Protein Design
Part 3: *de novo* design

Anastassia Vorobieva

📝 Workshop Schedule

TIME	WEDNESDAY 28 MAY	THURSDAY 29 MAY	FRIDAY 20 MAY
09:00 – 10:30	Protein modelling and structure prediction: Intro to key concepts, from physics-based modelling to AI	De novo protein design: introduction, minimal sequence design, structure-based design principles	De novo design with AI models: RFDiffusion, ProteinMPNN, and ColabFold
10:30 – 11:00	Break and questions	Break and questions	Break and questions
11:00 – 12:00	Introduction to protein design: predicting the effect of mutations on protein stability	Structure-based de novo design: How to generate new structures? The chicken-and-egg problem	Practical session: De novo design of a SARS-CoV-2 RBD binder using RFDiffusion and ProteinMPNN
12:00 – 13:30	Lunch	Lunch	Lunch
13:30 – 15:00	Practical session: AlphaFold hands-on	Practical session: Parametric design of alpha-helical bundles	Practical session: Data analysis and group slide preparation
15:15 – 17:45	Practical session: In silico mutational scanning and ΔΔG calculations	Practical session: Sequence design for parametric bundles with PyRosetta	Practical session: Group presentations and results discussion

De novo protein design

1. Introduction to protein design:
    - Definition and evolution in time

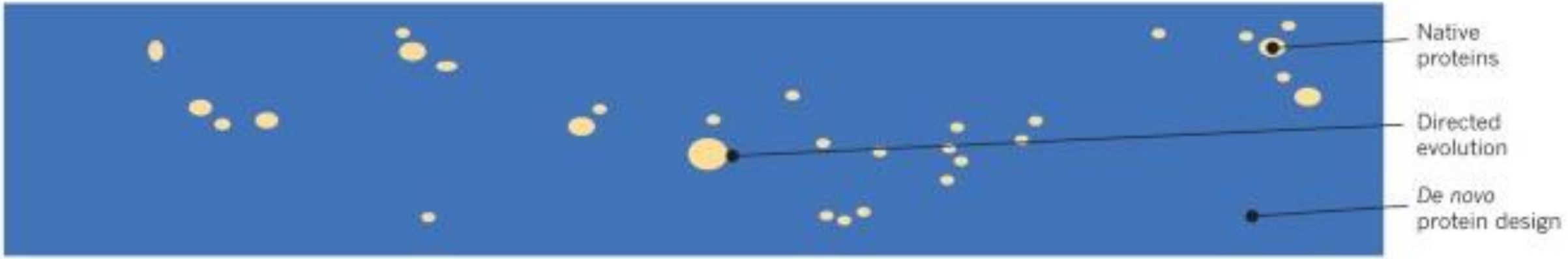
2. Minimal sequence design

3. Computational design
    - Principle
    - Pipeline
    - Designable backbones
    - Scoring and sampling
    - Parametric backbone generation
    - Fragment-based backbone generation

Introduction - protein design

Generating new-to-nature protein structure and sequences.

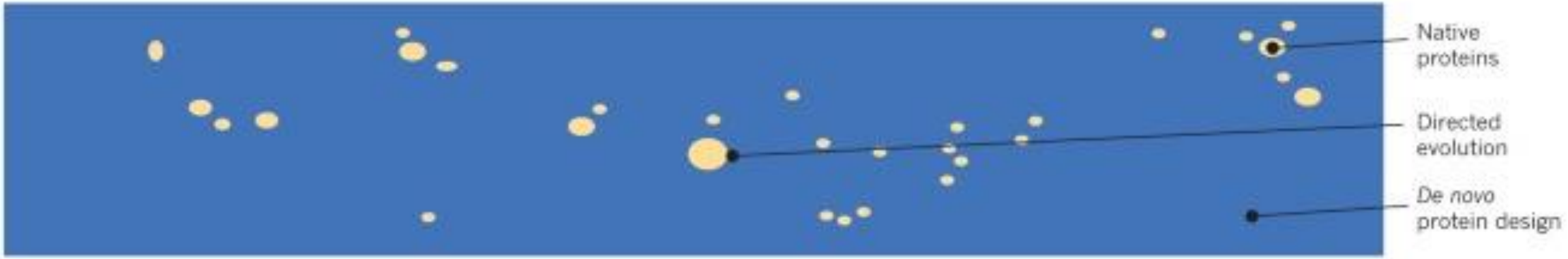
De novo protein design – exploring the dark protein sequence space



Single helix of 30 residues: $20^{30} = 10^{39}$ possible sequences

Which of these sequences encode foldable proteins?

De novo protein design – exploring the dark protein sequence space



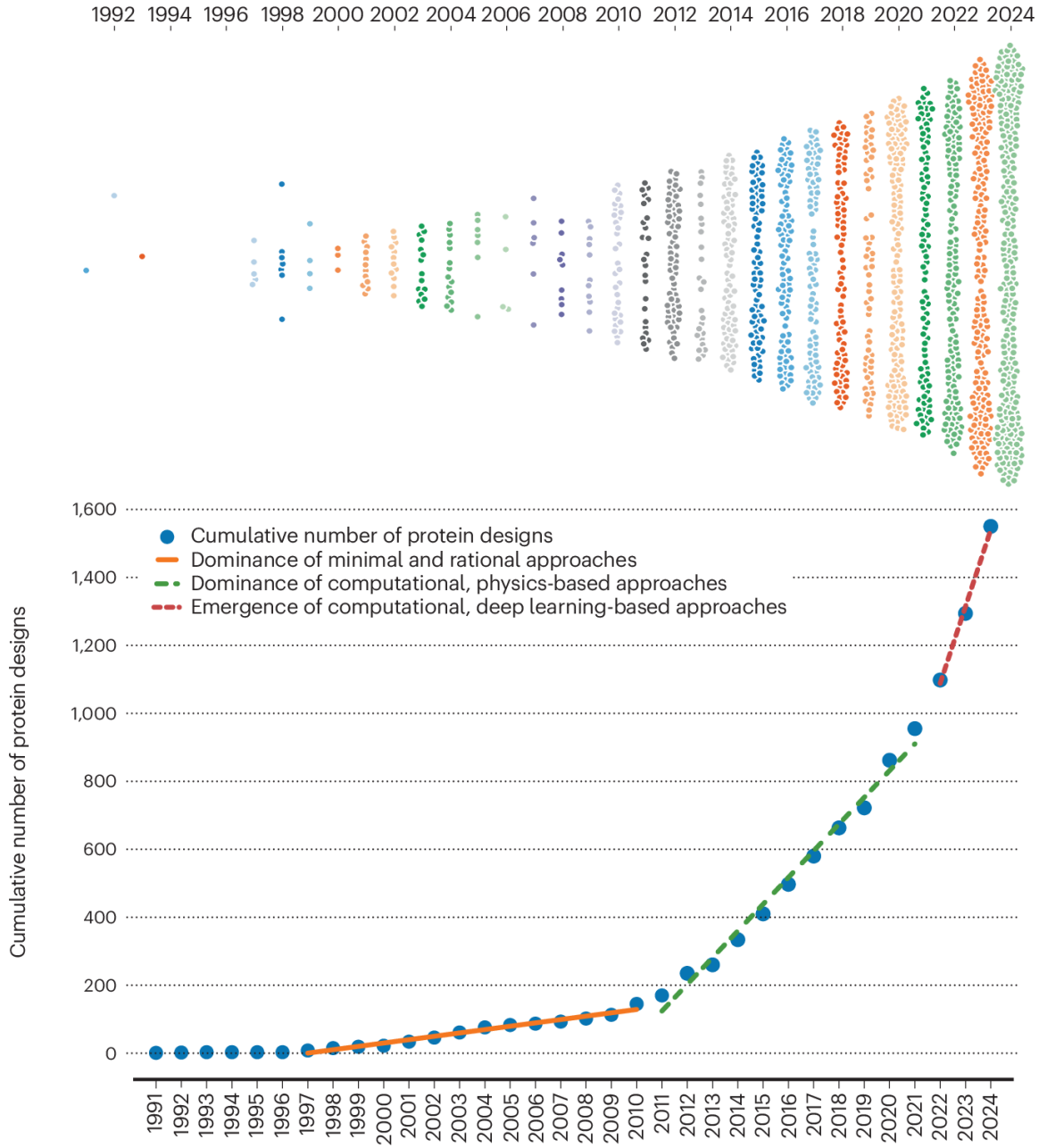
Single helix of 30 residues: $20^{30} = 10^{39}$ possible sequences

Which of these sequences encode foldable proteins?

Top7 – first time that a new-to-nature fold was designed from scratch

https://doi.org/10.1126/science.1089427

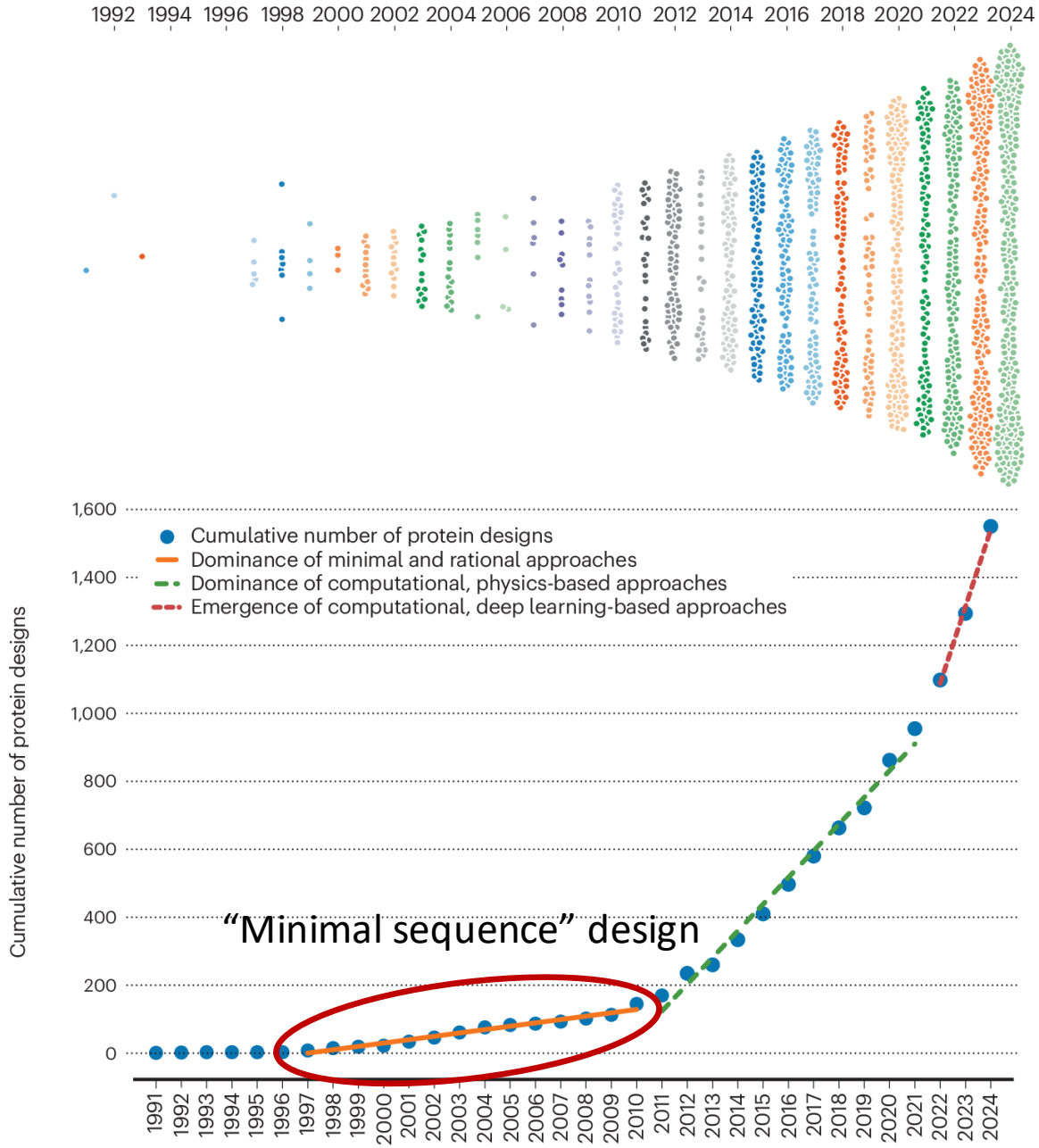
Evolution of *de novo* protein design



https://doi.org/10.1038/s41587-025-02607-x

Minimal sequence design

Evolution of *de novo* protein design

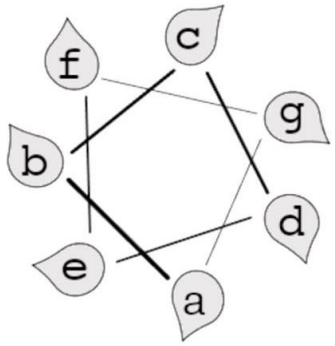


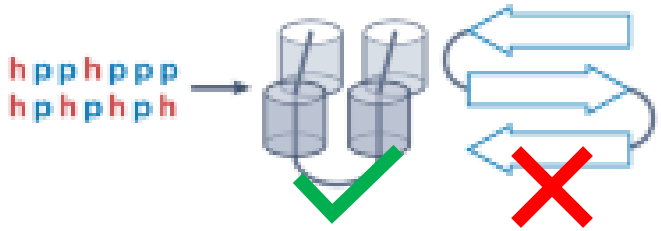
https://doi.org/10.1038/s41587-025-02607-x

Minimal sequence design

```
TM helix 1  ...MPYIYLAIAIAAEVVGTSALK...
TM helix 2  ...LIPSVGTLVGYGASFYLLSLT...
TM helix 3  ...YALWSGIGIVAISLVGVWILF...
TM helix 4  ...LDLMKIVGLALIVAGVVILNL...

Pattern     ...hhhhpGhGhhhhhphhGhhhhhp...
Idealized   ...LLLLSGLGLLLLSLLGLLLLS...
Position    ...abcdefgabcdefgabcdefg...
```





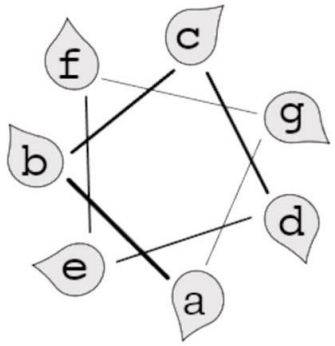
Typical design:

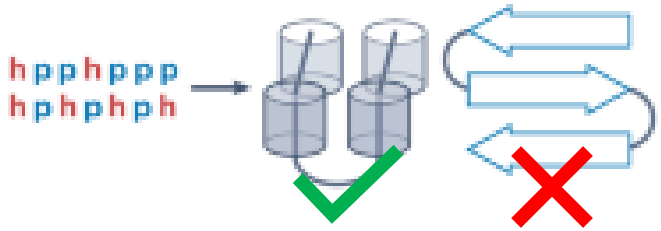
1. Glycine
2. Proline
3. One hydrophobic amino acid (Leucine or Valine)
4. One polar amino acid (Serine)

Minimal sequence design

```
TM helix 1  ...MPYIYLAIAIAAEVVGTSALK...
TM helix 2  ...LIPSVGTLVGYGASFYLLSLT...
TM helix 3  ...YALWSGIGIVAISLVGVWILF...
TM helix 4  ...LDLMKIVGLALIVAGVVILNL...

Pattern     ...hhhhpGhGhhhhphhGhhhhp...
Idealized   ...LLLLSGLGLLLLSLLGLLLLS...
Position    ...abcdefgabcdefgabcdefg...
```

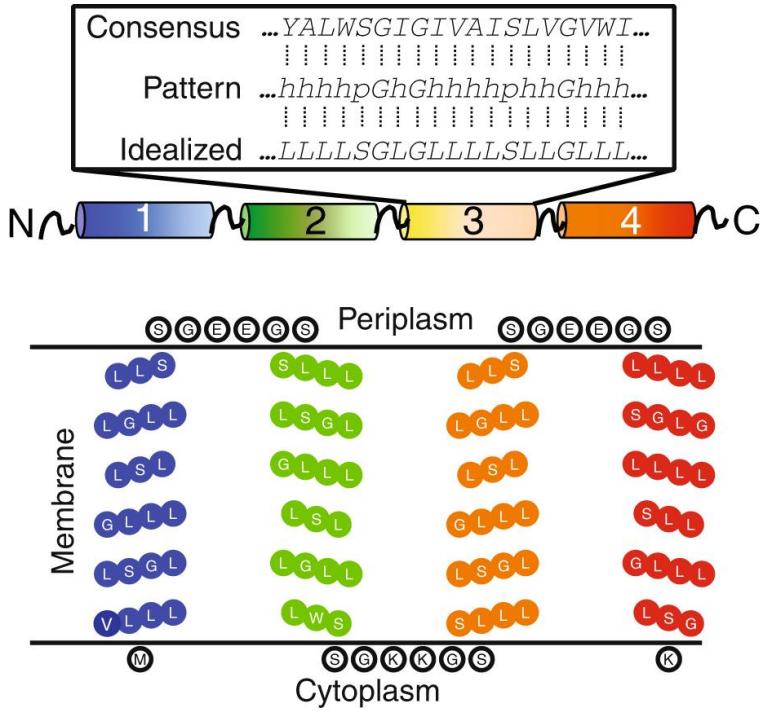


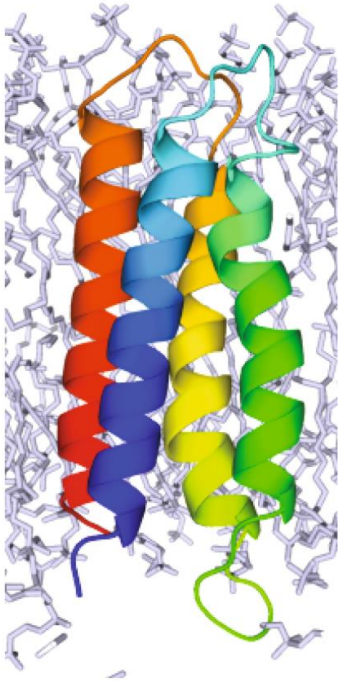


Typical design:

1. Glycine
2. Proline
3. One hydrophobic amino acid (Leucine or Valine)
4. One polar amino acid (Serine)

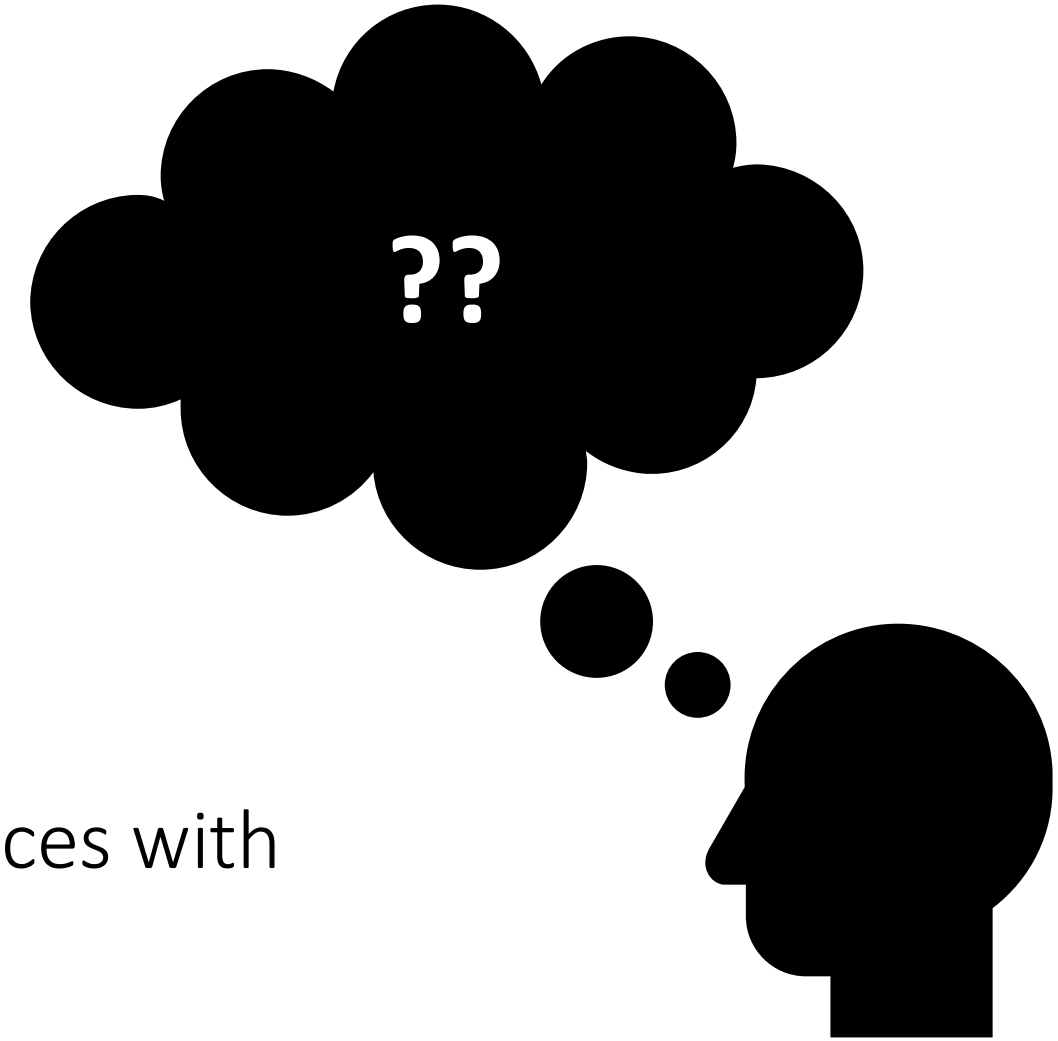
Example: Minimal membrane protein





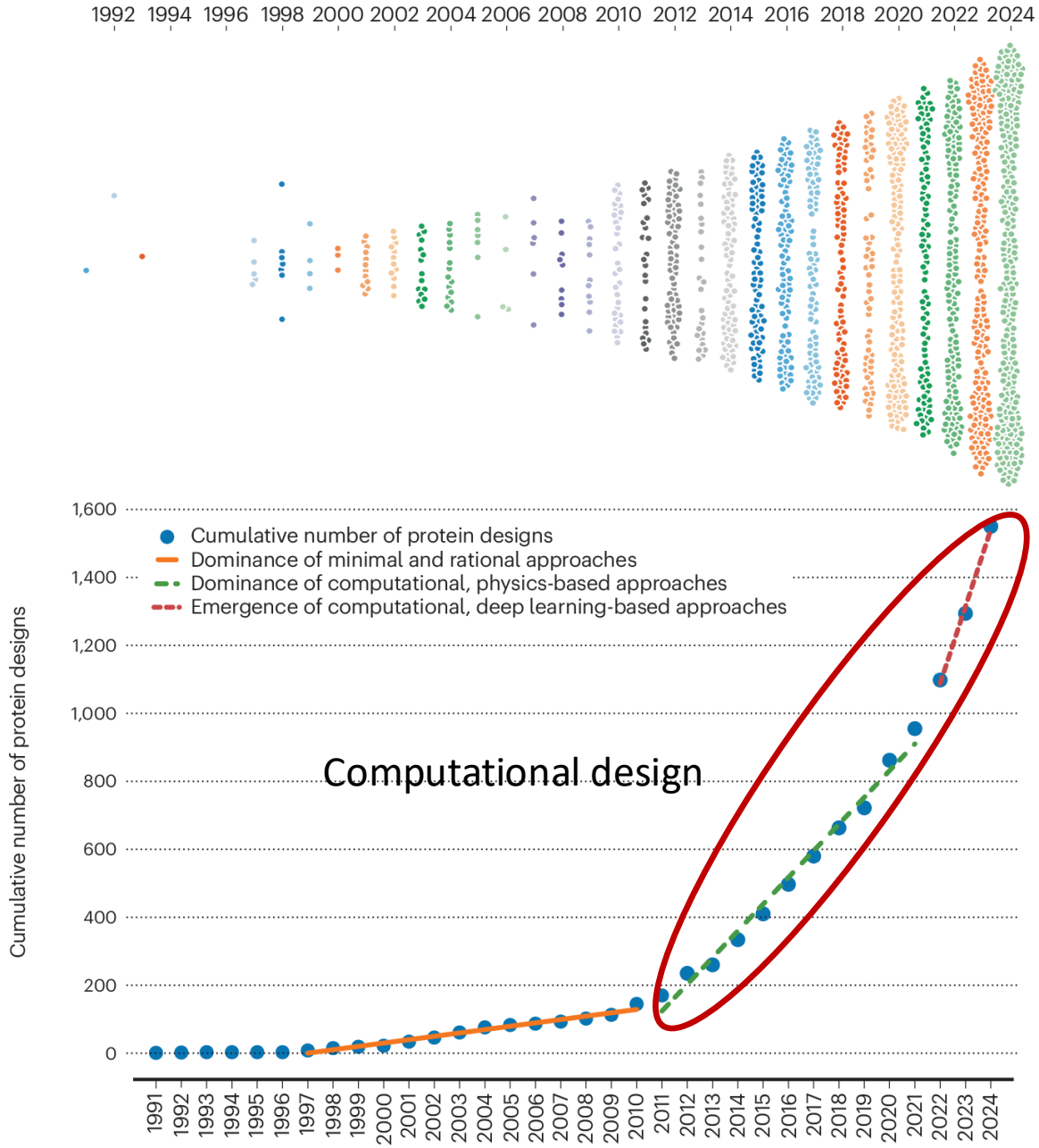
https://doi.org/10.1038/s41598-018-31964-8

How to generate diverse sequences with minimal design?



