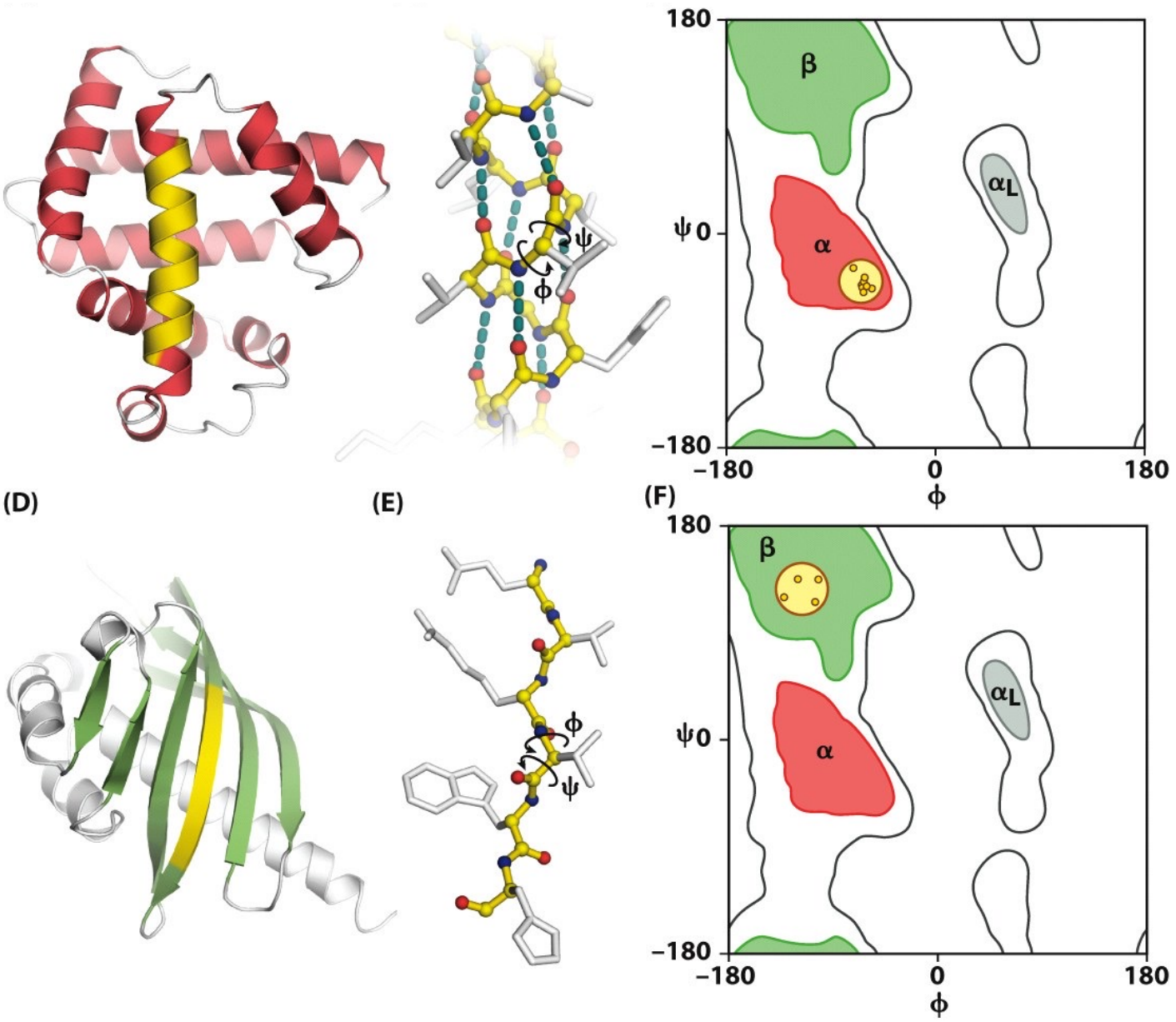
Secondary structure: local interactions



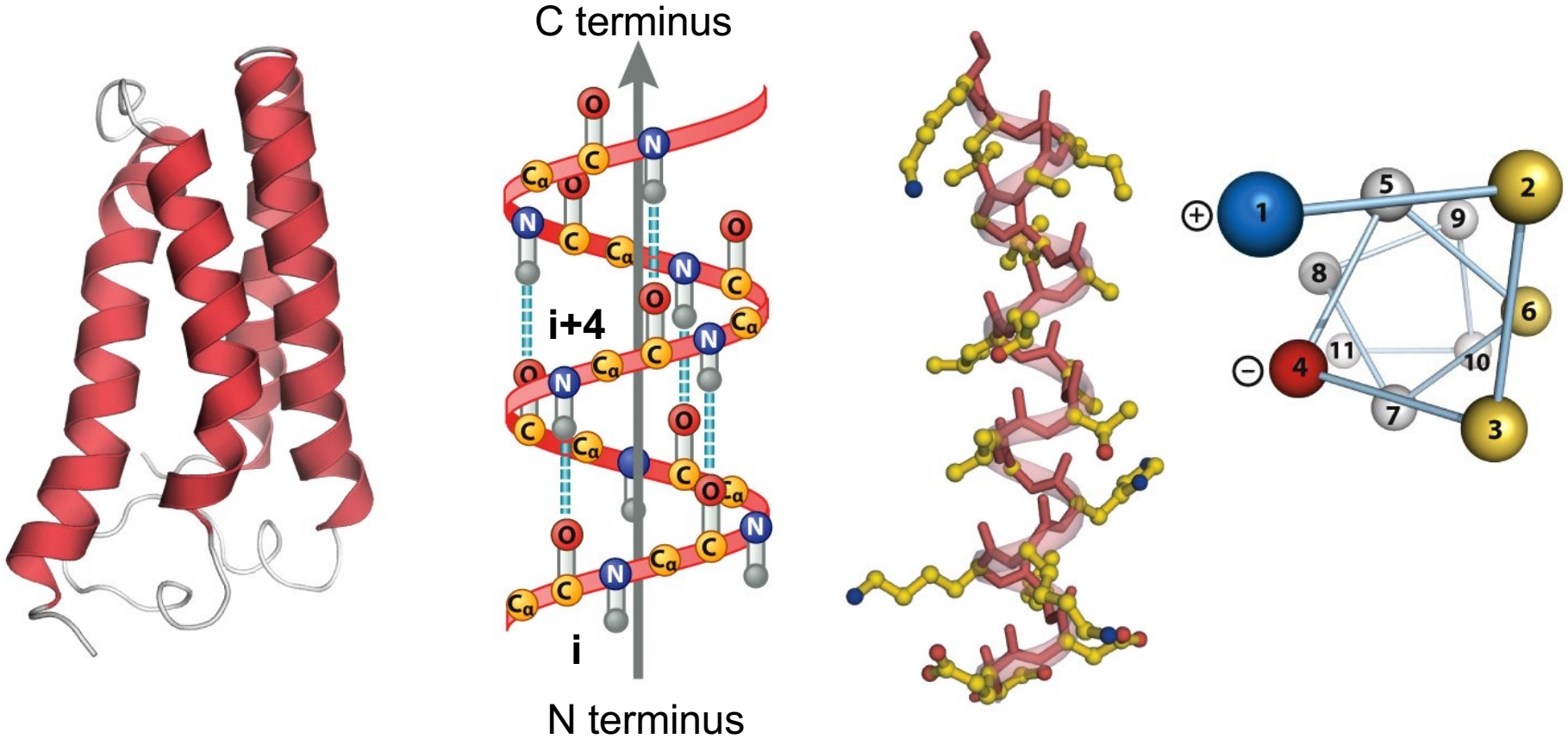
Dihedral subspaces define specific secondary structures

Alpha helix

Beta sheet



Helices built from short range backbone hydrogen bonds



- Alpha helix: the carbonyl oxygen of residue “i” forms a hydrogen bond with the amide of residue “i+4” in the same helix
- π helix: *i* - *i*+5
- 3_{10} helix, PolyProline II helix: *i* - *i*+3

Beta sheet built from long range backbone hydrogen bonds

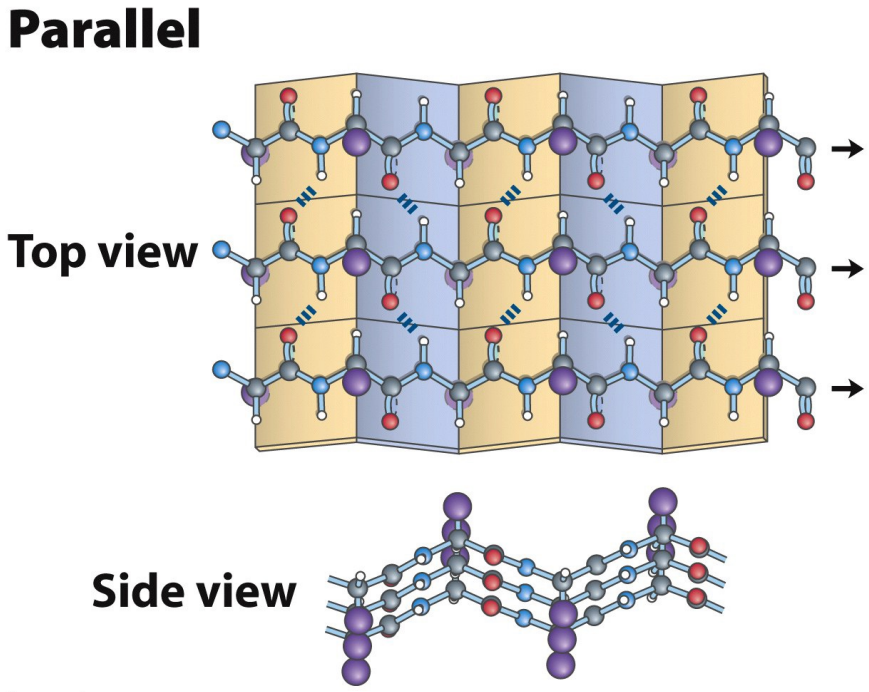


Figure 4-6b
Lehninger Principles of Biochemistry, Fifth Edition

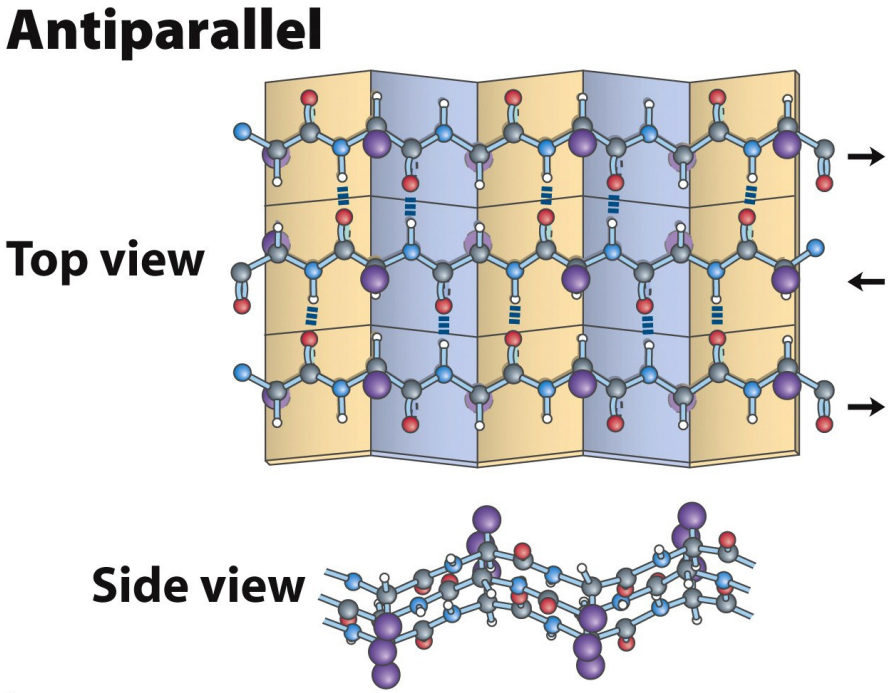
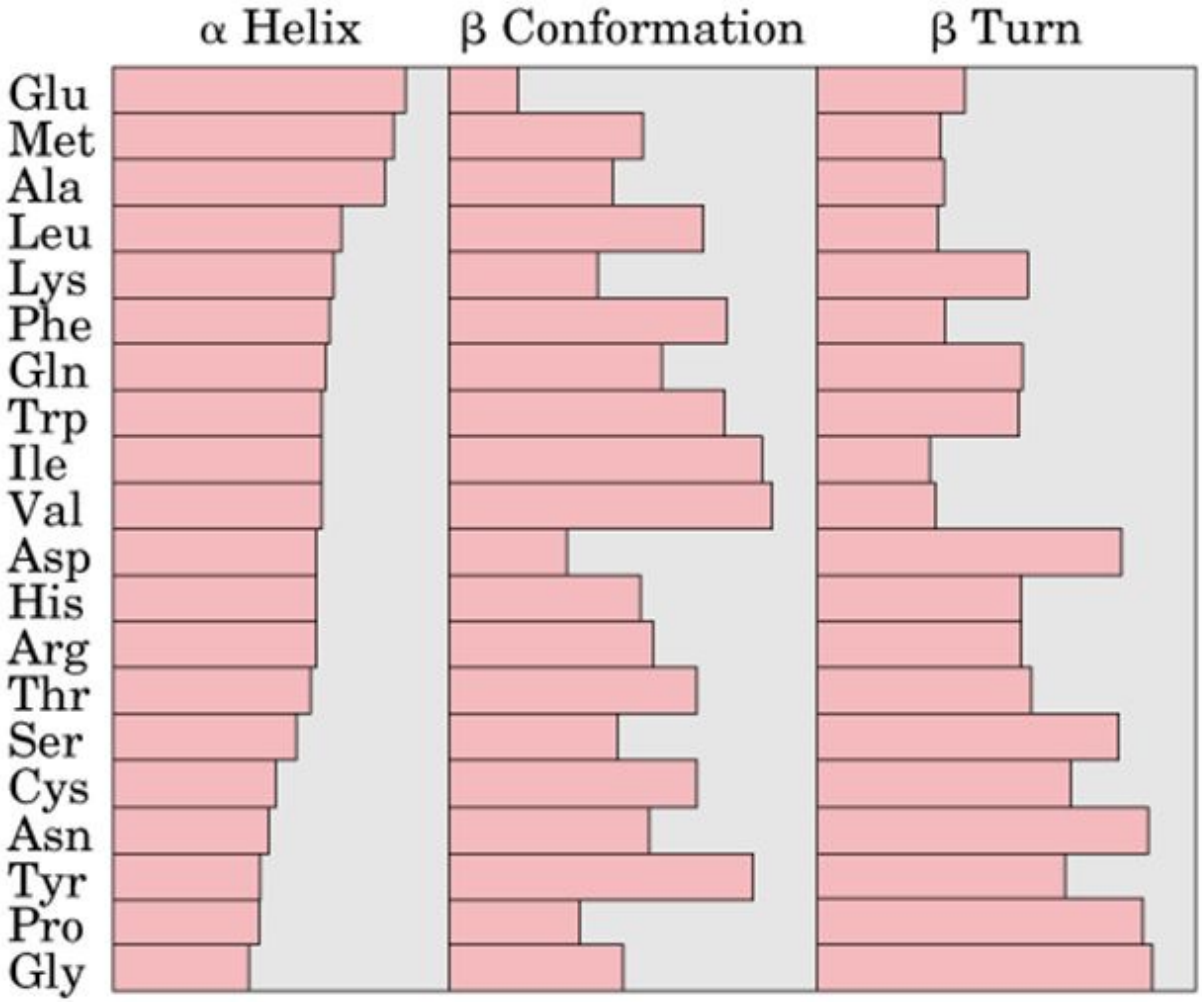


Figure 4-6a
Lehninger Principles of Biochemistry, Fifth Edition

In a b-sheet, carbonyl oxygens and amides form hydrogen bonds between the strands, i.e. between amino acids far away from each other in the primary sequence.

Amino acids have distinct secondary structure propensity



Glu, Met, Ala:
most frequent in a-helix

Val, Tyr, Ile:
most frequent in b-sheet

Pro, Gly, Asn:
Most frequent in b-turn

Conclusion:
Glu has a high a-helix propensity but
a low b-sheet propensity

Loops

- connect helices and strands
- at surface of molecule
- more flexible
- contain functional sites

Hairpin Loops (β turns)

- Connect strands in antiparallel sheet

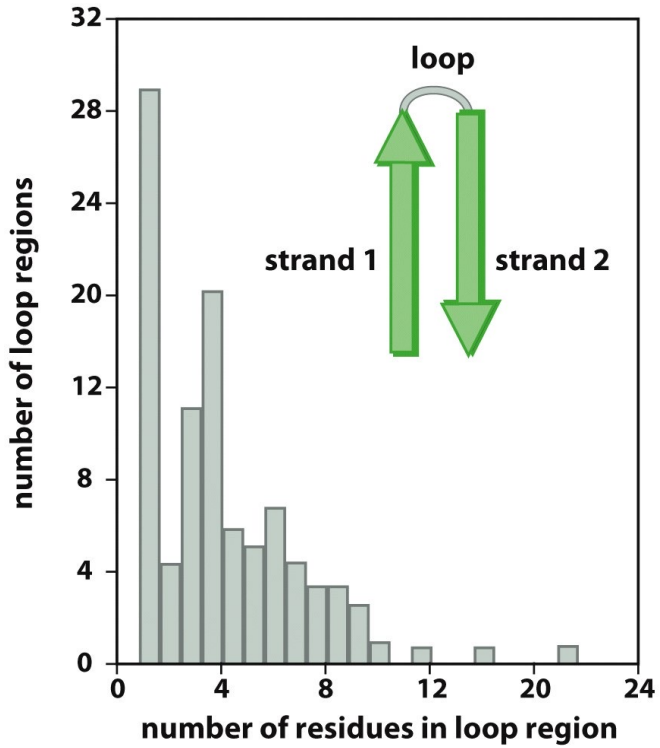


Figure 4.25 The Molecules of Life (© Garland Science 2013)

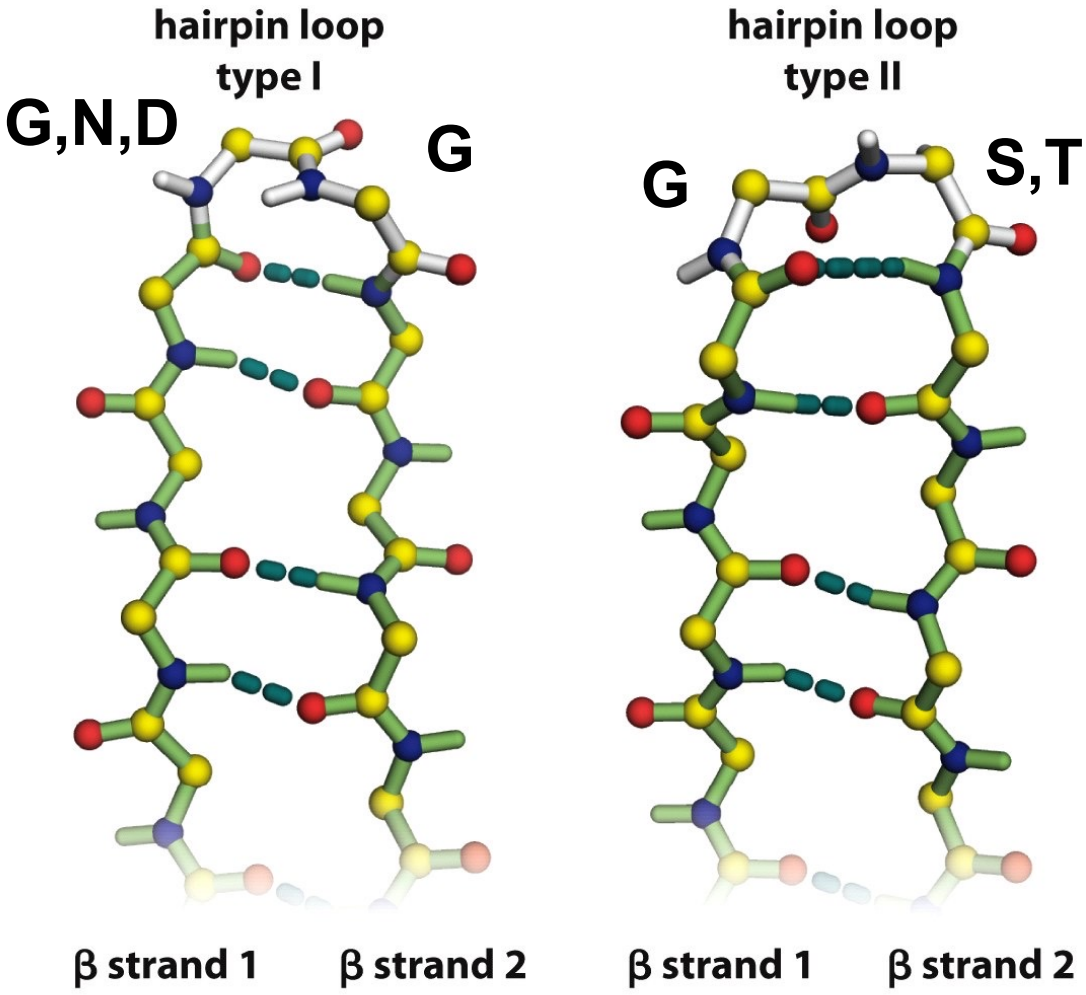
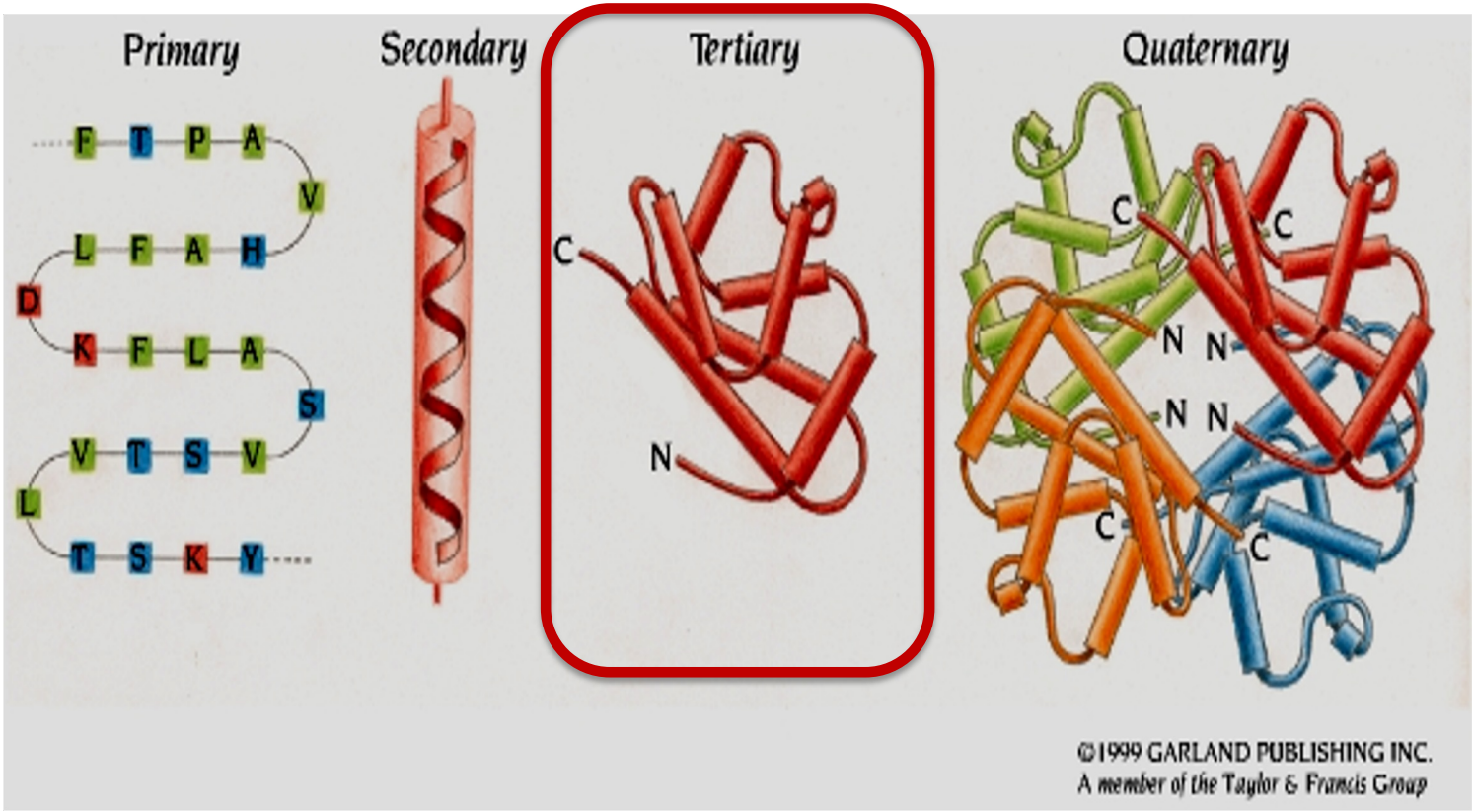


