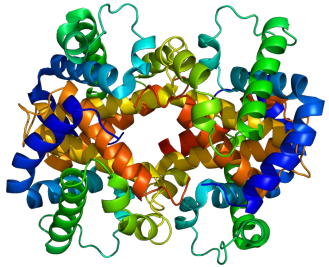
Modern Protein Design

Jeffrey Chang, Joyce Kang, Sunwoo Kang
1/25/2018

Background: *de novo* protein design

```
MVLSPADKTN VKAAWGKVGA HAGEYGAEAL ERMFLSFPTT
KTYFPHFDLS HGSAQVKGHG KKVADALTNA VAHVDDMPNA
LSALSDLHAH KLRVDPVNFK LLSHCLLVTL AAHLPAEFTP
AVHASLDKFL ASVSTVLTSK
```

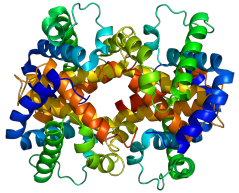
Protein structure prediction



Protein design

Architecture definition

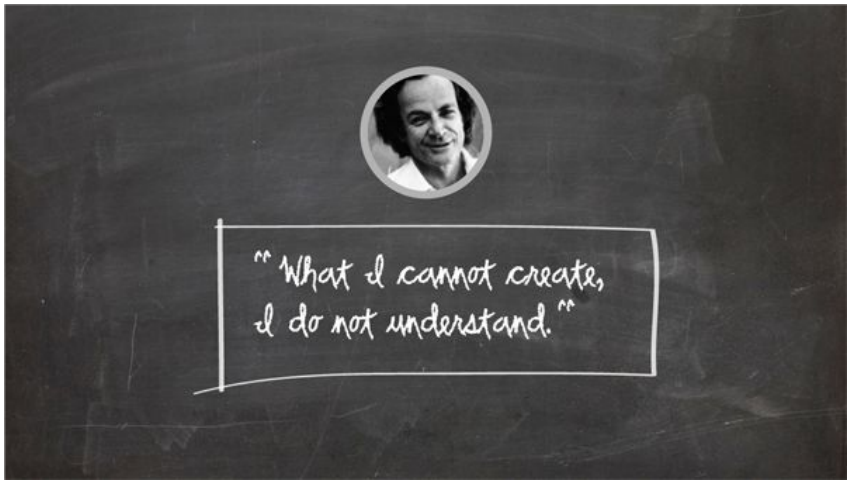
de novo protein design



```
MVLSPADKTN VKAAWGKVGA HAGEYGAEAL ERMFLSFPTT
KTYFPHFDLS HGSAQVKGHG KKVADALTNA VAHVDDMPNA
LSALSDLHAH KLRVDPVNFK LLSHCLLVTL AAHLPAEFTP
AVHASLDKFL ASVSTVLTSK
```

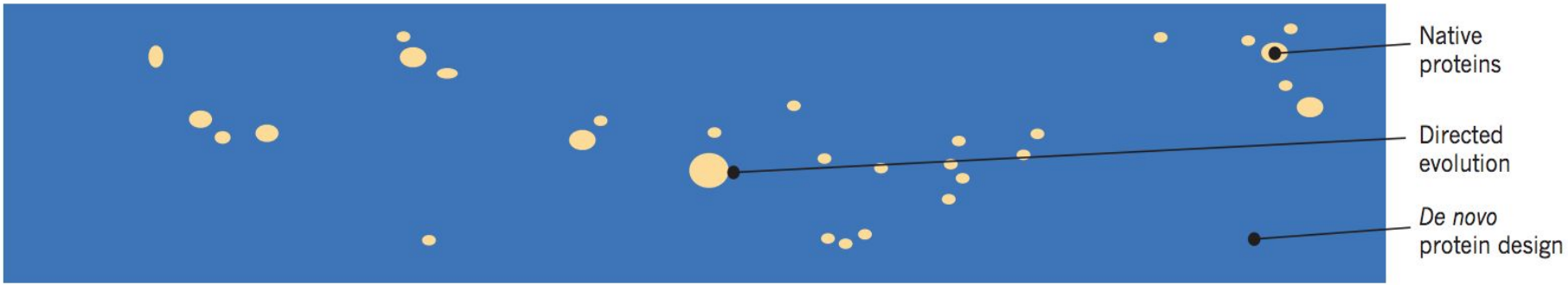
Why care about protein design?

- Explore protein-space
- Verify biophysical understanding
- Craft new functionality



Why is computation helpful for protein design?

1. Massive scale of parallel design, synthesis, and testing

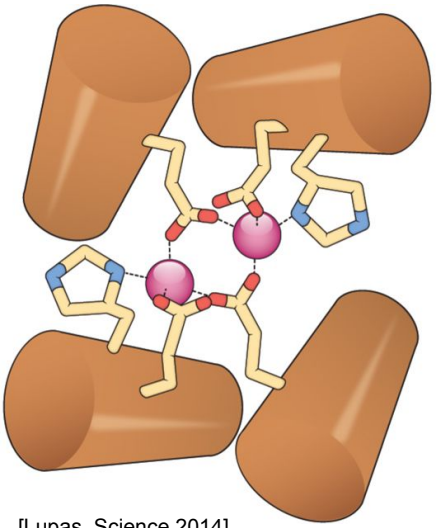


Not a graph

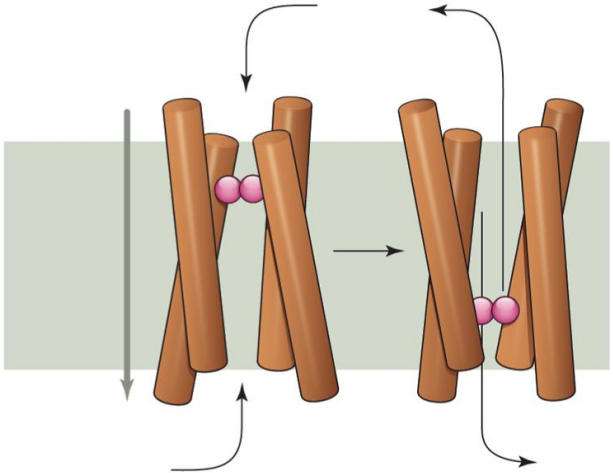
Supplementary materials: Huang *et al*, 2016

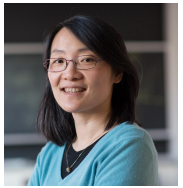
De novo design of a transmembrane Zn^{2+}-transporting four-helix bundle

Nathan H. Joh,[1] Tuo Wang,[2] Manasi P. Bhate,[1] Rudresh Acharya,[3] Yibing Wu,[1] Michael Grabe,[1*] Mei Hong,[2*] Gevorg Grigoryan,[4*] William F. DeGrado[1*]



[Lupas, Science 2014]

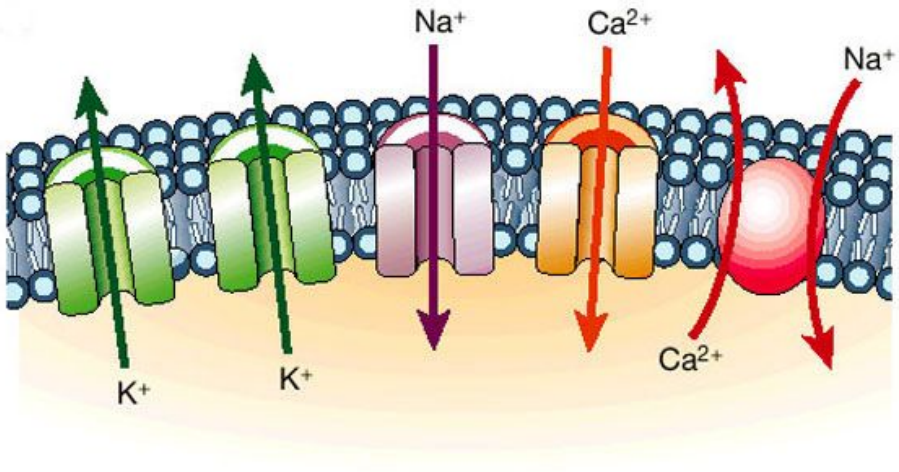




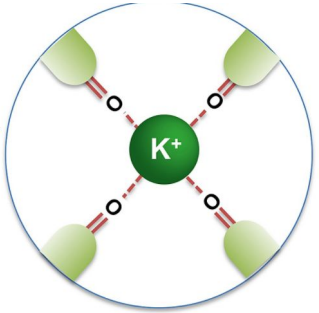
N. H. Joh et al., *Science* **346**, 1520 (2014)