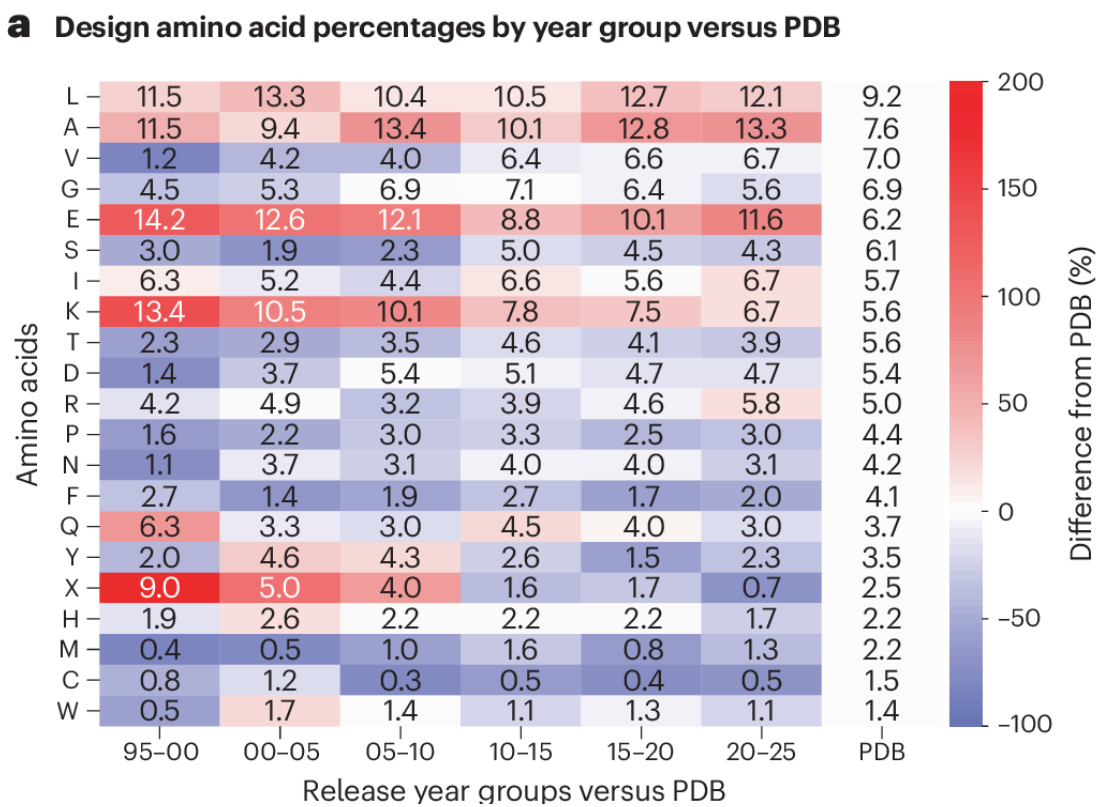
Computational design (structure-based)

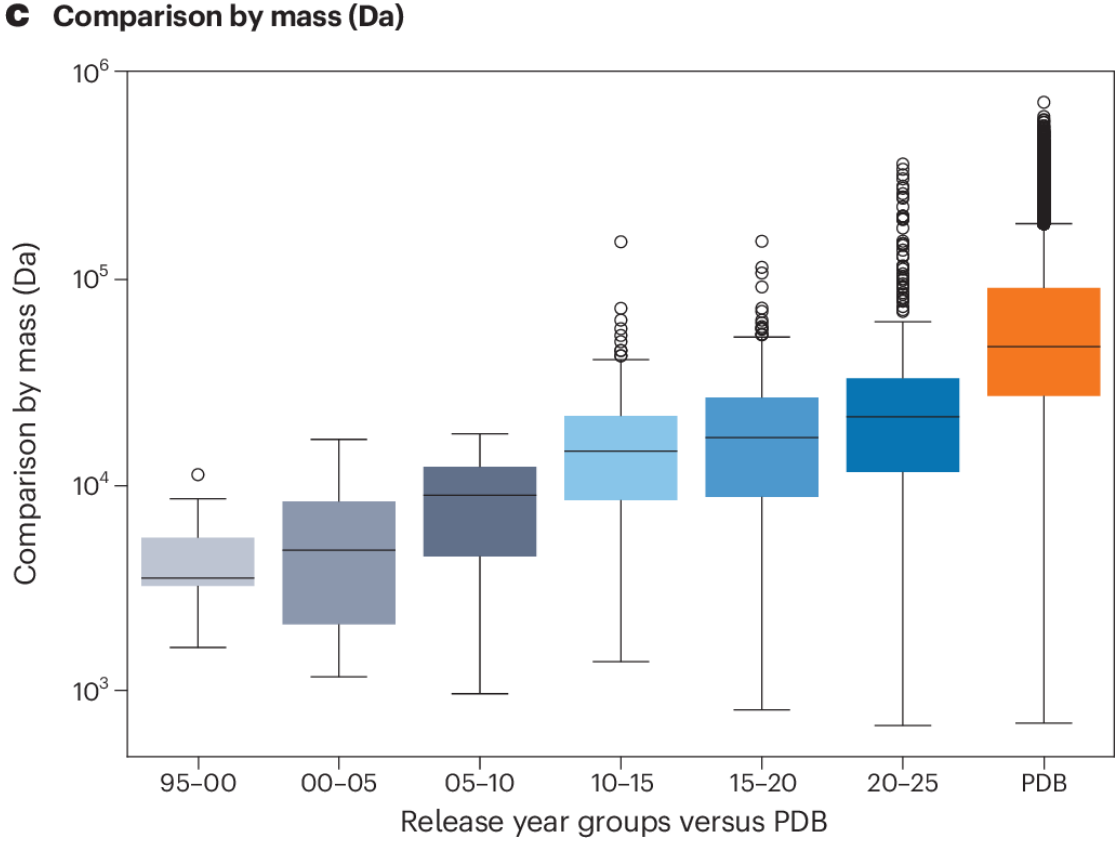
Evolution of *de novo* protein design



https://doi.org/10.1038/s41587-025-02607-x

Computational design enabled the generation of more diverse and longer sequences





https://doi.org/10.1038/s41587-025-02607-x

The folded states of proteins are likely global energy minima for their sequences (C. Anfinsen)

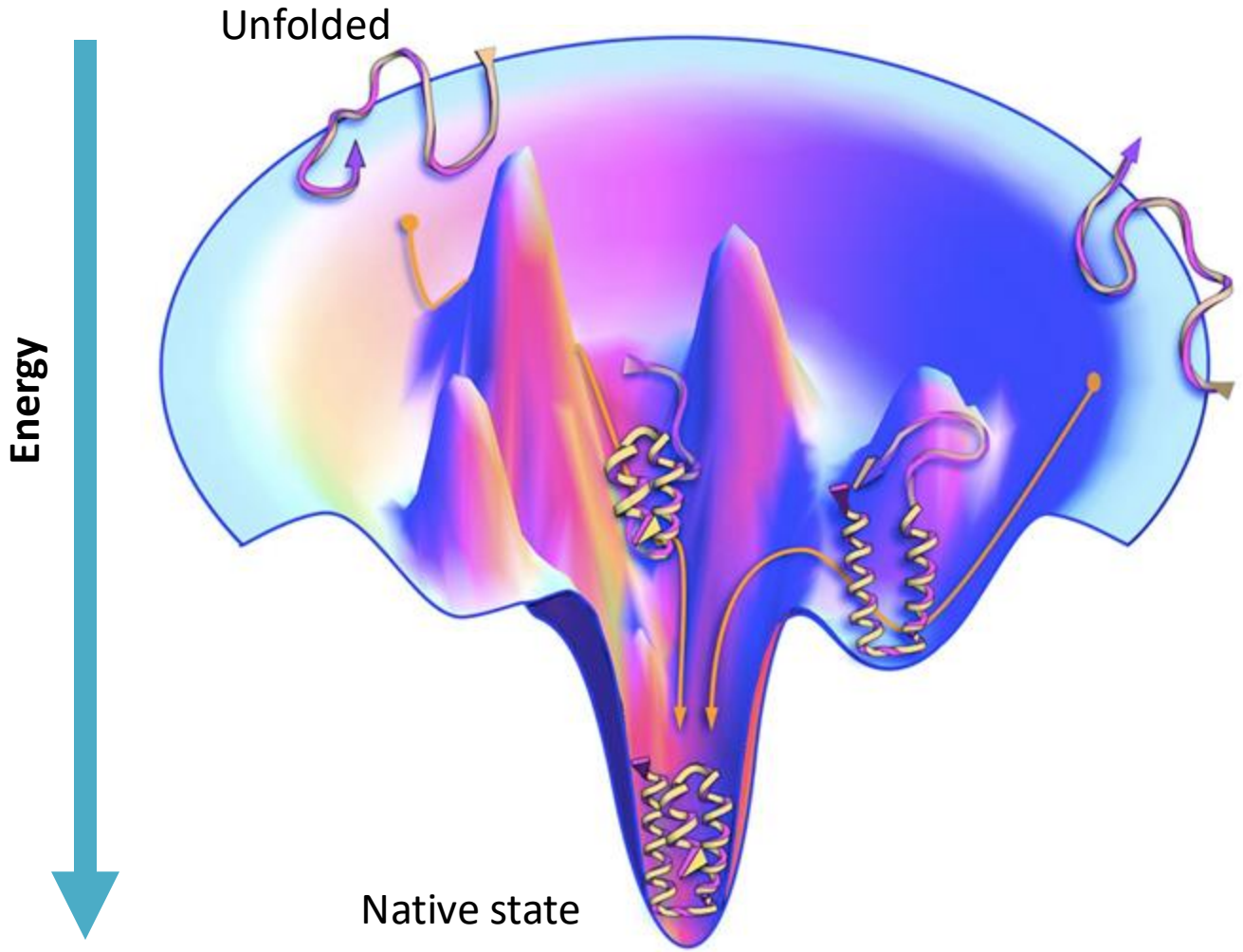
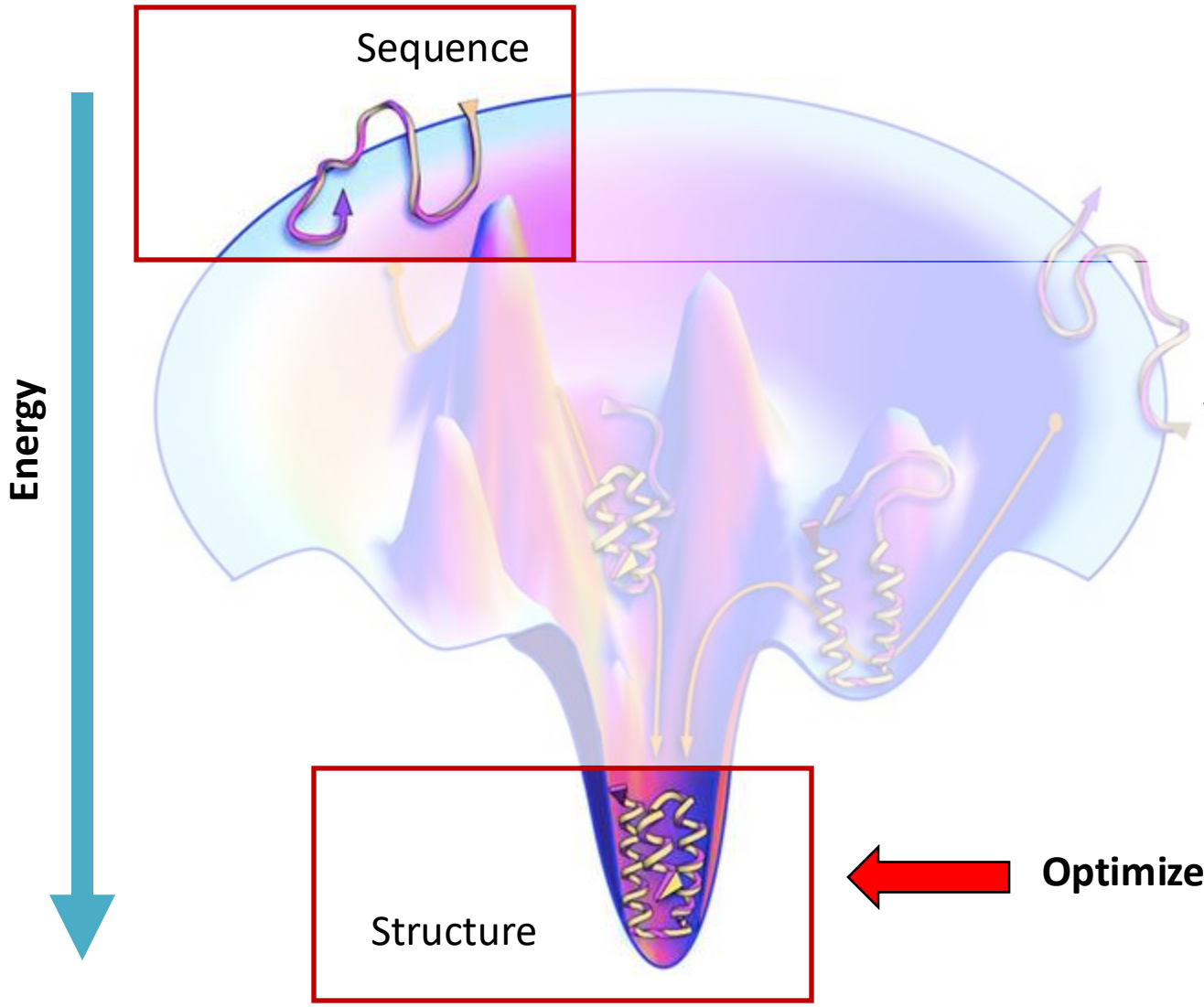


Image from: "The protein-folding problem, 50 years on." *science* 338, no. 6110 (2012): 1042-1046.

Structure prediction(*ab initio*) and design



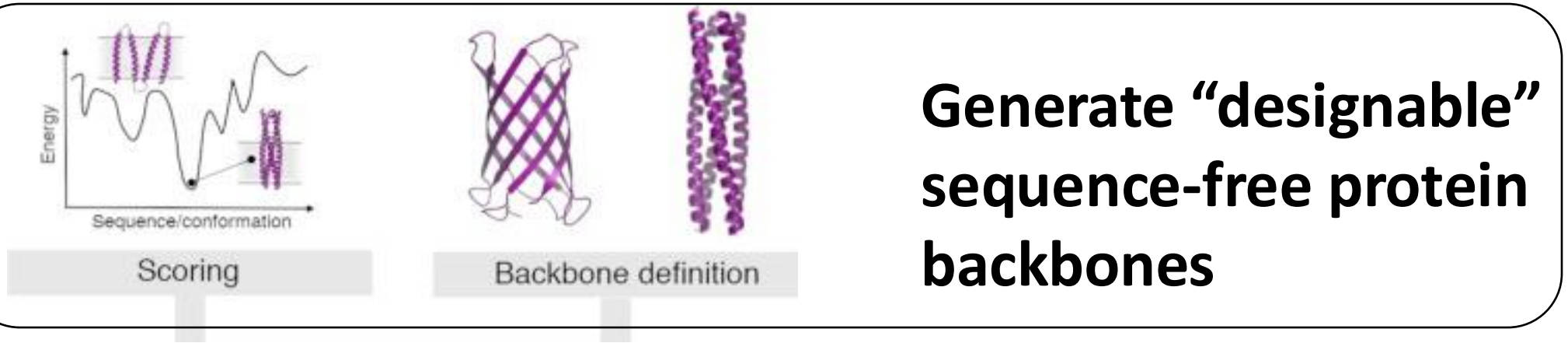
Usually, we don't know about those alternative local minima

Optimize the energy of the folded state

Image from: "The protein-folding problem, 50 years on." *science* 338, no. 6110 (2012): 1042-1046.

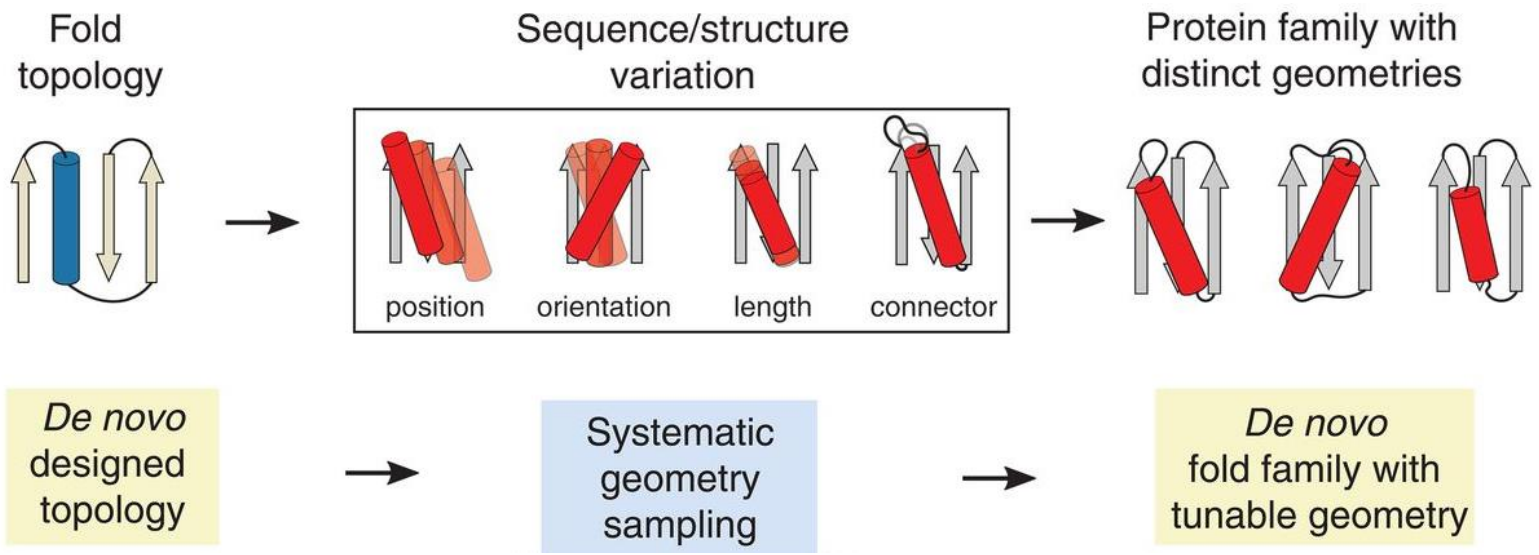
De novo design pipeline

Backbone generation



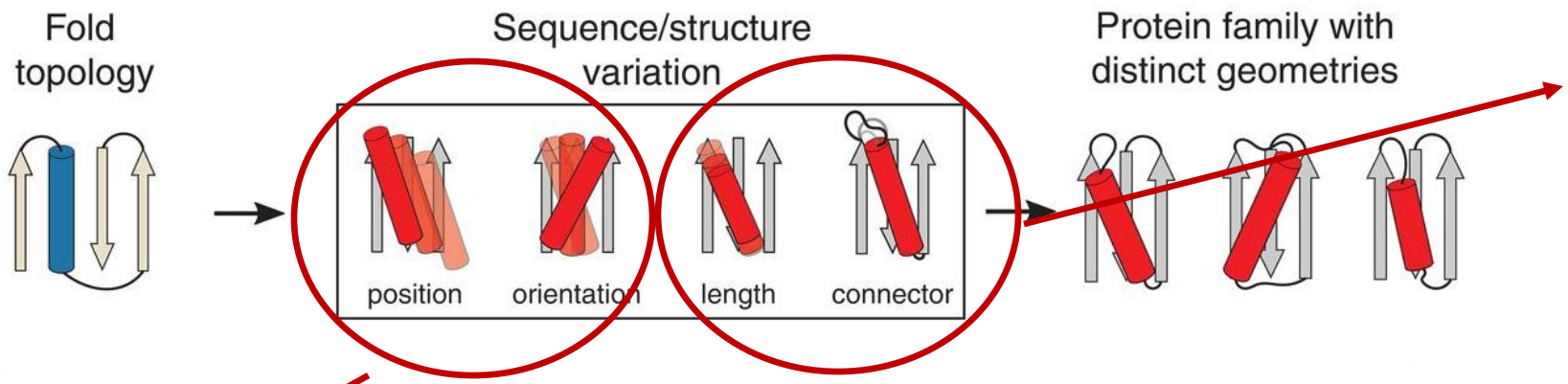
Generate “designable” sequence-free protein backbones

Designable backbones: new structures incorporating native-like features

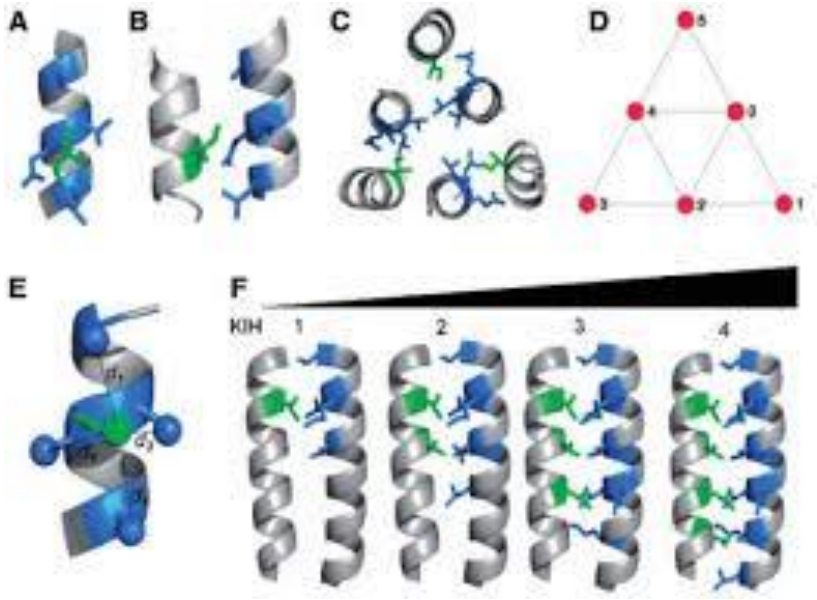


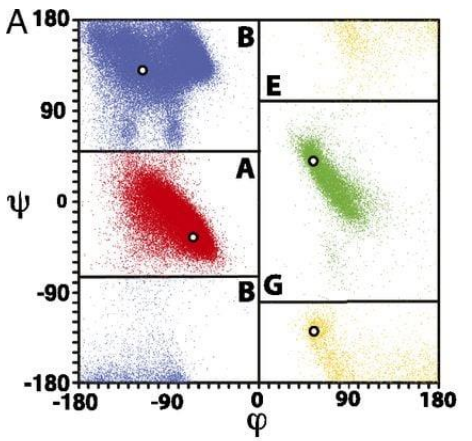
https://doi.org/10.1126/science.abc0881

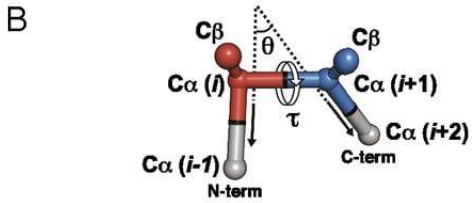
Designable backbones: new structures incorporating native-like features