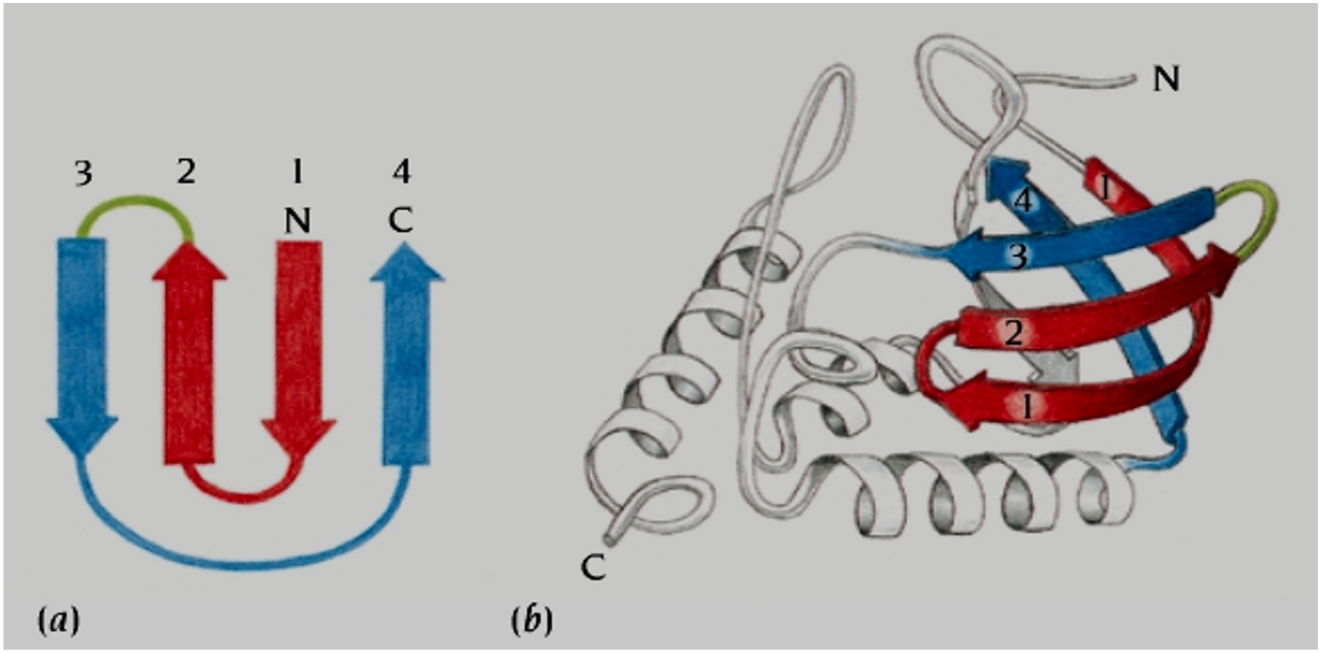
Figure 4.26 The Molecules of Life (© Garland Science 2013)

Connecting elements of secondary structure define tertiary structure



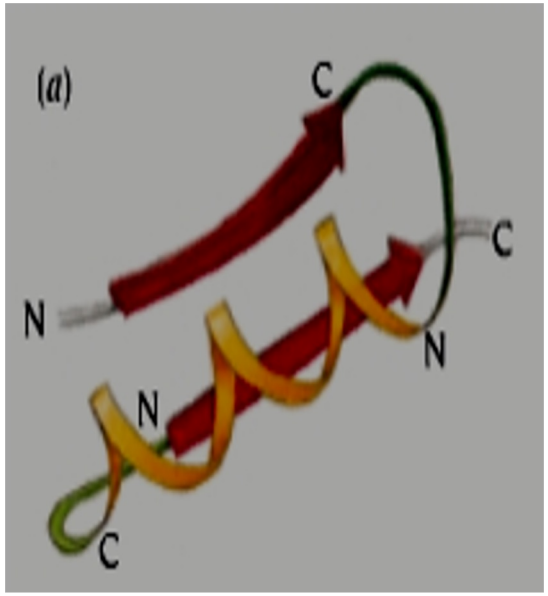
Super secondary structures – Greek Key Motif

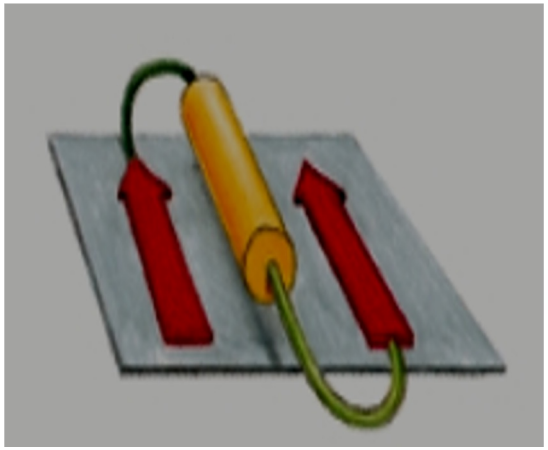
Most common topology for 2 hairpins



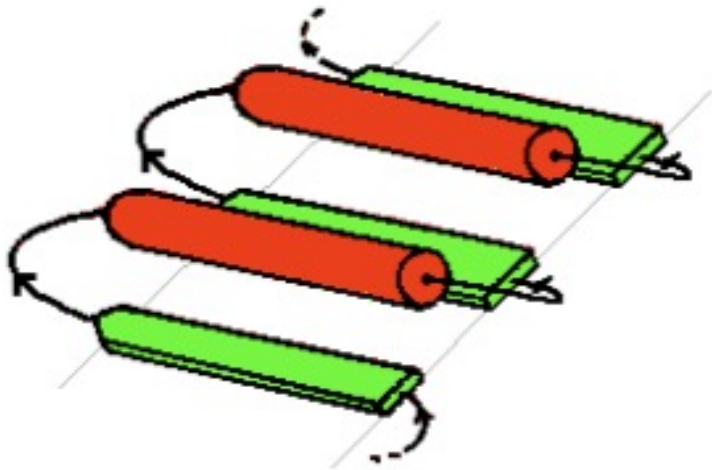
Super Secondary Structures- β–α–β Motif

- connects strands in parallel sheet

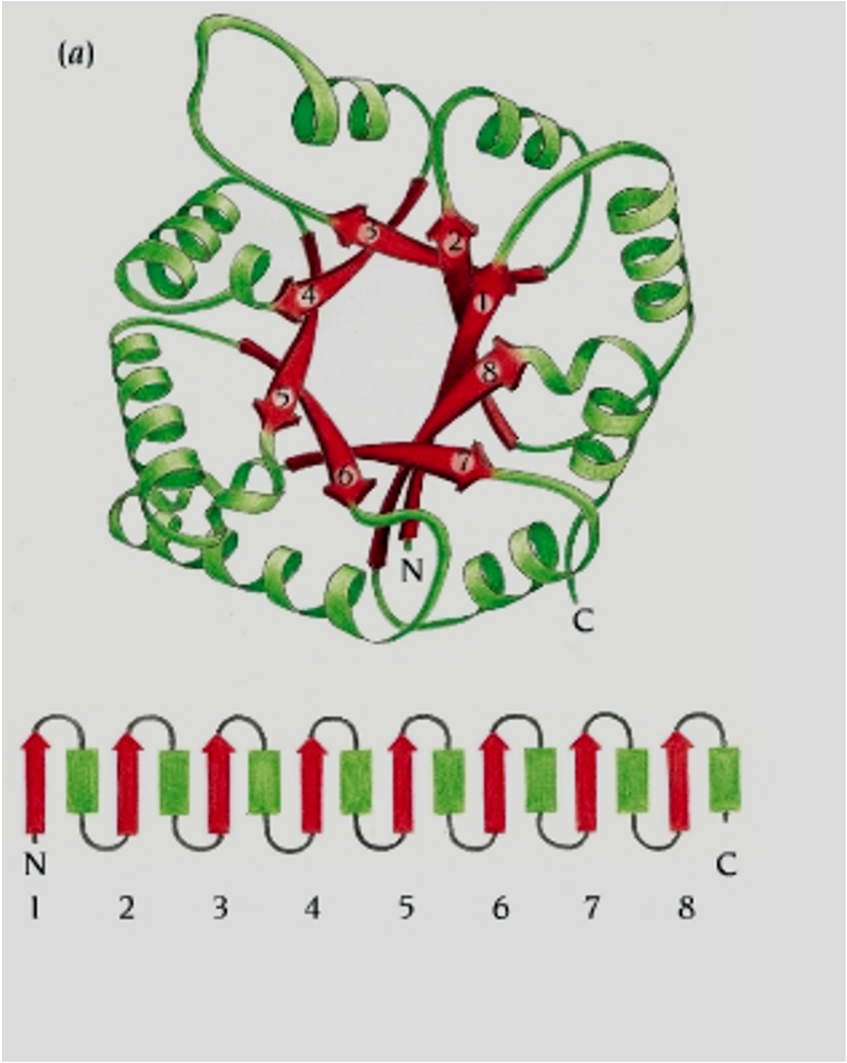




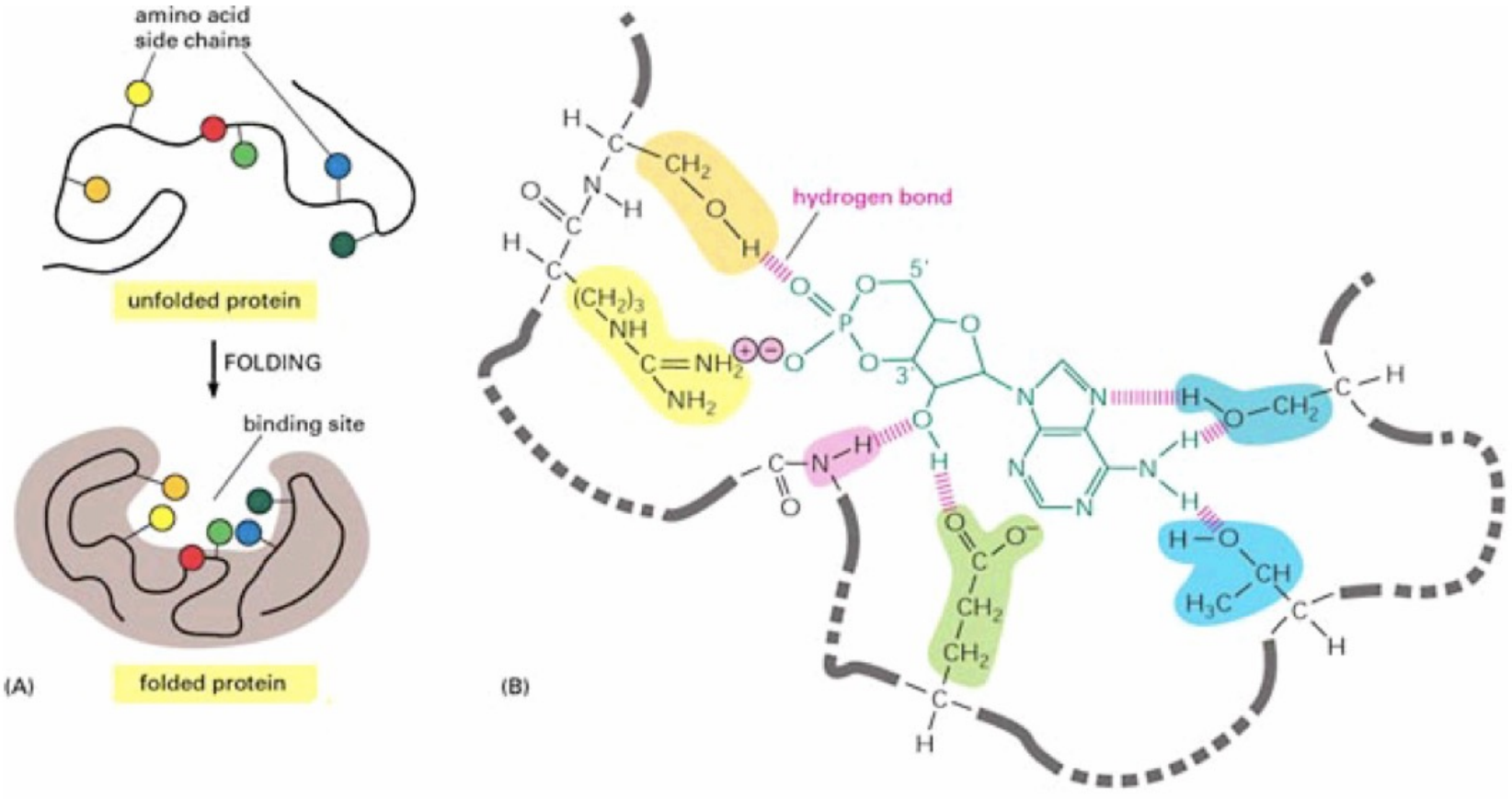
The Rossman fold



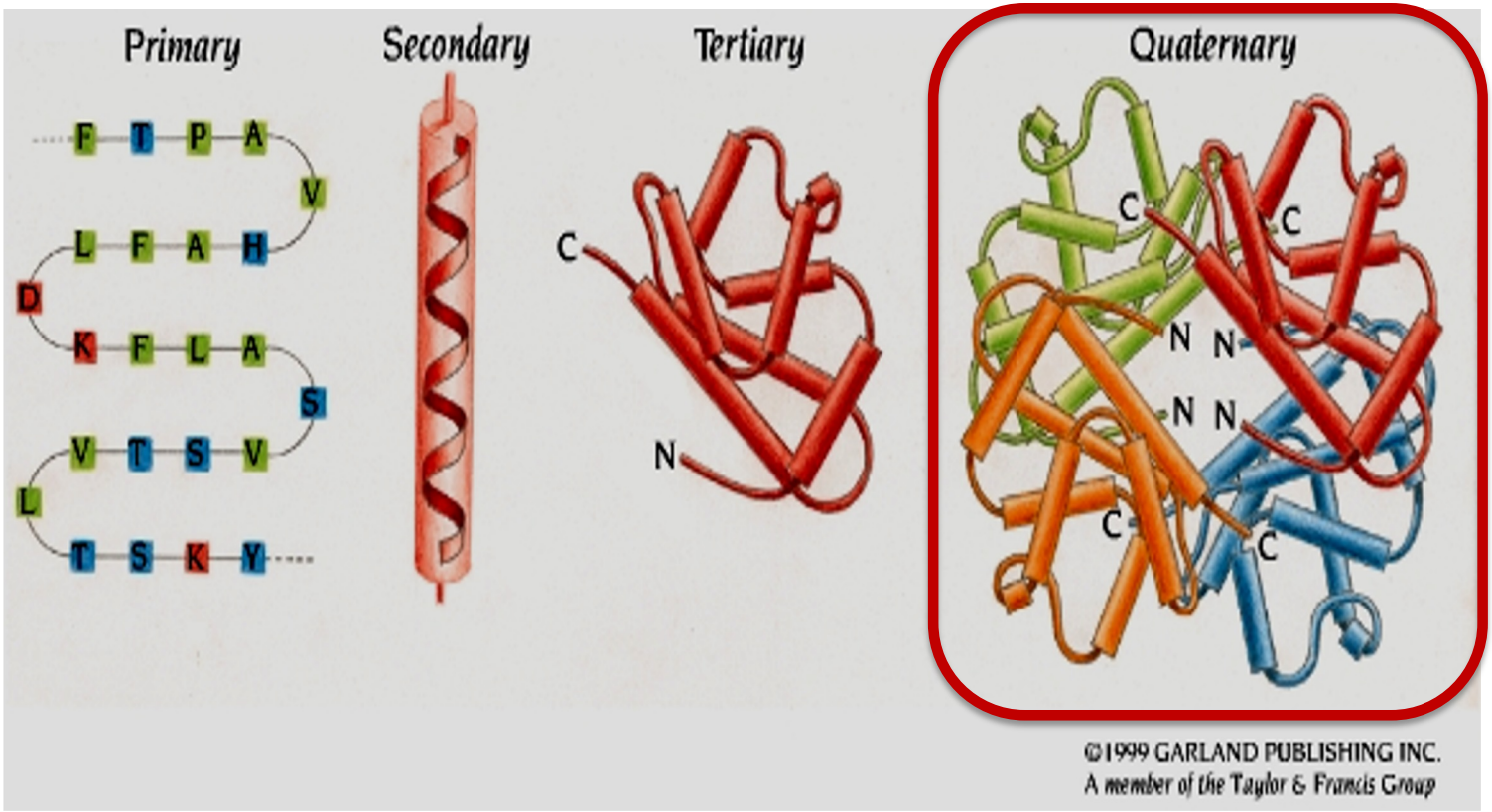
Repeated β–α–β motif creates β-meander: TIM barrel



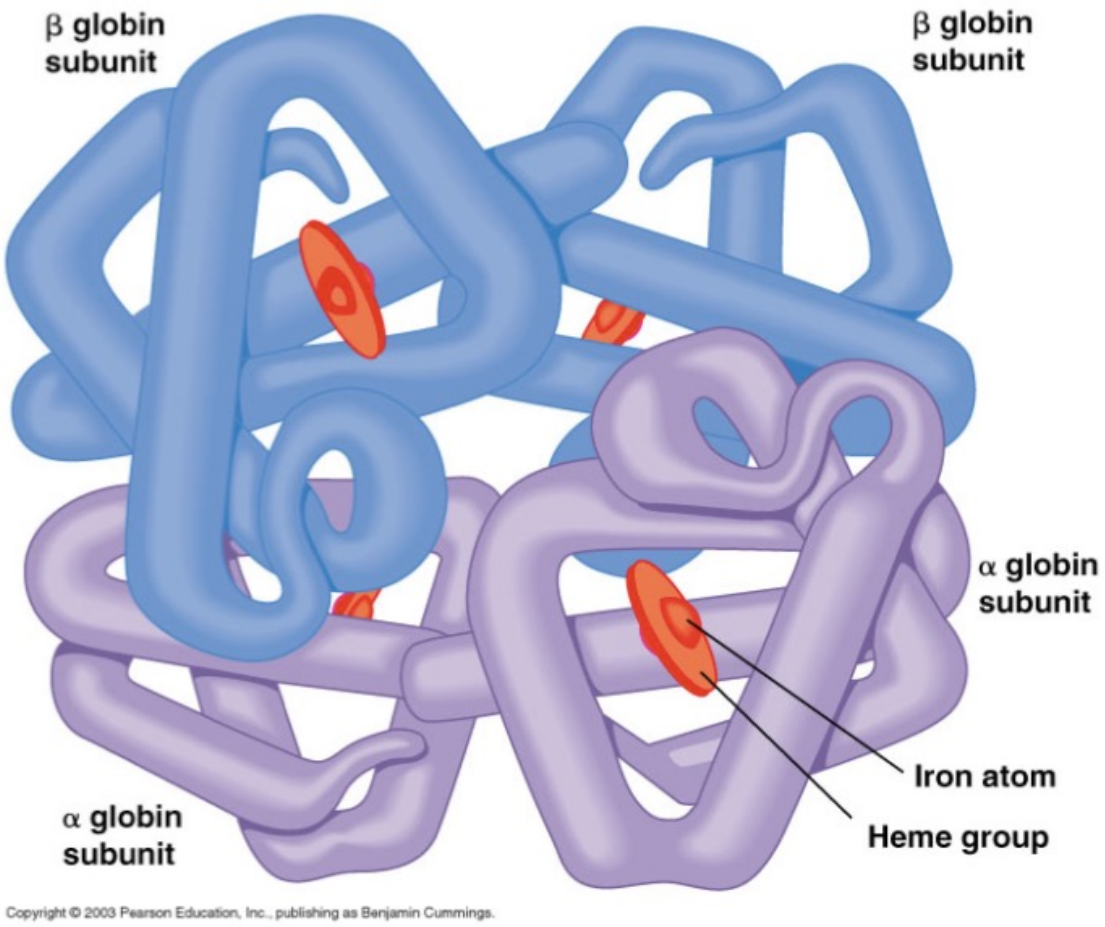
Tertiary structure defines protein function



The quaternary structure of a protein defines its biological functional unit



Quaternary structure: Hemoglobin consists of 4 distinct chains

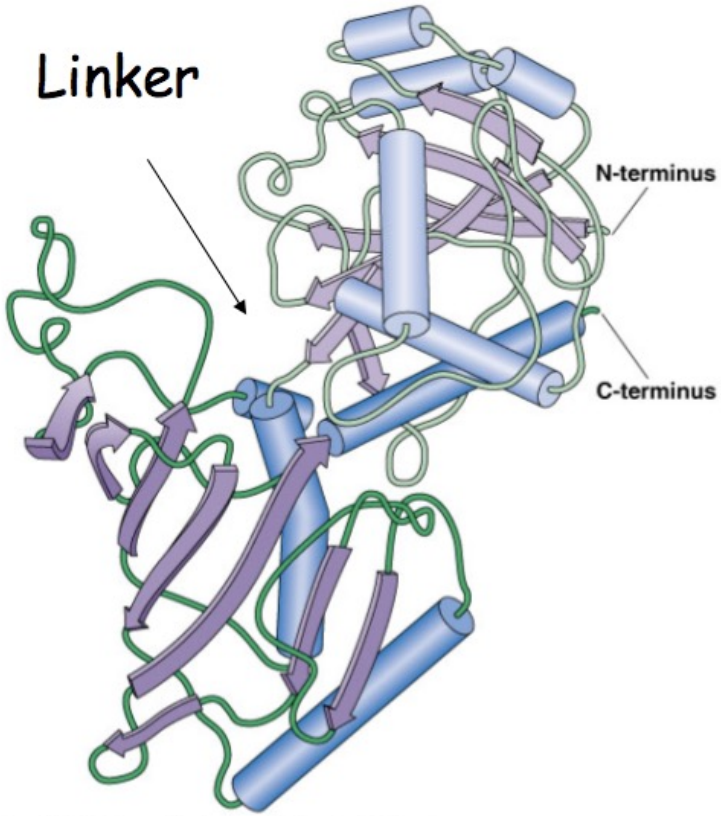


Quaternary structure: assembly of protein domains

(from two distinct protein chains, or two domains in one protein sequence)