Ion channels

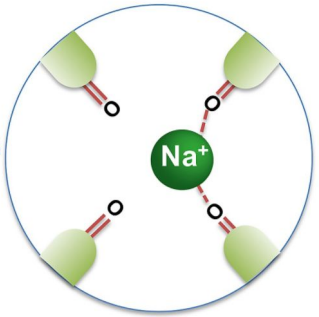
- Determine what goes in/out of cell
- Must be specific



[Rasband et al. Nature Education 2010]



1.5Å

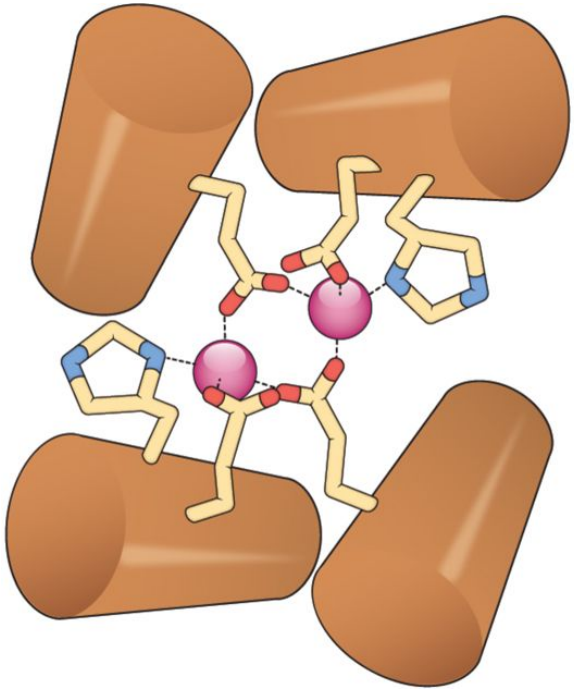


1.2Å

[Stavros, Fiona et. al., PLOS ONE (2014)]

Zn^{2+}-specific binding motif

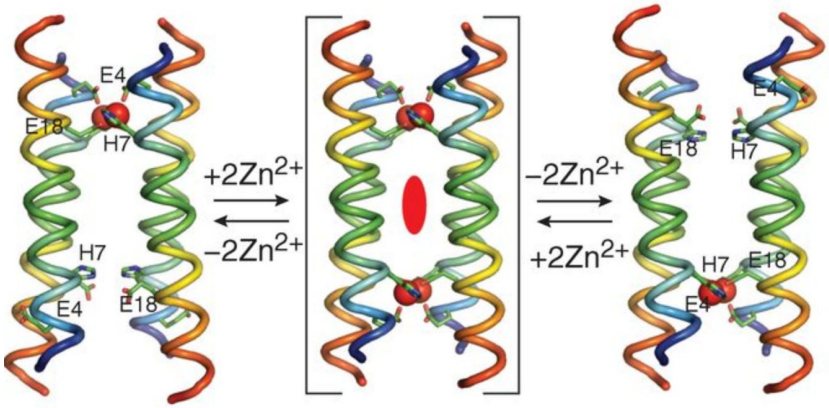
- “4Glu-2His-di-Zn^{2+}”
- Different from nature’s Zn^{2+} transporters



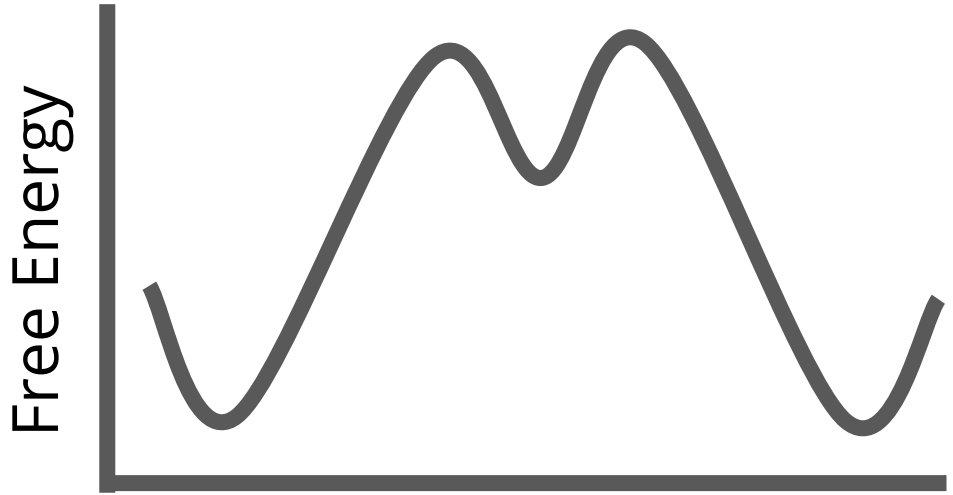
[Lupas, Science 2014]

Protein Dynamics

- “Rocks” between inward and outward
- Must engineer entire landscape

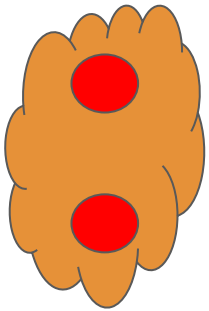


[Joh et. al., Science 2014]