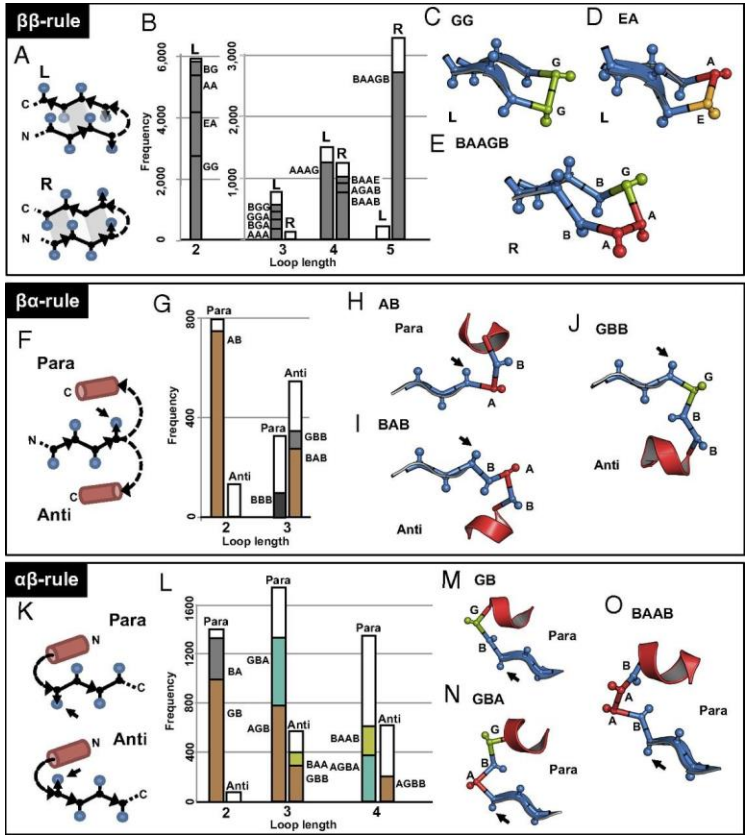
1. Side-chain packing







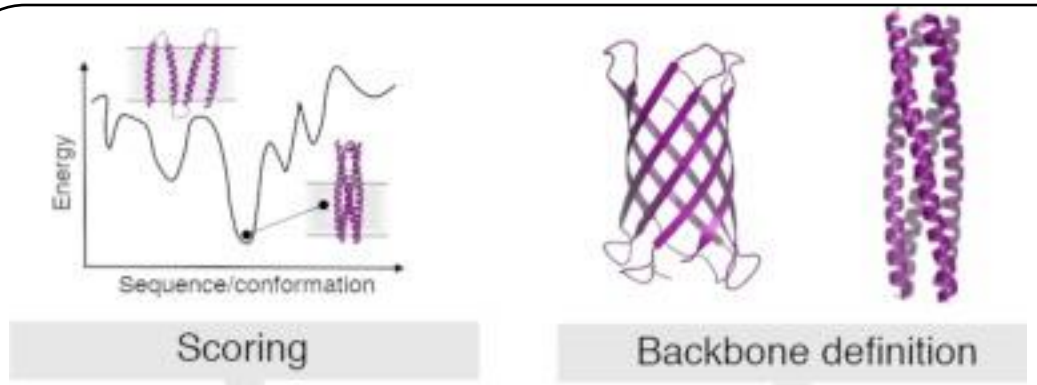
2. Connector geometry: packing and strain



https://doi.org/10.1073/pnas.1509508112

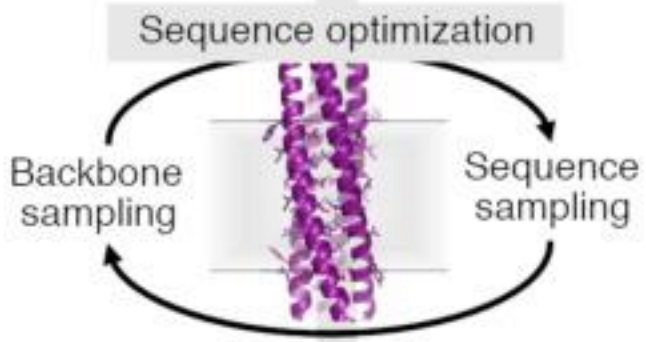
De novo design pipeline

Backbone generation



Generate “designable” sequence-free protein backbones

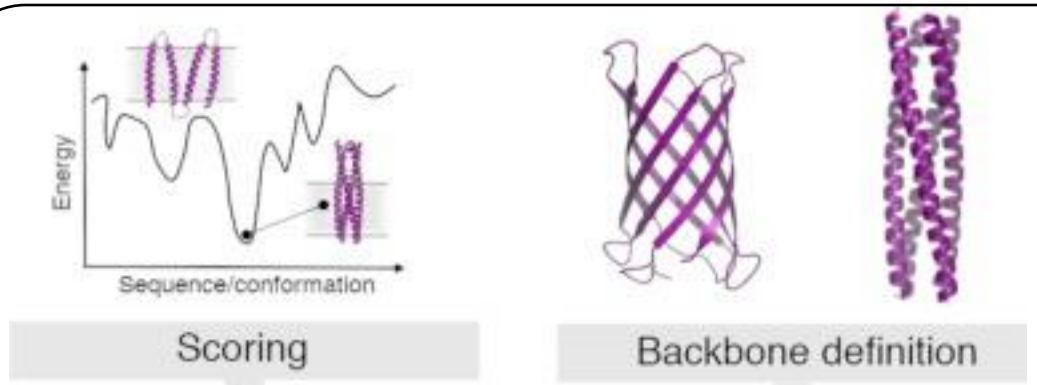
Sequence design



Design new sequences compatible with the backbones

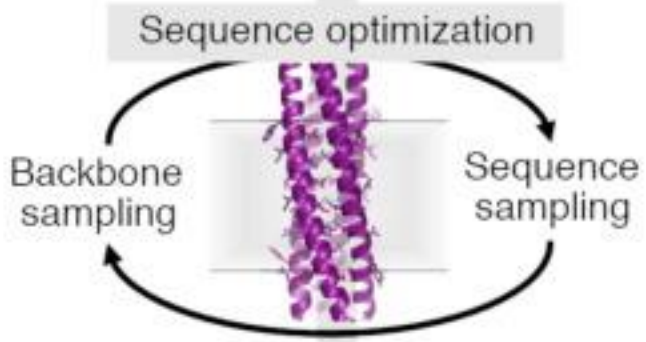
De novo design pipeline

Backbone generation



Generate "designable" sequence-free protein backbones

Sequence design



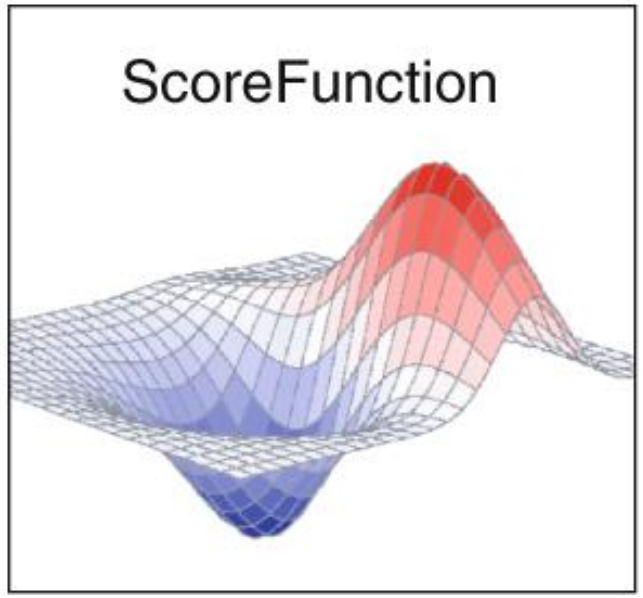
Design new sequences compatible with the backbones

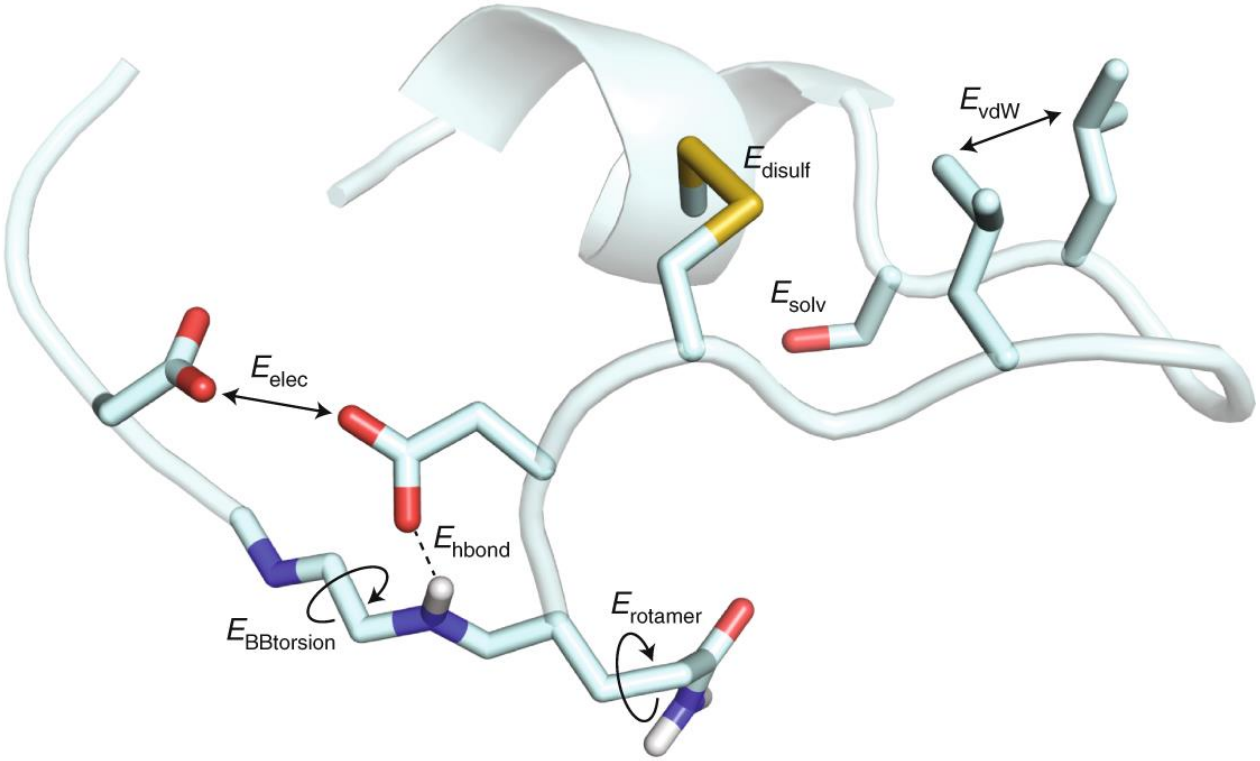
Reminder: Two major components of physics based models?

Scoring – energy functions

weight

$$\Delta E_{\text{total}} = \sum_i w_i E_i(\Theta_i, \text{aa}_i)$$





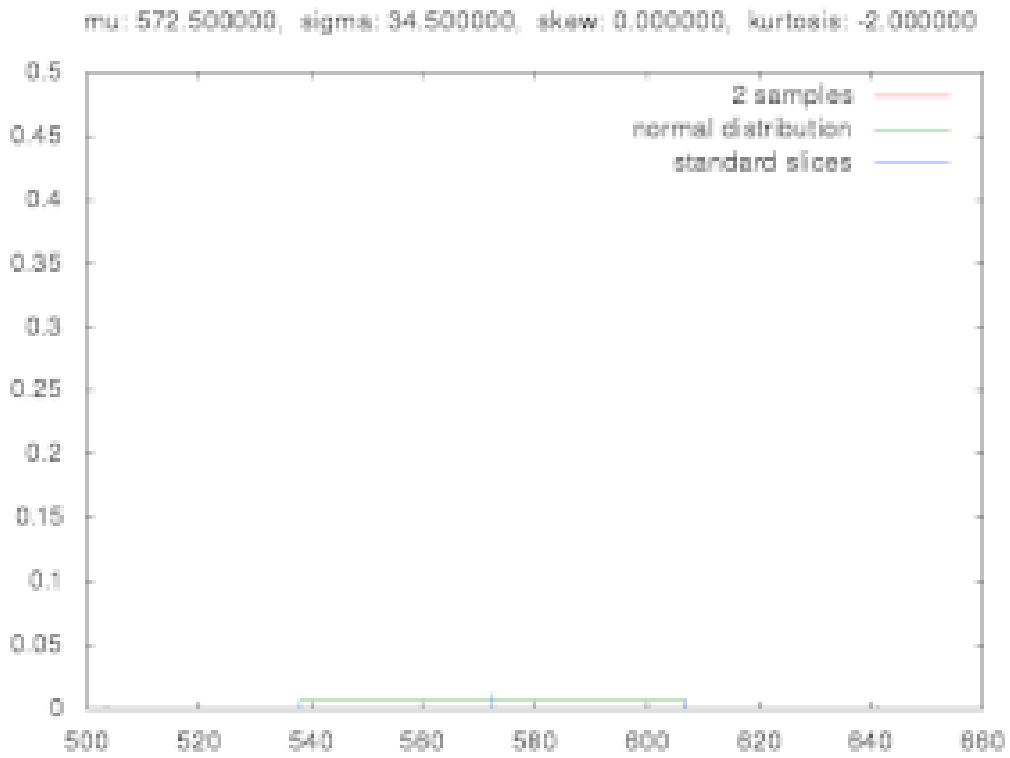
E_{vdW}	Lennard–Jones for attractive or repulsive interaction
E_{hbond}	Hydrogen bonding allows buried polar atoms
E_{elec}	Electrostatic interaction between charges
E_{disulf}	Disulfide bonds between cysteines

E_{solv}	Implicit solvation model penalizes buried polar atoms
$E_{BBtorsion}$	Backbone torsion preferences from main-chain potential
$E_{rotamer}$	Side-chain torsion angles from rotamer library
E_{ref}	Unfolded state reference energy for design

Sampling methods

Sampling residue type and rotamers

- Monte Carlo algorithm to simulate random sampling of a normal distribution



Sampling methods

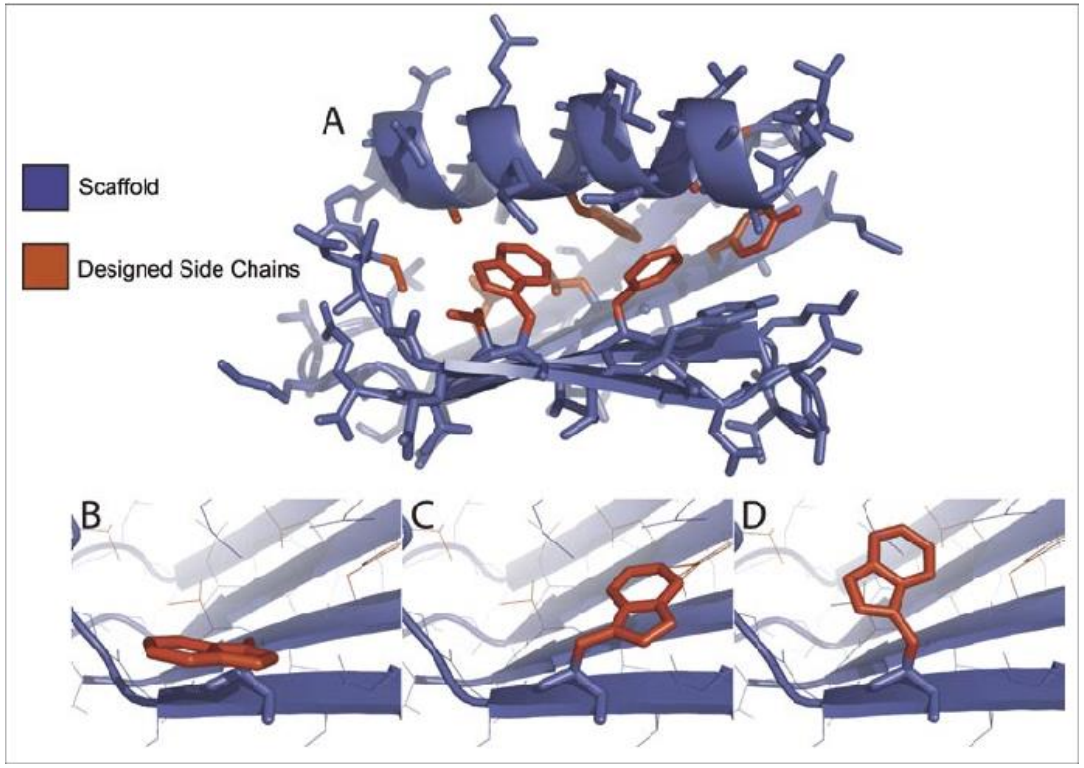
Sampling residue type and rotamers

- Simulated **annealer**

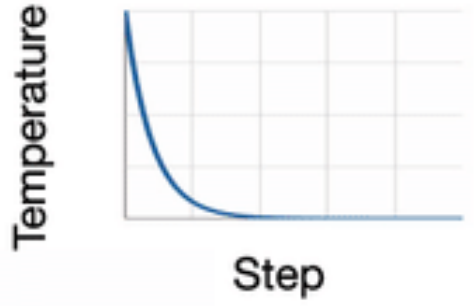
Sampling methods

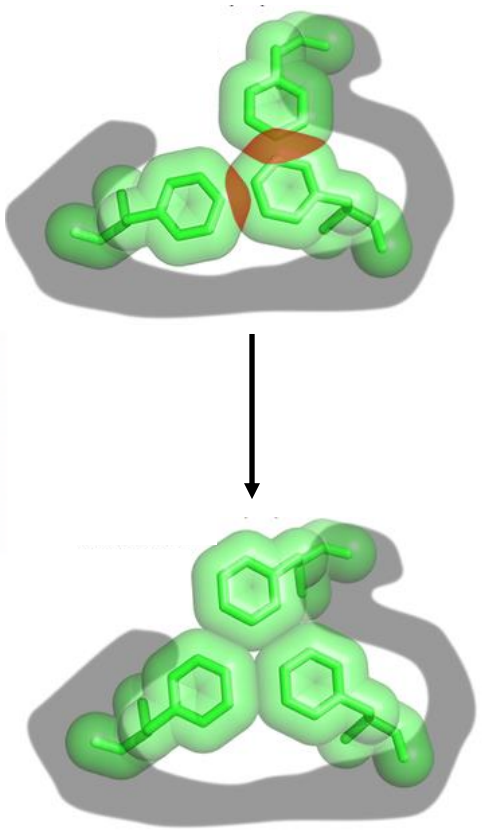
Sampling residue type and rotamers

- Simulated **annealer – Rosetta design**



Low repulsive energy – many moves are allowed





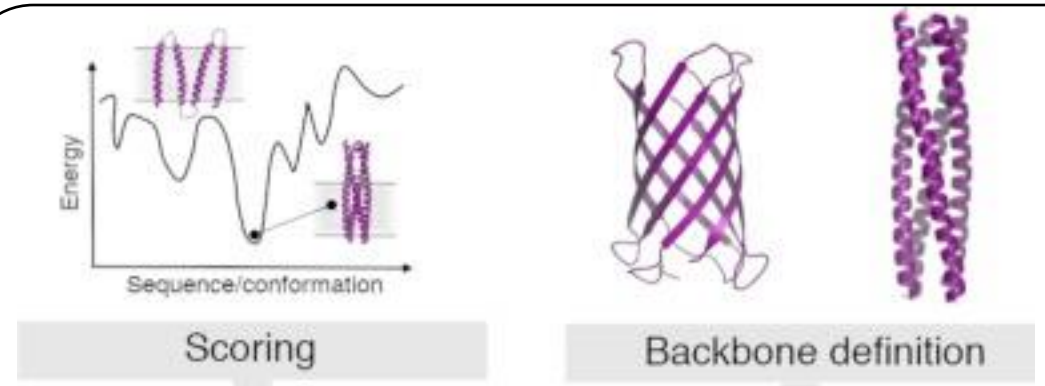
High repulsive energy – less moves are allowed