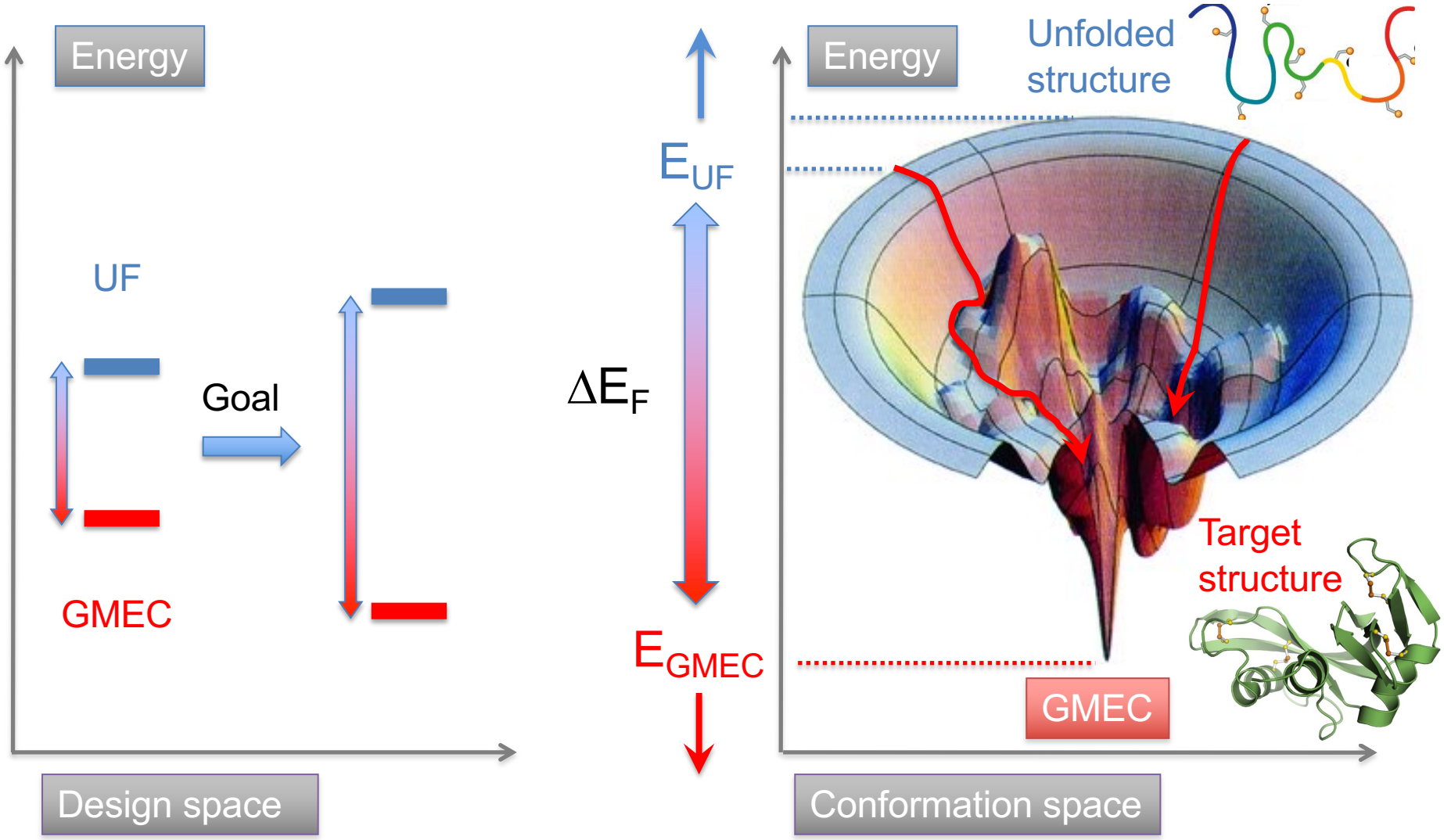
Glyceraldehyde phosphate dehydrogenase:

- domain 1 binds the substance for being metabolized,
- domain 2 binds a cofactor



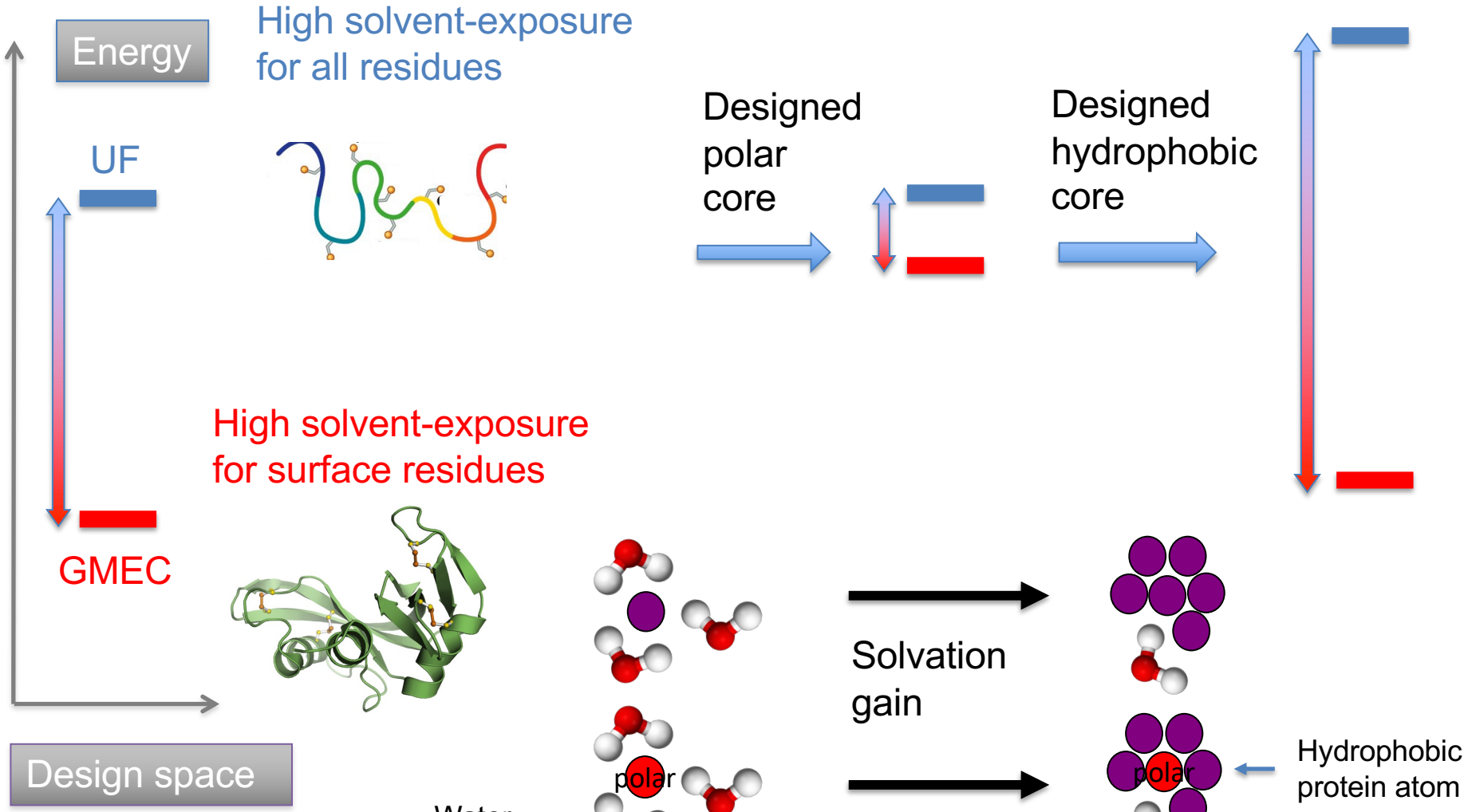
Protein Design – the folding problem

Goal #1: Design a sequence that folds into a given structure



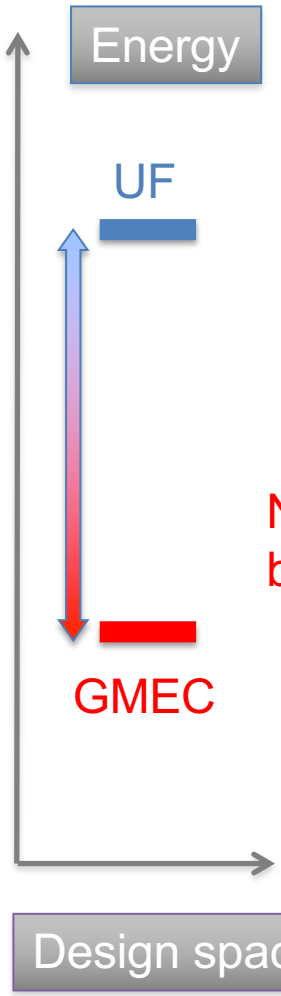
Protein Design – the solvation problem

Goal #1: Maximizing ΔE_F



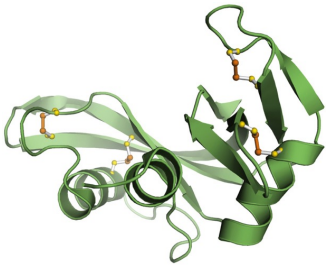
Protein Design – the interaction problem

Goal #1: Maximizing ΔE_F

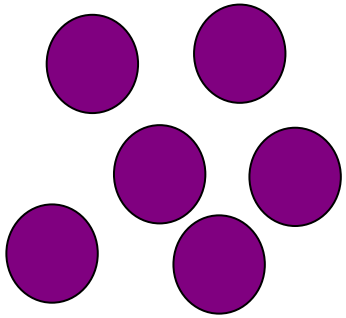


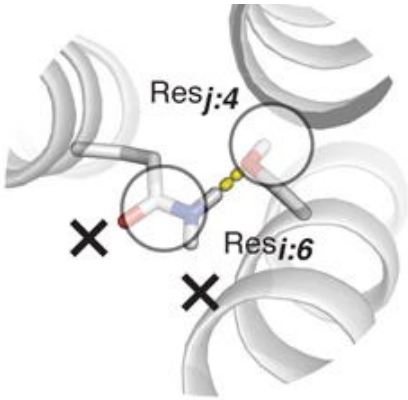
Little interactions between residues

Numerous interactions between core residues



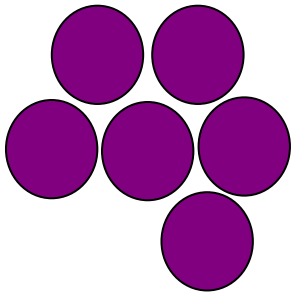
Weak interactions in protein cores



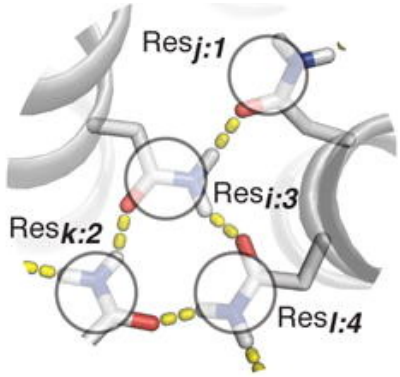


Optimal interactions in protein cores

hydrophobic packing

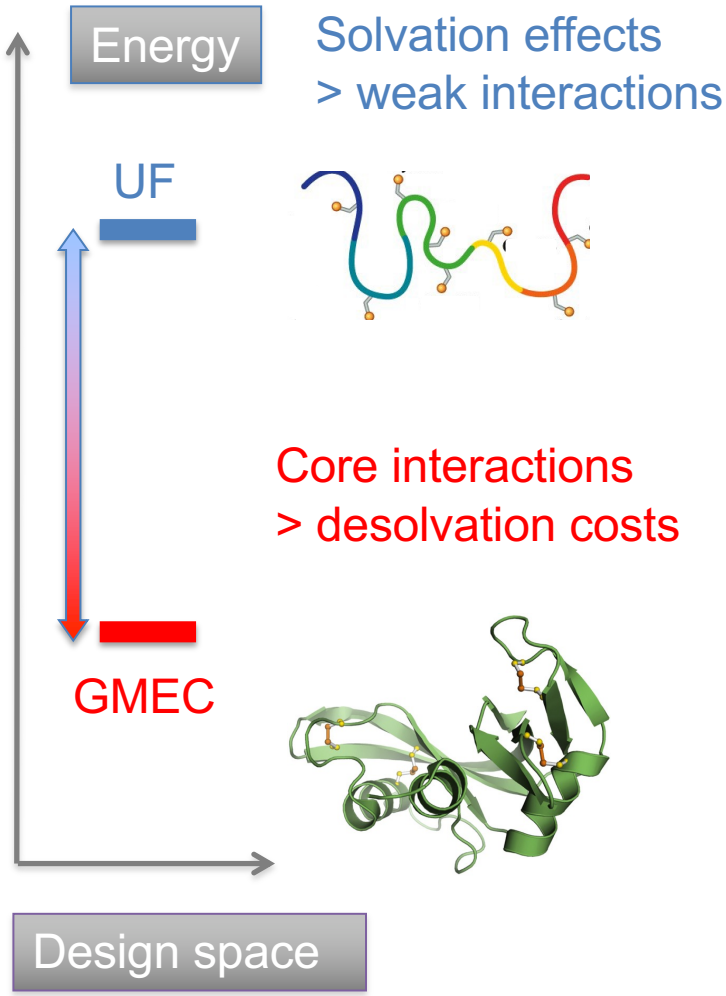


Fully connected polar interaction networks



Protein Design – the interaction problem

Goal #1: Maximizing ΔE_F



The scoring function calculates the balance between opposite energy terms (e.g. polar interactions vs solvation)

$$\textbf{\textit{Score}} = S_{LJ(atr+rep)} + S_{solvation} + S_{hb(srbb+lrbb+sc)} + S_{elec} + S_{dunbrack} + S_{pair} - S_{ref} + S_{prob1b} + S_{intrares} + S_{gsolt} + S_{h2o}(solv + hb) + S_{\_plane}$$

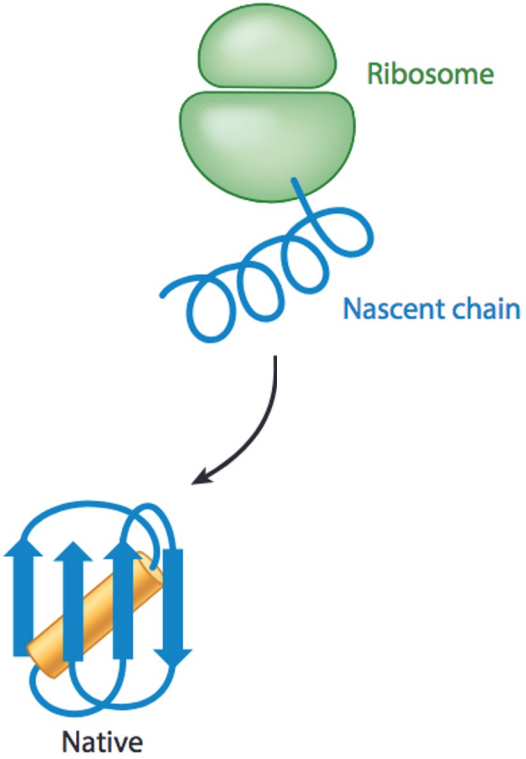
Protein Design – Examples overview

Protein design: Design a sequence that fits to a given structure

1. Design protein stability

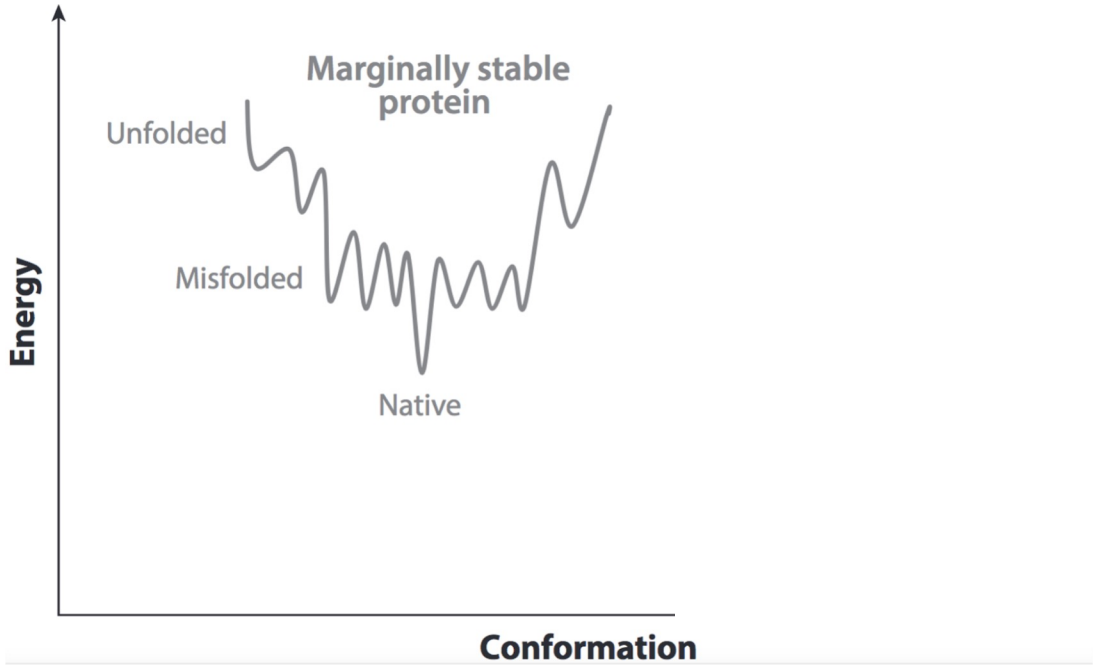
2. Design new protein folds (protein chimera; de novo design ; ANNs)

Protein stability & misfolding are serious challenges

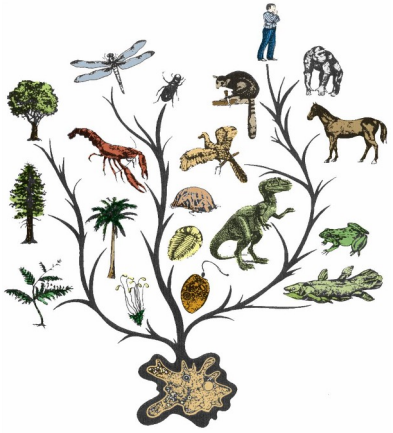


Goldenzweig *Annu Rev Biochem* (2018)

The design goal: improve native-state stability relative to unfolded & misfolded states

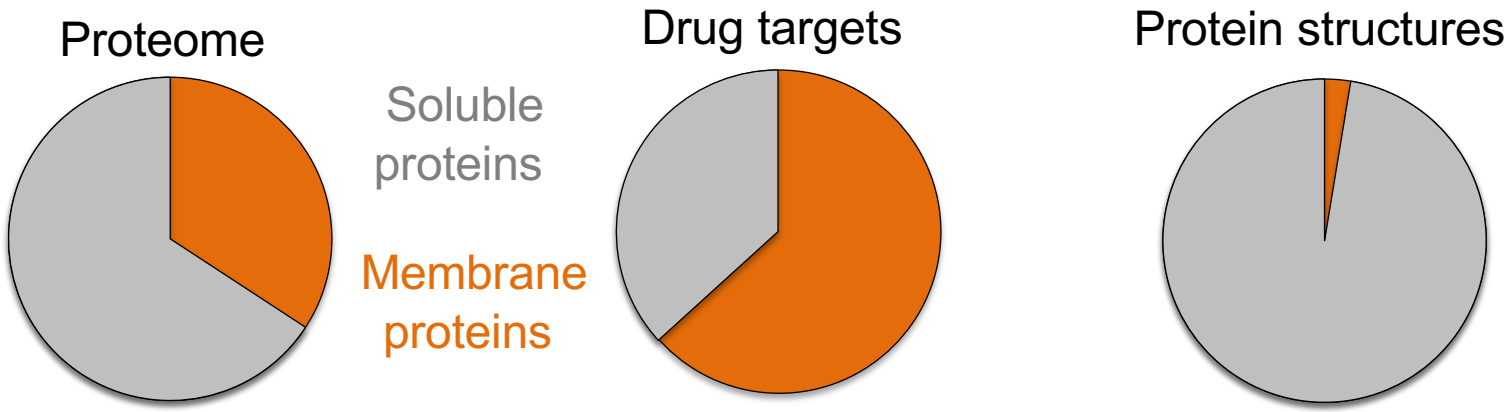


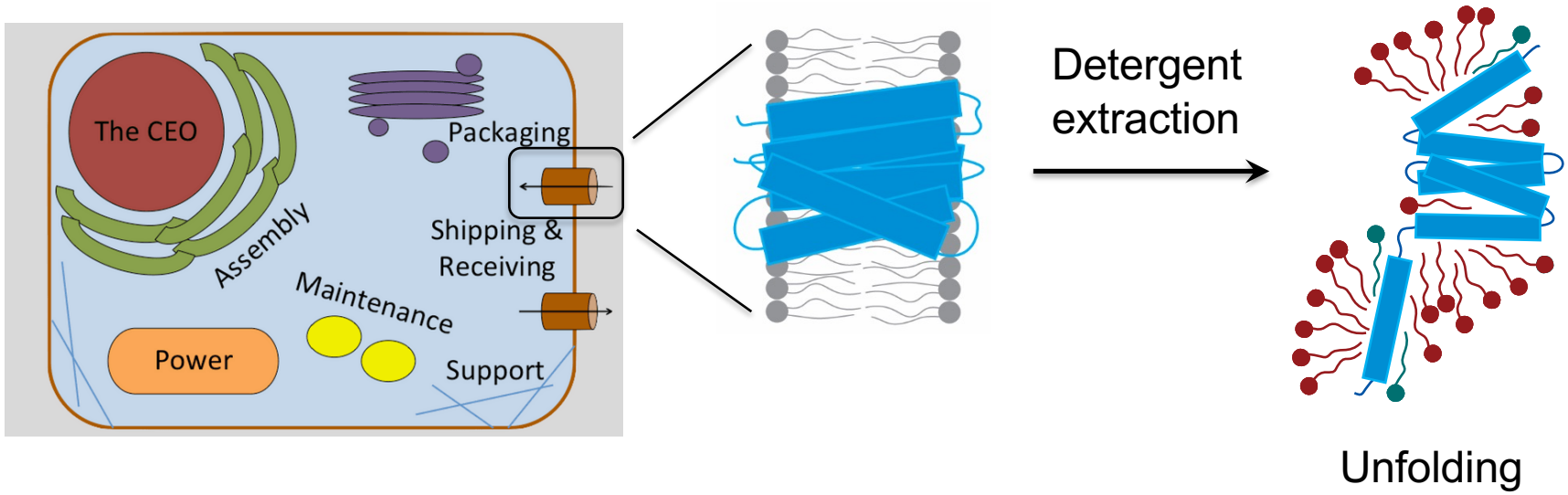
Evolutionary data may counter undesireable states



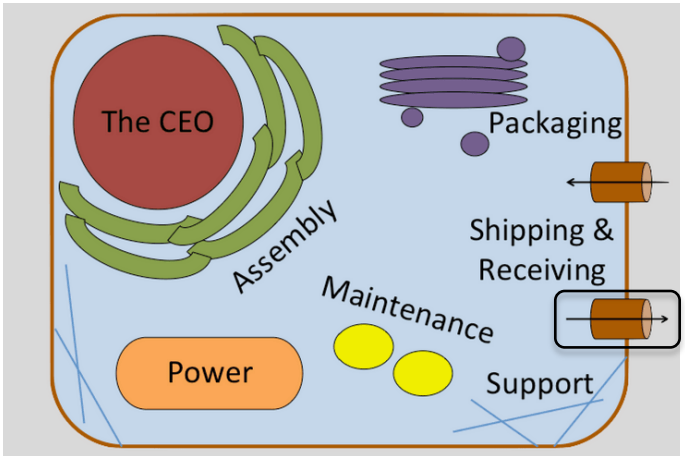
Goldenzweig *Annu Rev Biochem* (2018)

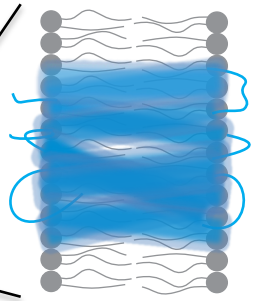
Membrane protein challenges: metastability



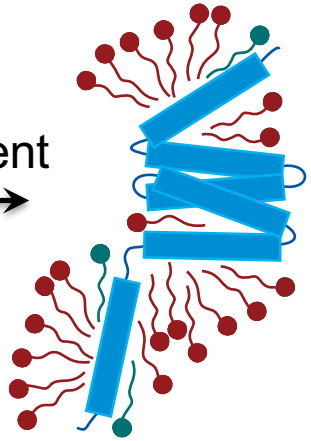


Membrane protein stabilization by design



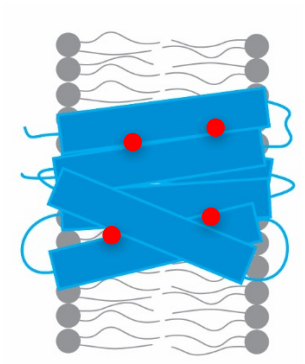


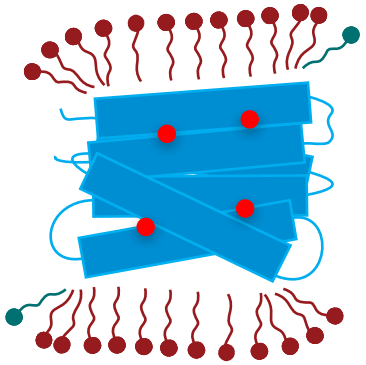
Detergent

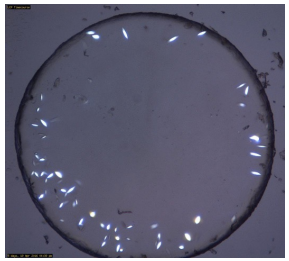


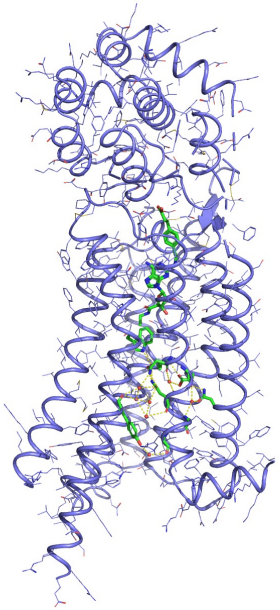
Unfolding

In silico
design







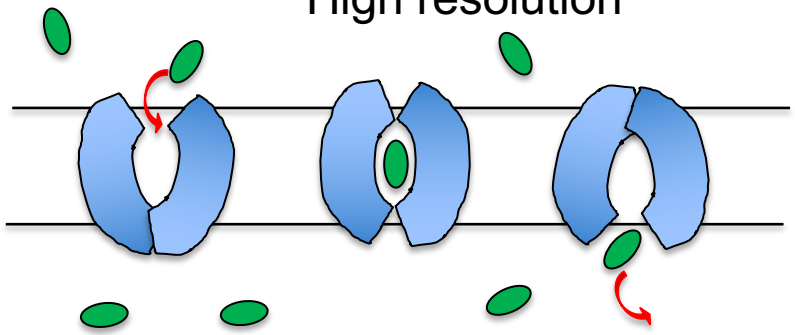


● mutation

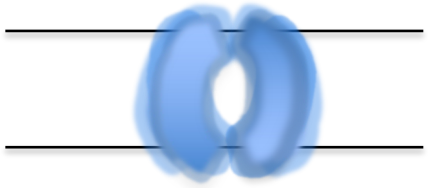
(PNAS 2012; Nature 2020)