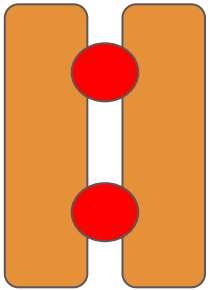
Design Strategy

Zn^{2+} binding site



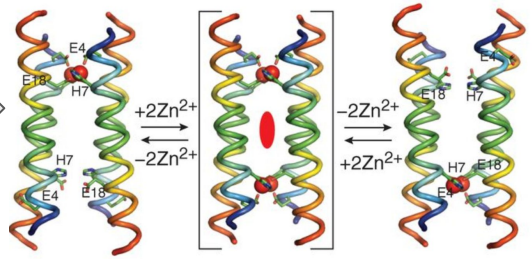
Optimize Backbone

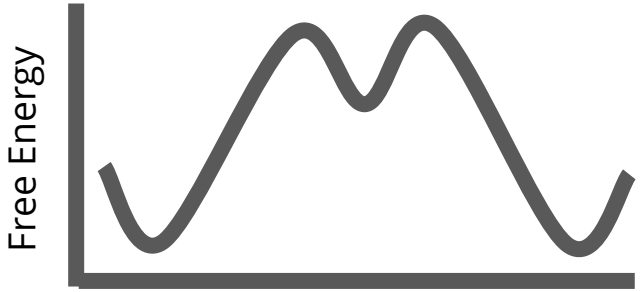
Backbone Shape



Optimize Residues

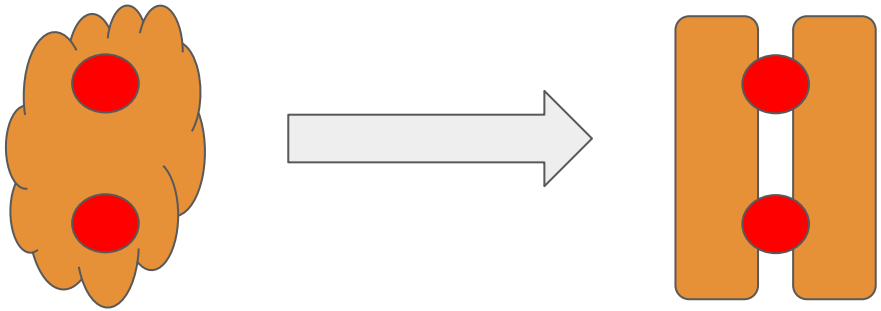
Full landscape

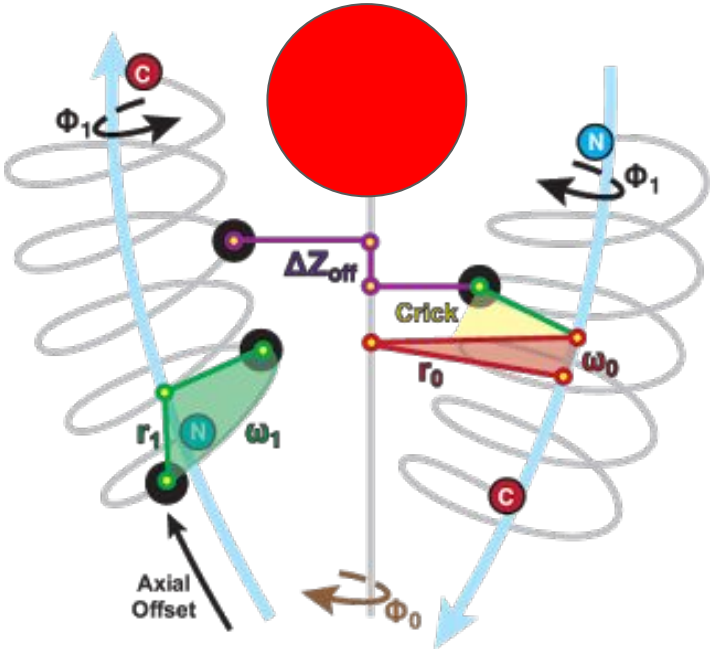




Backbone design: “Coiled-coil”

- Find best backbone that accommodates active site
  - Search for best “Crick parameters” to describe coiled-coil geometry

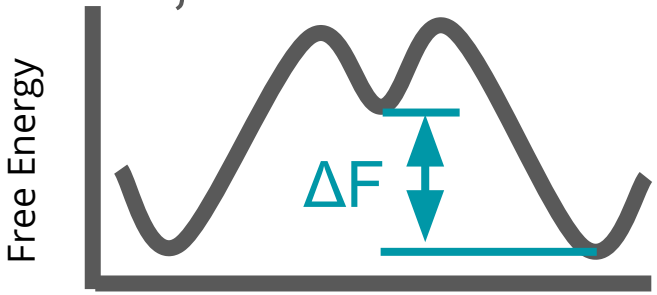


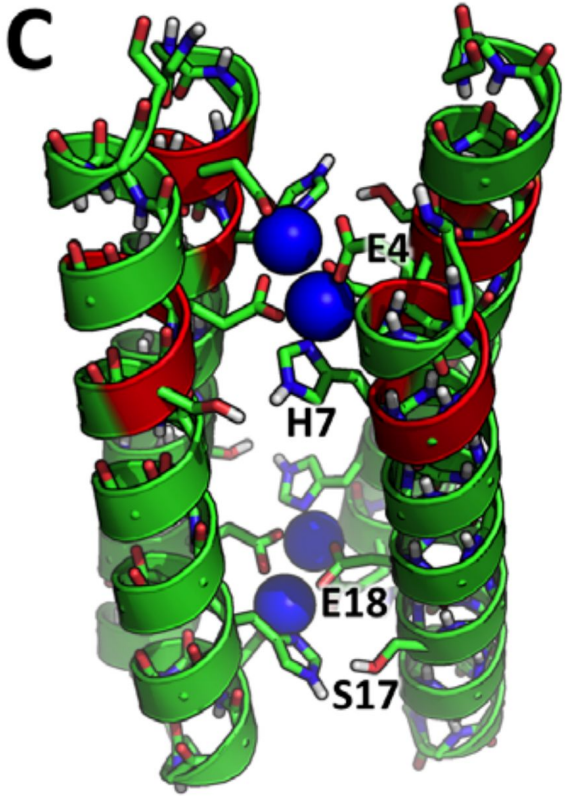


[Homepage of Chris MacDermaid]

Residue design

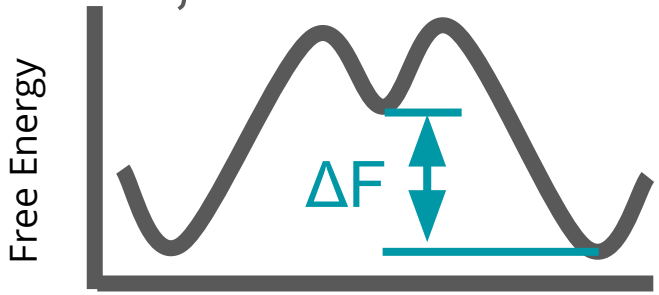
- Find a sequence that
  - Prefers open state over coiled-coil
  - Prefer membranes over water
- Monte Carlo search
- 1,008 final candidates chosen

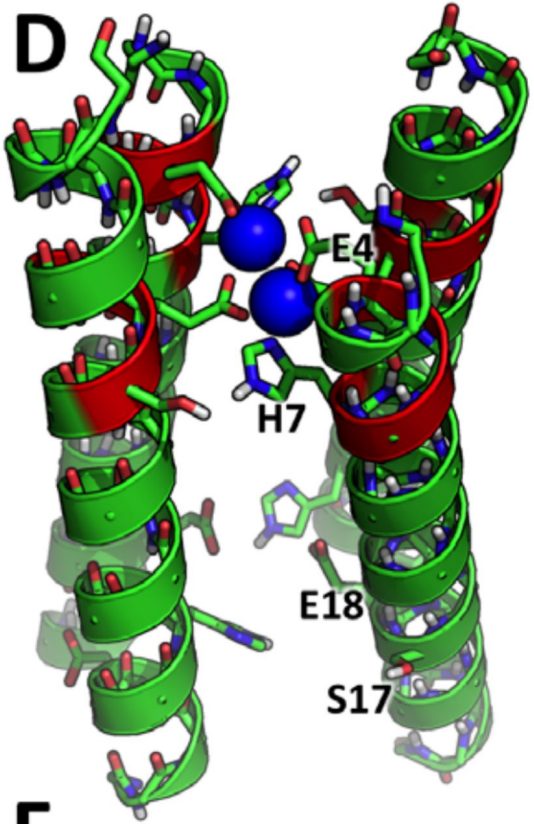




Residue design

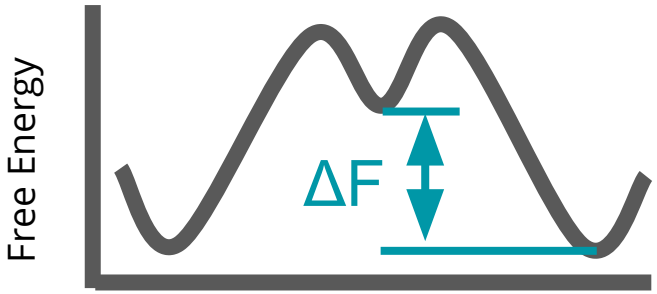
- Find a sequence that
  - Prefers open state over coiled-coil
  - Prefer membranes over water
- Monte Carlo search
- 1,008 final candidates chosen