Temperature

https://www.youtube.com/watch?v=lkrVokd5Mis

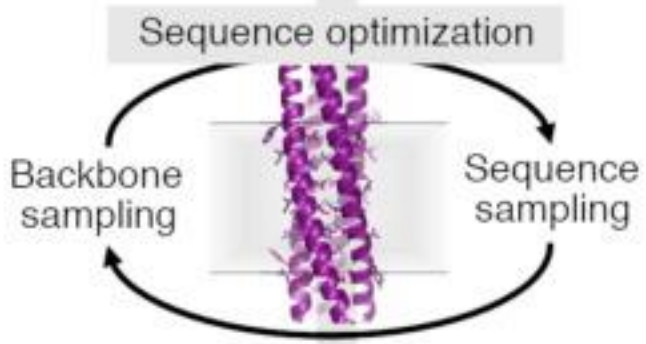
De novo design pipeline

Backbone generation



Generate "designable" sequence-free protein backbones

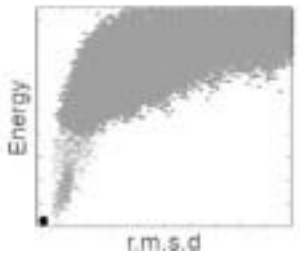
Sequence design



Design new sequences compatible with the backbones

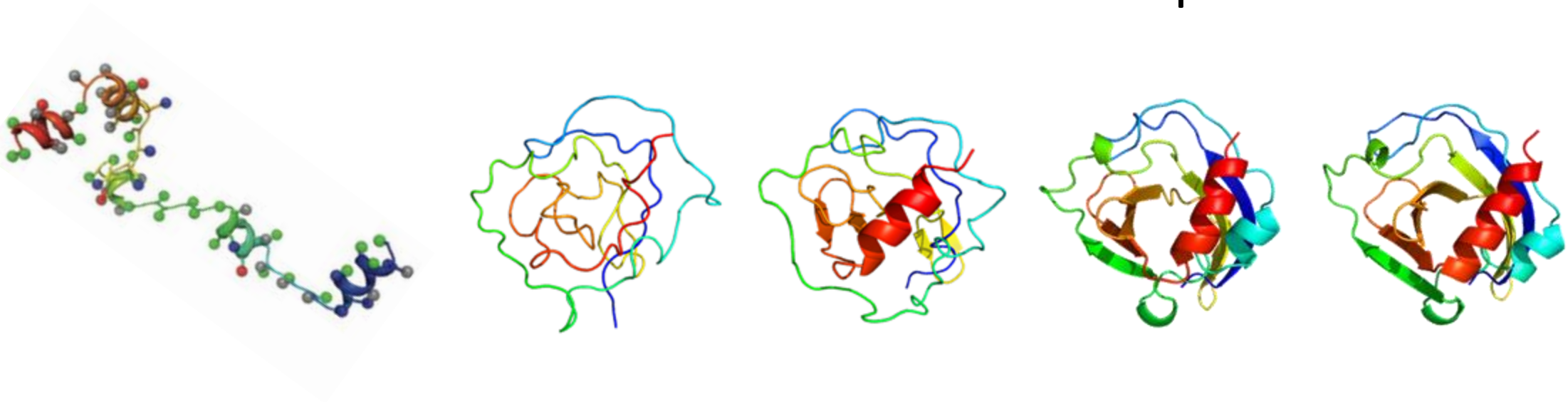
Validation

Energy landscape of the designed sequence converges on target structure



Verify sequence/structure compatibility with *ab initio* structure prediction

Reminder: *Ab initio* structure prediction

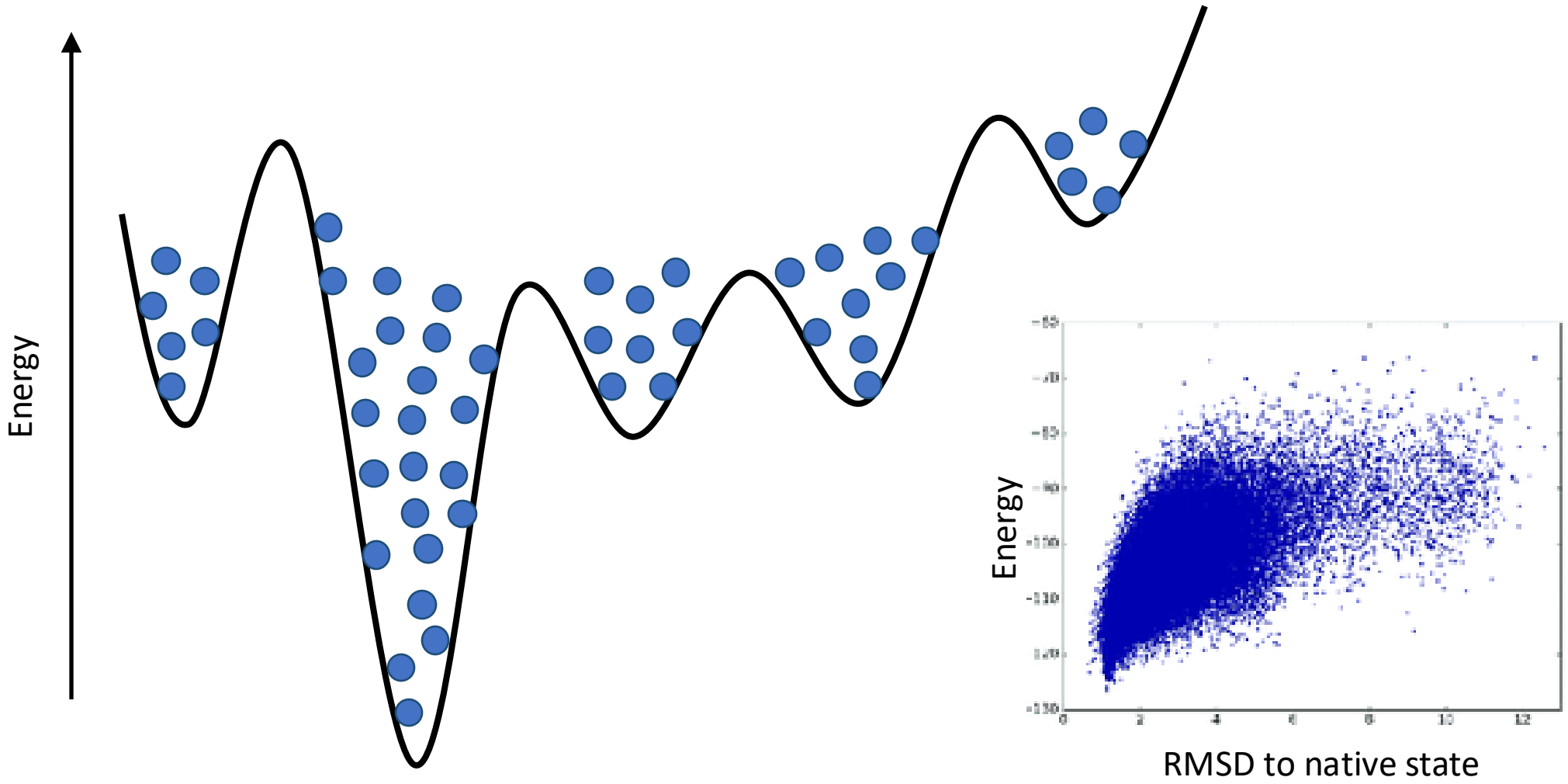


Sequence

Structure

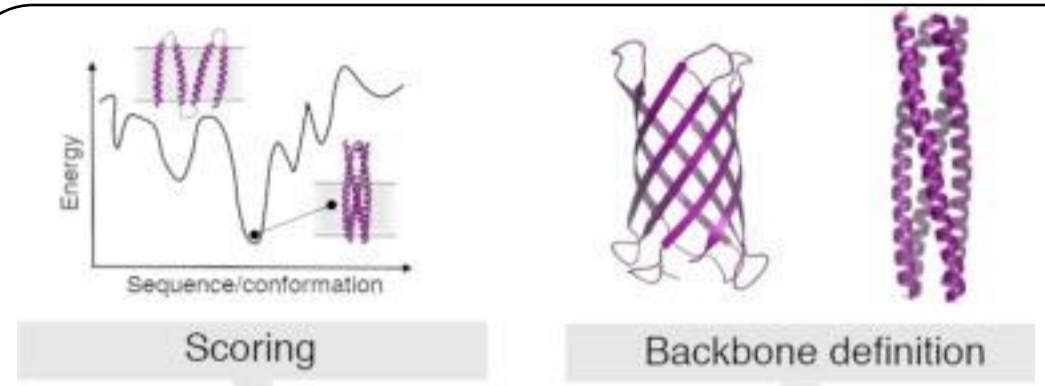
Stage1	Stage2	Stage 3a	Stage 3b	Stage4	Relax
uniform: 9 (25)	uniform: 9 (25)	uniform: 9 (25)		uniform: 3 (200) smooth: 3 (200)	pack minimize
score0 vdw	**score1** vdw, pair env, hs_pair ss_pair, sheet	**score2** vdw, cenpack pair, env, cbeta, hs_pair, ss_pair, sheet	**score5** vdw, cenpack pair, env, cbeta, hs_pair, ss_pair, sheet	**score3** vdw, cenpack pair, env, cbeta, **rg**, hs_pair, ss_pair rsigma, sheet	**ref2015**

Lowest energy structures converge to same solution



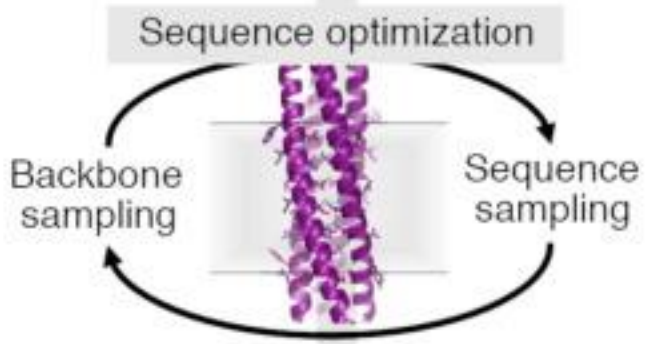
De novo design pipeline

Backbone generation



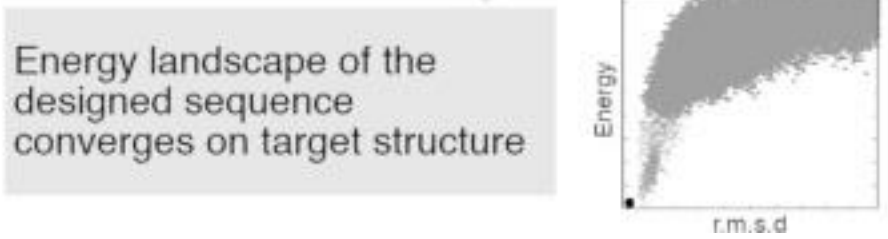
Generate "designable" sequence-free protein backbones

Sequence design



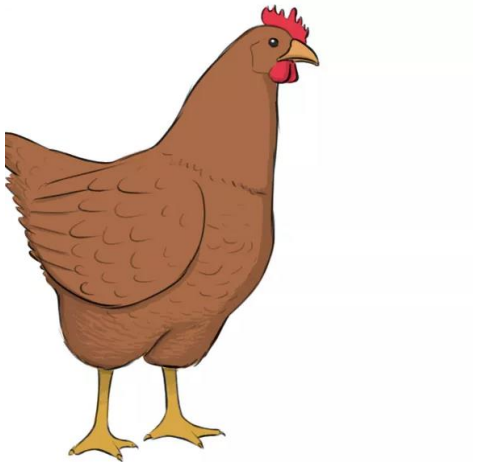
Design new sequences compatible with the backbones

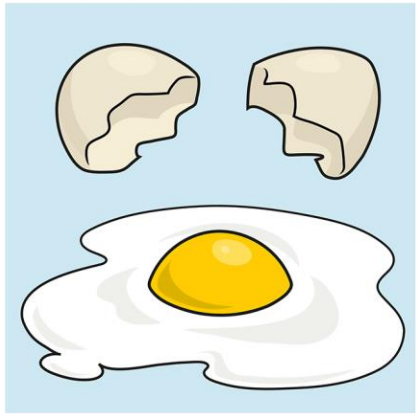
Validation



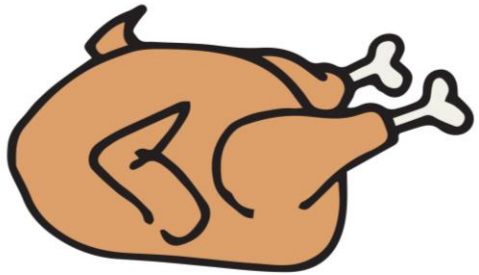
Verify sequence/structure compatibility with *ab initio* structure prediction

De novo protein design: chicken/egg problem?



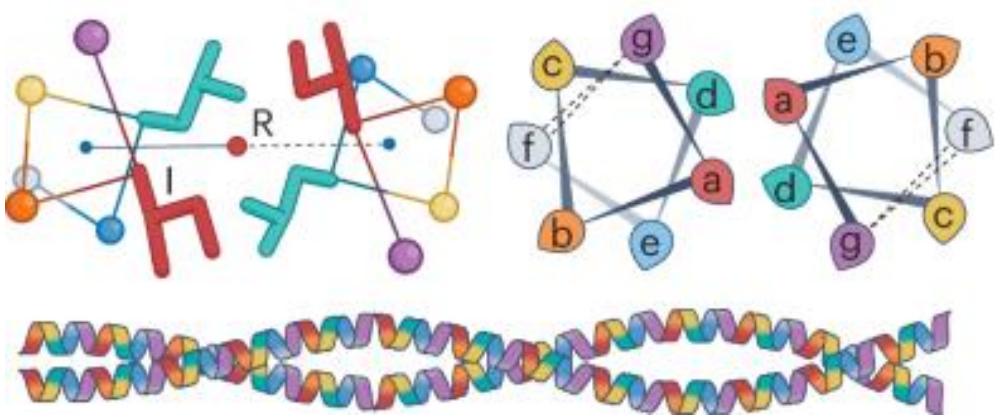


De novo protein design: chicken/egg problem?



Parametric design of alpha-helical bundles

Knob-into-hole packing of side-chains



Crick coiled coil equations

For a helix in a coiled coil with supercoil axis along z, the Cartesian coordinates of the Calpha of residue *t* are given by:

$$x = R_0\cos(\omega_0 t+\varphi_0') + R_1\cos(\omega_0 t+\varphi_0')\cos(\omega_1 t+\varphi_1) - R_1\cos(\alpha)\sin(\omega_0 t+\varphi_0')\sin(\omega_1 t+\varphi_1)$$

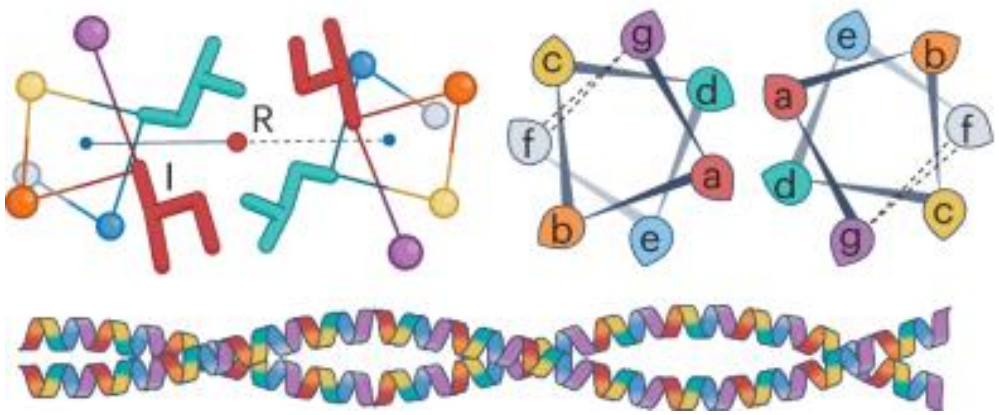
$$y = R_0\sin(\omega_0 t+\varphi_0') + R_1\sin(\omega_0 t+\varphi_0')\cos(\omega_1 t+\varphi_1) + R_1\cos(\alpha)\cos(\omega_0 t+\varphi_0')\sin(\omega_1 t+\varphi_1)$$

$$z = (\omega_0 R_0/\tan(\alpha))t - R_1\sin(\alpha)\sin(\omega_1 t+\varphi_1) + \Delta z$$

https://doi.org/10.1016/j.jbc.2023.104579

Parametric design of alpha-helical bundles

Knob-into-hole packing of side-chains



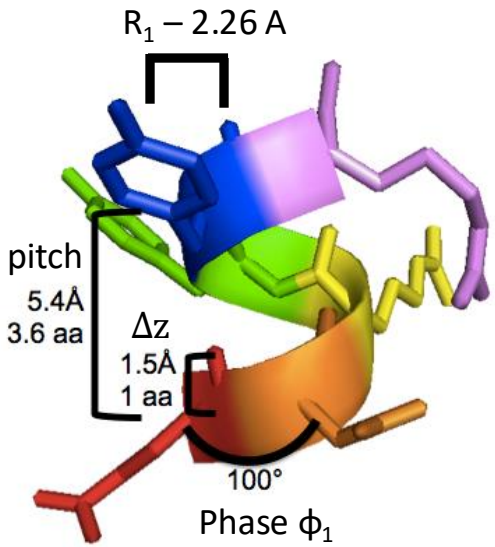
Crick coiled coil equations

For a helix in a coiled coil with supercoil axis along z, the Cartesian coordinates of the Calpha of residue *t* are given by:

$$x = R_0\cos(\omega_0 t+\varphi_0') + R_1\cos(\omega_0 t+\varphi_0')\cos(\omega_1 t+\varphi_1) - R_1\cos(\alpha)\sin(\omega_0 t+\varphi_0')\sin(\omega_1 t+\varphi_1)$$

$$y = R_0\sin(\omega_0 t+\varphi_0') + R_1\sin(\omega_0 t+\varphi_0')\cos(\omega_1 t+\varphi_1) + R_1\cos(\alpha)\cos(\omega_0 t+\varphi_0')\sin(\omega_1 t+\varphi_1)$$

$$z = (\omega_0 R_0/\tan(\alpha))t + R_1\sin(\alpha)\sin(\omega_1 t+\varphi_1) + \Delta z$$

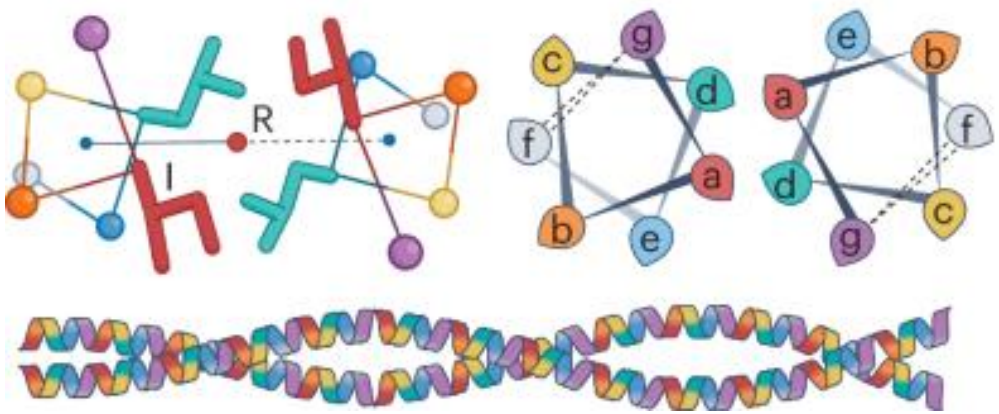


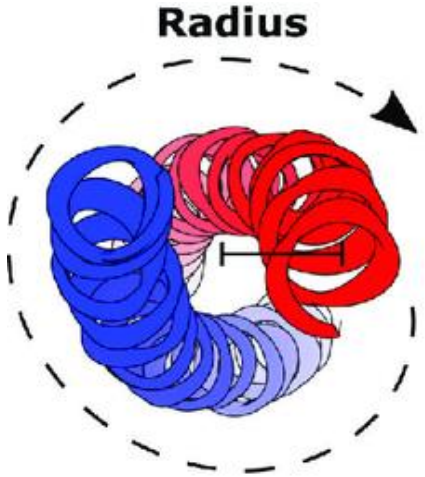
ω_1 – helical twist

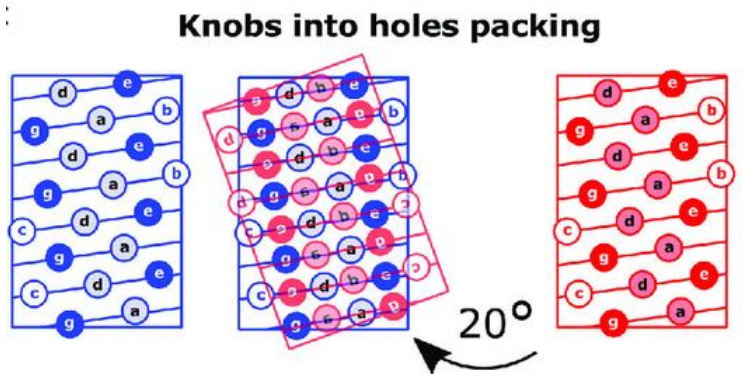
https://doi.org/10.1016/j.jbc.2023.104579

Parametric design of alpha-helical bundles

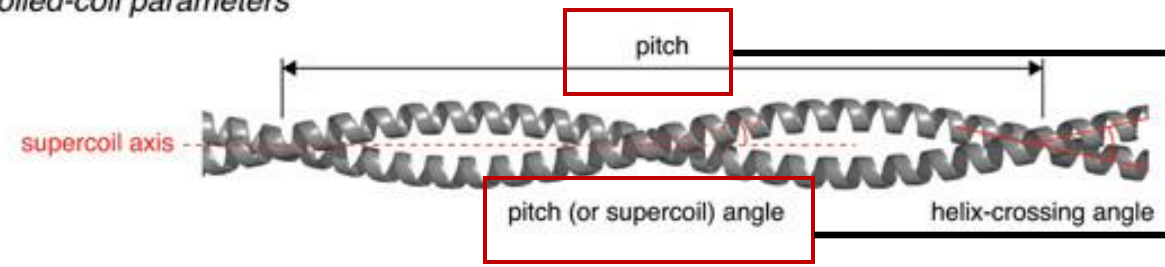
Knob-into-hole packing of side-chains







Coiled-coil parameters



Number of residues to bring the helices back to sync.

Angle between helix and supercoil axis

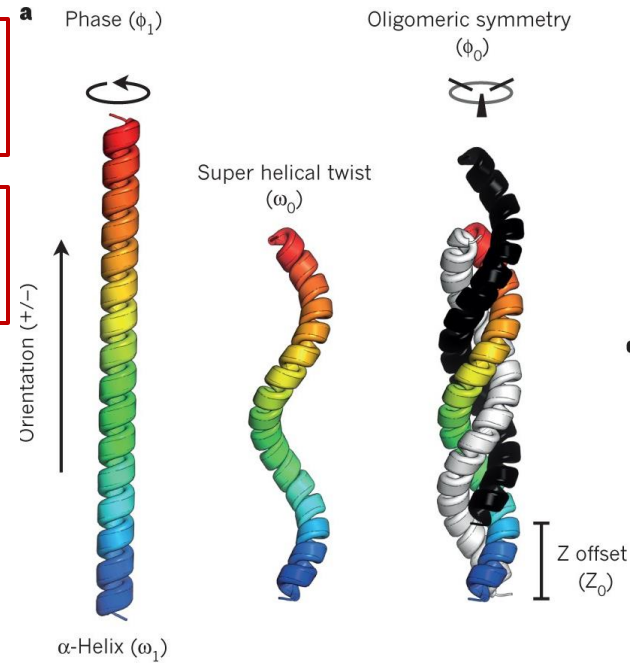
Crick coiled coil equations

For a helix in a coiled coil with supercoil axis along z, the Cartesian coordinates of the Calpha of residue *t* are given by:

$$x = R_0\cos(\omega_0 t+\varphi_0') + R_1\cos(\omega_0 t+\varphi_0')\cos(\omega_1 t+\varphi_1) - R_1\cos(\alpha)\sin(\omega_0 t+\varphi_0')\sin(\omega_1 t+\varphi_1)$$

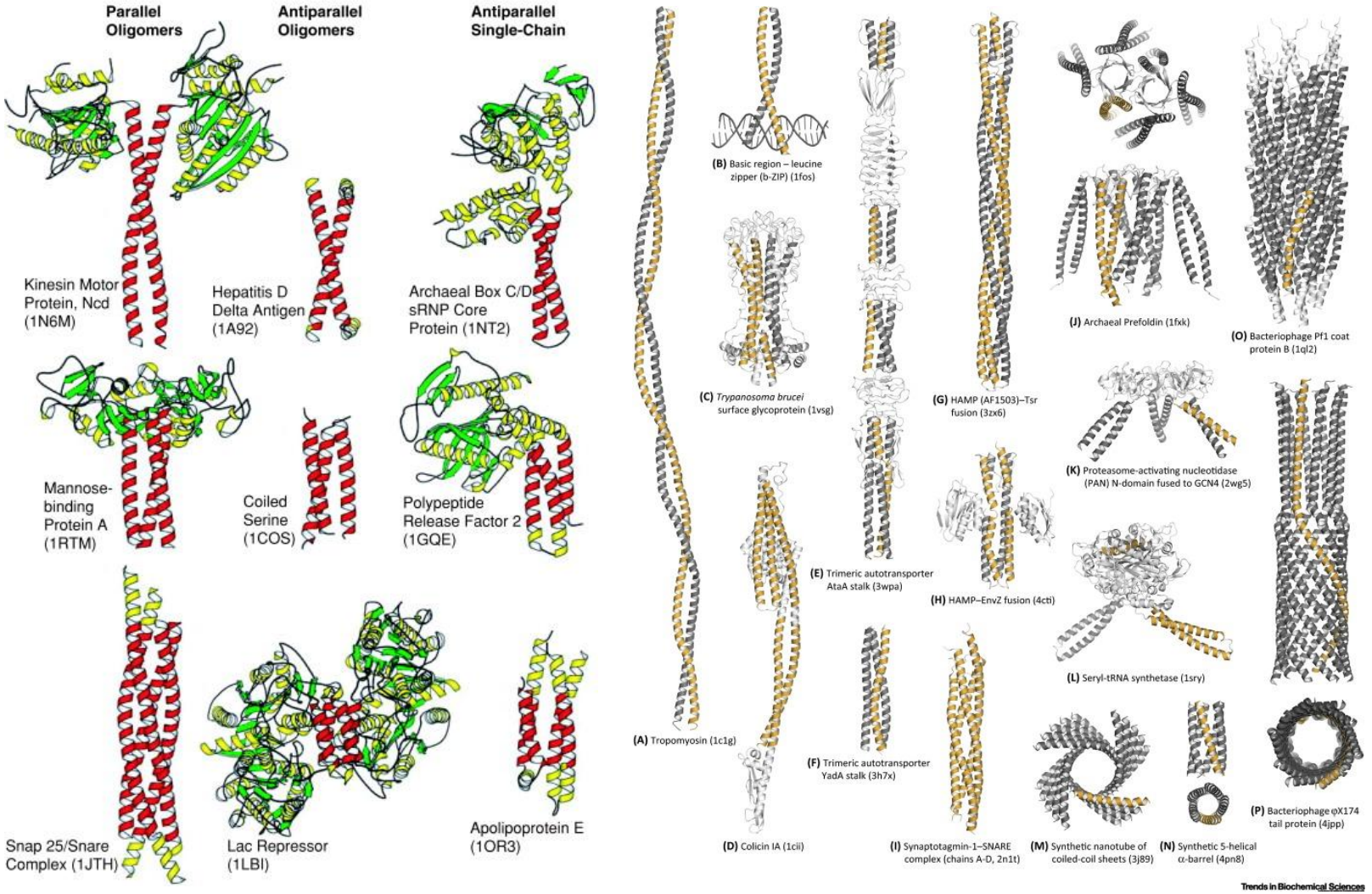
$$y = R_0\sin(\omega_0 t+\varphi_0') + R_1\sin(\omega_0 t+\varphi_0')\cos(\omega_1 t+\varphi_1) + R_1\cos(\alpha)\cos(\omega_0 t+\varphi_0')\sin(\omega_1 t+\varphi_1)$$

$$z = (\omega_0 R_0/\tan(\alpha))t - R_1\sin(\alpha)\sin(\omega_1 t+\varphi_1) + \Delta z$$