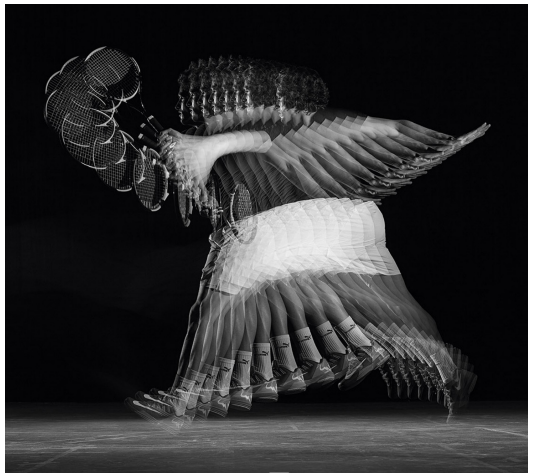
Membrane protein challenges: motions

High resolution

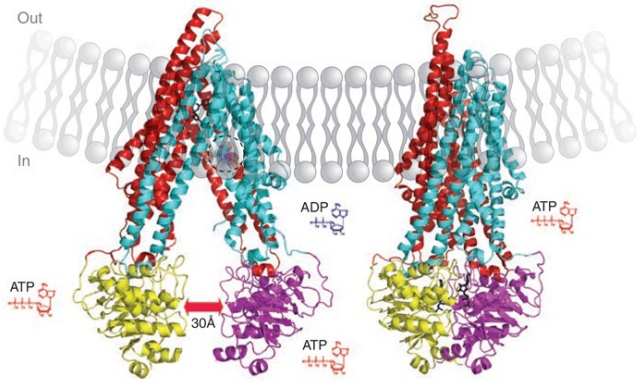


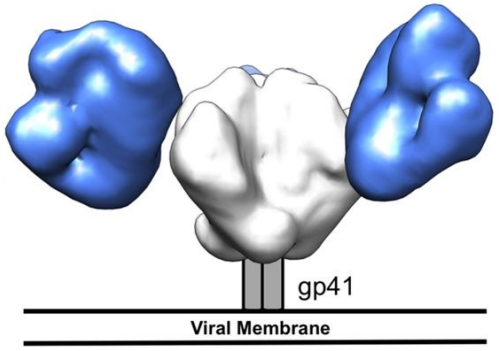
Low resolution



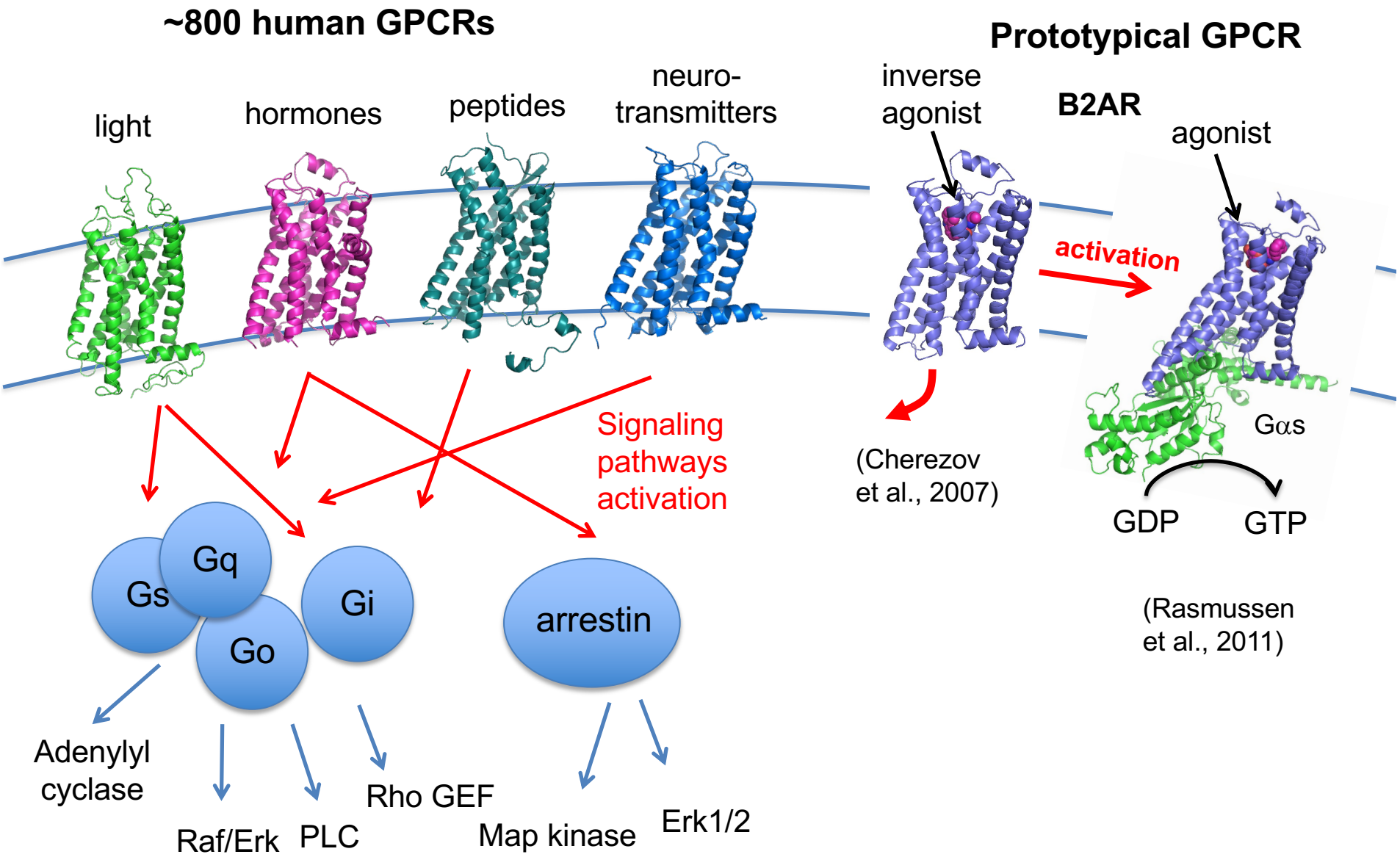


J-Y Lemoigne

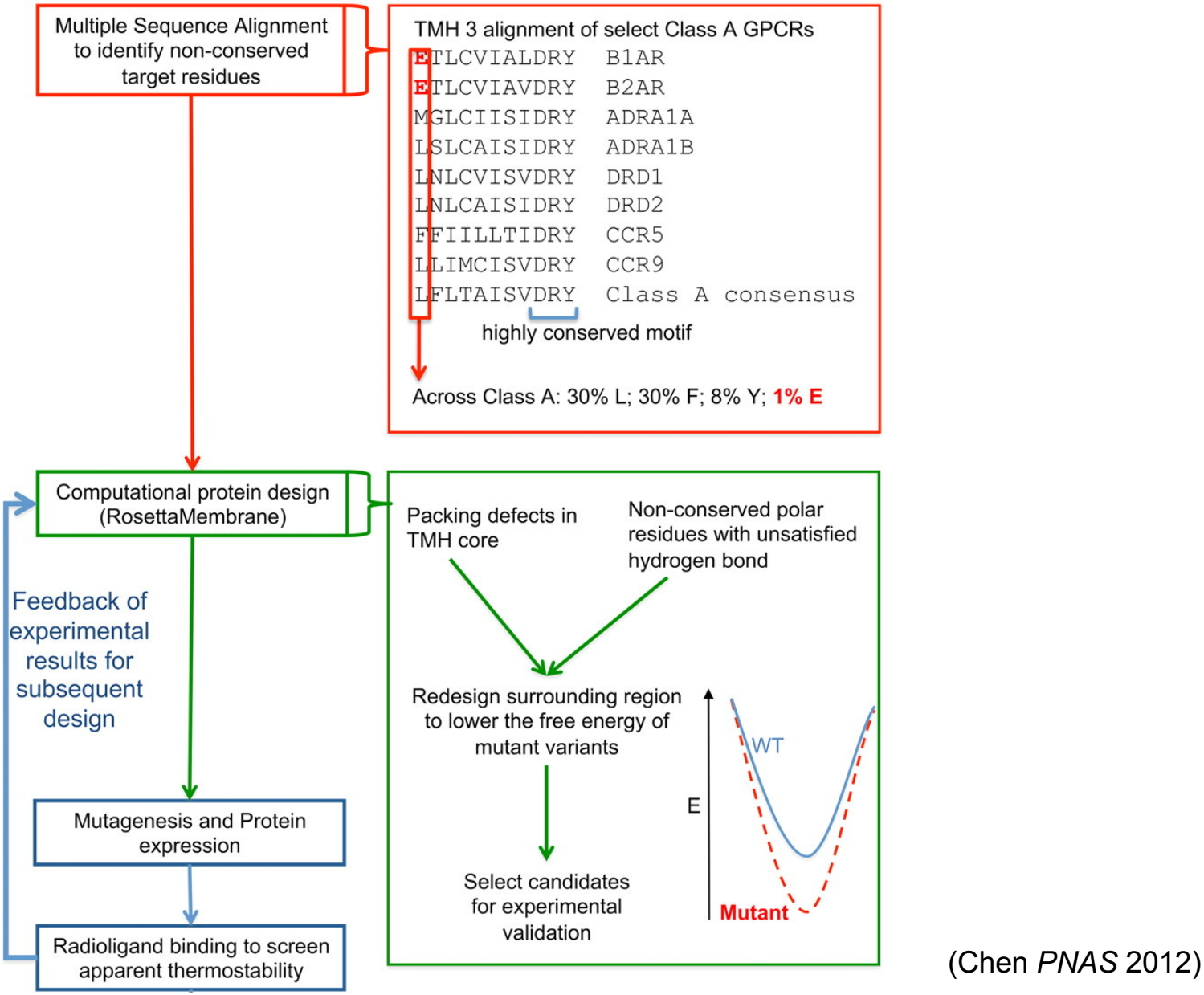




G protein coupled receptors are largest family of signaling receptors and drug targets

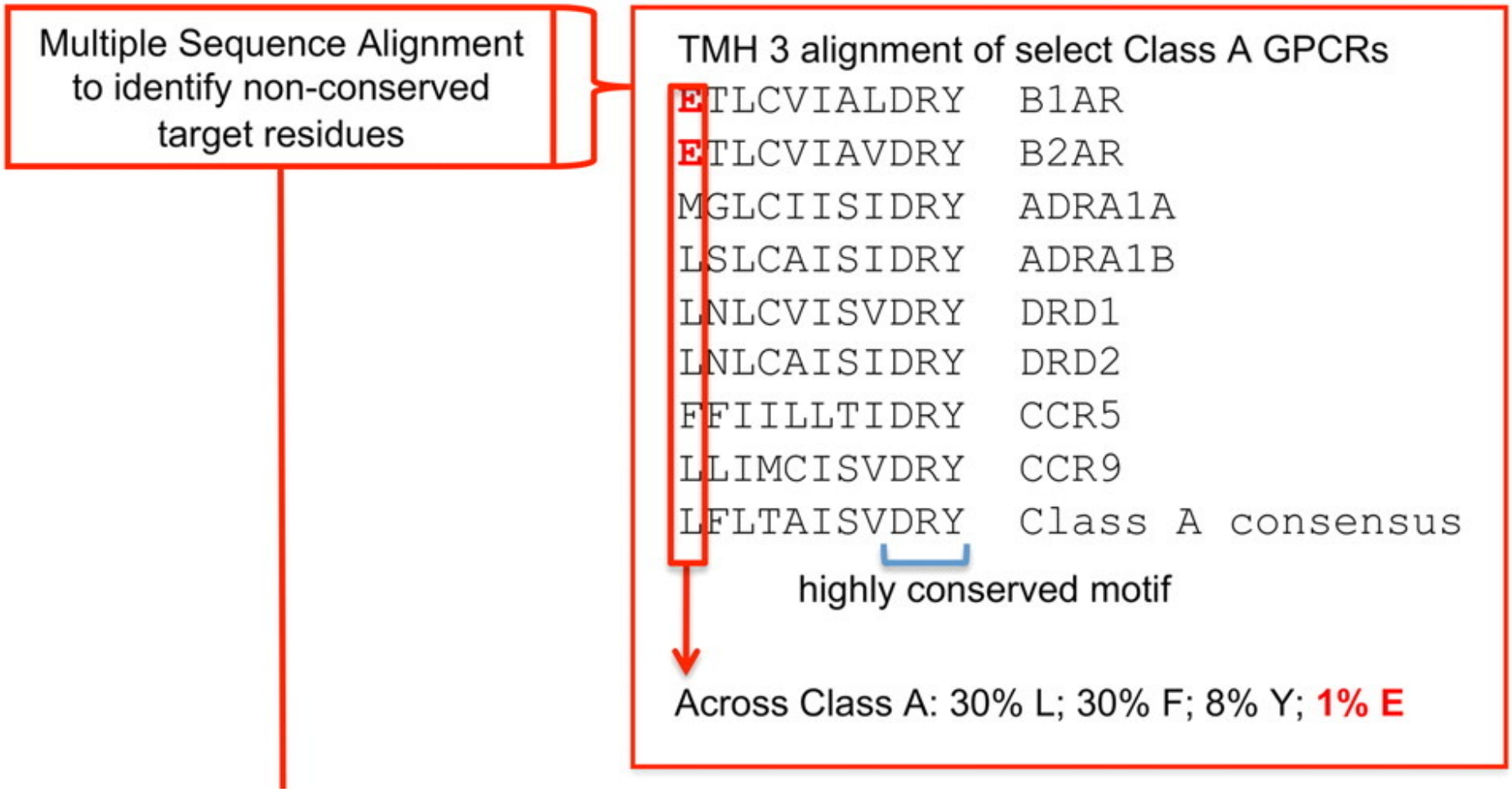


Integrated computational / experimental approach to design stabilized membrane proteins



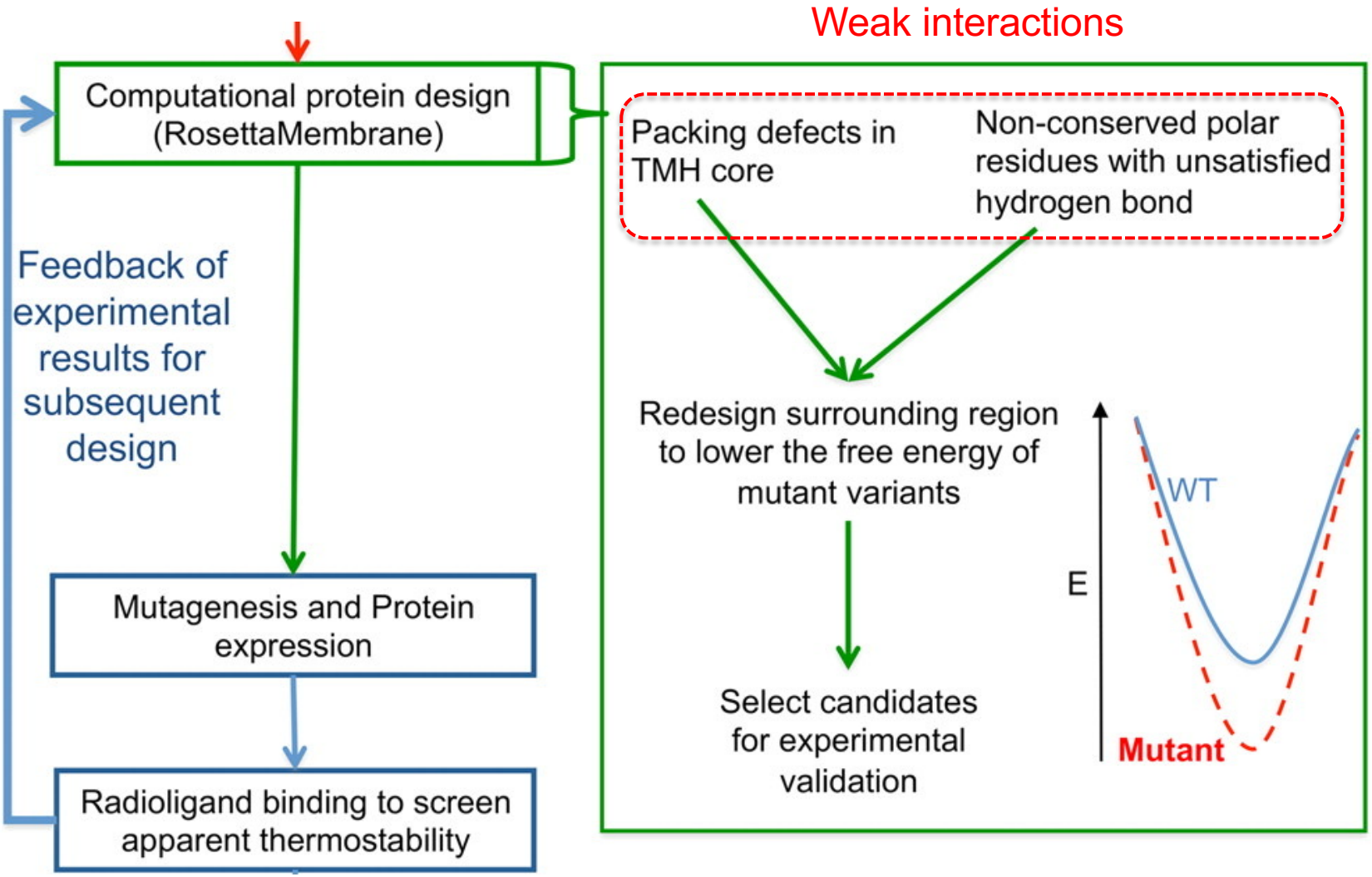
(Chen *PNAS* 2012)

Integrated computational / experimental approach to design stabilized membrane proteins



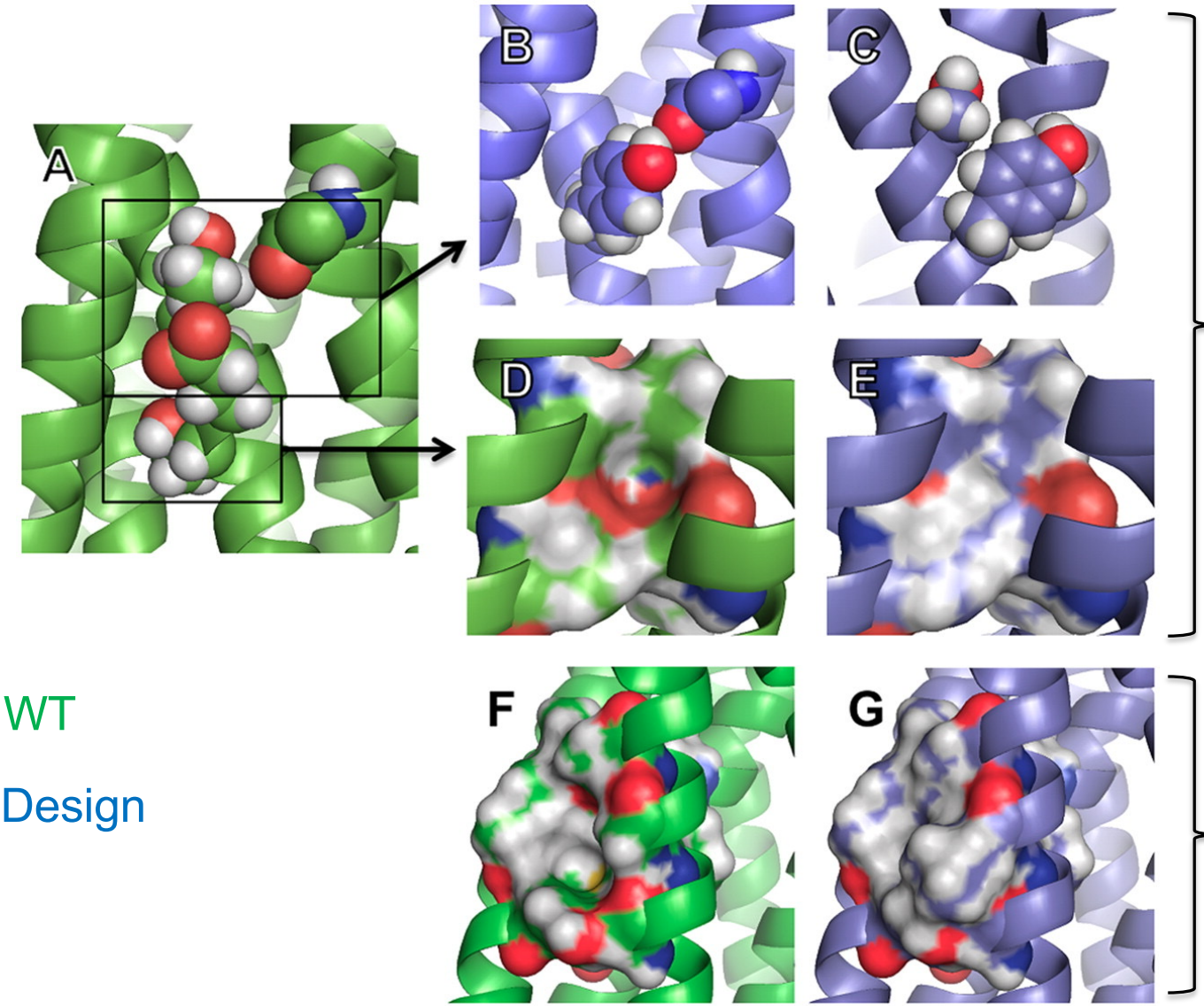
(Chen *PNAS* 2012)

Integrated computational / experimental approach to design stabilized membrane proteins



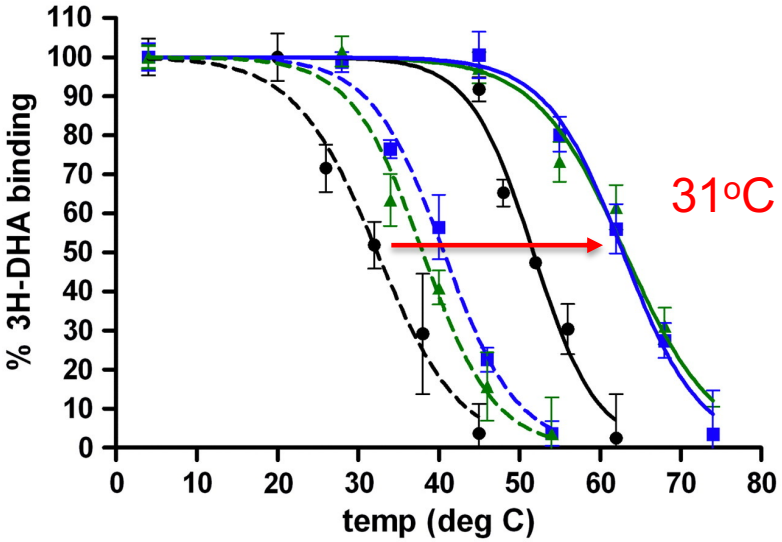
(Chen *PNAS* 2012)