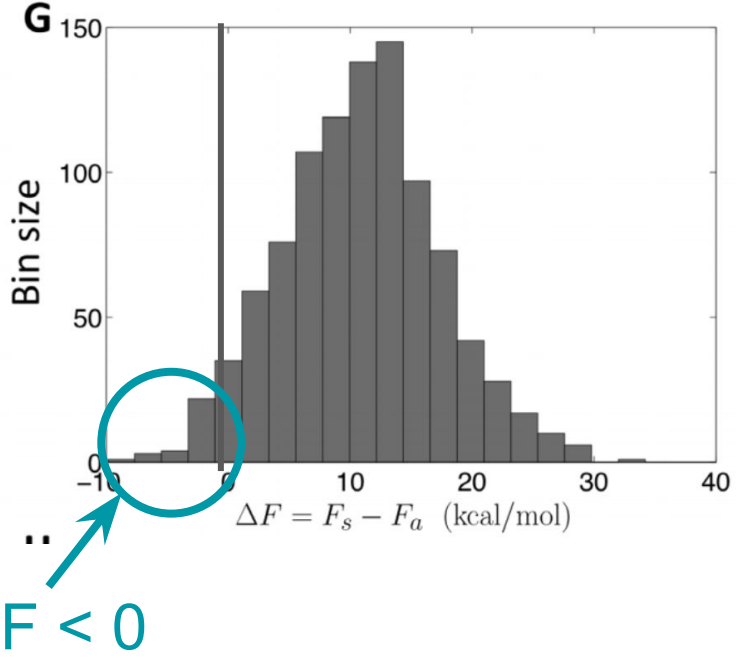




Final Residue optimization

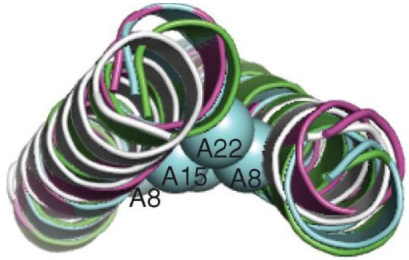
- VALOCIDY calculations to estimate free energy diffs
- Hand-picked final protein

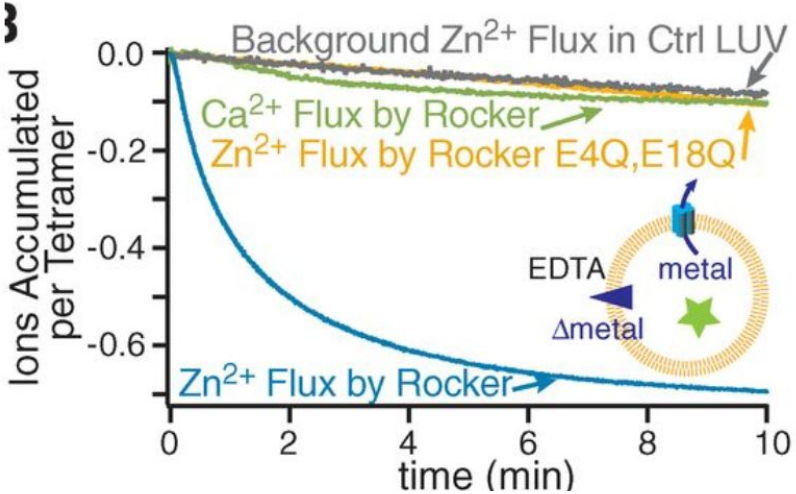




Experimental Validation

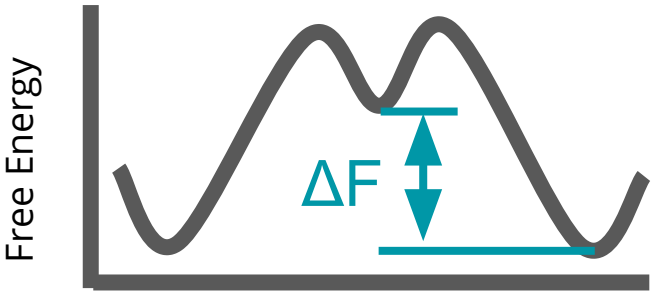
- Crystal structure: Cα RMSD 2.3-2.6Å
  - 2.8Å resolution, non-Zn^{2+} form crystallized as dimers
- Selective for Zn^{2+} and Co^{2+} over Ca^{2+}
- Proton antiporter
- And more
  - Titration: 2 Zn^{2+} per tetramer
  - Binding loosens helix packing
  - Antiparallel association of monomers
  - K_D of dimer/tetramer/octamer formation
  - M-M kinetics of ion transport

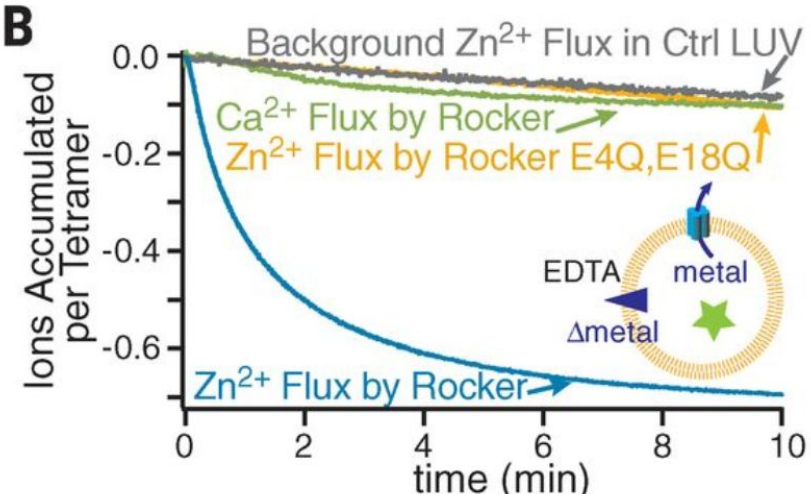




Further work

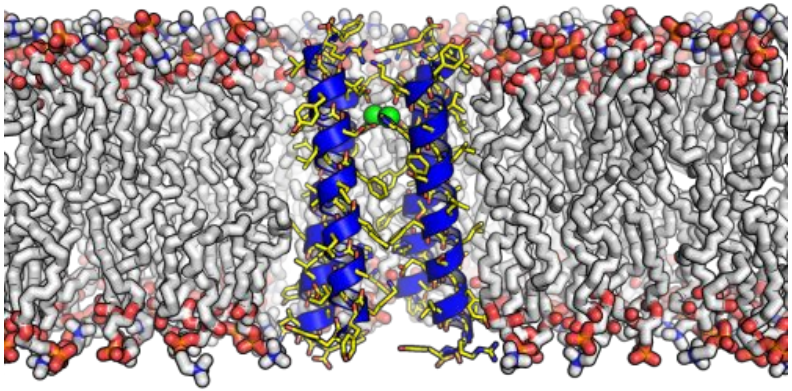
- Improve kinetics
- Crystallize the tetramer
- Simulate full transport cycle
- Better cheap scoring metrics
- Explore lesser-known territory



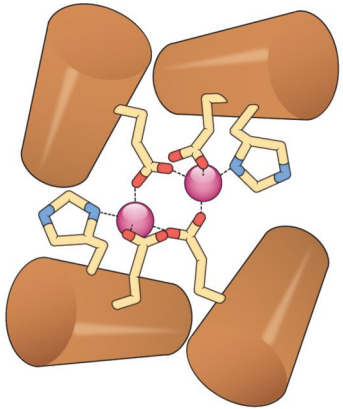


De novo design of a transmembrane Zn^{2+}-transporting four-helix bundle

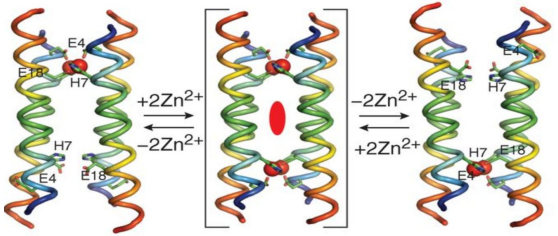
Questions?