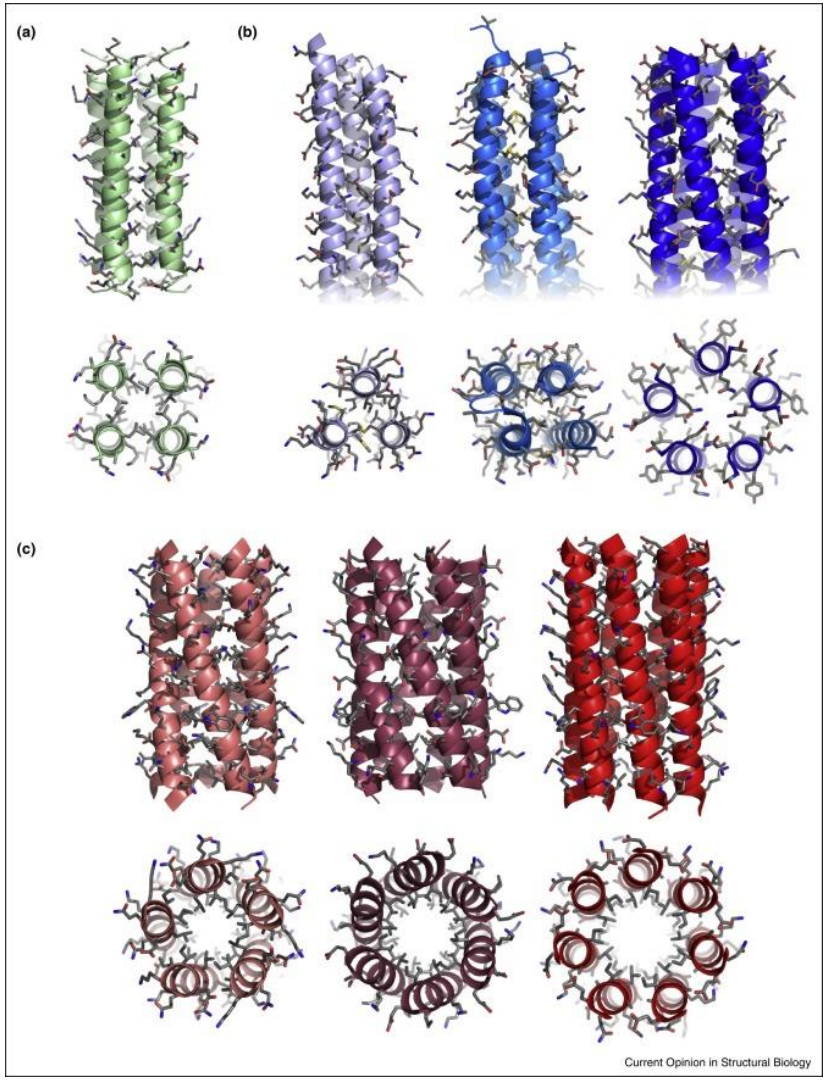


https://doi.org/10.1002/bies.201600062

Natural *vs* designed alpha-helical bundles

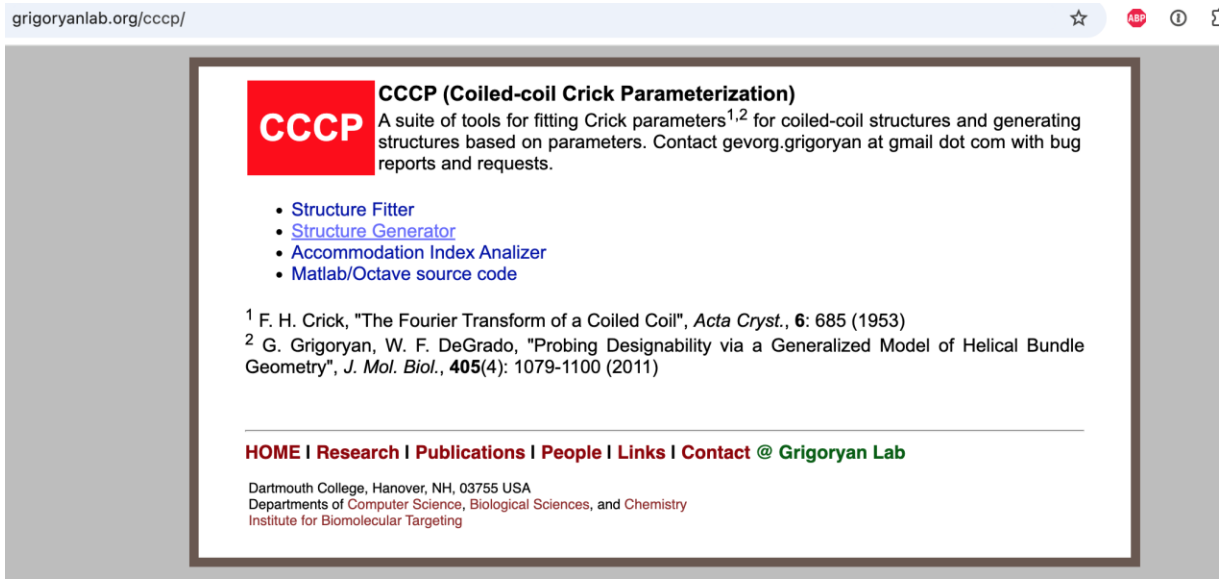


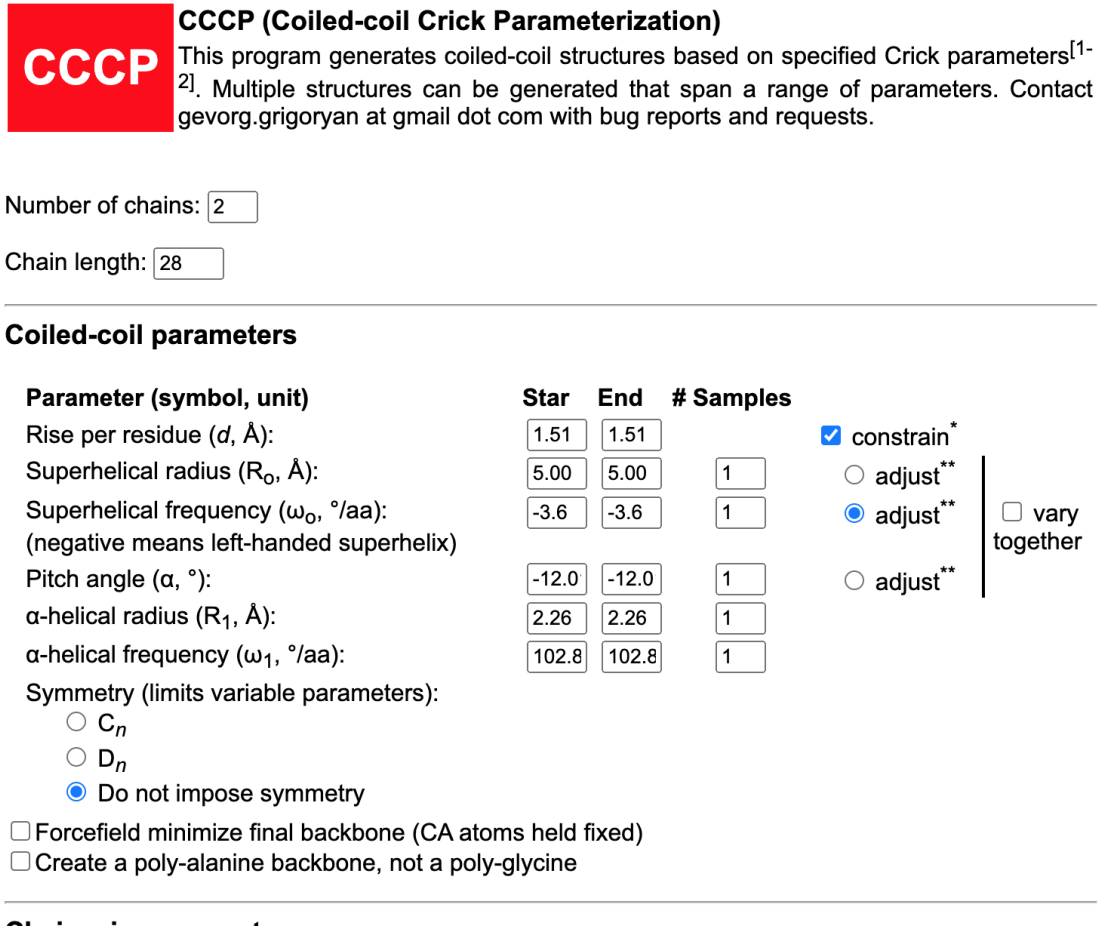
Natural helical bundles and coiled-coil domains hidden in proteins



De novo designed coiled coils and bundles

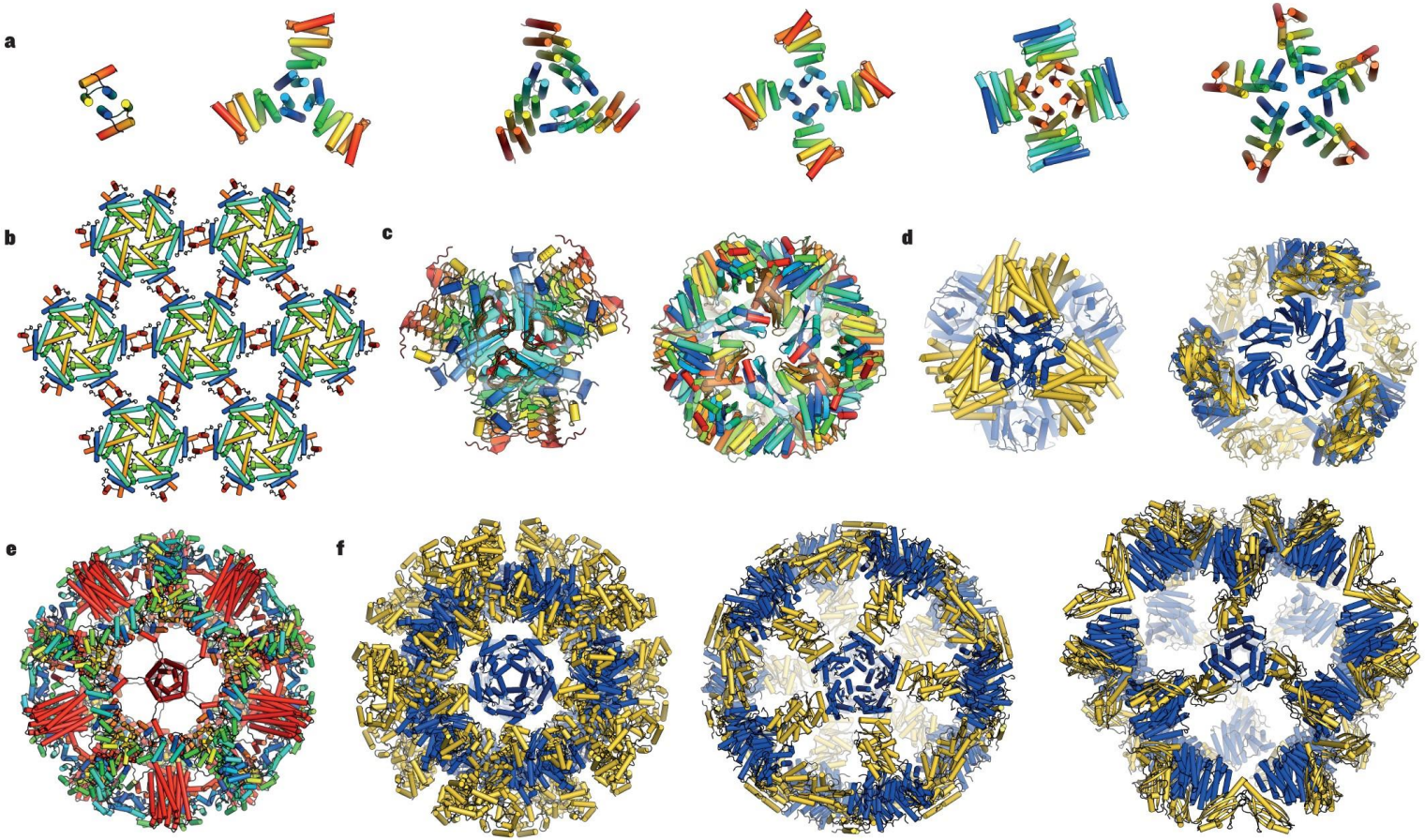
Generating coiled coils with ideal parameters - CCCP

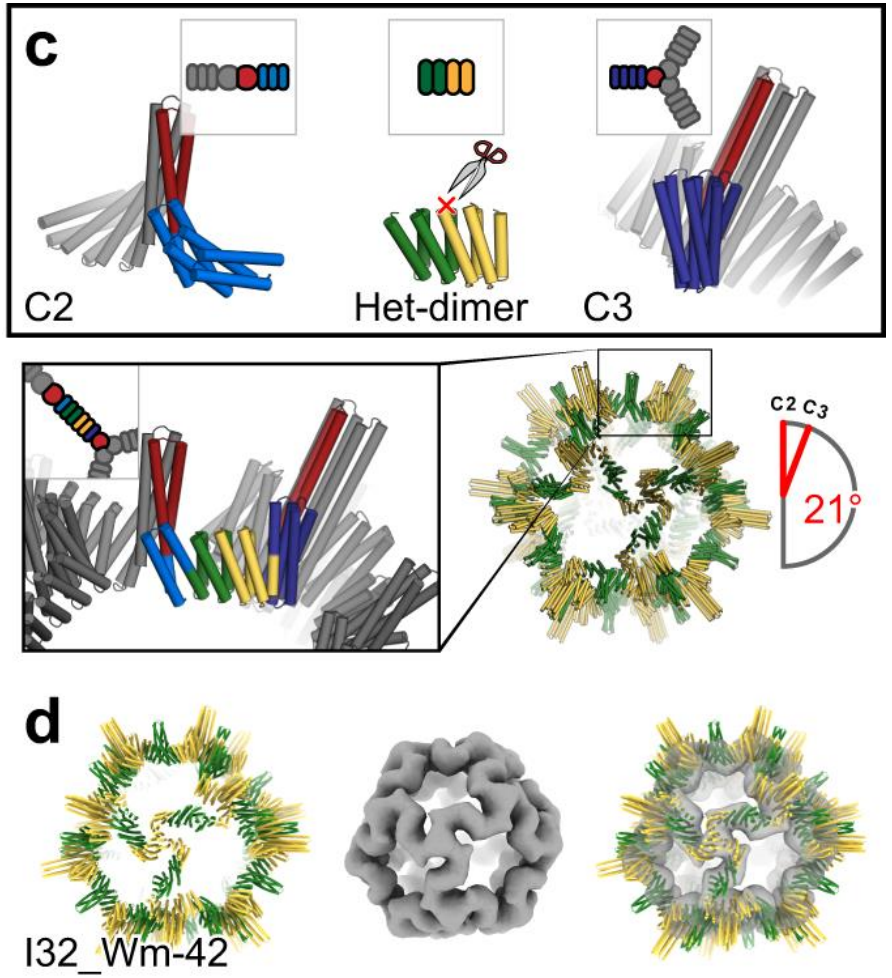




Server: https://grigoryanlab.org/cccp/

Example 1: alpha-helical bundles as building units



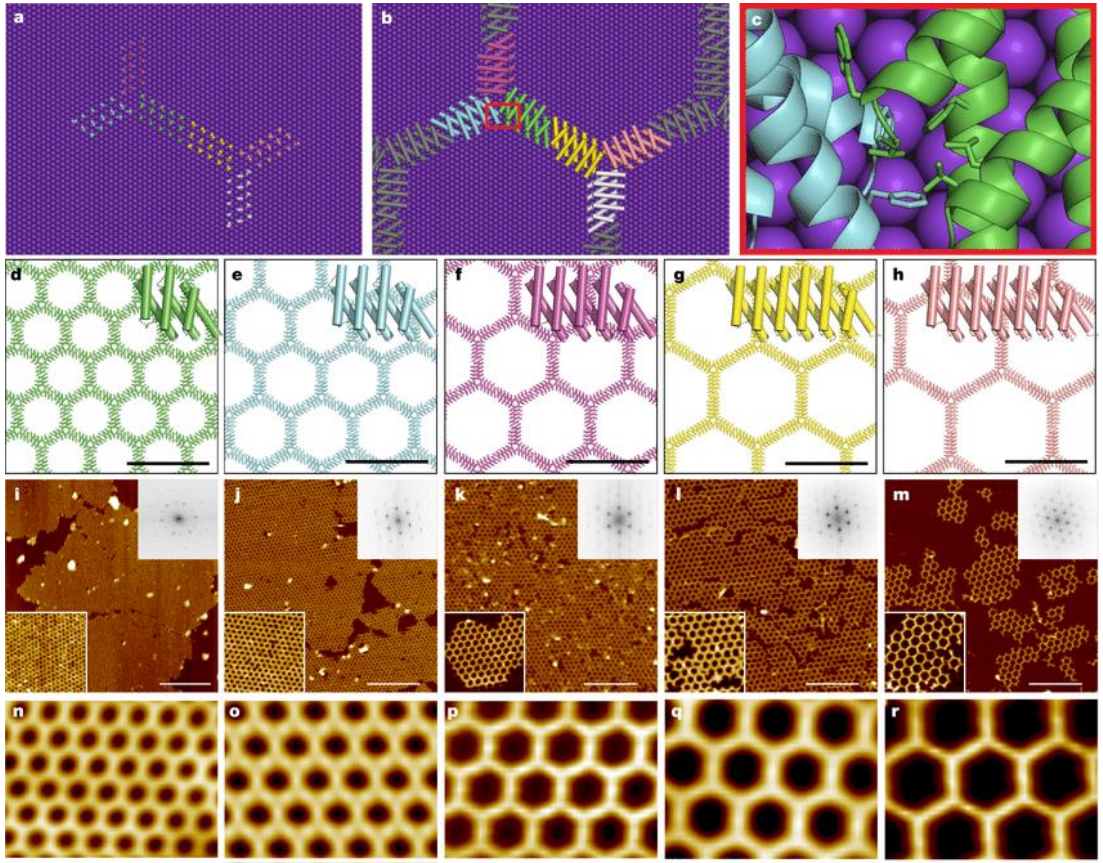


https://doi.org/10.1038/nature19946

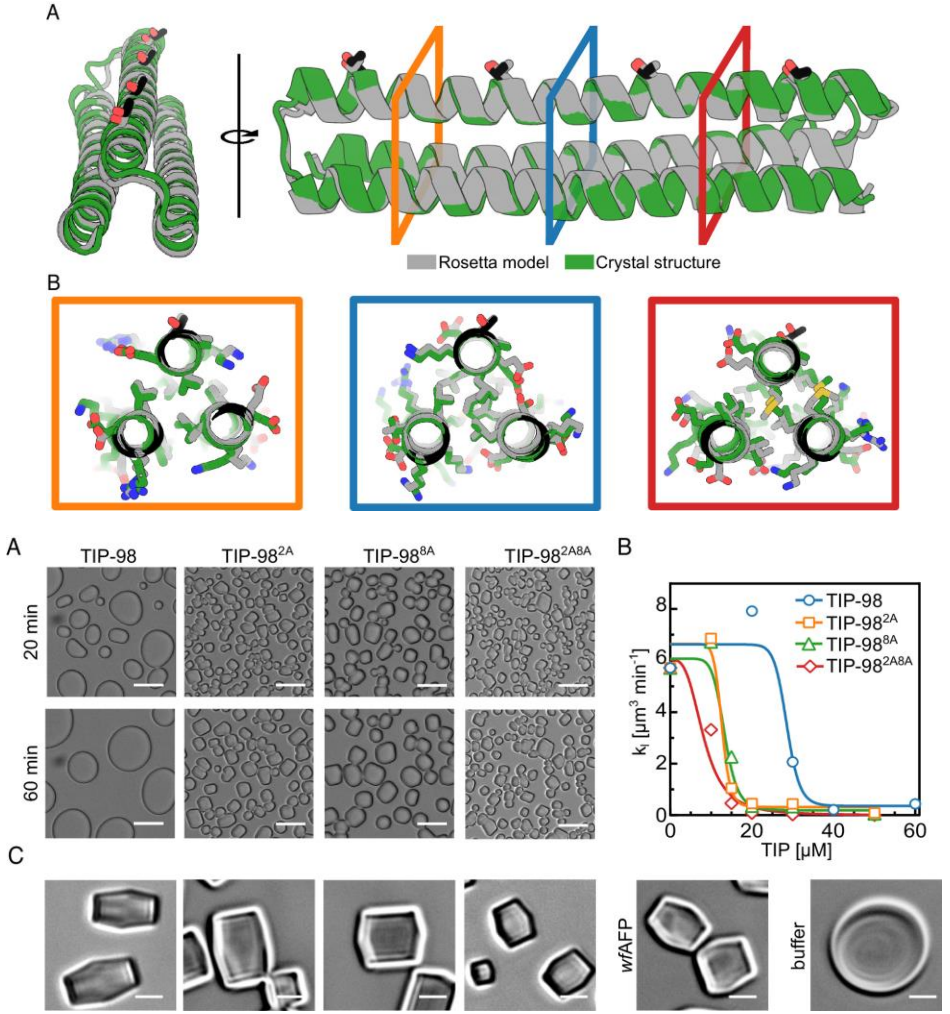
https://doi.org/10.1038/s41467-021-22276-z