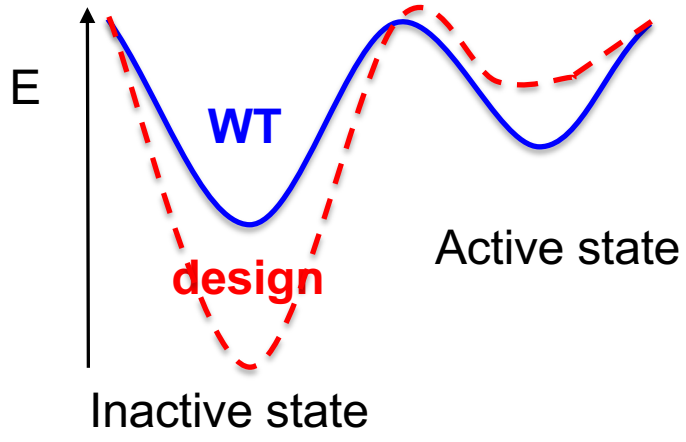
Stabilizing Designs Target Nonconserved Polar Residues and Packing Defects

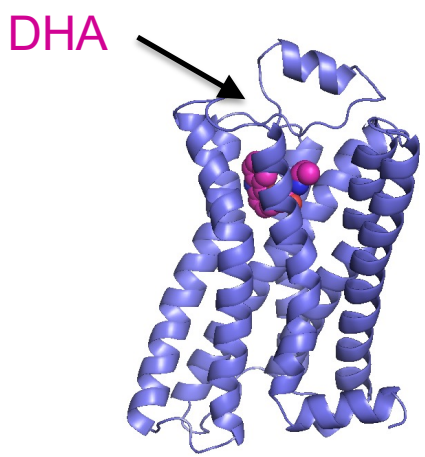


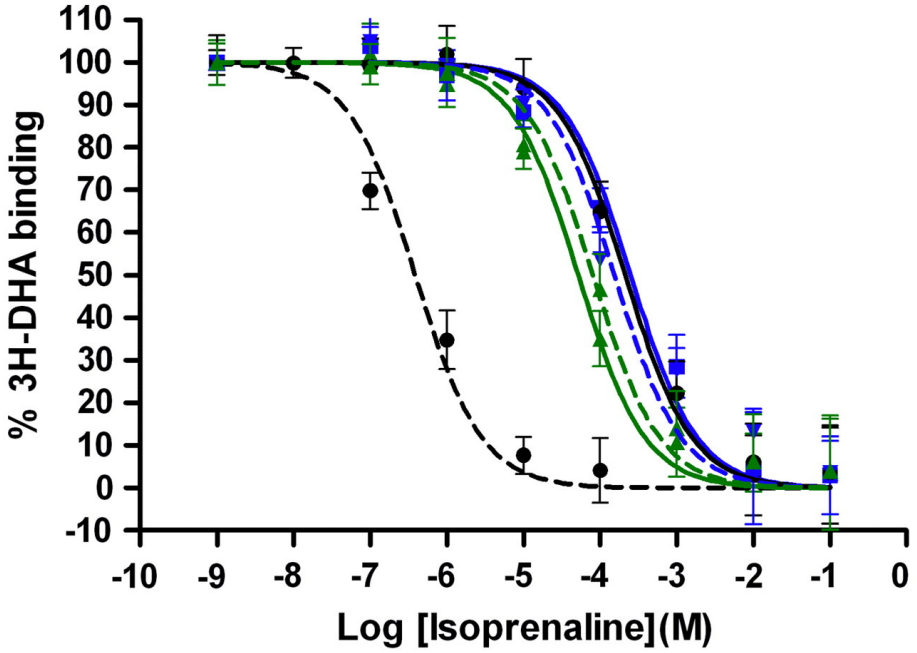
(Chen *PNAS* 2012)

Designed GPCR stabilized and specifically locked in the ligand bound state









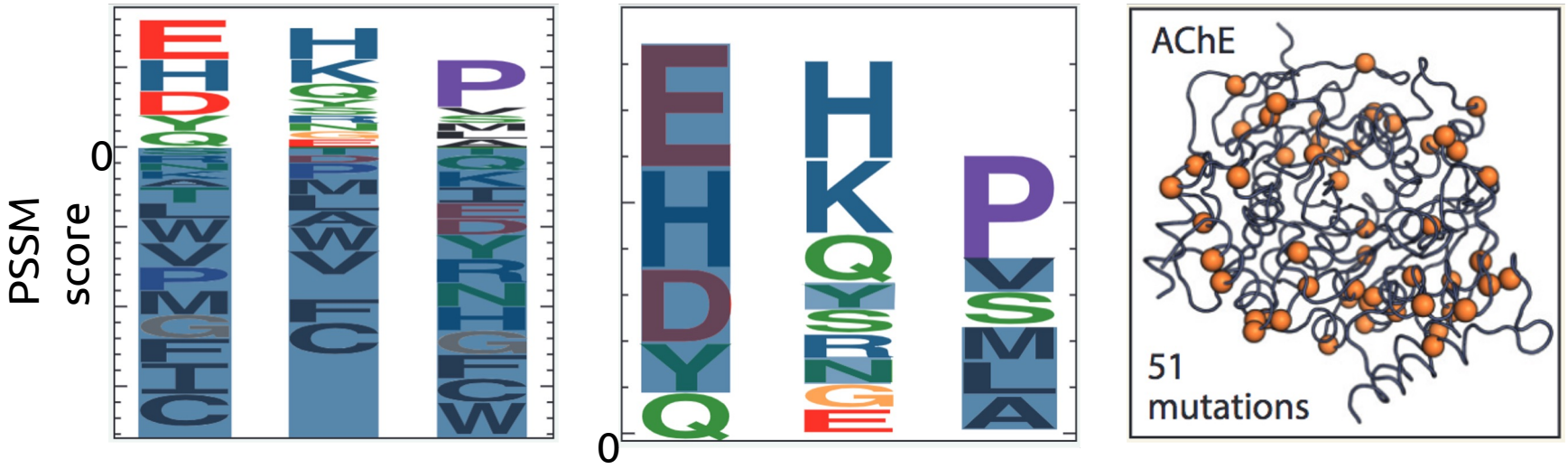
(Chen *PNAS* 2012)

The Protein Repair One-stop Shop (PROSS) algorithm

Step 1
Conservation analysis

Step 2
Atomistic modeling

Step 3
Combinatorial design

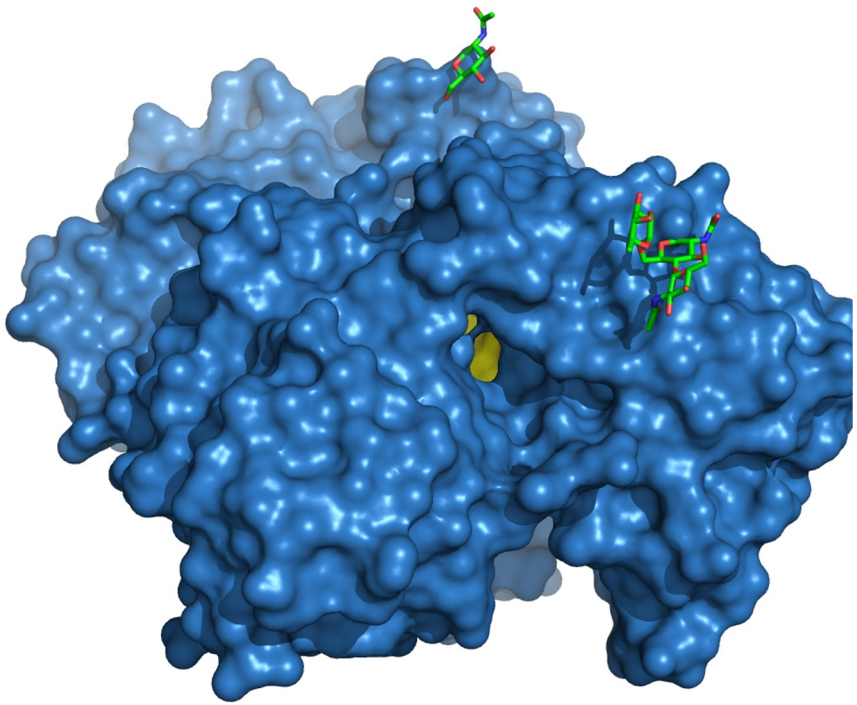


Eliminated identities

Goldenzweig *Mol Cell* 2016

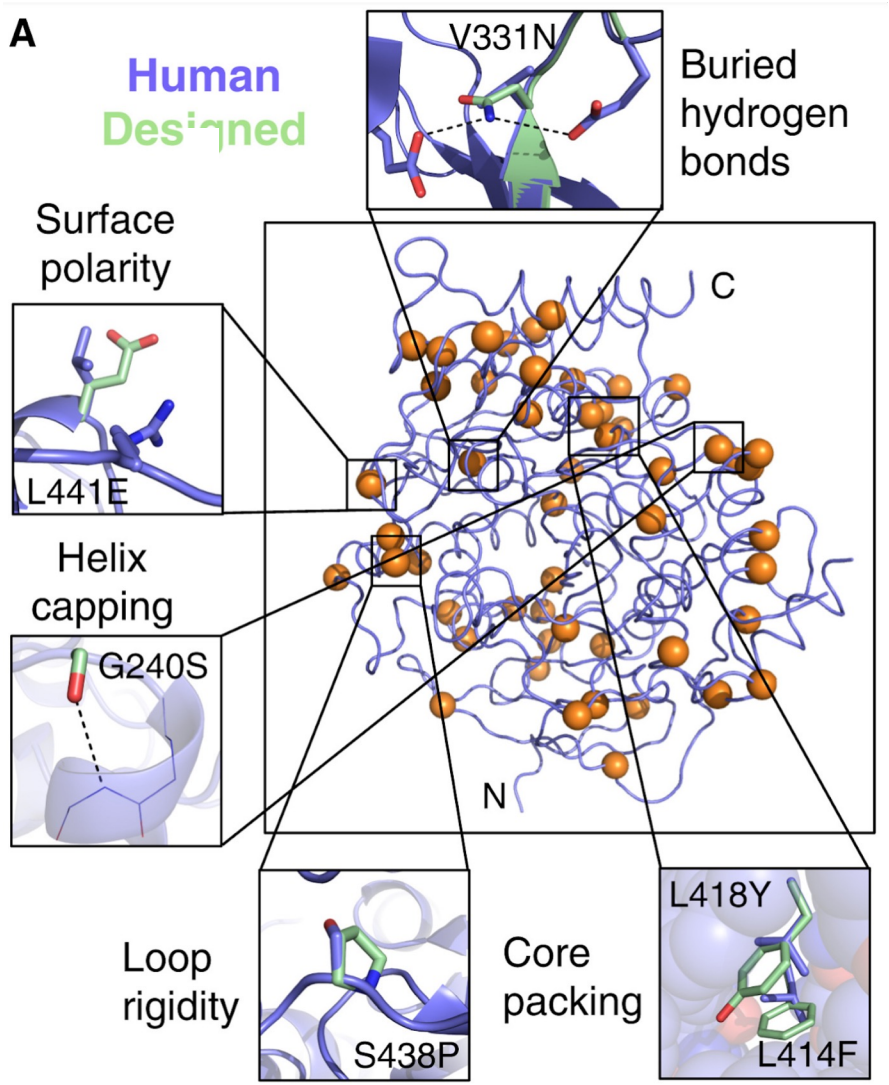
hAChE is an essential & very challenging enzyme

- Essential role in neuromuscular junctions
- Target of nerve agents
- >500 aa
- Multiple disulphides & glycosylation sites
- Active site buried 20 Å from surface

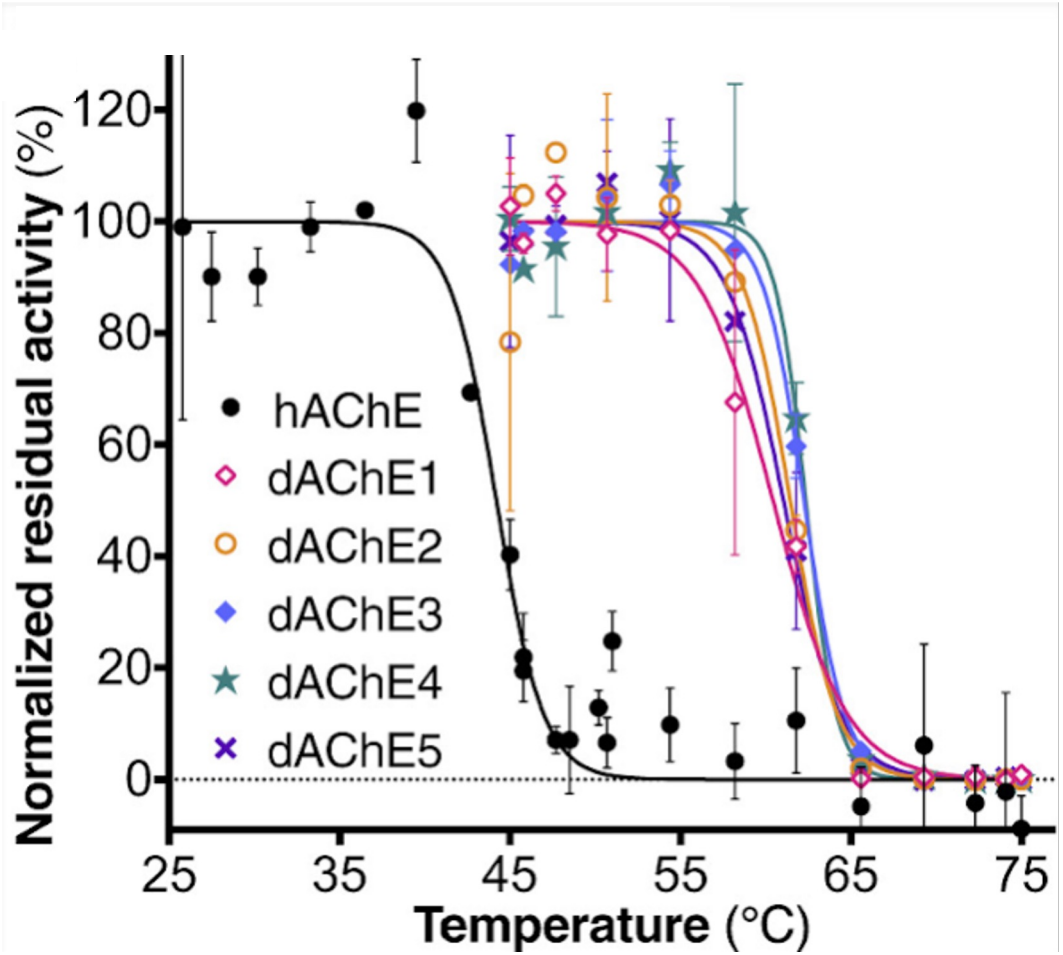


Active site
Glycosylation sites
Disulphides

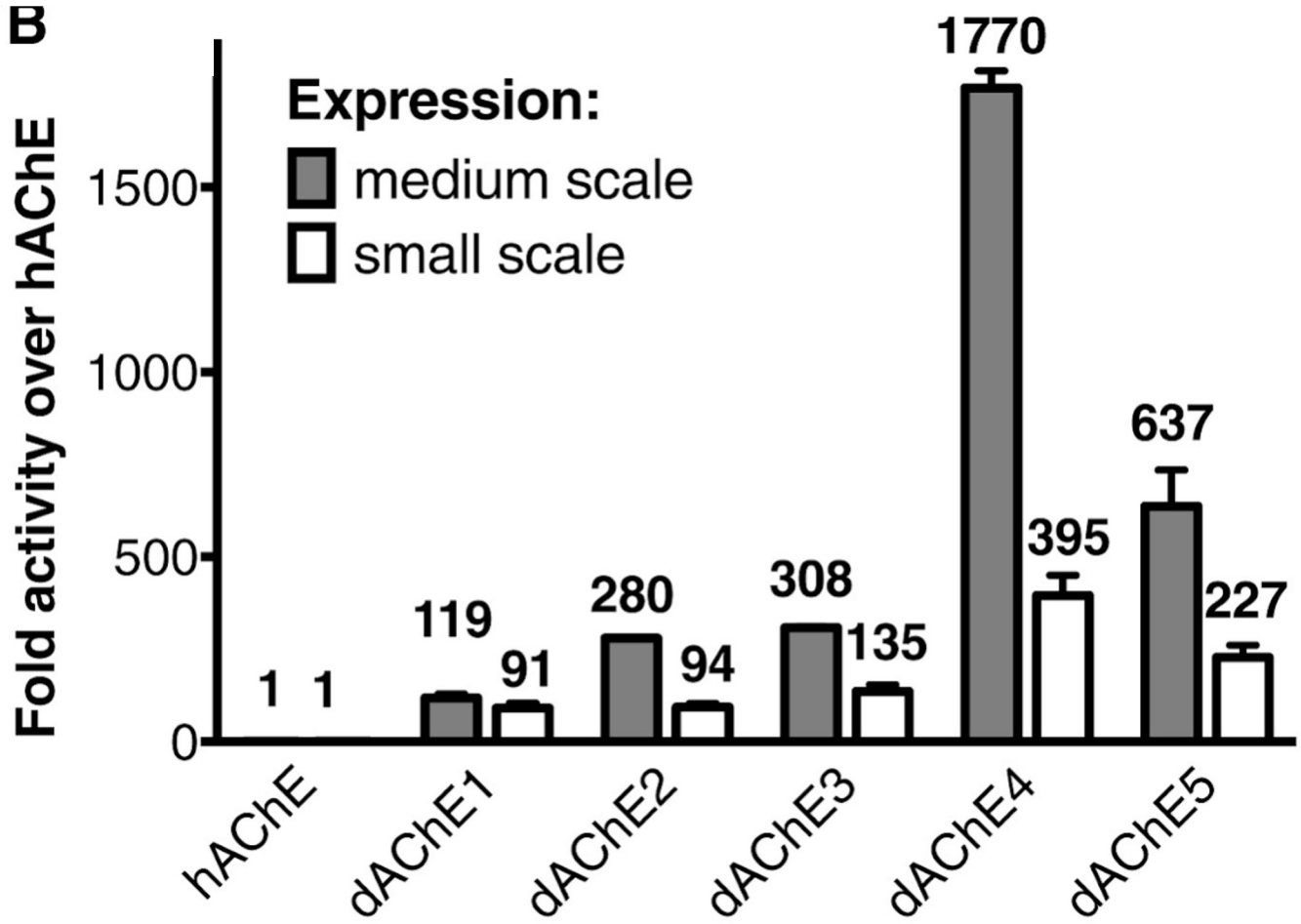
Best design: 51 simultaneous mutations relative to hAChE



All designs are fully functional & more stable



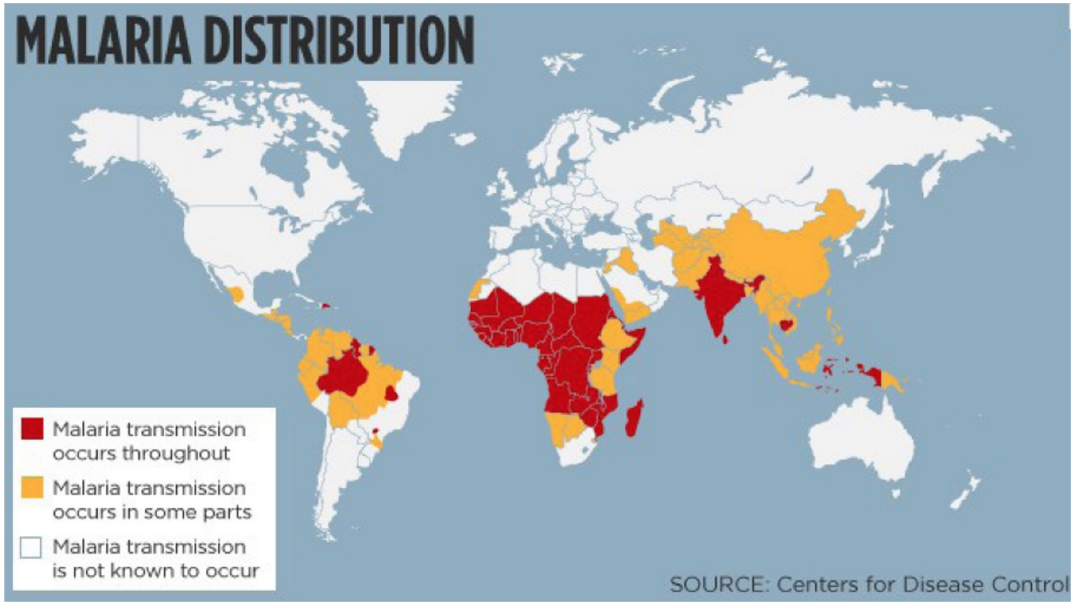
Dramatic improvement in bacterial expression levels



Collaboration with Dan Tawfik (Weizmann)

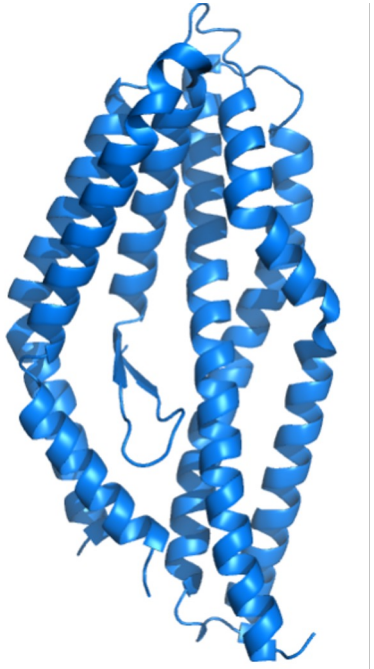
Malaria is the most virulent parasitic disease; no effective vaccine

- >3 billion people at risk
- >200 million clinical cases per year
- ~500,000 deaths per year, mostly of children



*Pf*RH5: The prime vaccine candidate for the blood-stage

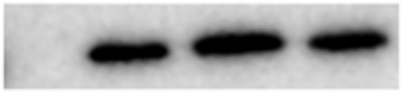
- Challenging to develop:
- Unstable & requires expensive insect-cell expression
- Vaccine requirements: cost-effective microbial expression; stability > 40°C



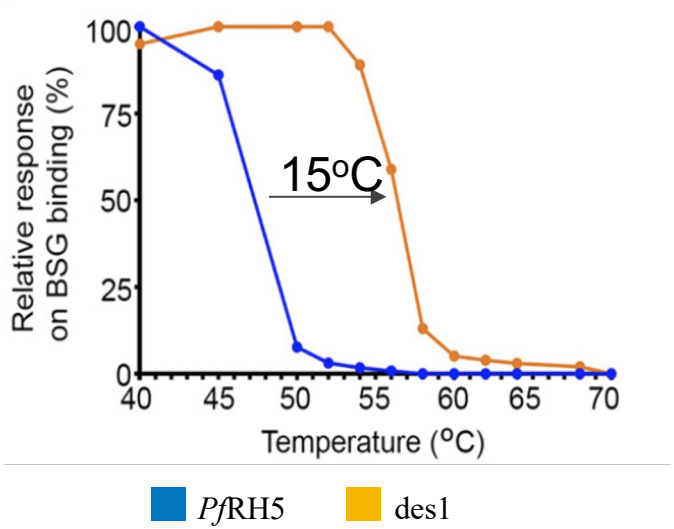
Wright..Higgins *Nature* 2015

Design is efficiently produced in bacteria & functionally identical to *Pf*RH5

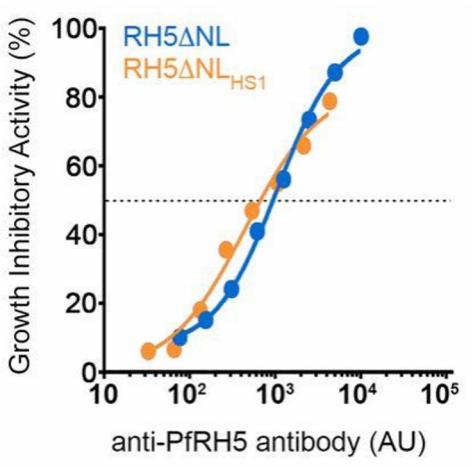
RH5 des1 des2 des3



Improved heat tolerance



Identical neutralising titers



Collaboration with Matthew Higgins & Simon Draper (Oxford)