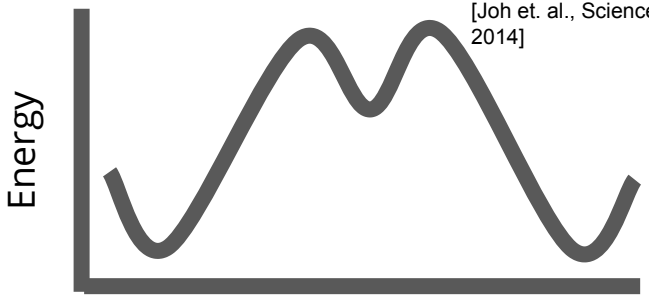


[Dartmouth]



[Lupas, Science 2014]





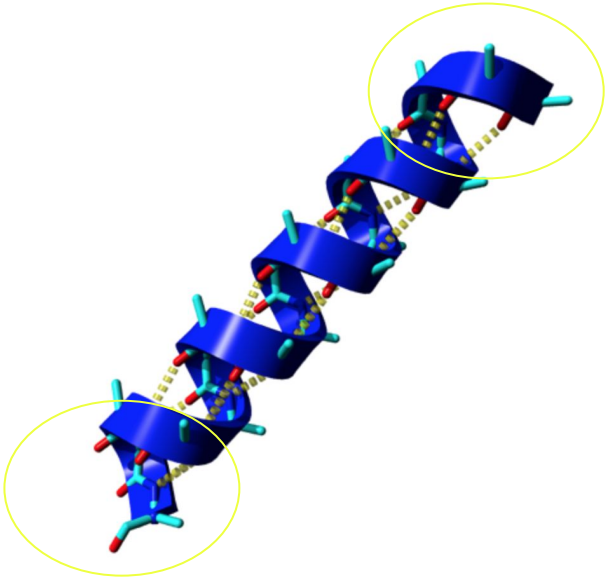
[Joh et. al., Science 2014]

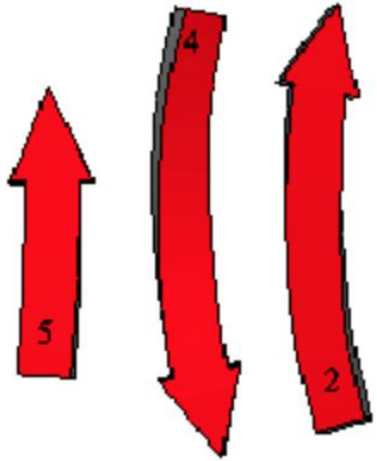
Global analysis of protein folding using massively parallel design, synthesis, and testing

Rocklin *et al.*, Science (2017)

Problem

- Finding global determinants of stability



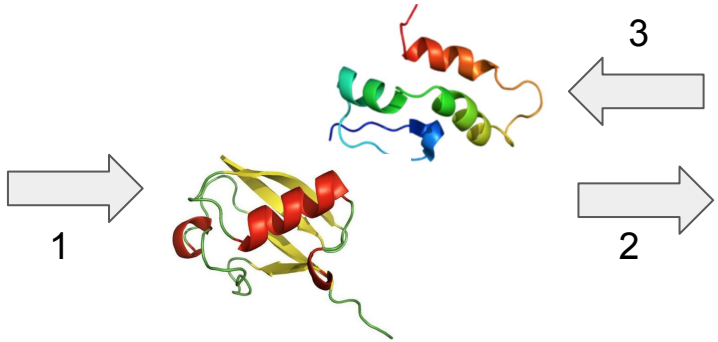


Solution Summary

- Massive parallel design, synthesis, and testing of miniprotein

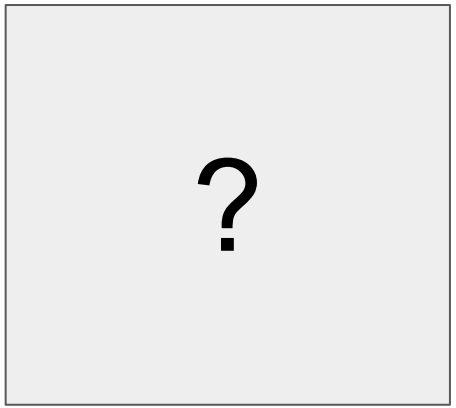
de novo sequences

Sutterstock Pic



stable structure

Biological molecules, Weebly



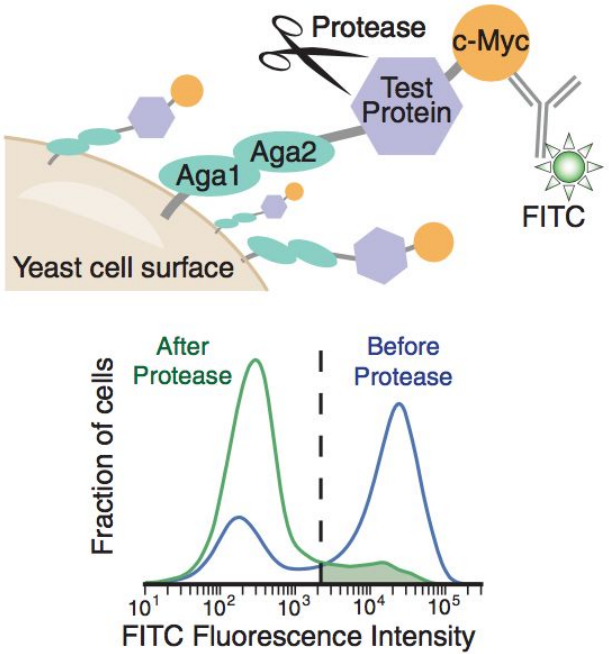
global determinants

1. Massively Parallel Design & Testing

- Four protein topologies (**ααα**, βαββ, αββα, ββαββ)
  1. 5,000 - 40,000 de novo proteins
  2. 1000 design by ranking
  3. 2 Negative controls
     a. fully scrambled
     b. patterned scrambled