Example 2: latice-matching proteins

Mica-binding proteins and self-assembling material



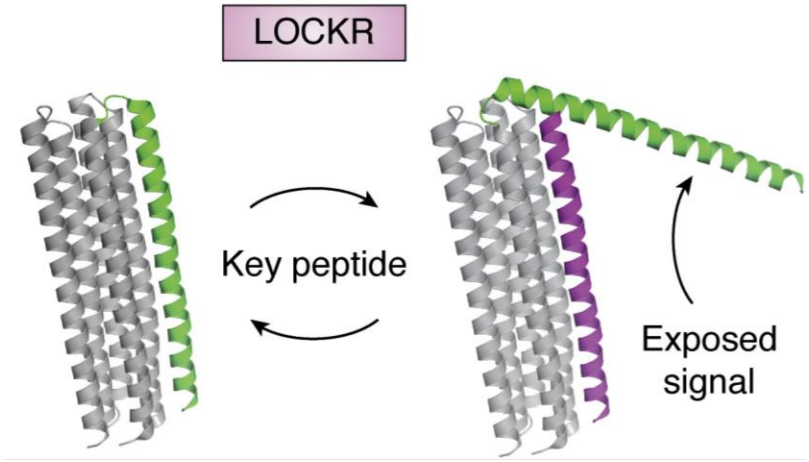
Ice crystal-shaping proteins



https://www.nature.com/articles/s41586-019-1361-6

https://doi.org/10.1073/pnas.2220380120

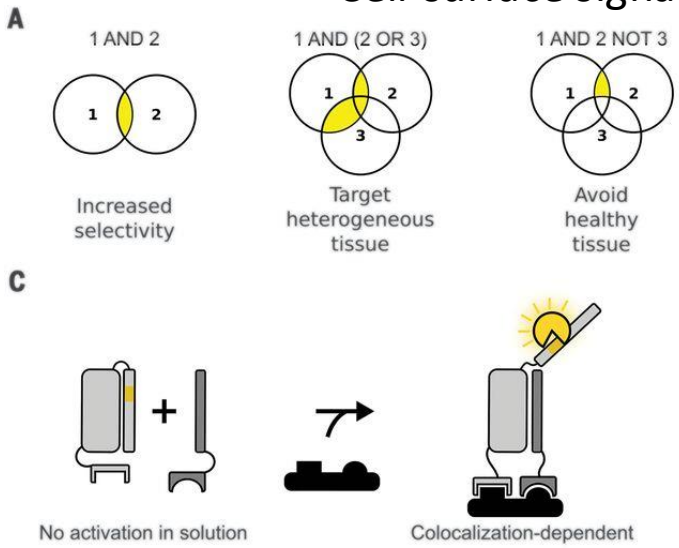
Example 3: protein logic and switches

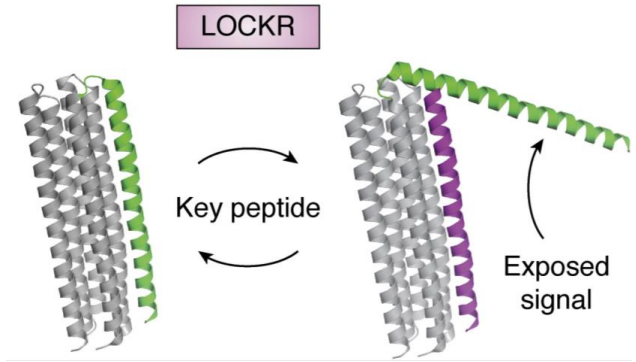


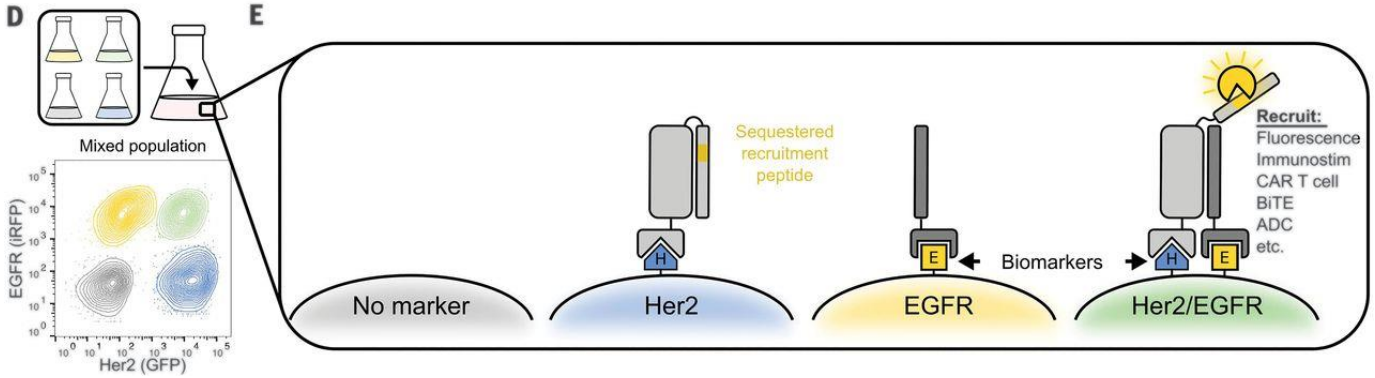
https://www.nature.com/articles/s41586-019-1361-6

Example 3: protein logic and switches

Cell-surface signalling logic





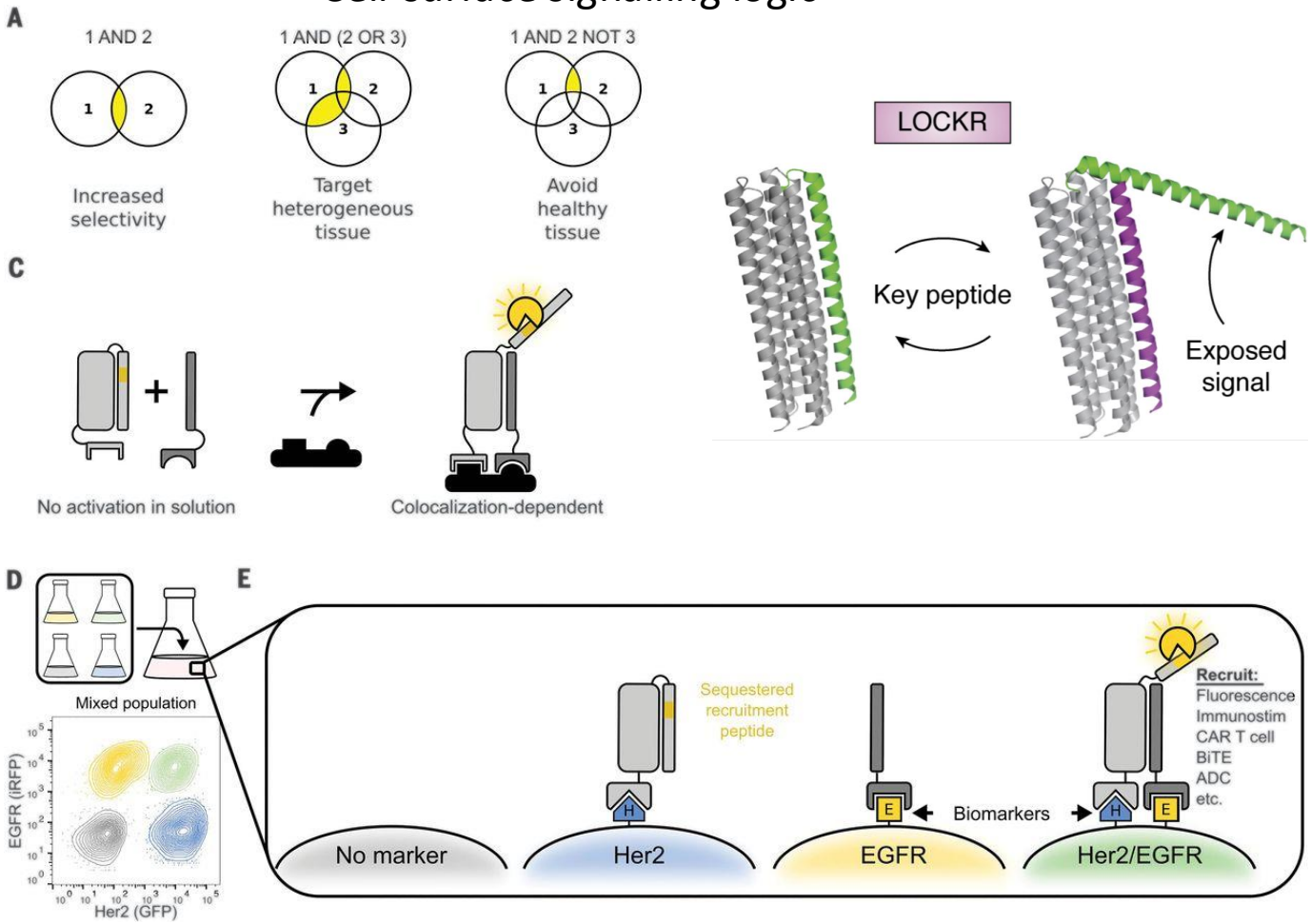


https://doi.org/10.1038/s41586-021-03258-z

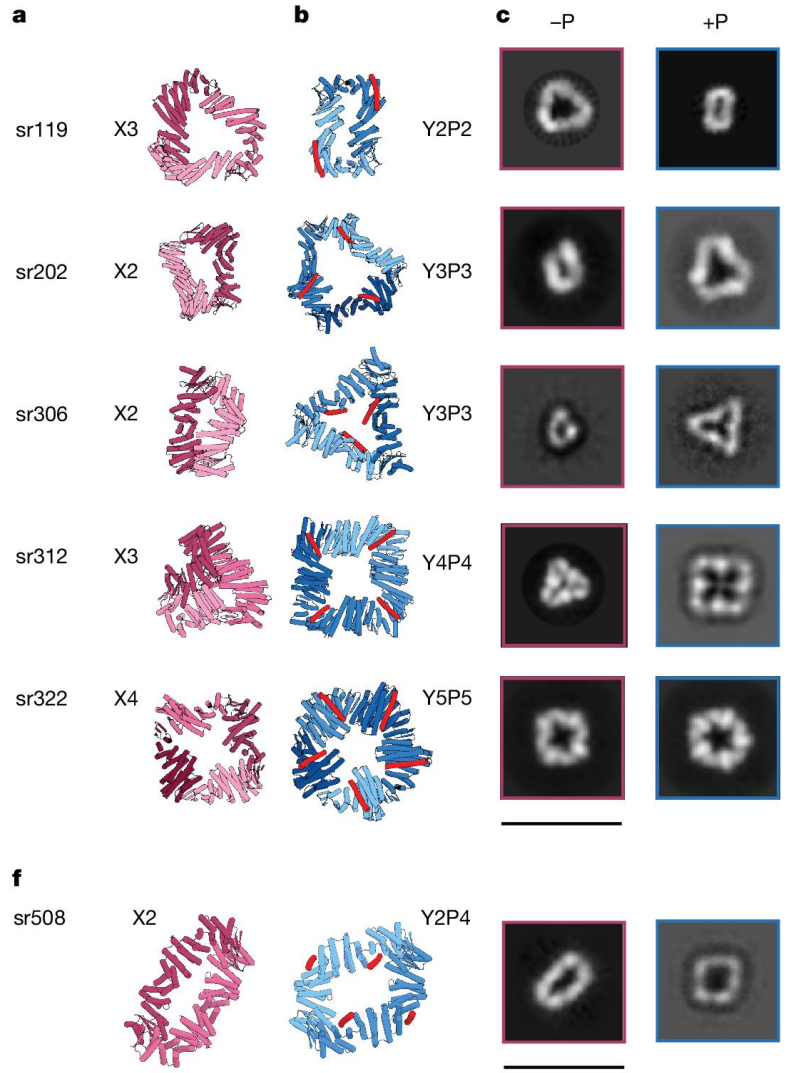
https://doi.org/10.1126/science.aba6527

Example 3: protein logic and switches

Cell-surface signalling logic



Switchable assemblies

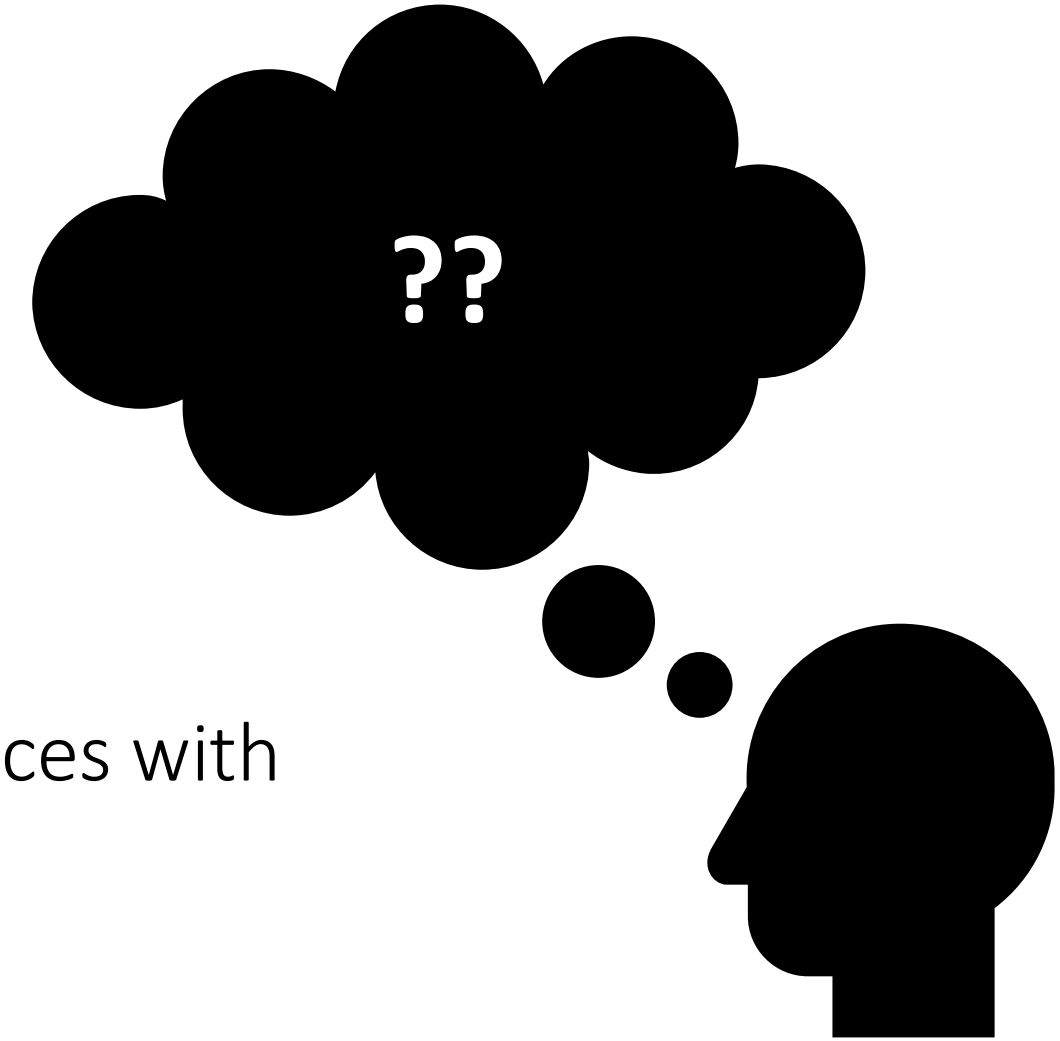


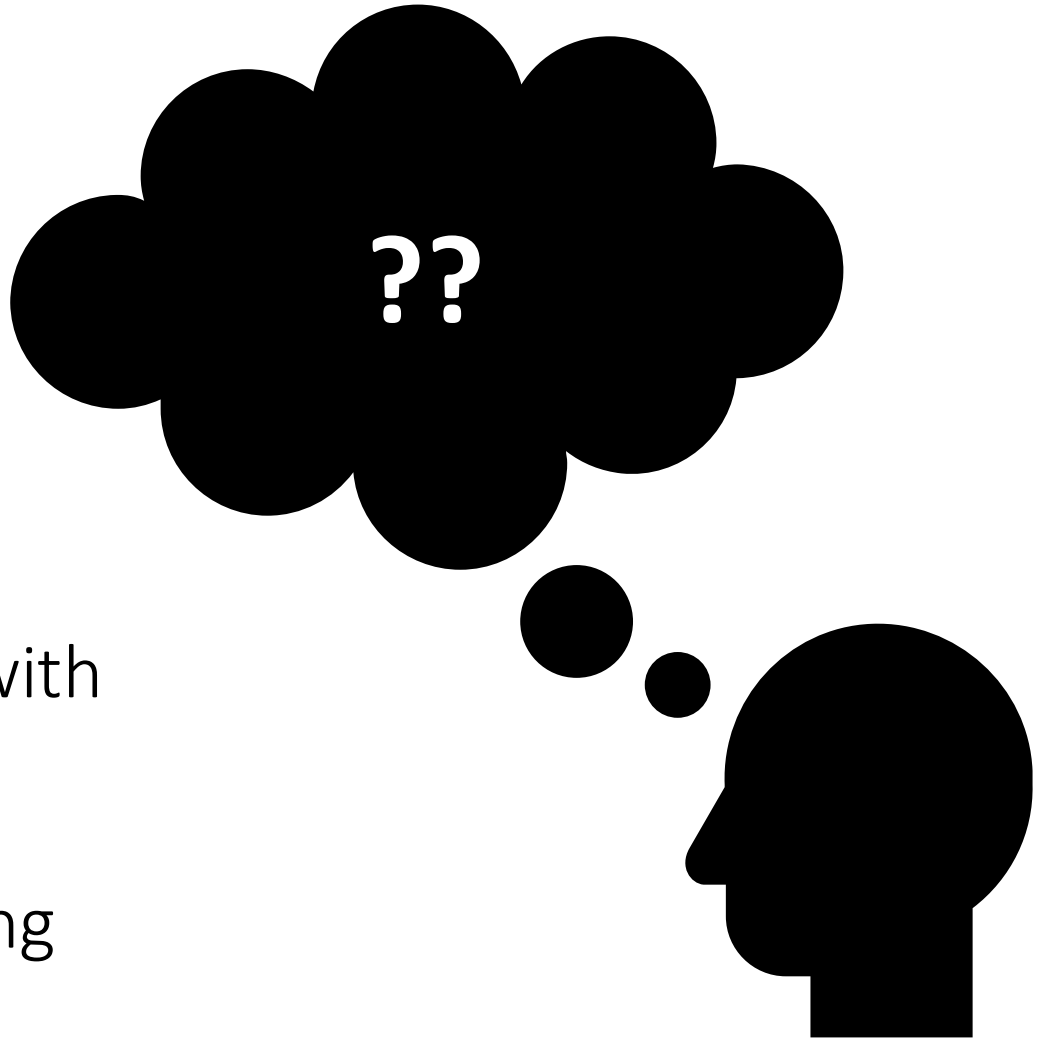
https://doi.org/10.1038/s41586-021-03258-z

https://doi.org/10.1126/science.aba6527

https://doi.org/10.1038/s41586-024-07813-2

How to generate diverse sequences with parametric design?



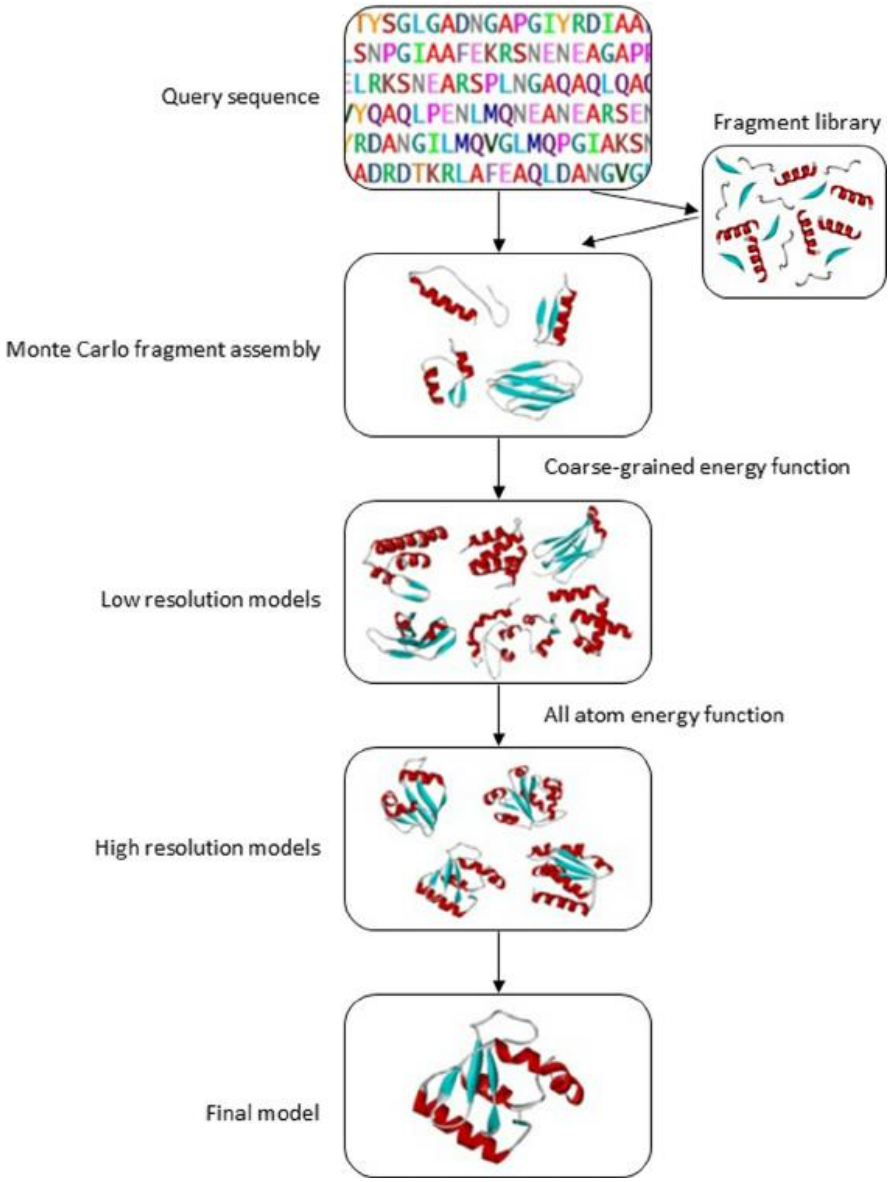


How to generate diverse sequences with parametric design?