Campeotto *PNAS* 2017

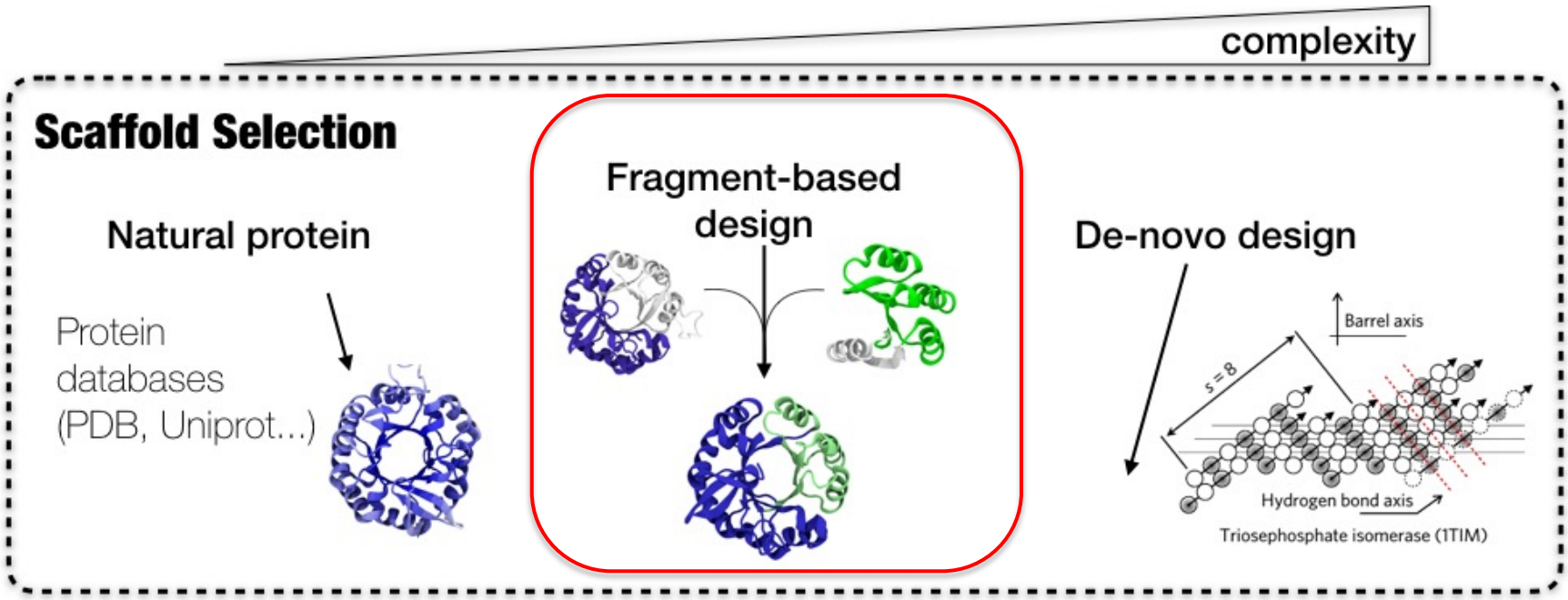
Protein Design – Examples overview

Protein design: Design a sequence that fits to a given structure

1. Design protein stability (membrane proteins)

2. Design new protein folds (protein chimera; de novo design; ANNs)

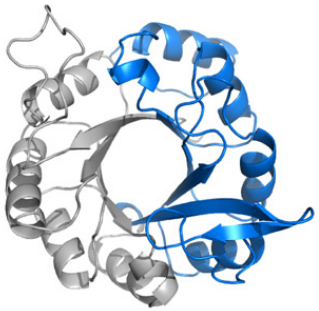
Protein fold design approaches



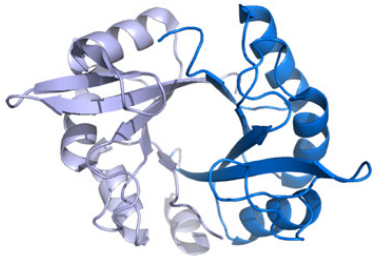
Lechner, Ferruz, Höcker (2018) *Curr Opin Struc Biol*

Design by mimicking natural evolution of proteins through duplication & recombination

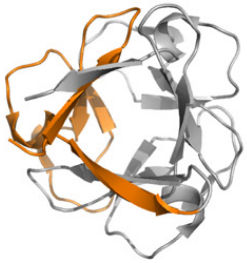
Duplication



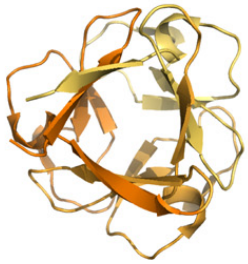
duplicated & optimized



Höcker *et al.* (2001) *Nat Struc Biol*
Höcker *et al.* (2009) *Biochemistry*

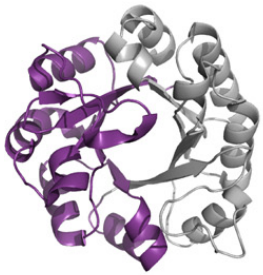


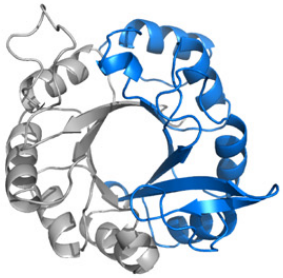
consensus triplicate



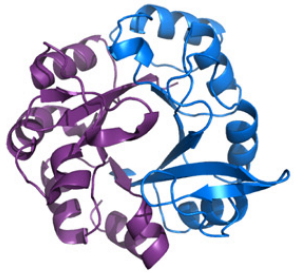
Broom *et al.* (2012) *Structure*
Lee & Blaber (2011) *JMB*

Recombination within one fold





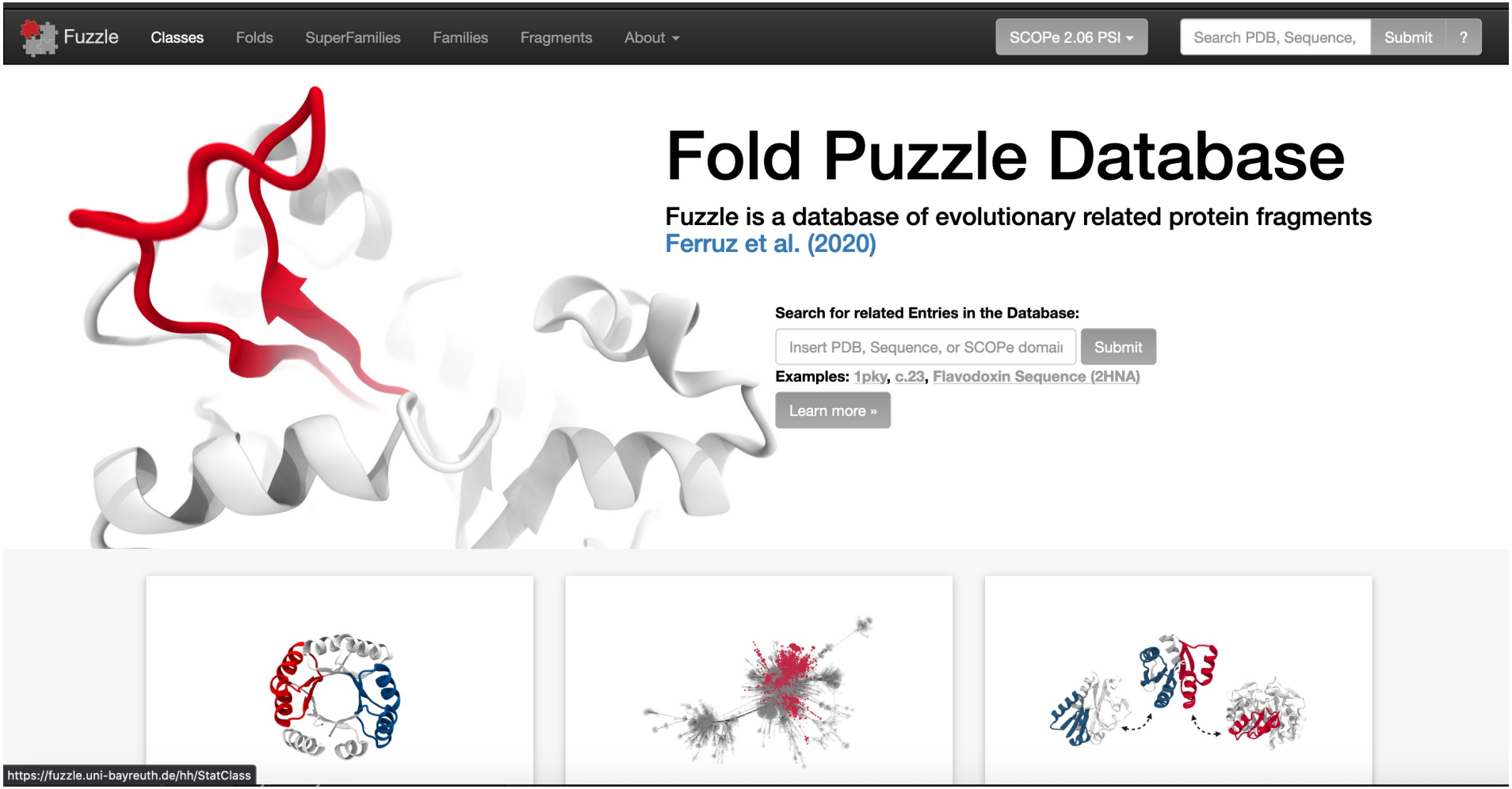
recombined & functionalized



Höcker *et al.* (2004) *PNAS*
Claren *et al.* (2009) *PNAS*

Höcker (2014) *Curr Opin Struc Biol*

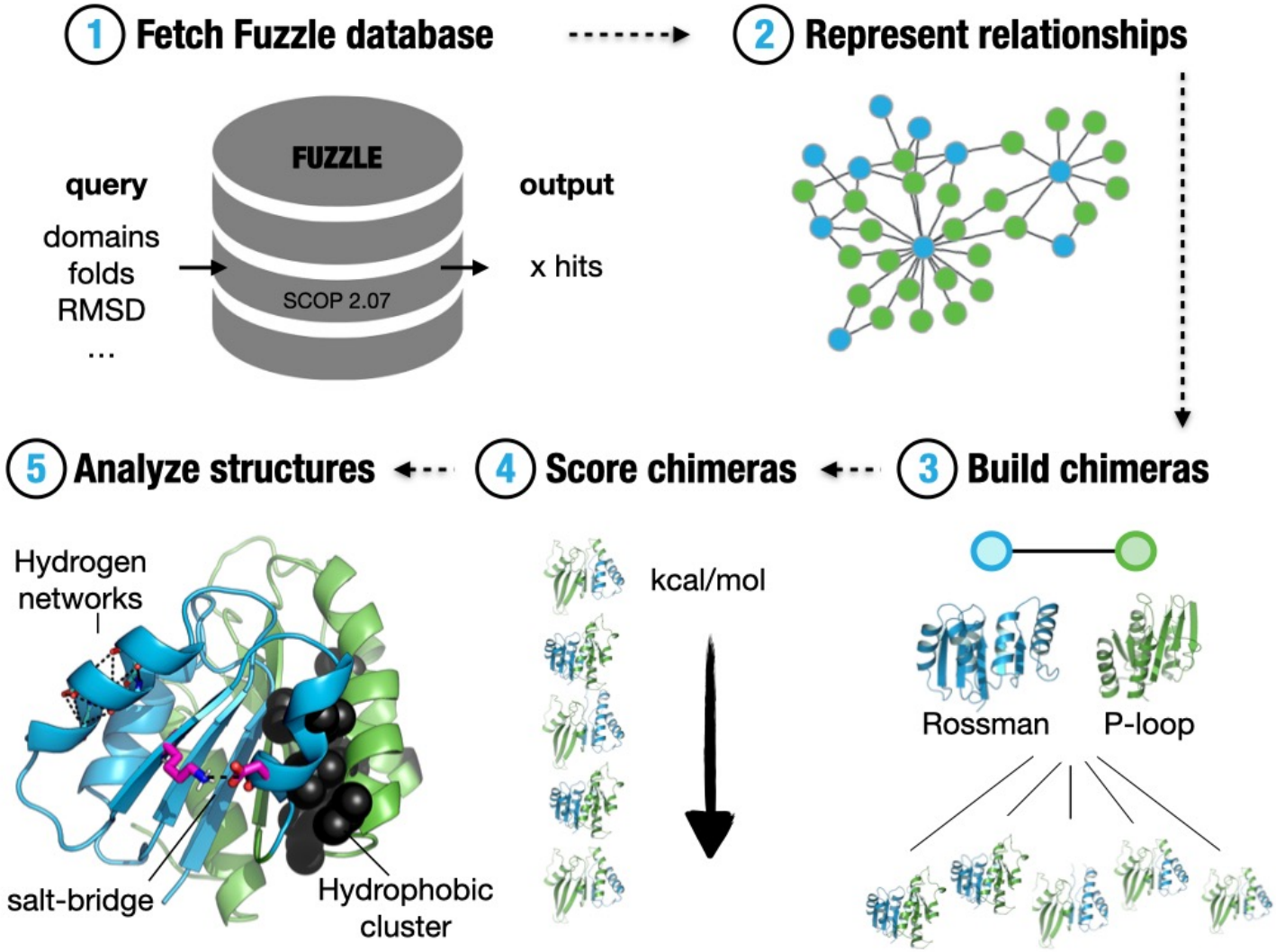
https://fuzzle.uni-bayreuth.de/



Ferruz, Lobos, Lemm, Toledo-Patino, Farias-Rico, Schmidt, Höcker, (2020) *J Mol Biol 432*: 3898-914

PROTLEGO

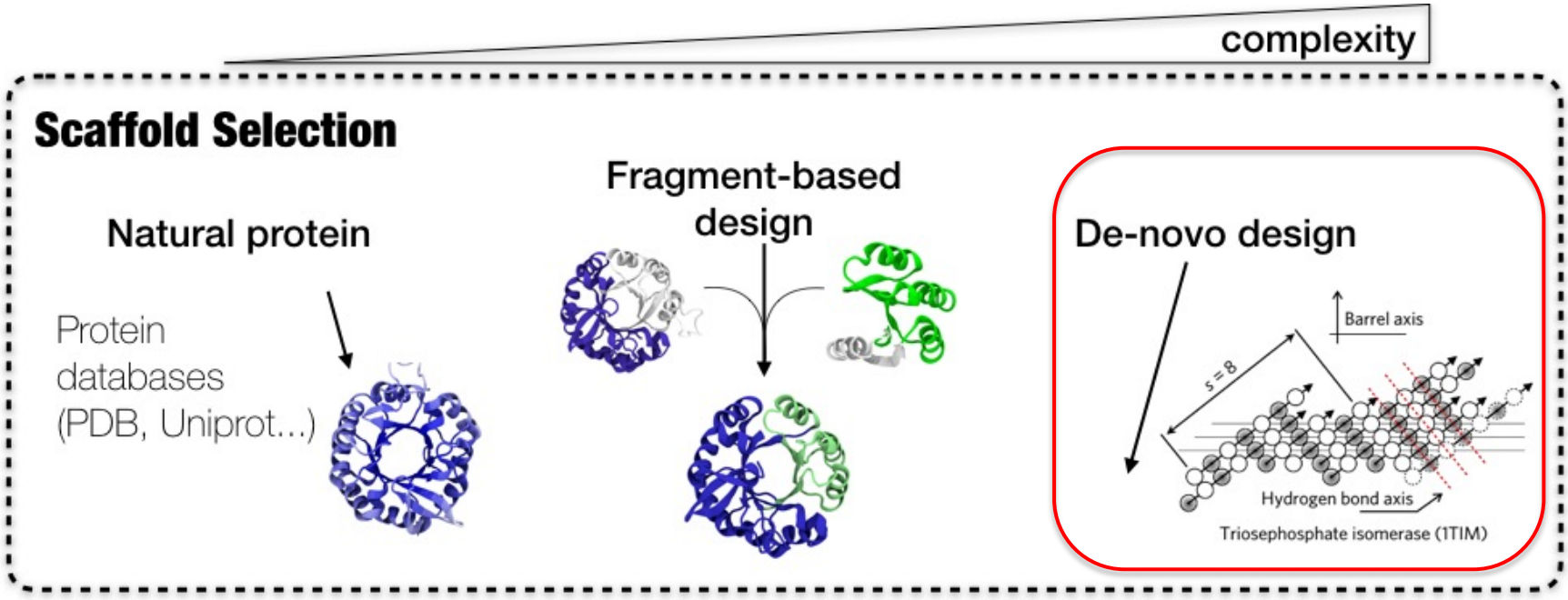
Design of protein chimeras with ProtLego



https://hoecker-lab.github.io/protlego/

Ferruz, Noske, Höcker, (2020) *bioRxiv* / in revision

Protein fold design approaches



Lechner, Ferruz, Höcker (2018) *Curr Opin Struc Biol*

TOP7 – Design of a new fold

Kuhlman, Dantas, … & Baker Science, 2003

1. Define new scaffold not observed in Nature

2. Find sequence that will fold into scaffold

Approach: Iterate between

Sequence design (with fixed backbone structure) and

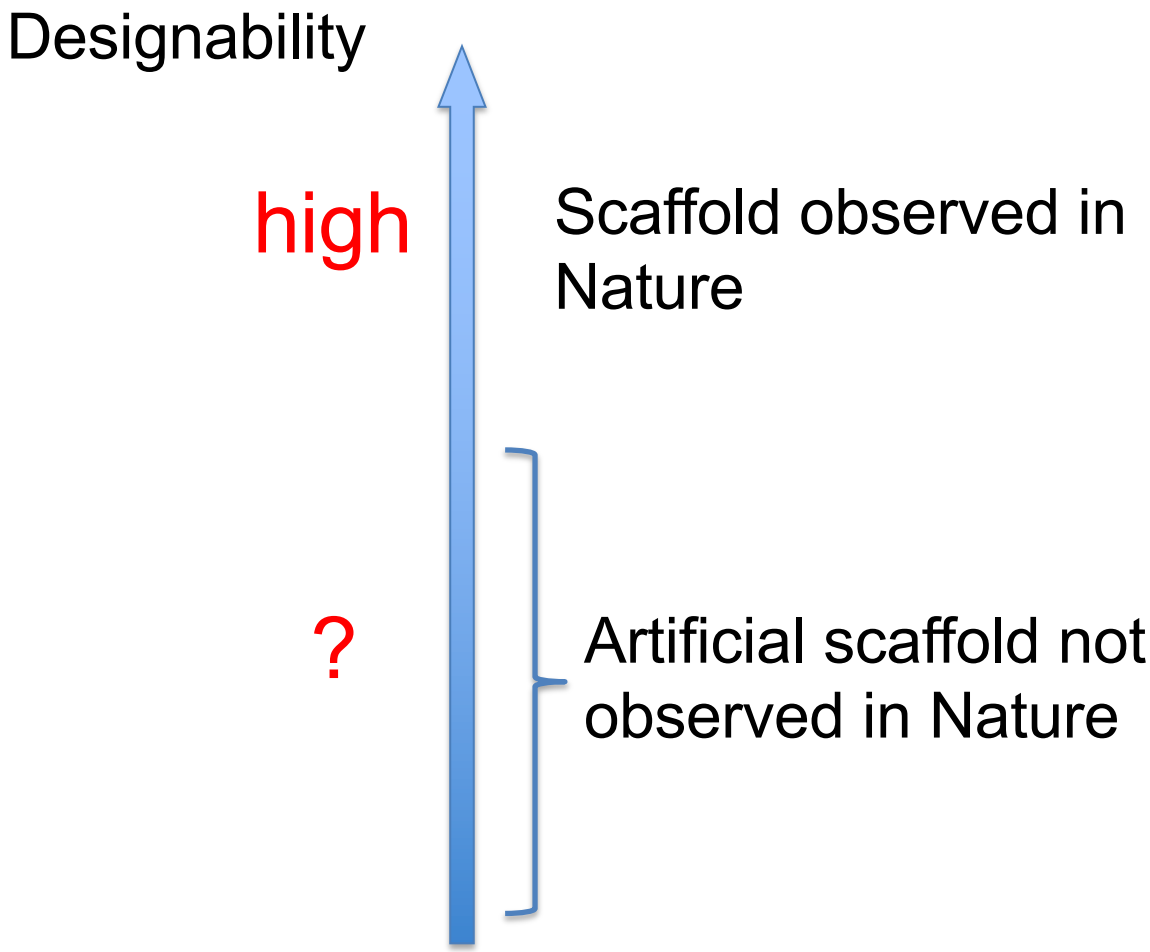
Structure prediction (with fixed sequence)

Why do we need a structure prediction step?

Because we are starting with a synthetic scaffold that is a very low resolution guess

Design of a new fold: the designability problem

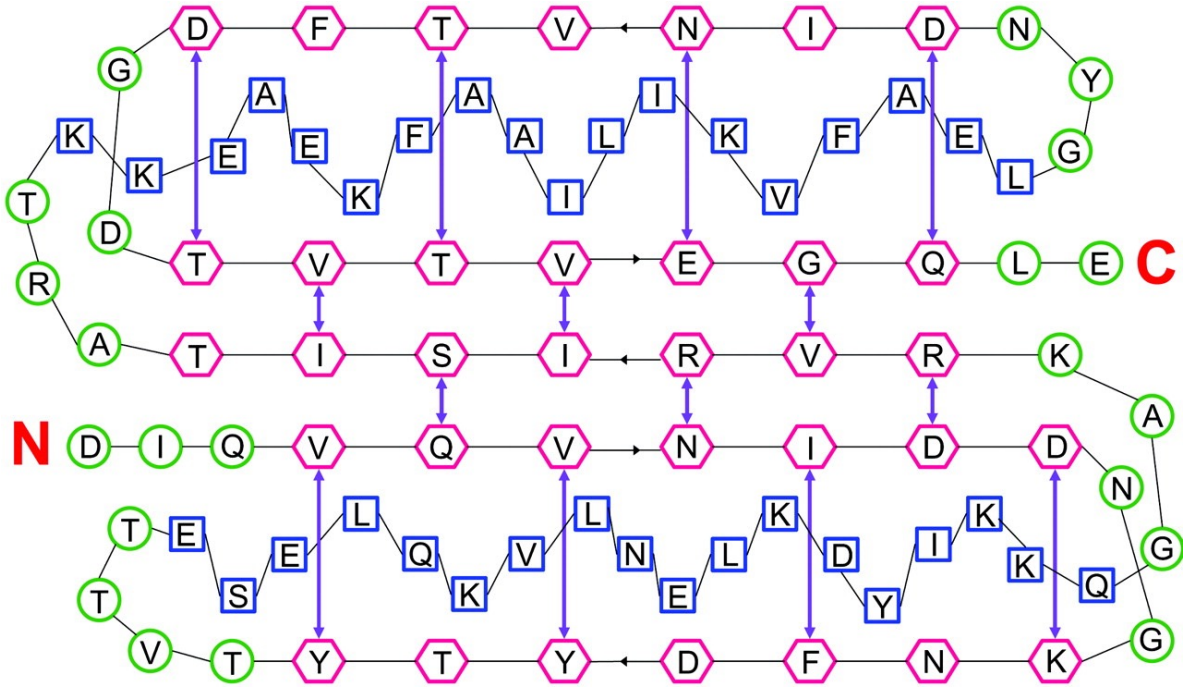
Designability: the probability to find a (# of) sequence folding into a specific scaffold