1. Massively Parallel Design & Testing

- Proteolysis assay to measure stability



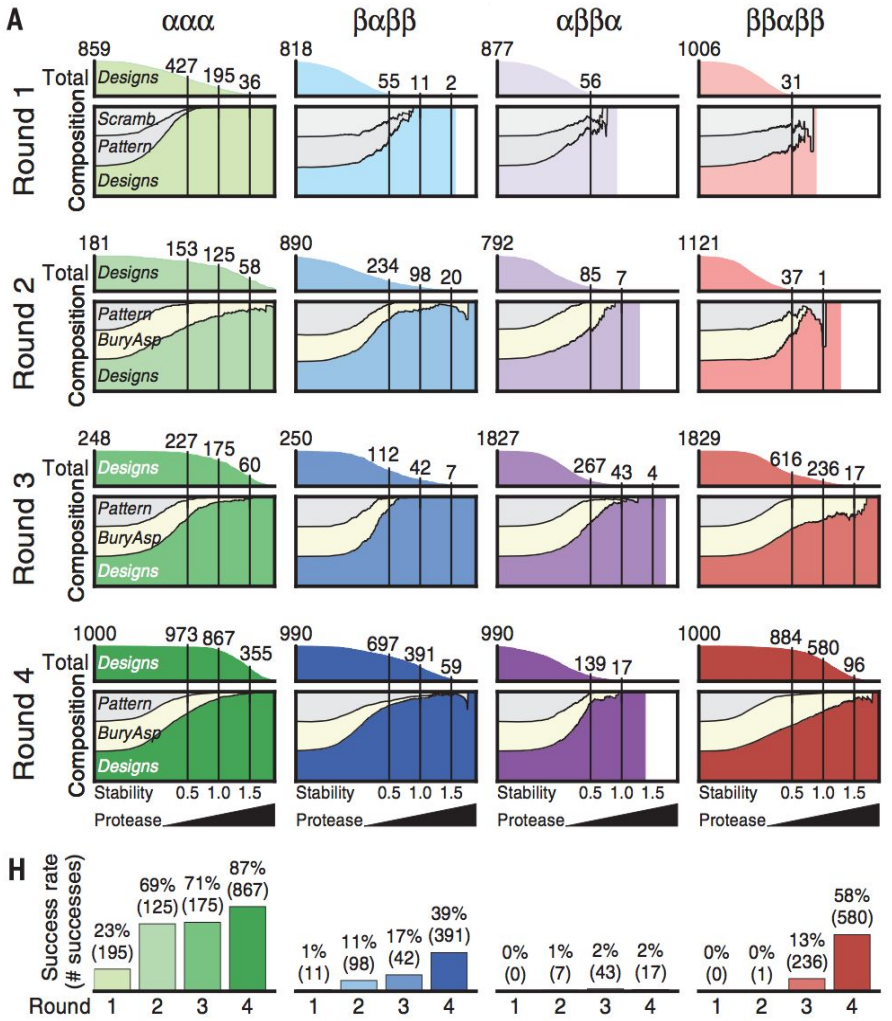
Protease EC_{50} value:

"protease concentration at which one half of the cells pass the collection threshold "

Rocklin *et al.*, Science (2017)

1. Massively Parallel Design

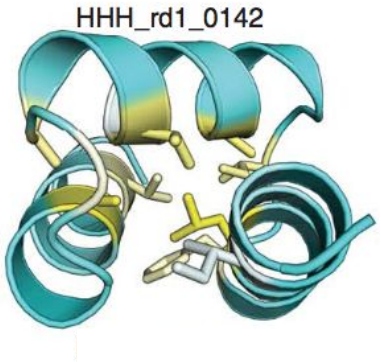
- 206 stable structures
  - 195 ααα
  - 11 βαββ
  - High stability relative to neg. control sequences
- Experimental verification
  - High melting point above 70℃
  - Structure characterization through NMR



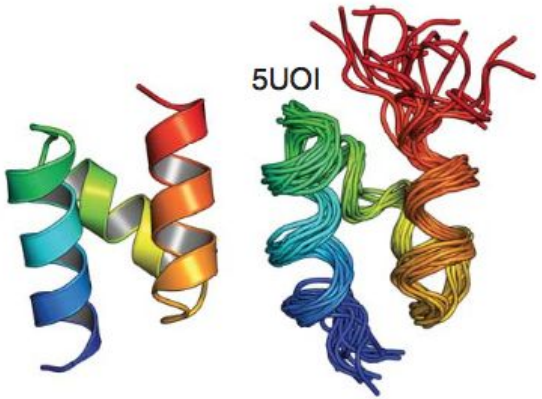
Rocklin *et al.*, Science (2017)

2. Global Determinants of Stability

- 60 structural and sequence-based metrics
- Looked for difference between stable ααα vs. unstable ααα



1. Buried nonpolar surface area

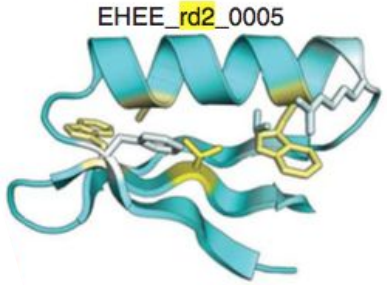


2. Local sequence structure agreement

Rocklin *et al.*, Science (2017)

3. Iterative data driven protein design

- Updated metrics weighting
- Increase in success rate
  - ααα: 23% to 69%
  - βαββ: 1% to 11%
  - αββα: 7 structure
  - ββαββ: 1 structure
- Even greater NPSA
  - Limit: decrease in solubility
  - αββα: 17% to 39%
  - ββαββ: 13% to 58%