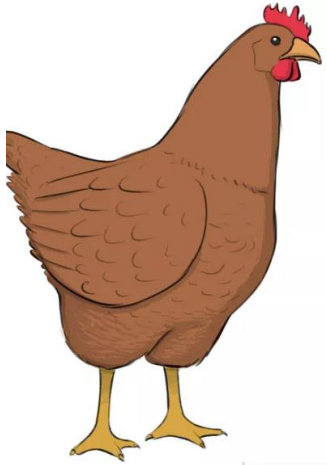
➔ Superhelical parameters fine-tuning

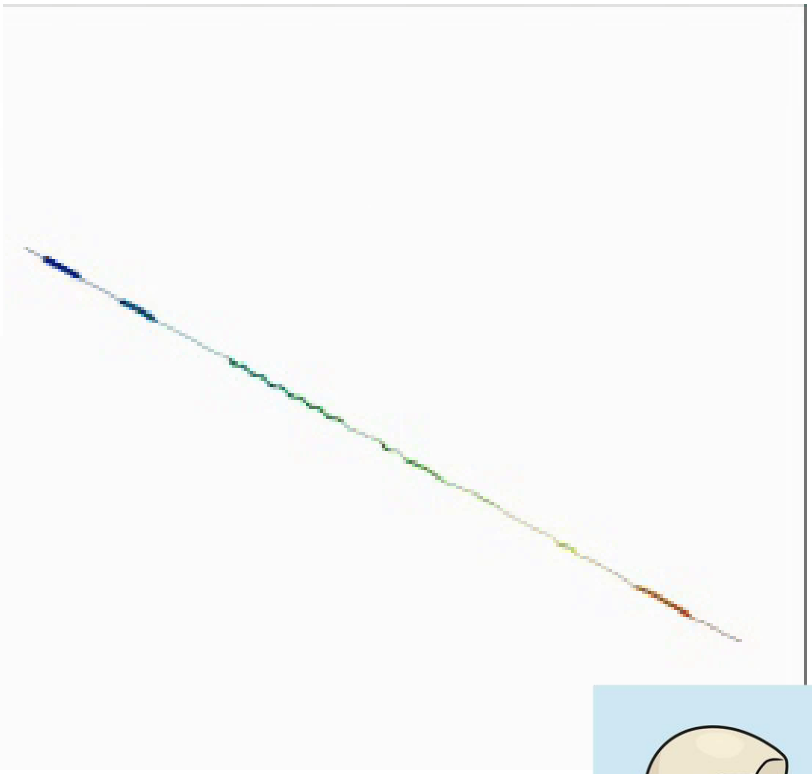
BUT … parameteric designs mostly unsuccessful for beta-sheet proteins

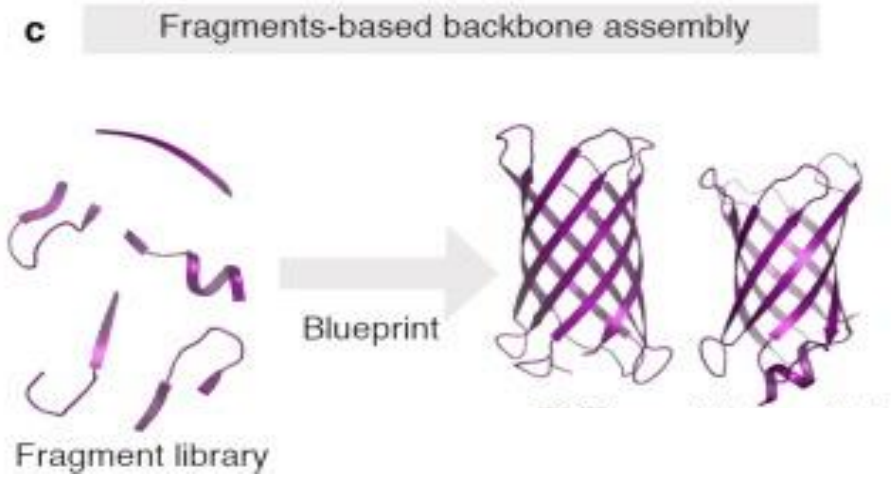
“Stealing” concepts from structure prediction

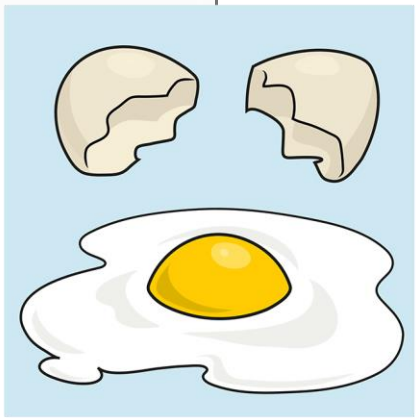


“Stealing” concepts from structure prediction



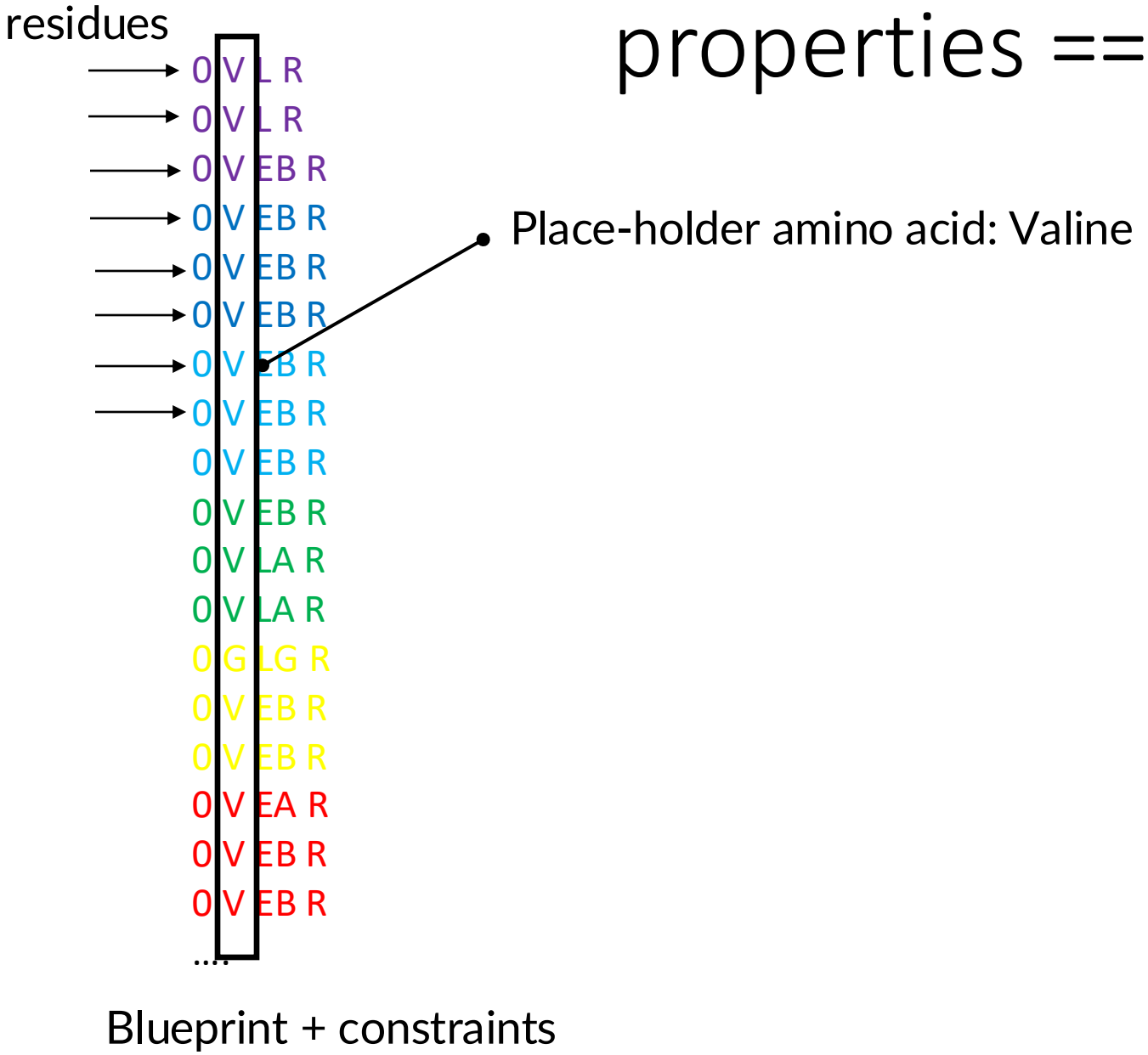






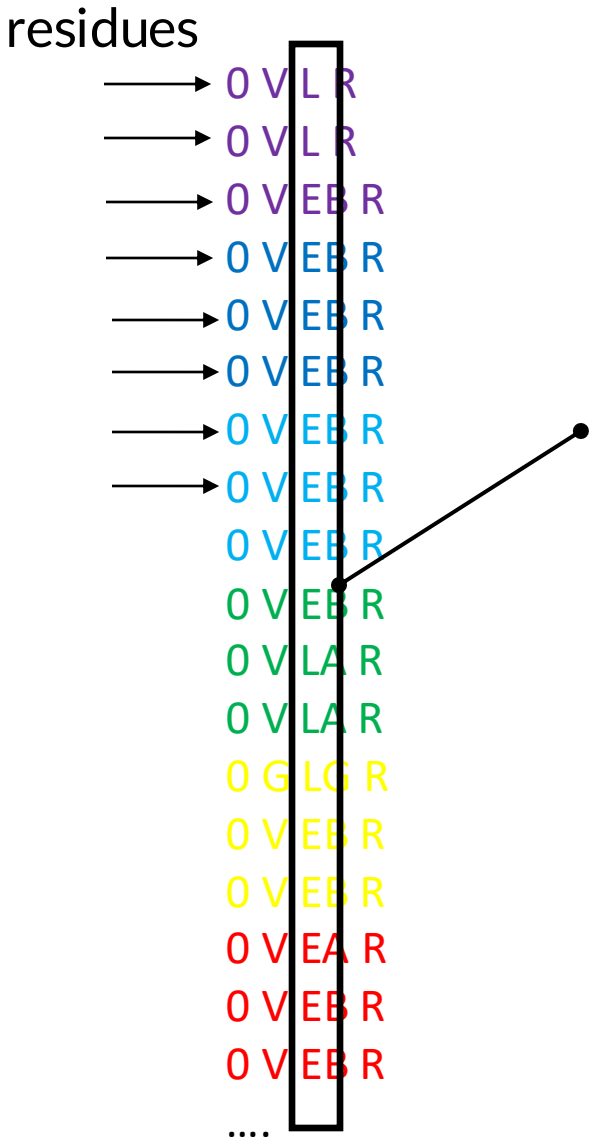
How can we pick fragments if the sequence is unknown?

Picking fragments based on local structure properties == Blueprint



Blueprint + constraints

Picking fragments based on local structure properties == Blueprint



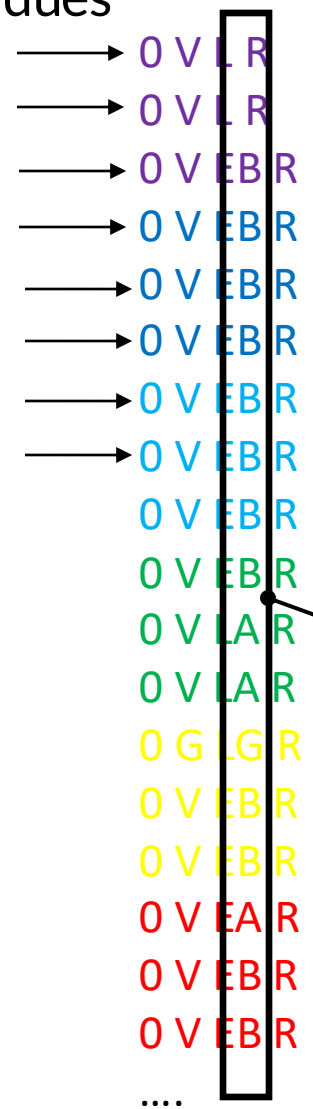
Blueprint + constraints

Place-holder amino acid: Valine

Secondary structure:
E = Beta-strand
H = Alpha-helix
L = Loop

Picking fragments based on local structure properties == Blueprint

residues

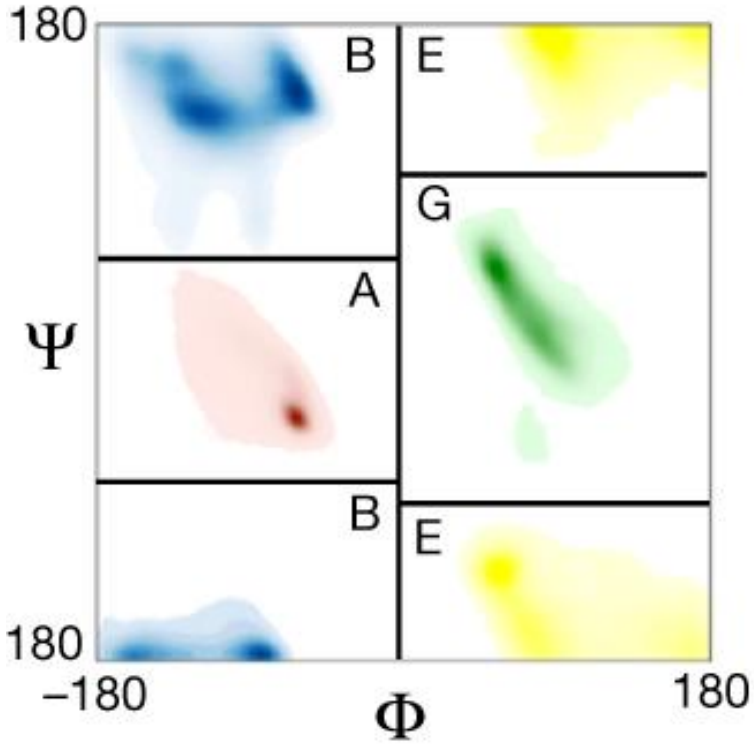


Blueprint + constraints

Place-holder amino acid: Valine

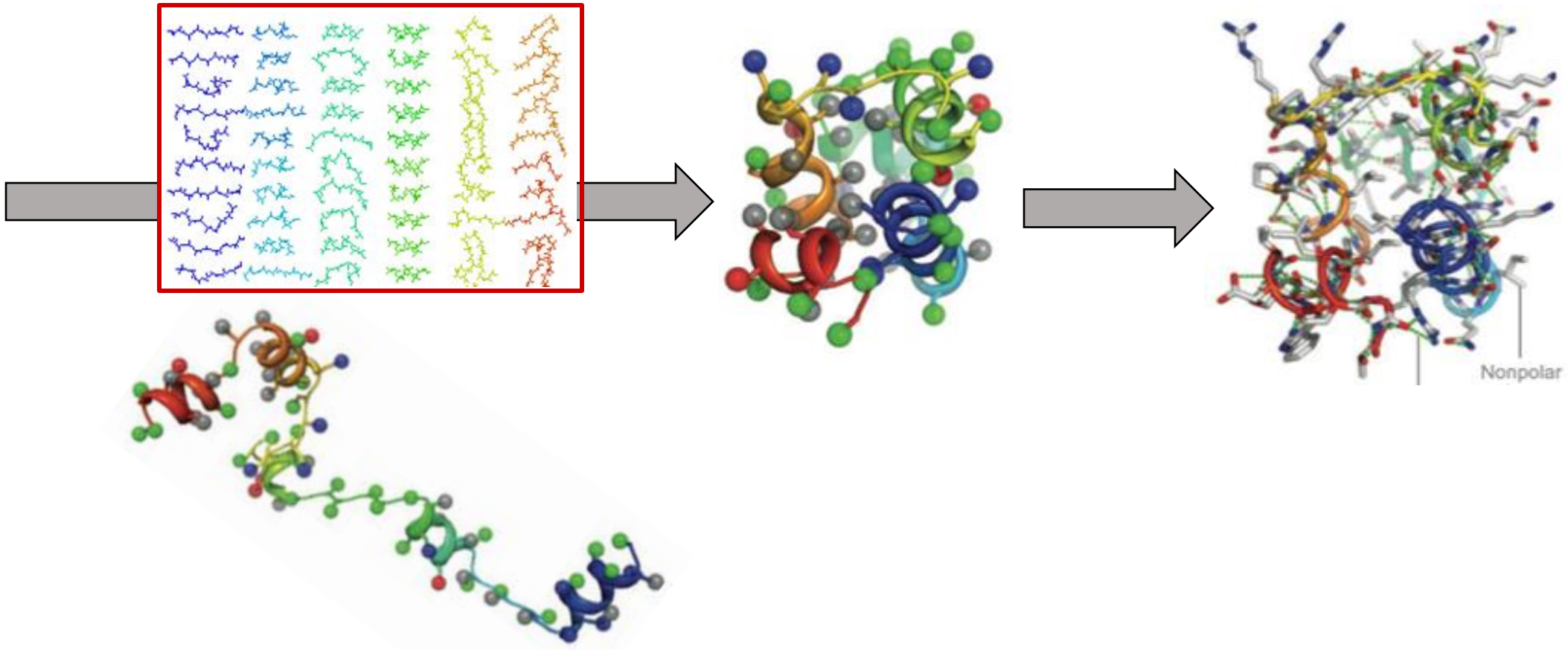
Secondary structure:
E = Beta-strand
H = Alpha-helix
L = Loop

ABEGO type
"O" = cis-peptide bond



Picking fragments based on local structure properties == Blueprint

0 V L R
0 V L R
0 V EB R
0 V EB R
0 V EB R
0 V EB R
0 V EB R
0 V EB R
0 V EB R
0 V EB R
0 V LA R
0 V LA R
0 G LG R
0 V EB R
0 V EB R
0 V EA R
0 V EB R
0 V EB R
....