TOP7 – Design of a new fold

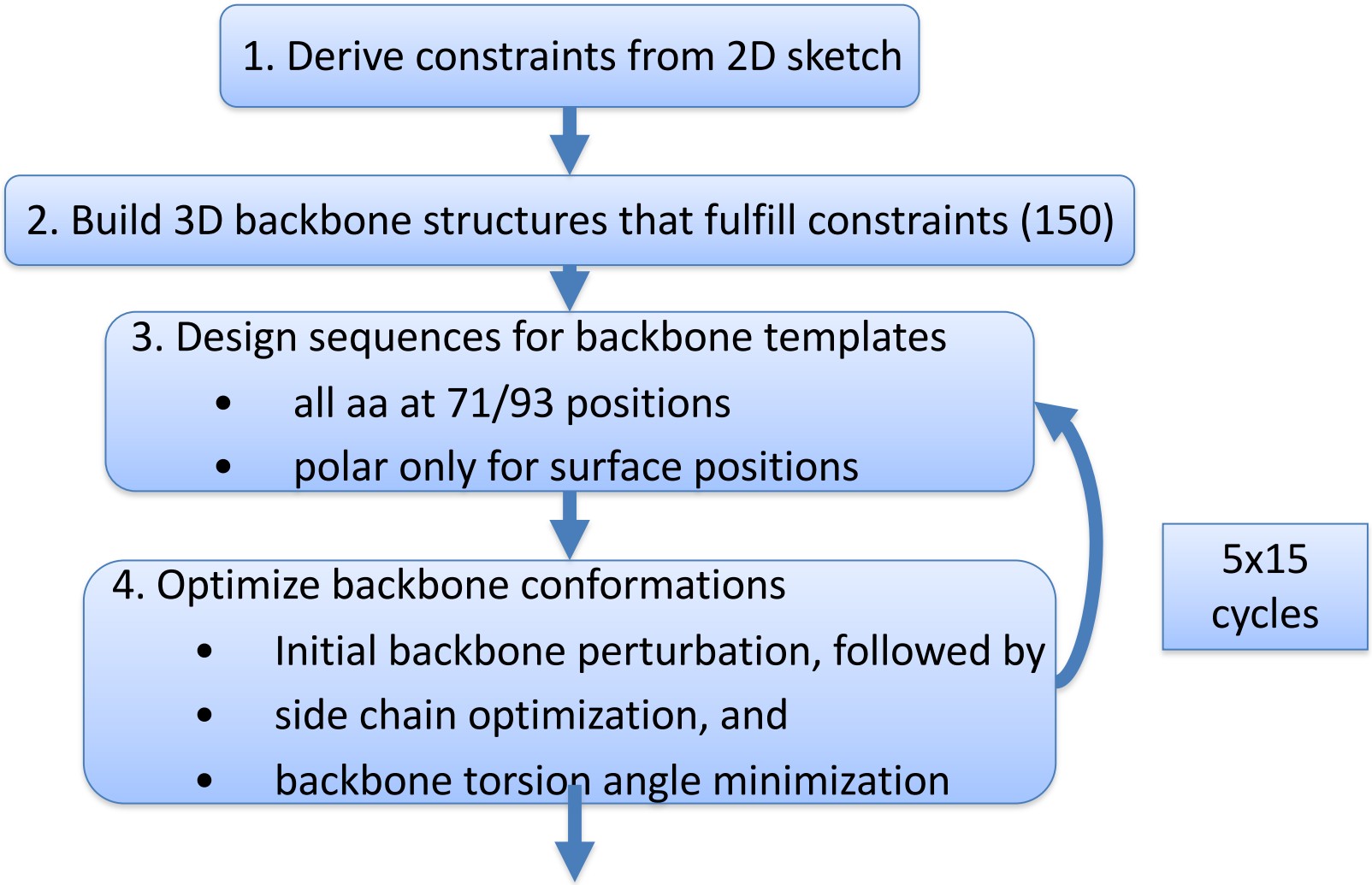
Kuhlman, Dantas, … & Baker Science, 2003

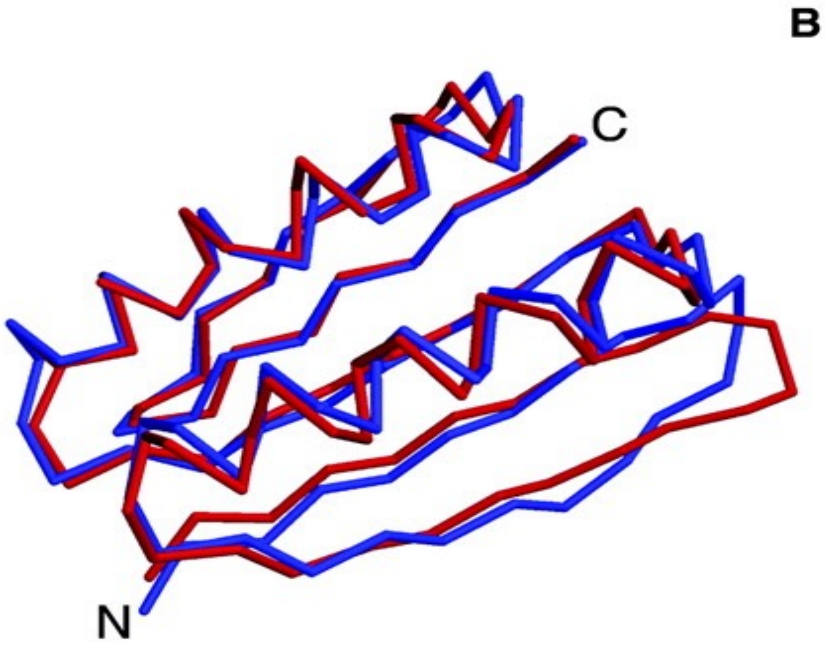
2D sketch of a novel fold

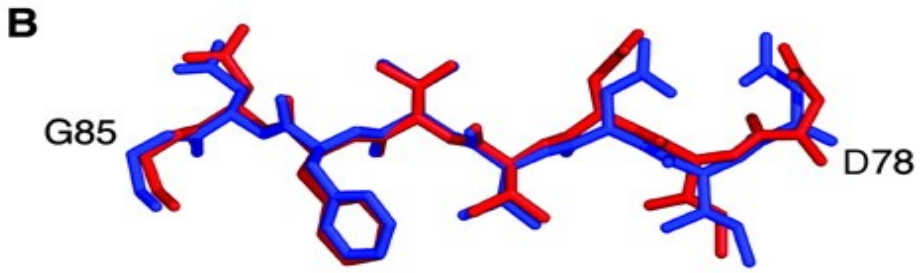


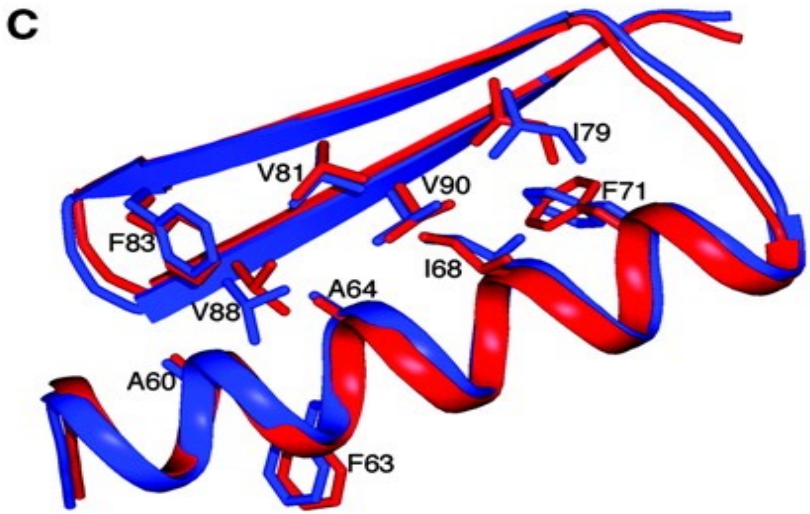
Legend: (hexagon, strand; square, helix; circle, other). Hydrogen bond partners are shown as purple arrows. The amino acids shown are those in the final designed (Top7) sequence

Creation scheme of TOP7









Blue: model; Red: xray

Assessment of Design

(1) Structure

- 1.17Å backbone rmsd
- highly accurate!

(2) Stability

- stable at 98^0C!
- stable at ~5M Gu-Hcl!

TOP7

- No sequence memory → more stringent test of force field and minimization procedure
- Optimized steric packing prevents molten globules
- No similarity to natural sequences (psiblast)

→ What can we learn from a protein that did not undergo natural selection??

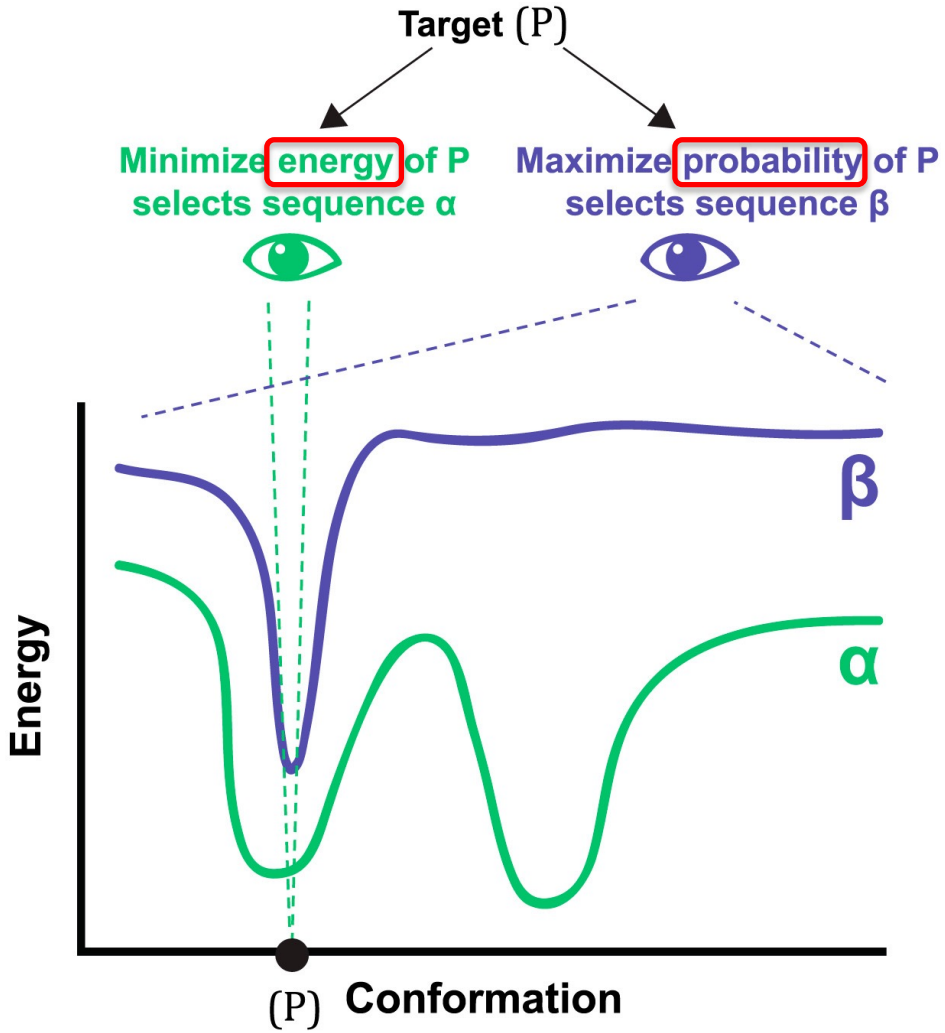
Protein Design – Examples overview

Protein design: Design a sequence that fits to a given structure

1. Design protein stability (membrane proteins)

2. Design new protein folds (protein chimera; de novo design; ANNs)

Protein sequence design by conformational landscape optimization to prevent alternative conformations



Using trRosetta: predicts the probability of residue–residue distances (Q) and orientations for a sequence (X).

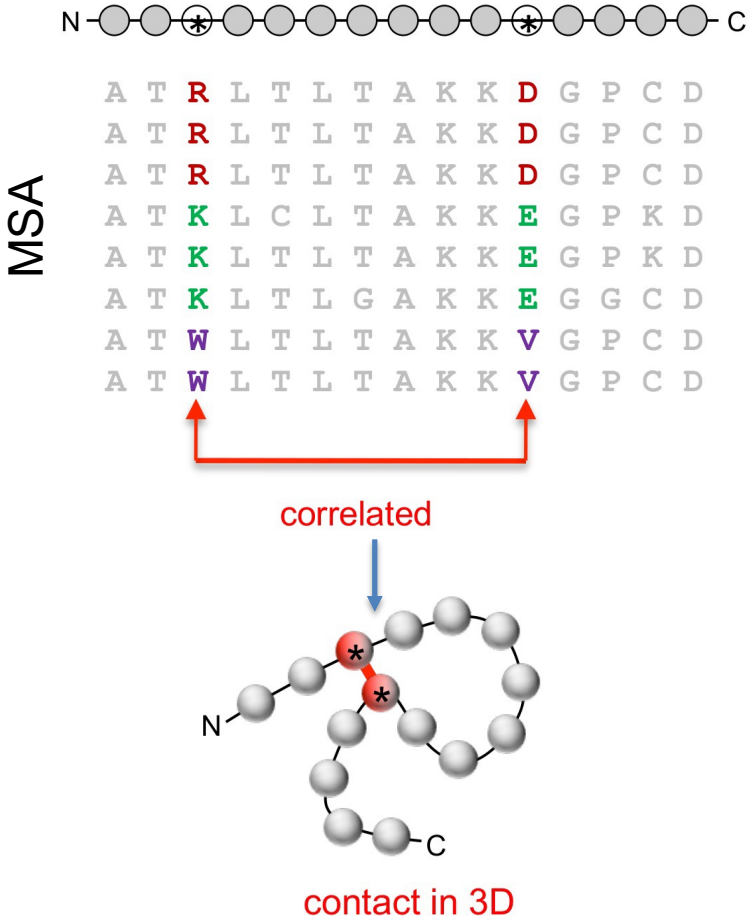
Rationale: Probability distributions over possible distances and orientations should contain information about alternative conformations

(Yang, PNAS 2020)

Convolutional neural network trRosetta: predicting interresidue geometries and protein 3D structure from a multiple sequence alignment

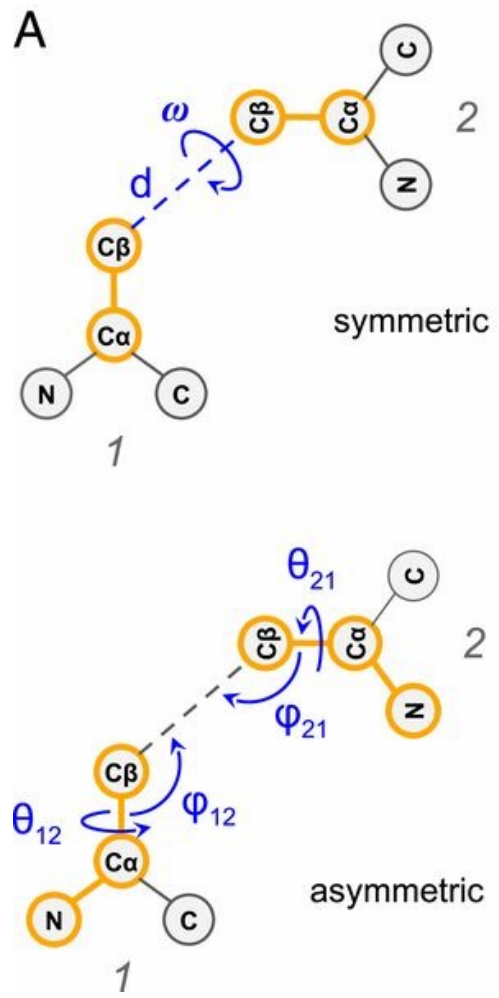
Protein sequence database:

Contact prediction from co-evolution



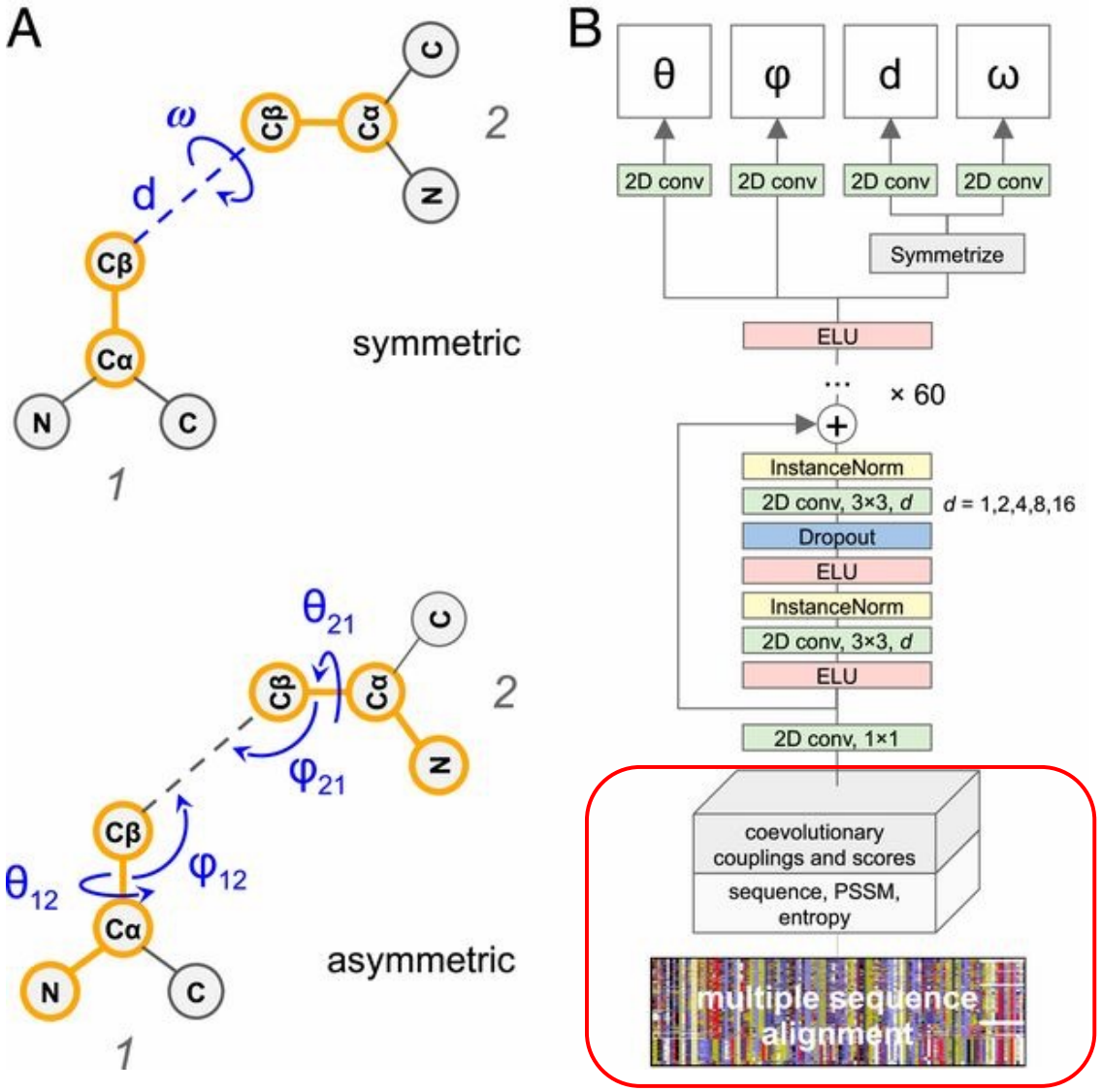
Protein structure database

?



d, ω, θ12, φ12, θ21, and φ21 fully define the relative positions of the backbone atoms of 2 residues

Convolutional neural network trRosetta: predicting interresidue geometries and protein 3D structure from a multiple sequence alignment

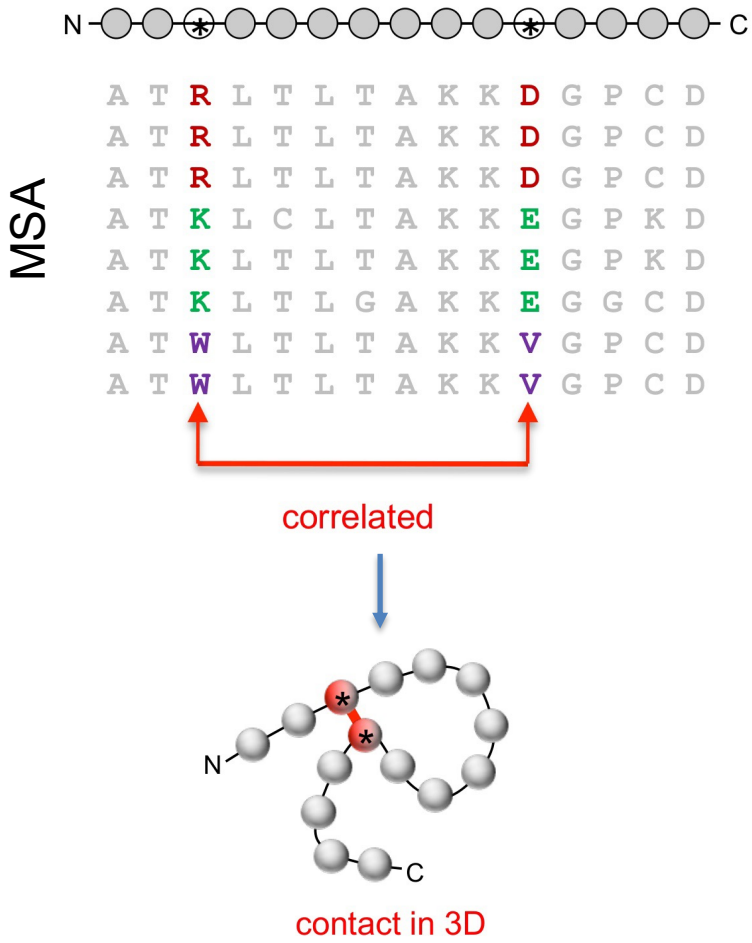


Can we train an ANN to learn distance and geometry from multiple sequence alignments?

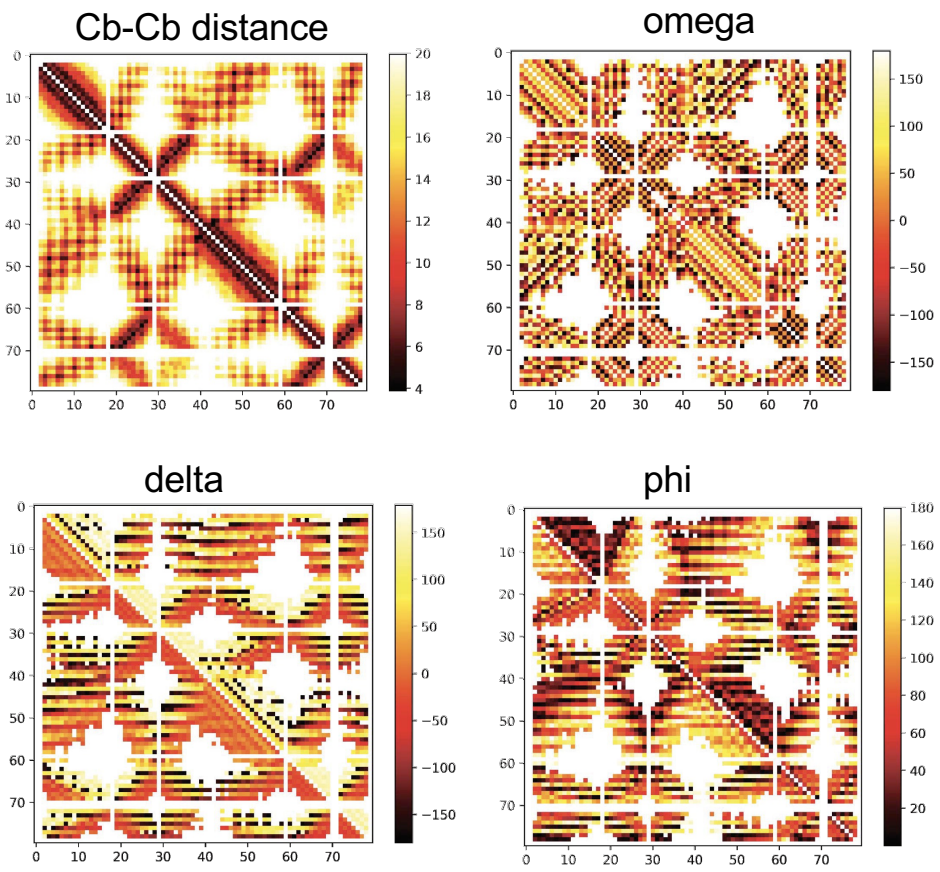
(Yang, PNAS 2020)

Convolutional neural network trRosetta: predicting interresidue geometries and protein 3D structure from a multiple sequence alignment