1) Increase buried NPSA

2) local sequence agreement

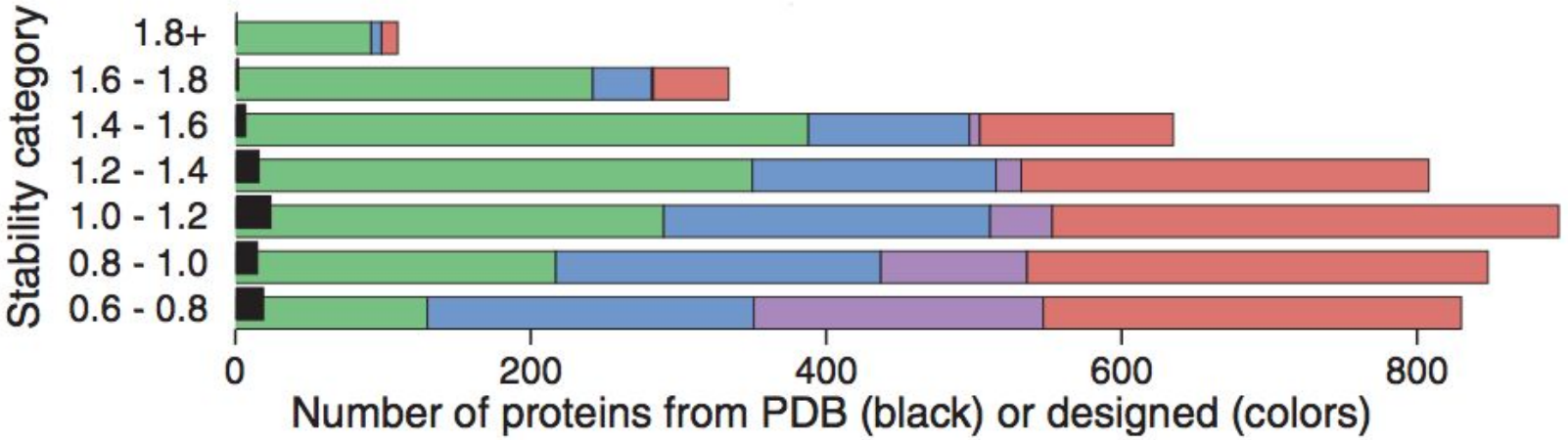
1) amino acid packing density

2) total Rosetta Energy

4. Sequence determinant of stability

- Average stability of each amino acids

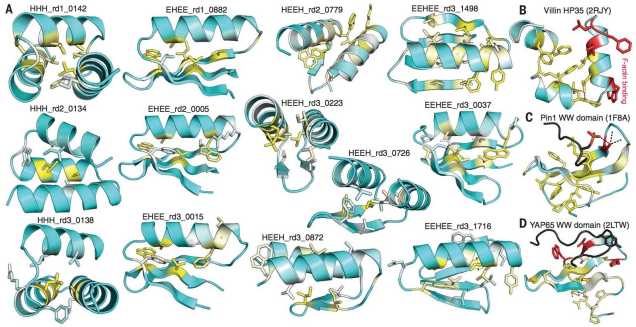
4. Stability measurement & Comparison



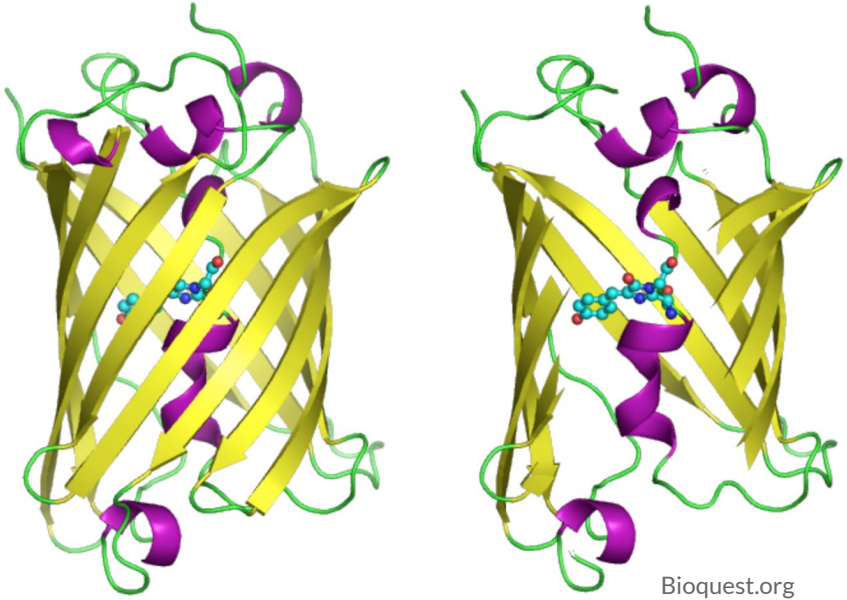
Rocklin *et al.*, Science (2017)

Limitations

- Unclear which metrics were used, and what weighting was applied
- Differences between each round of designs
- “Global Analysis” of miniprotein
- *In vivo* effect of these miniprotein



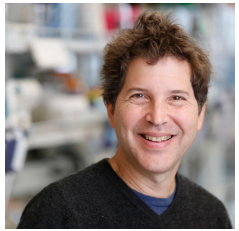
Rocklin *et al.*, Science (2017)



Bioquest.org

Massively parallel *de novo* protein design for targeted therapeutics

Aaron Chevalier[1,2*], Daniel-Adriano Silva[1,2*], Gabriel J. Rocklin[1,2*], Derrick R. Hicks[1,2,3], Renan Vergara[1,2,4], Patience Murapa[5], Steffen M. Bernard[6,7], Lu Zhang[8,9], Kwok-Ho Lam[10], Guorui Yao[10], Christopher D. Bahl[1,2], Shin-Ichiro Miyashita[11,12], Inna Goreshnik[1], James T. Fuller[5], Merika T. Koday[5,13], Cody M. Jenkins[5], Tom Colvin[1], Lauren Carter[1,2], Alan Bohn[5], Cassie M. Bryan[1,2], D. Alejandro Fernández-Velasco[4], Lance Stewart[2], Min Dong[11,12], Xuhui Huang[9], Rongsheng Jin[10], Ian A. Wilson[6,7], Deborah H. Fuller[5] & David Baker[1,2]

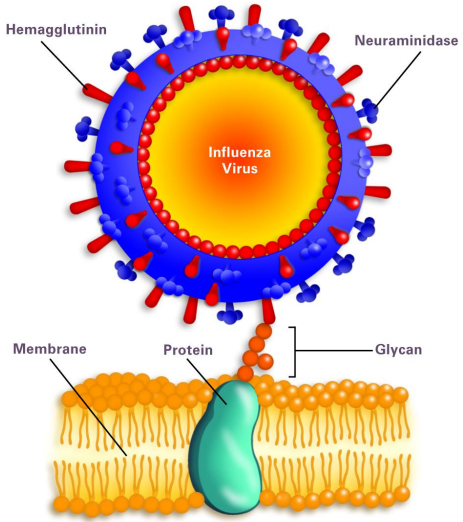


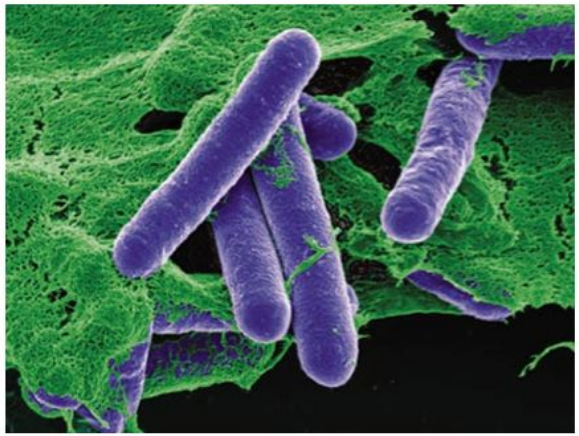
David Baker

NATURE | VOL 550 | 5 OCTOBER 2017

Goal: generate binders for a given target

Influenza A H1 haemagglutinin (HA)

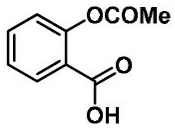




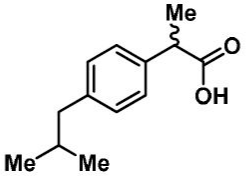
Botulinum neurotoxin B (BoNT/B)

Image creds: https://www.ncbi.nlm.nih.gov/pubmedhealth/PMHT0025751/
https://microbewiki.kenyon.edu/index.php/Clostridium_botulinum_Neurotoxins

Why small binding proteins?



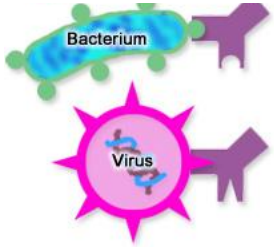
Aspirin (1)



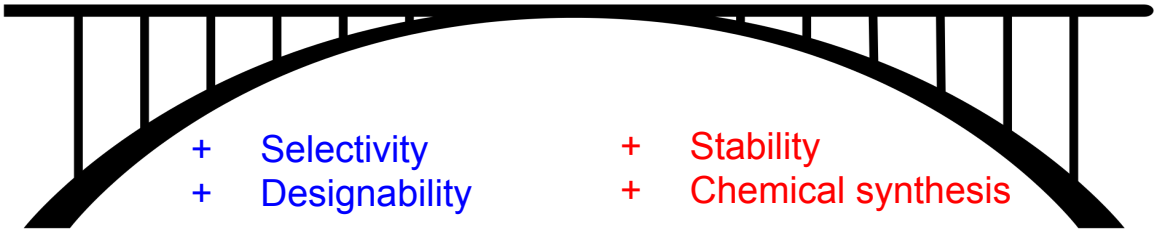
Ibuprofen (2)

Small molecules

Small (4–12 kDa) binding proteins



Antibodies



“Massively parallel”

- Integrated computational and experimental approach
- Rapid design and parallel testing of 10,000+ mini-protein binders
- Advances in DNA manufacturing and protein design

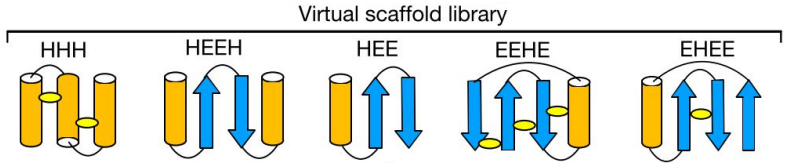
Size of genetically encodable computationally designed proteins (~40 AAs)

Size of oligonucleotides (230 bp) that can be made as batches of 10,000 or larger

Computational methodology

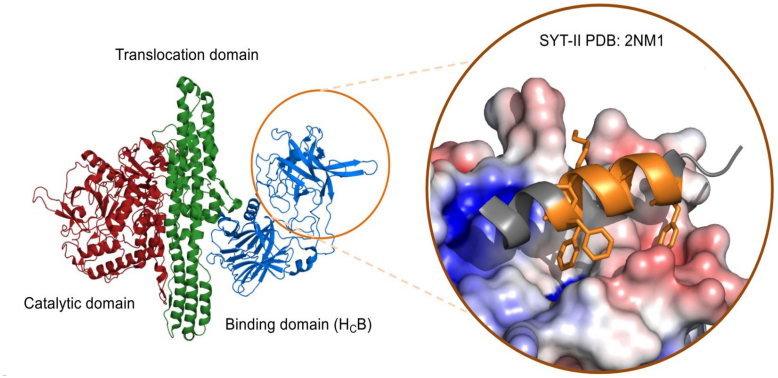
Step 1: Generate virtual scaffold libraries