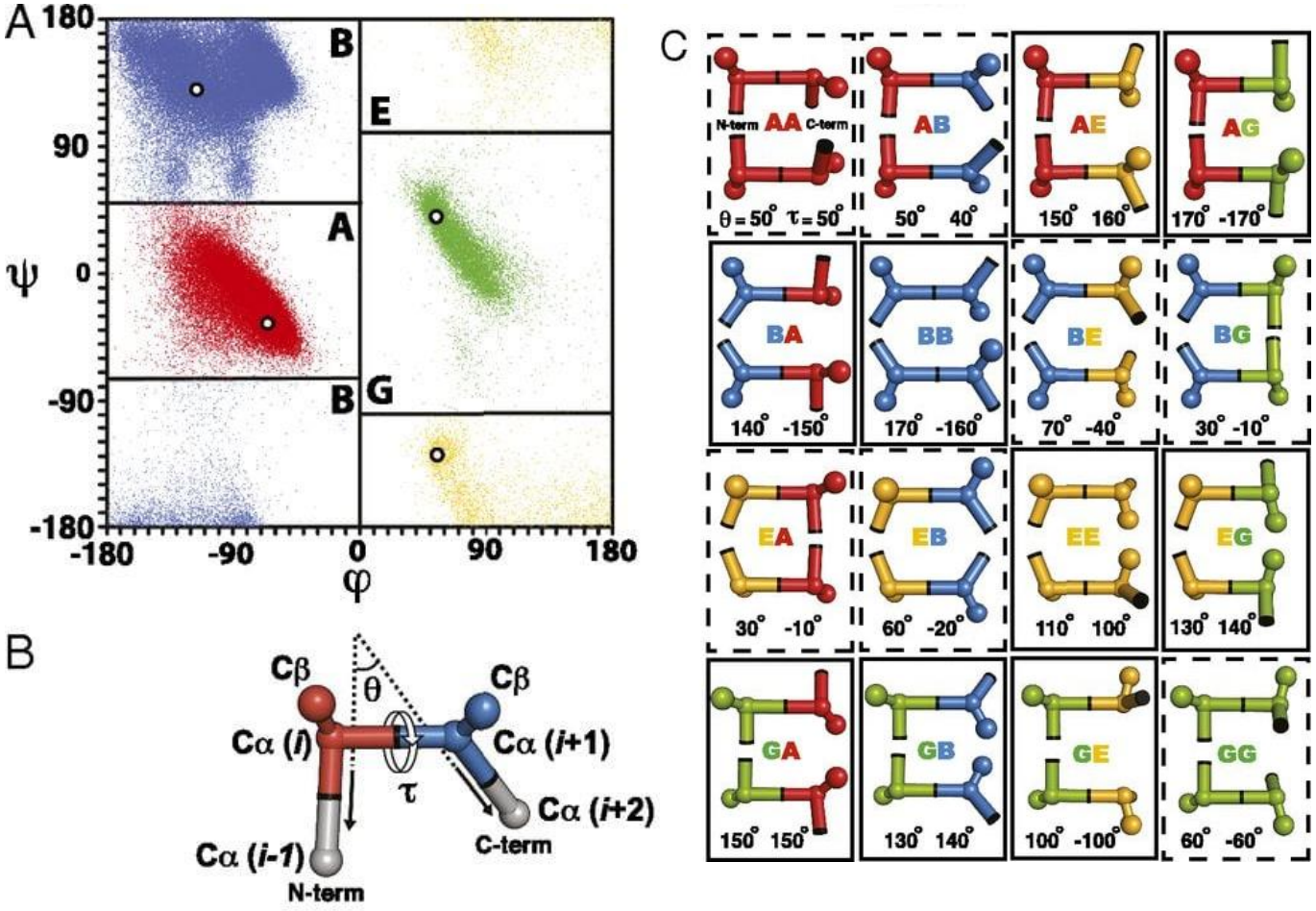
Blueprint + constraints

Ab initio folding

Sequence design

Generating complex protein folds

Controlling loop structures …

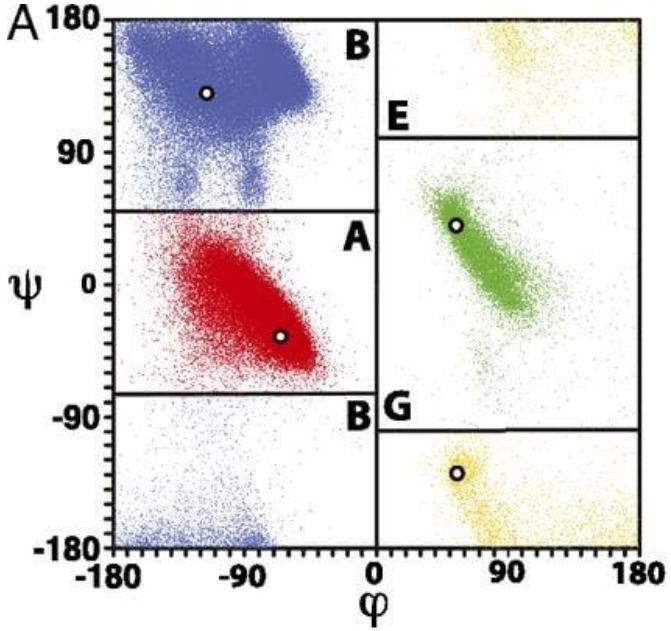


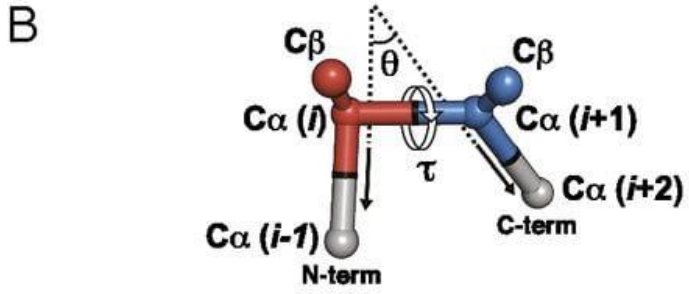
https://doi.org/10.1073/pnas.1509508112

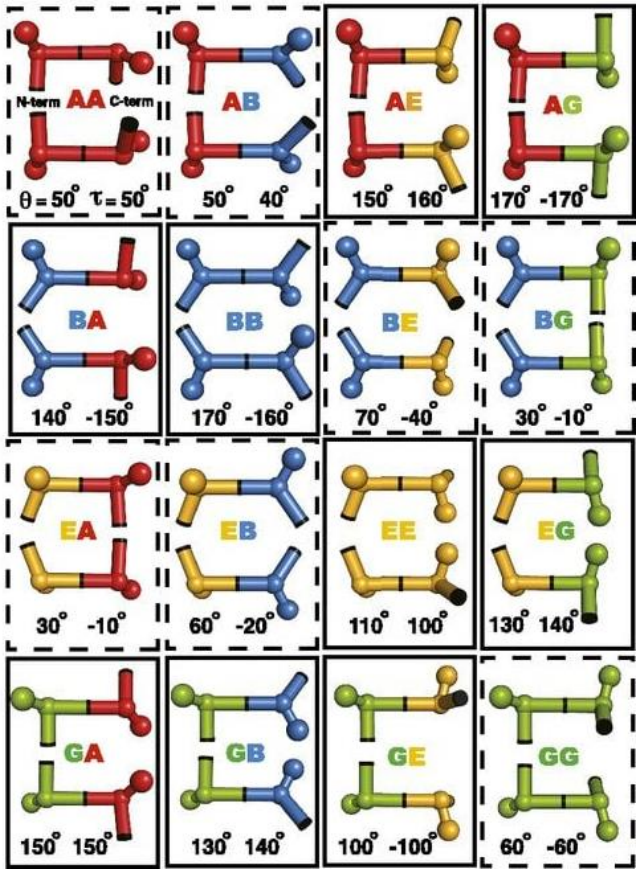
Generating complex protein folds

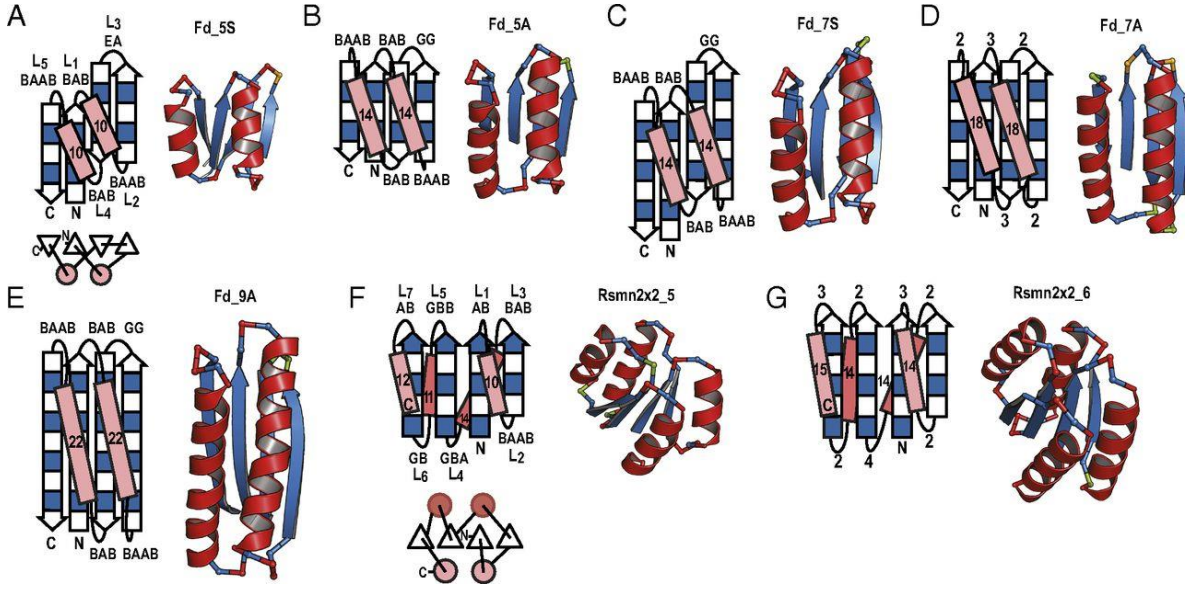
Controlling loop structures …

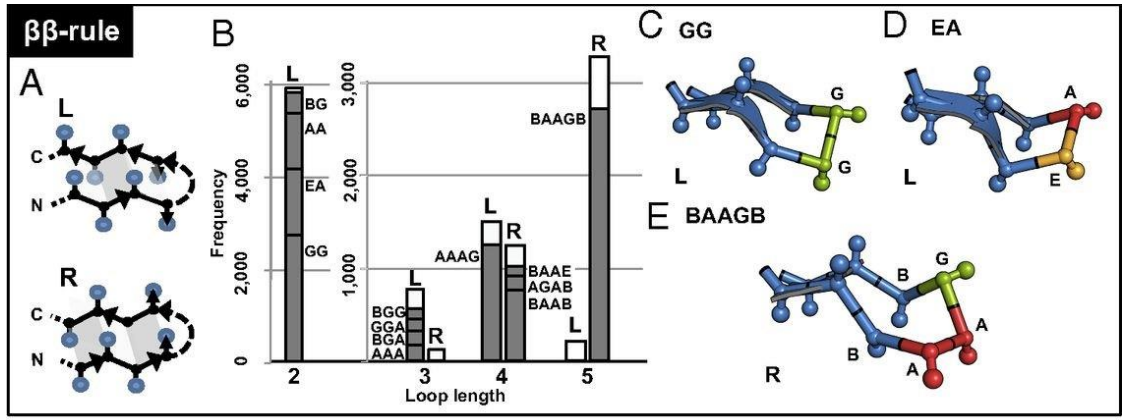
… to generate more designable backbones





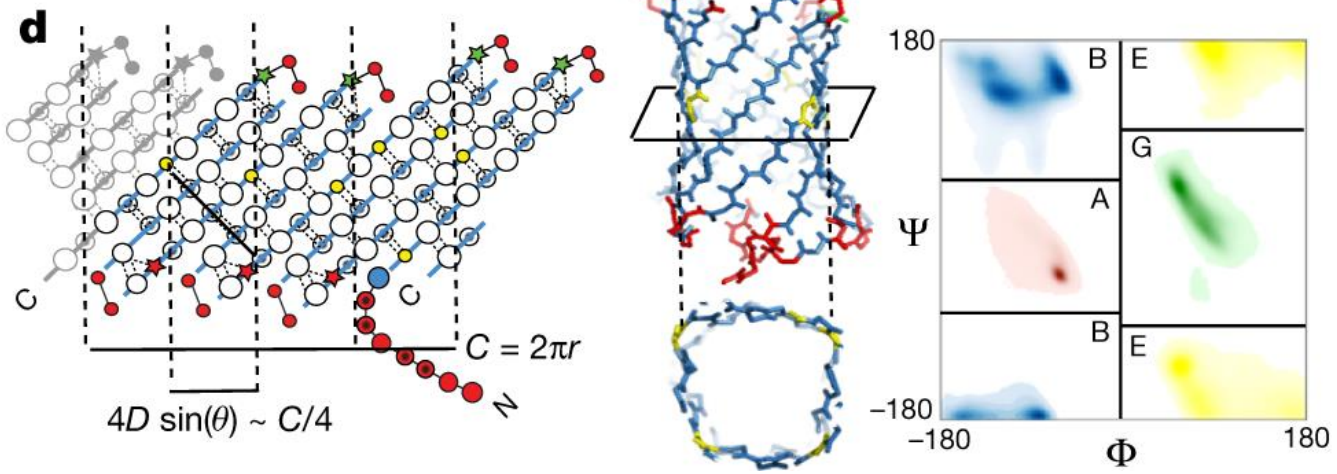


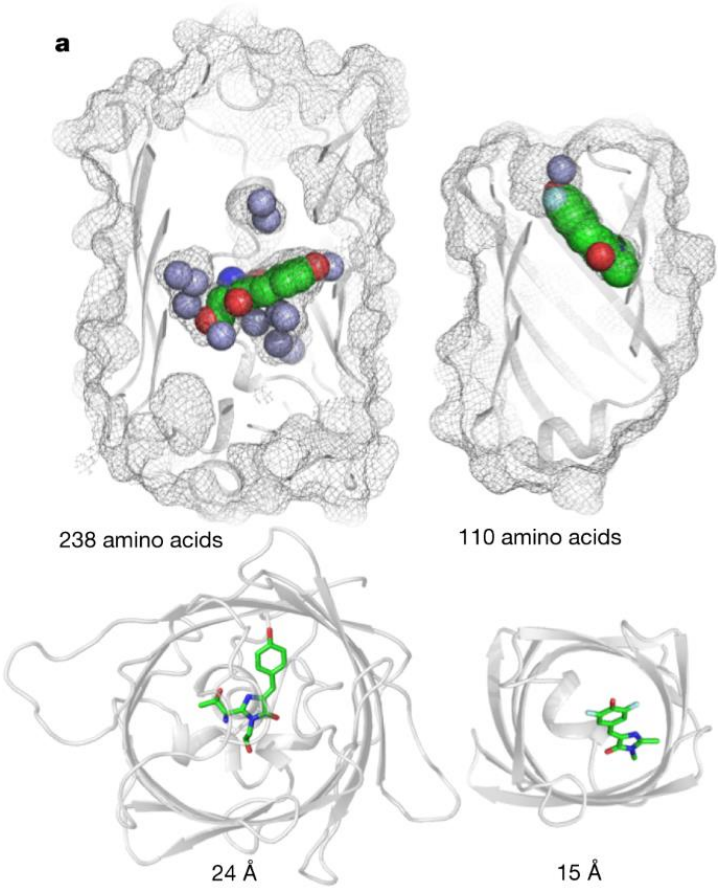


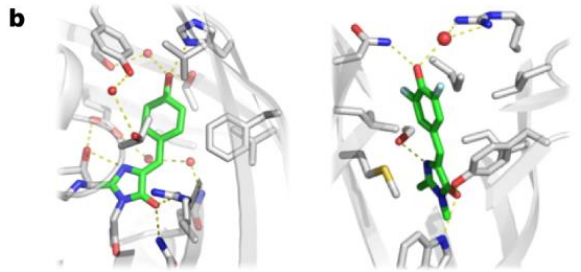


https://doi.org/10.1073/pnas.1509508112

Generating complex protein folds

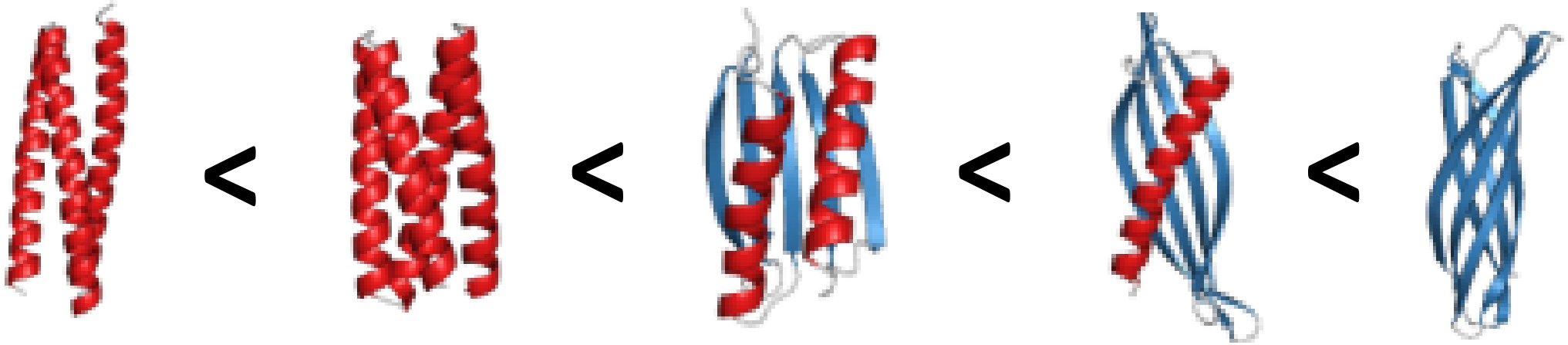






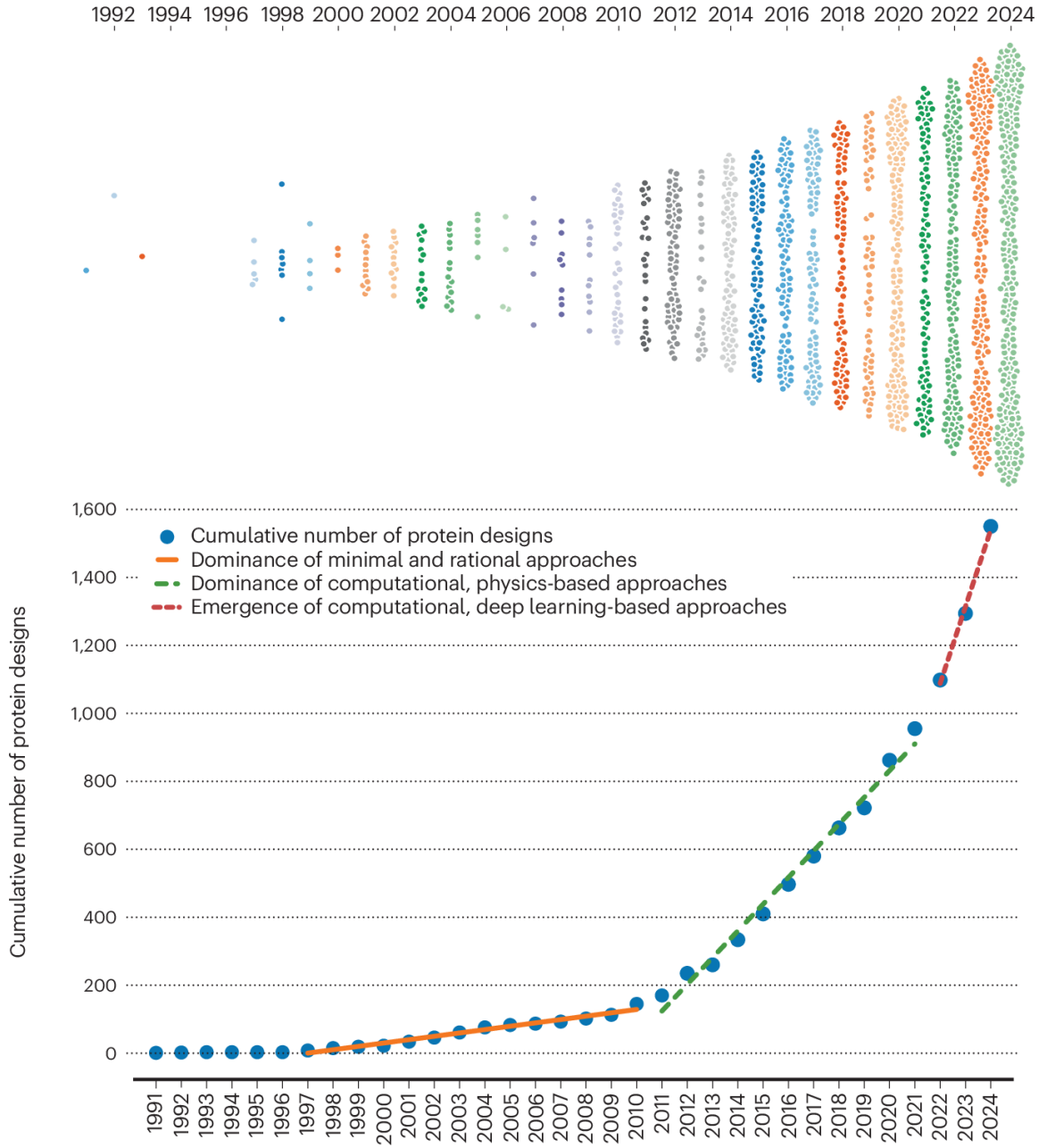
https://doi.org/10.1038/s41586-018-0509-0

De novo design: remaining challenges



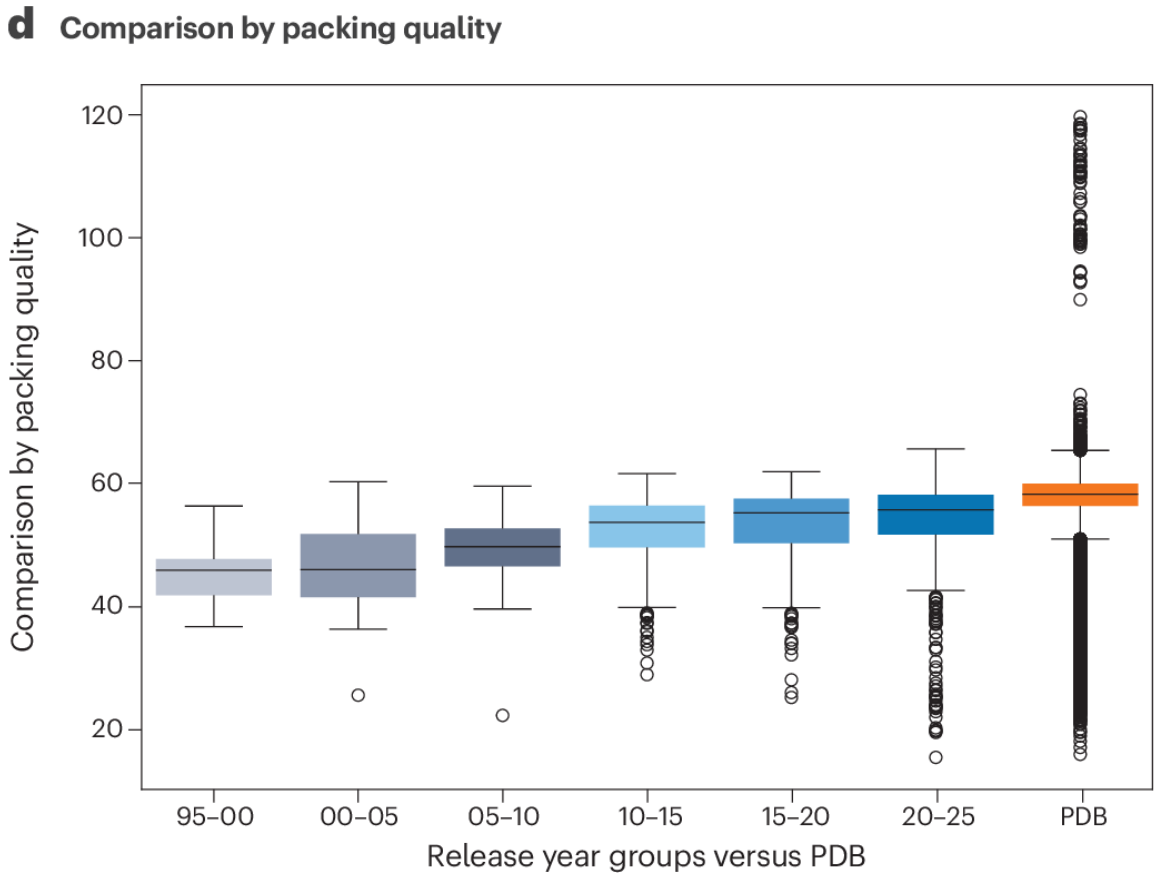
Increases with:
- Size
- Long-range interactions
- Beta-sheet content
- Membrane proteins require specific energy functions

Evolution of *de novo* protein design



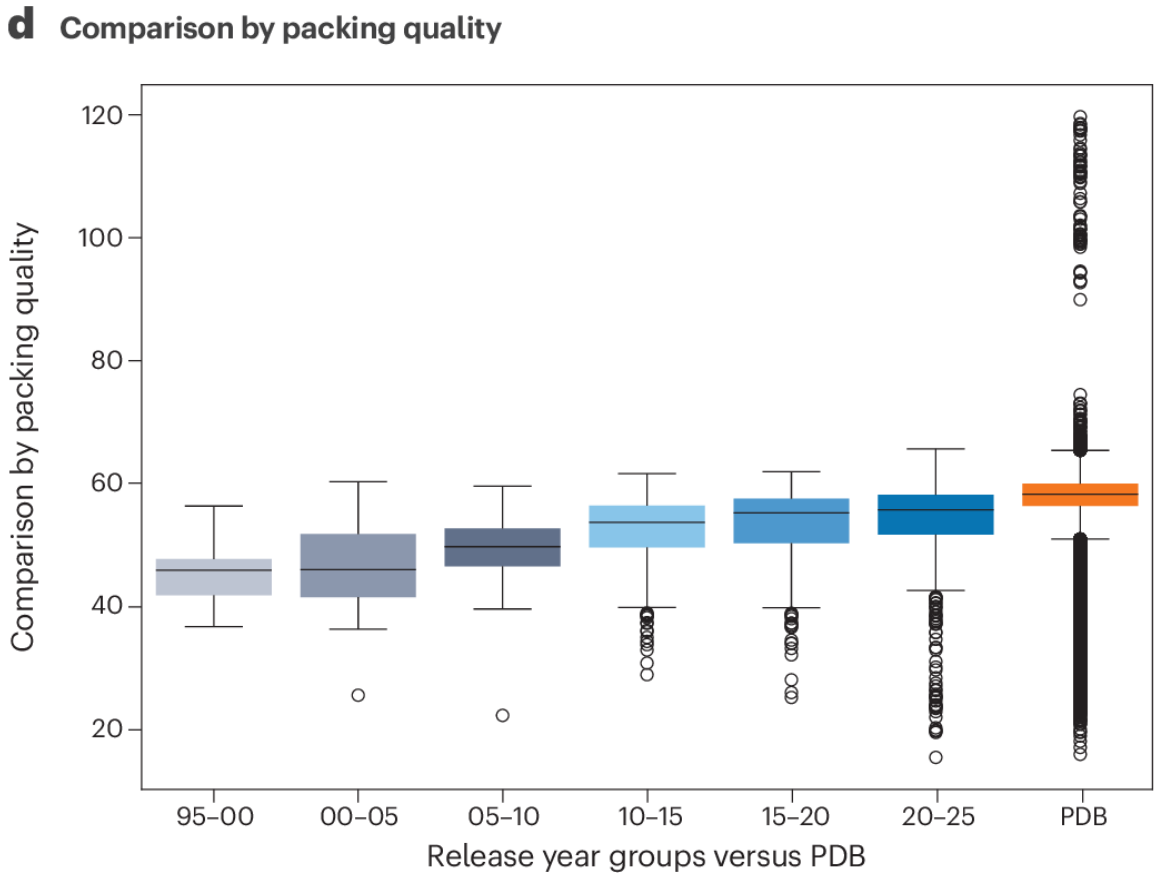
https://doi.org/10.1038/s41587-025-02607-x

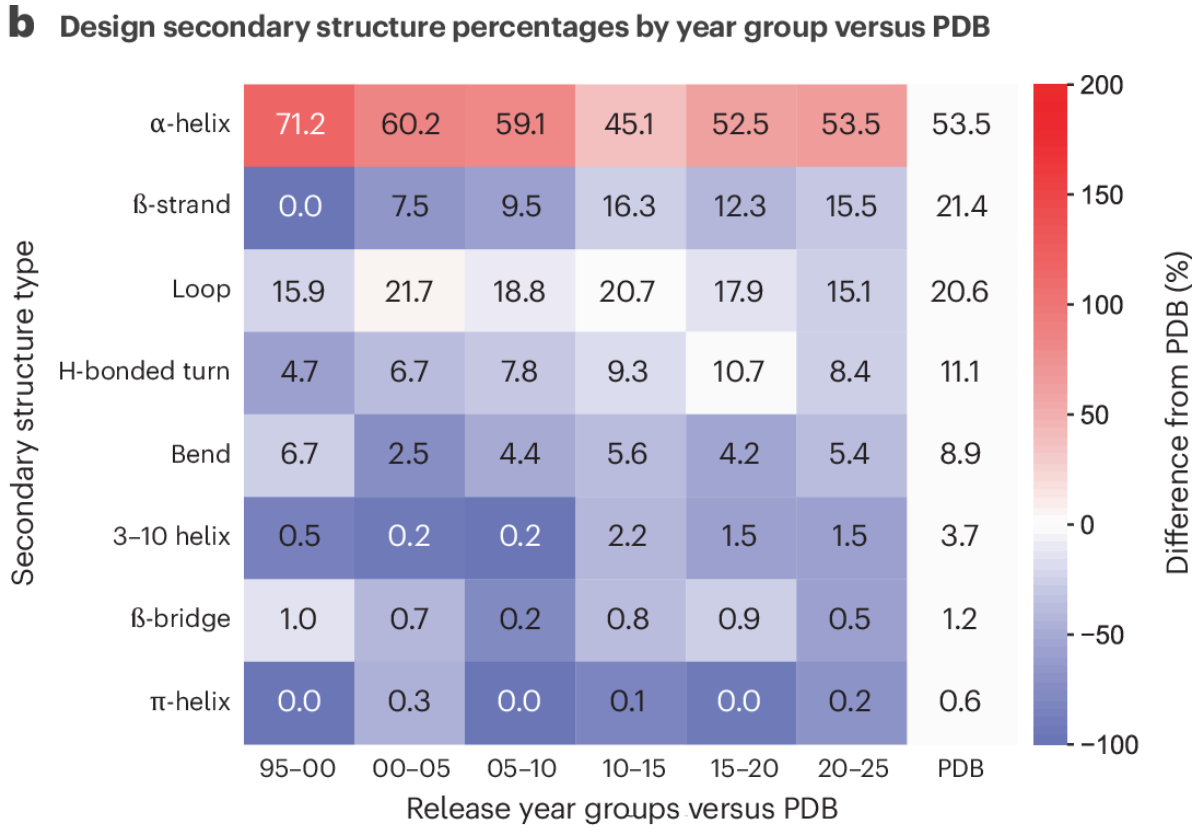
De novo designed folds have better packing quality (more *designable*) …



https://doi.org/10.1038/s41587-025-02607-x

… but remain pre-dominantly alpha-helical





https://doi.org/10.1038/s41587-025-02607-x

Questions?

Thank you!