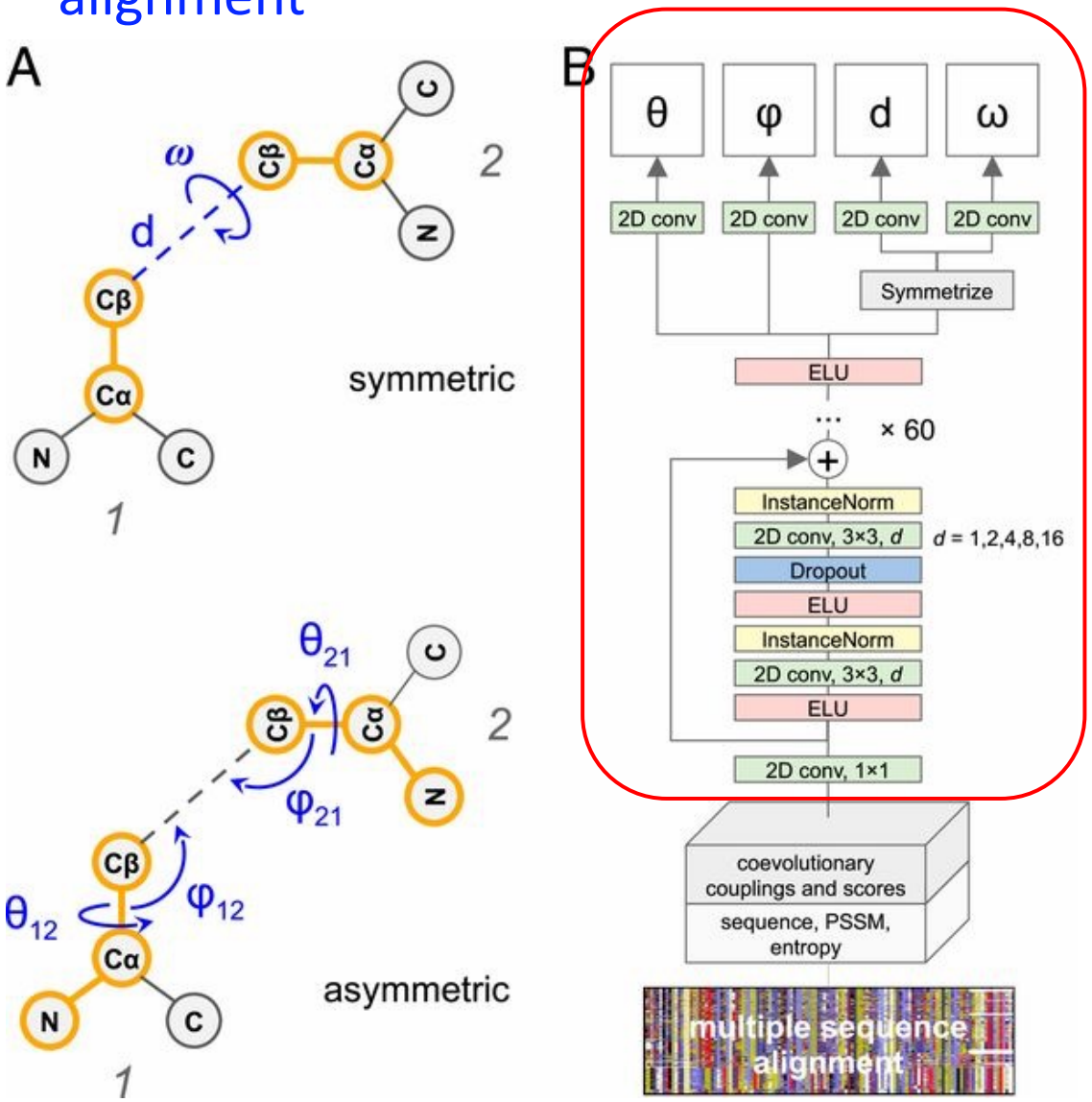
Contact prediction from co-evolution



All of the coordinates show characteristic patterns => ideal for training a deep neural network to predict them



Convolutional neural network trRosetta: predicting interresidue geometries and protein 3D structure from a multiple sequence alignment

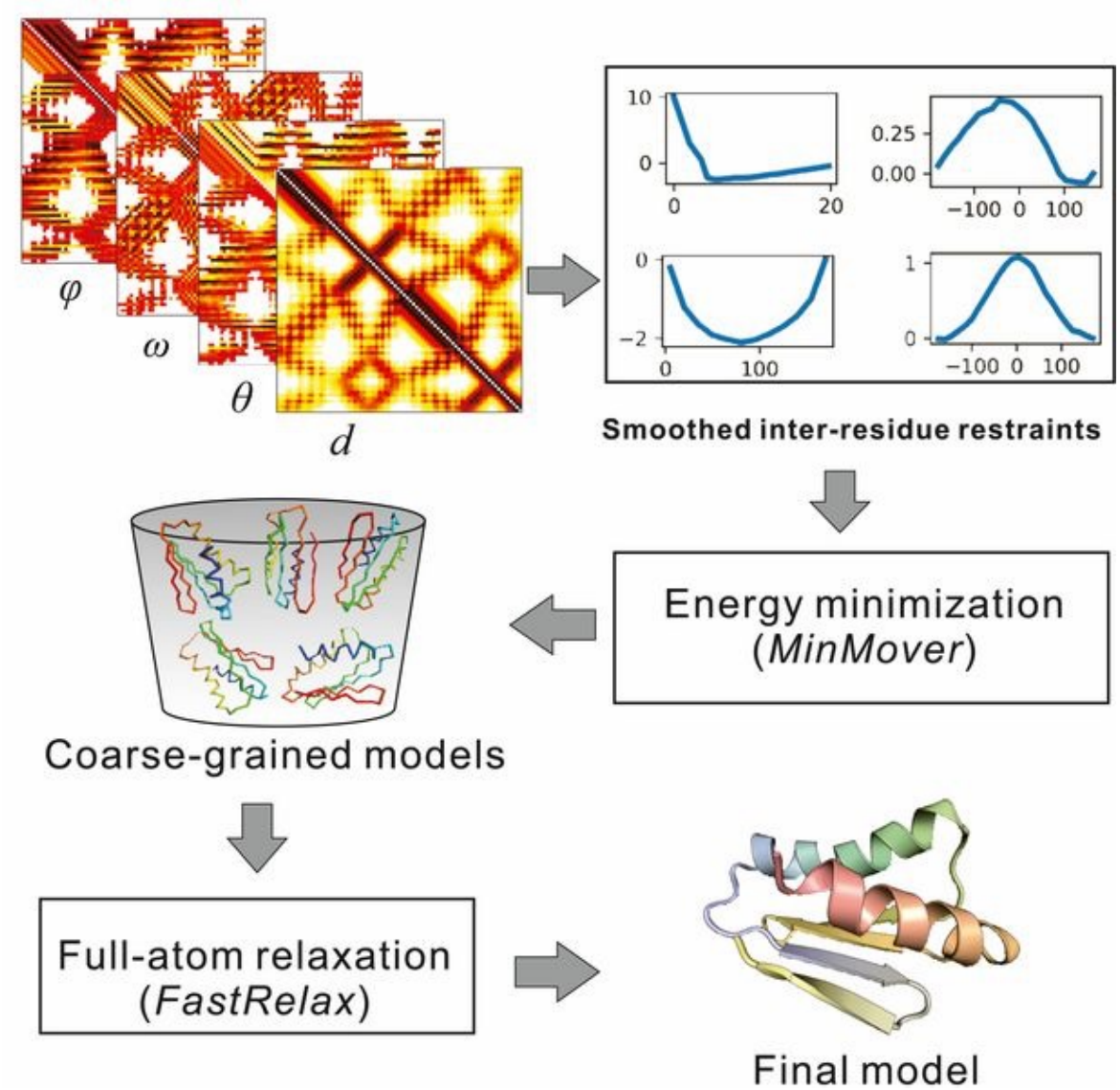


Stack of dilated residual-convolutional blocks that gradually transforms 1- and 2-site features derived from the MSA

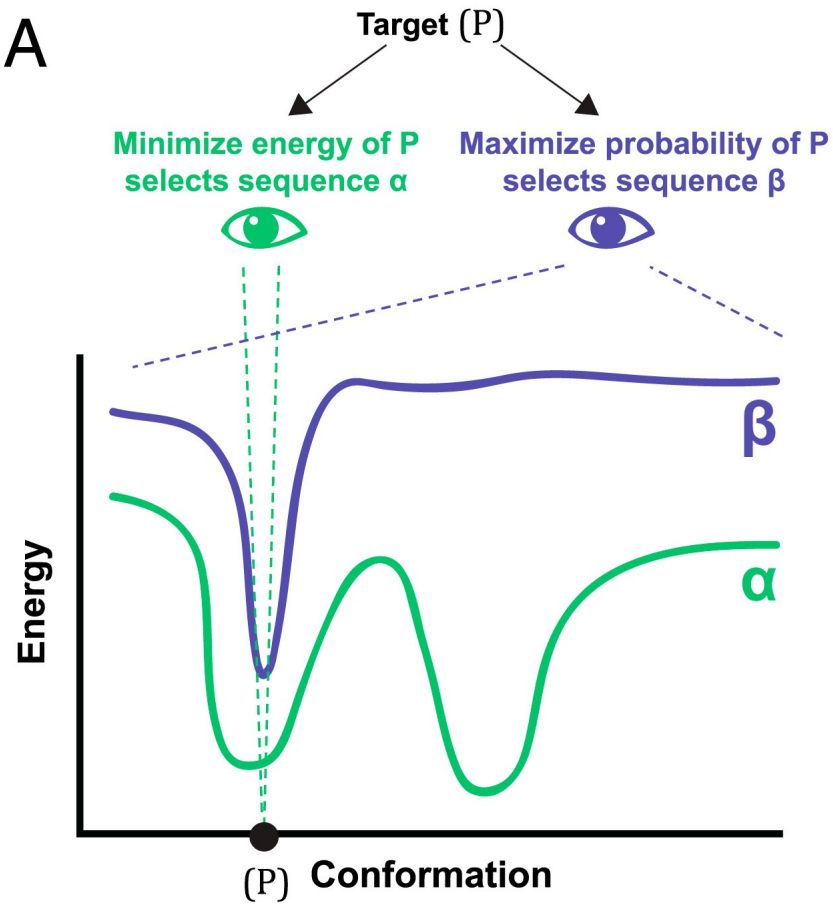
Training: simultaneous prediction of the 4 objectives on 16,047 protein chains with the average length of 250 amino acids for whose MSAs can be constructed

Loss over the 4 objectives

3D structure reconstruction using trRosetta



Protein sequence design by conformational landscape optimization



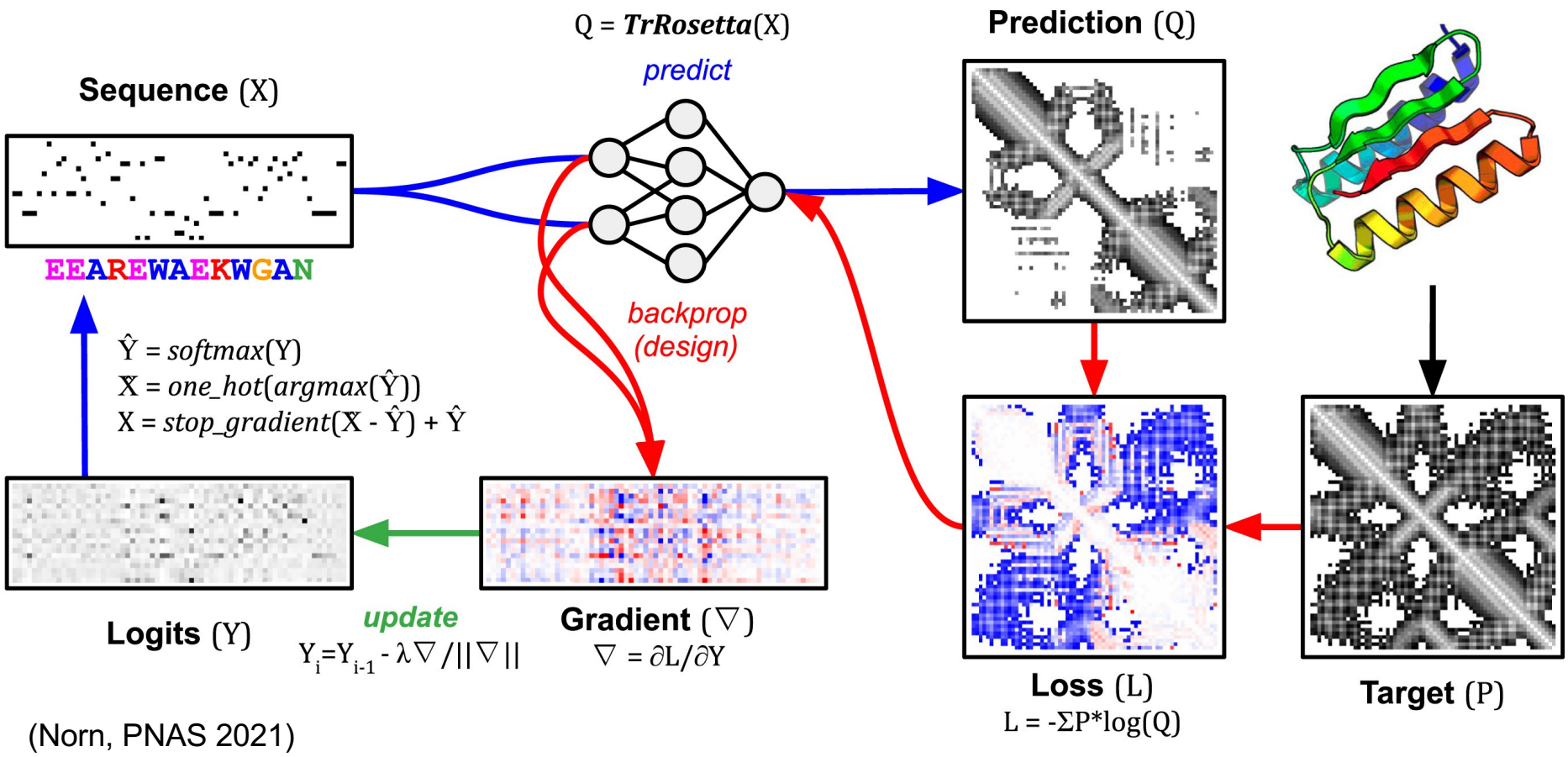
Solution:

directly optimize over all possible amino acid sequences and all possible structures in a single calculation by backpropagating gradients through trRosetta from the desired structure to the input amino acid sequence

Protein sequence design by conformational landscape optimization

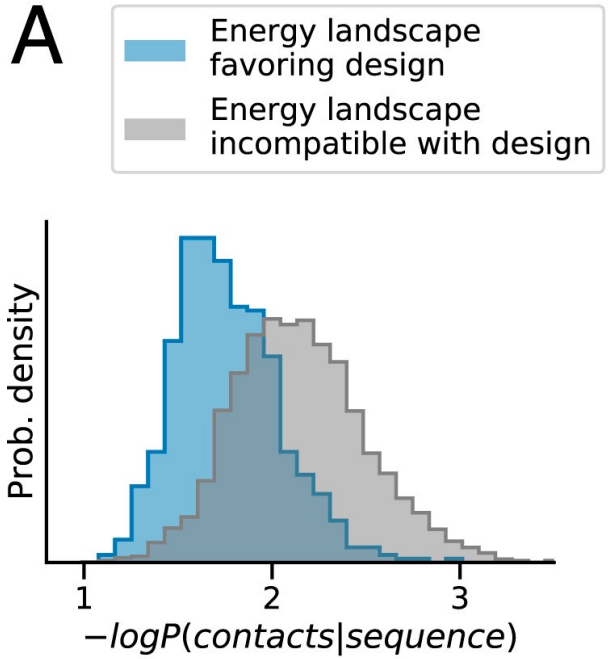
Rationale: trRosetta predicts the probability of residue–residue distances (Q) and orientations for a sequence (X). Probability distributions over possible distances and orientations should contain information about alternative conformations

Overview of trRosetta fixed backbone sequence design method

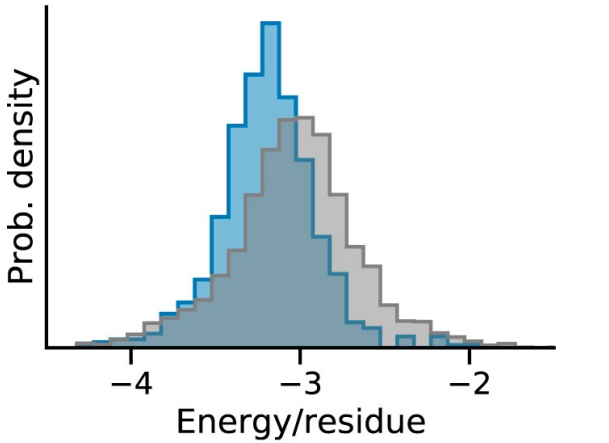


(Norn, PNAS 2021)

trRosetta predicts properties of the folding energy landscape

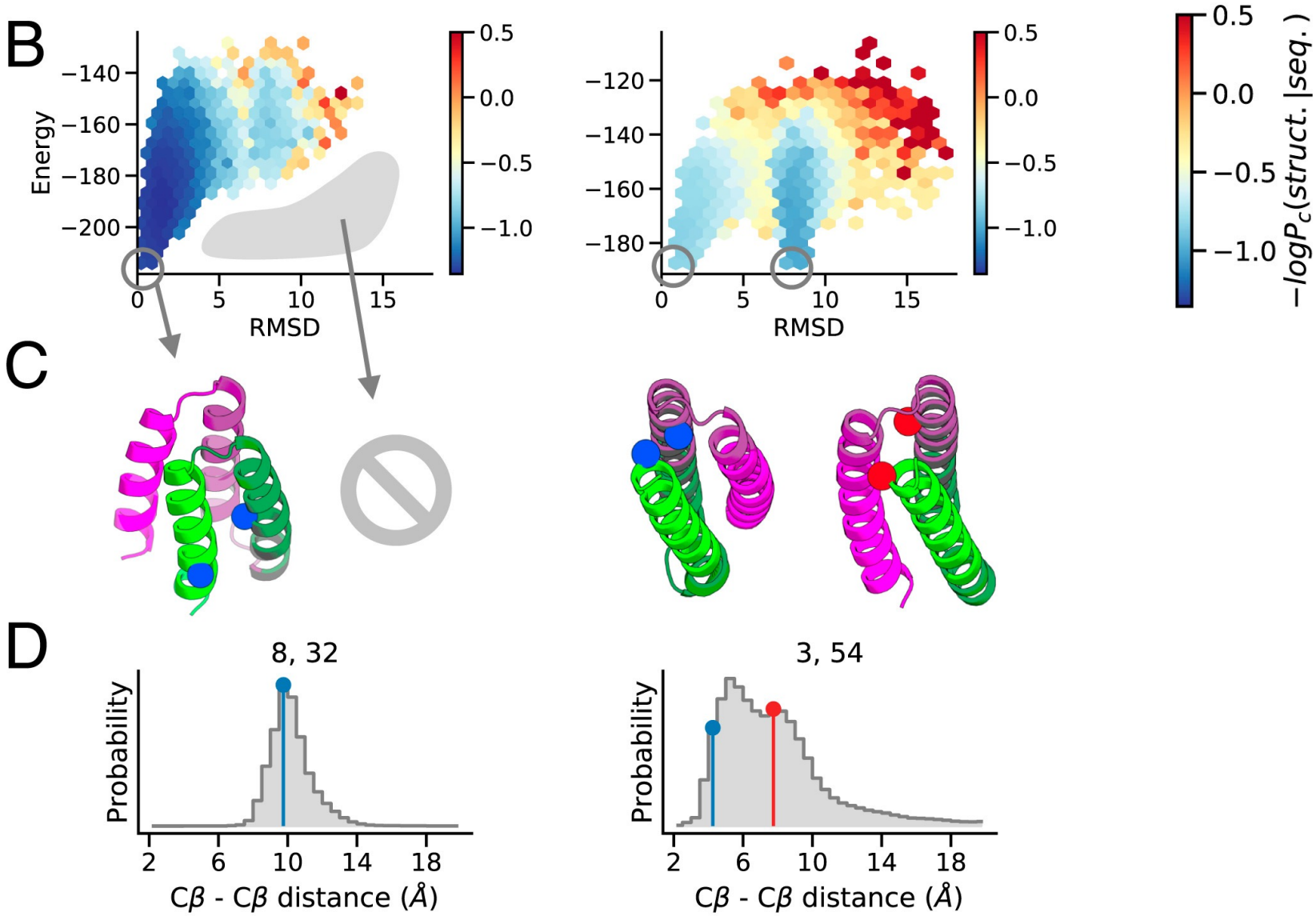


Probability distribution of distance & orientation by trRosetta

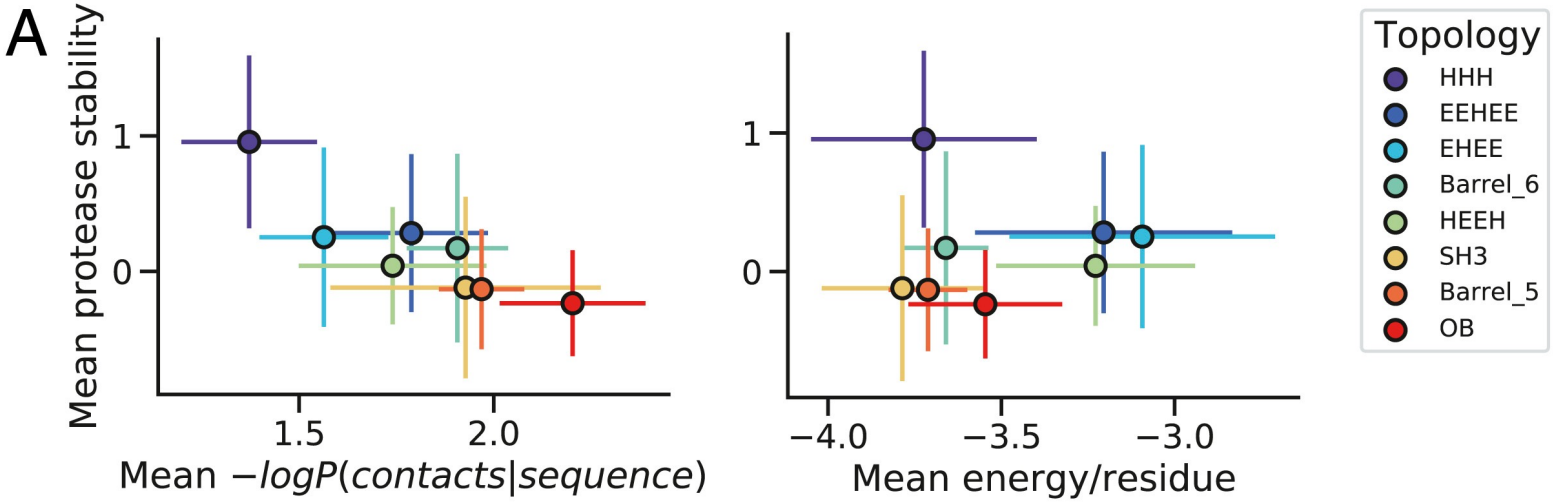


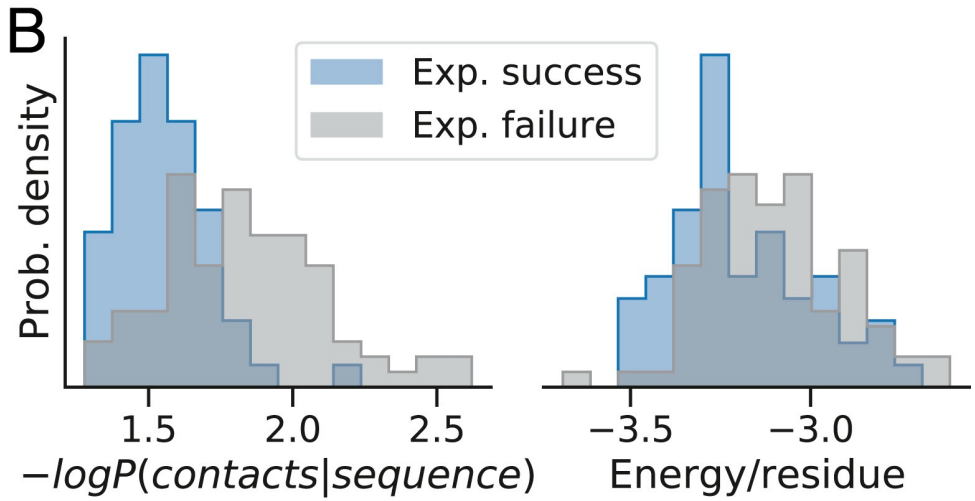
Energy prediction by Rosetta

trRosetta predicts alternative low-energy conformations



trRosetta predicts scaffold designability and experimental success





Take home messages

Fixed backbone trRosetta design outperforms traditional Rosetta in generating sequences that fold with high probability into a target structure

Fixed backbone trRosetta design procedure converged for a variety of ~100-residue protein structures after ~25 iterations, requiring only a few minutes of GPU time. (compared to CPU hours for Rosetta)

Major points to remember

1. Design sequence for a target structure: maximixing ΔE folding
2. ΔE folding is a compromise between opposite interactions
3. Which features to target for optimizing stability
4. De novo design challenges: designability
5. De novo design requires sequence-structure exploration
6. ANNs trained on protein sequences and structures can automatically optimize the design for a target structure