- 37-43 residue mini-protein backbones



Step 2: Design binding interfaces

- Superimpose helical segments of the scaffolds on interface helices in previously solved HA and BoNT/B complexes
- Seed the newly formed interfaces with hotspot residues from these helices
- Discard candidates with protein/target backbone clashes

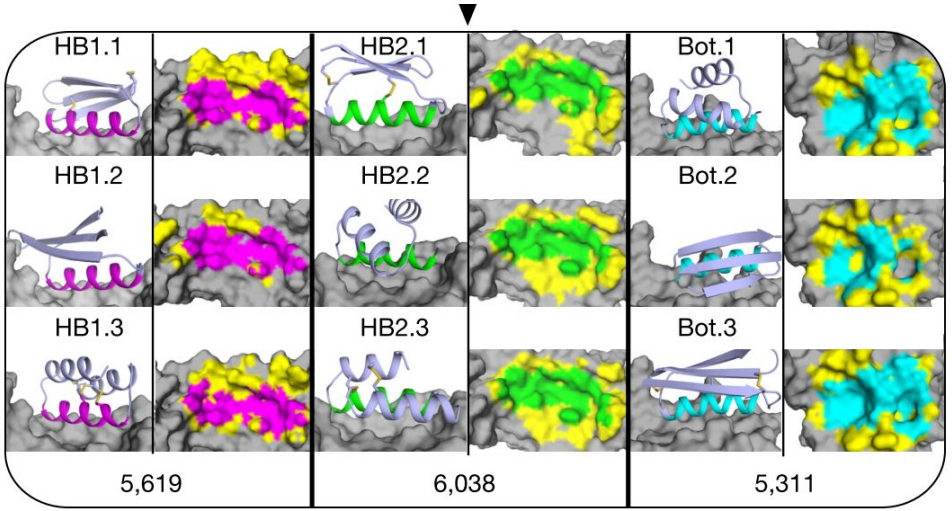


Chevalier et al. (2017)

Computational methodology

Step 3: Optimize other residues for high-affinity binding and stability

- Optimize monomer and interaction energies with Rosetta sequence design



Yellow = new contact areas generated by Rosetta sequence design

Chevalier et al. (2017)

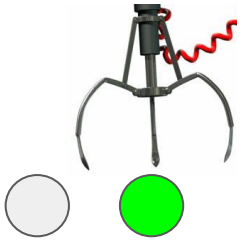
Experimentally test top ~10,000 candidates

- 7,276 designs against HA and 3,406 designs against BoNT
- Included variety of control sequences
  - Randomly permute AAs outside helical interface, core residues randomly permuted, loops mutated to Gly-Ser, designed binding sites omitted

Genetically encode 16,968 mini-proteins and amplify oligo pools

Incubate yeast libraries with a range of concentrations of fluorescently labelled target

FACS sorting to retrieve cells displaying designs that bound the target



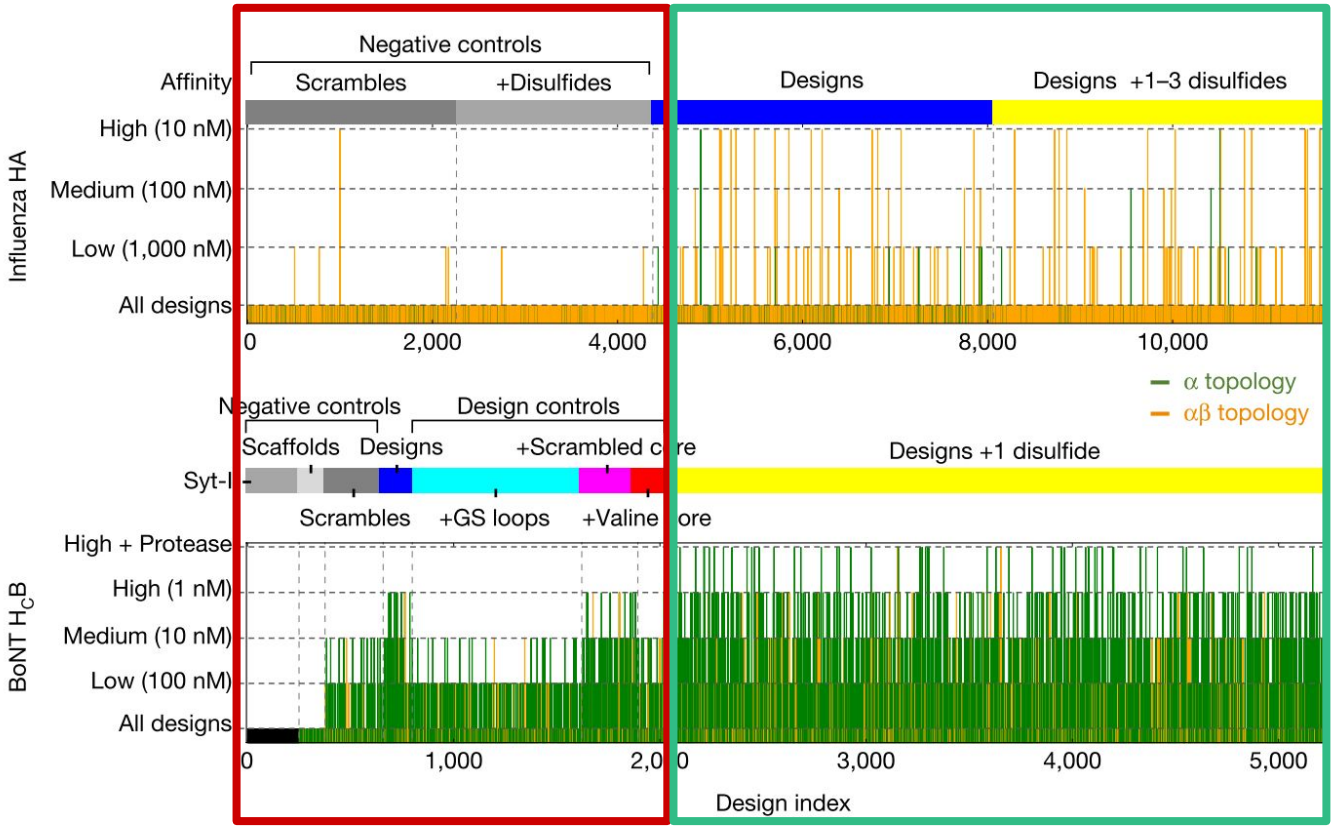
Next-gen sequencing

to determine frequency of each design and control sequence in each pool

Chevalier et al. (2017)

Controls = bad binders

Designed candidates = good binders



Chevalier et al. (2017)

Key insights

- Lower computed folding and binding energies → observed binding activity
- Features most strongly associated with binding are local sequence-structure compatibility and # of contacts across interface
- Disulfide bonds do not guide folding but provide stability against proteolysis
- MD simulations comparing binders / non-binders showed that binders had less fluctuations in interface hotspot residues
- Loops may play an underappreciated, instructive role in folding
- No single protein topology or shape is the best fit for all interfaces

Chevalier et al. (2017)

Testing *in vitro* and *in vivo*

- Little to no antibody response
- Protected mice from lethal dose of influenza → 100% survival

Chevalier et al. (2017)

Limitations

- Proteins are much shorter (37-43 residues) than many in nature
- Did not design molecular machines with changing conformation (e.g. transporters)
- Use of previously solved HA and BoNT/B complexes to seed key residues in binding interface

Chevalier et al. (2017)