

# Protein design

CS/CME/BioE/Biophys/BMI 279

Oct. 12, 2021

Ron Dror

# Outline

- What is protein design, and why do it?
- Overall approach: Simplifying the protein design problem
- Protein design methodology
  - Designing the backbone
  - Select sidechain rotamers: the core optimization problem
  - Optional: giving the backbone wiggle room
  - Optional: negative design
  - Optional: complementary experimental methods
- Examples of recent designs
- How well does protein design work?

What is protein design, and why do it?

# Problem definition

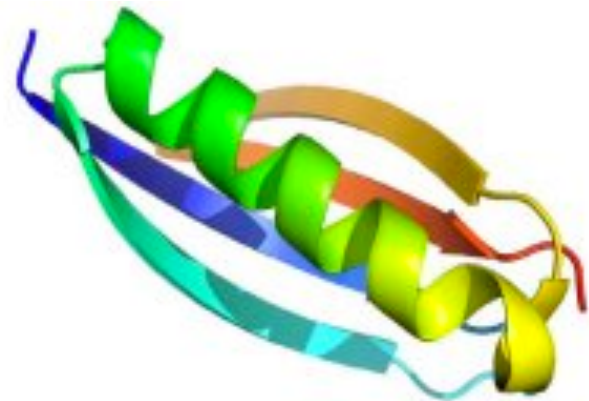
- Given the desired three dimensional structure of a protein, design an amino acid sequence that will assume that structure.
  - Of course, a precise set of atomic coordinates would determine the sequence. Usually we start with an *approximate* desired structure.
  - Alternatively, we may want to design for a particular function (e.g., the ability to bind a particular ligand). “Function” gets translated to specific desired structural properties

EEVTIKANLIFAN  
GSTQTAEFKGTKE  
KALSEVLAYADTL  
KKDNGEWTIDKRV  
TNGVIILNIKFAG

**Protein Folding**



**Protein Design**



# Sample applications

- Designing enzymes (proteins that catalyze chemical reactions)
  - Useful for production of industrial chemicals and drugs
- Designing proteins that bind specifically to other proteins
  - Potential for HIV, cancer, Alzheimer's treatment
  - Special case: antibody design
- Designing sensors (proteins that bind to and detect the presence of small molecules—for example, by lighting up or changing color)
  - Calcium sensors used to detect neuronal activity in imaging studies
  - Proteins that detect TNT or other explosives, for mine detection
- Making a more stable variant of an existing protein
  - Or a water-soluble variant of a membrane protein

Overall approach: simplifying the  
protein design problem

# The “direct” approach (doesn't work in practice!)

- Given a target structure, search over all possible protein sequences
- For each protein sequence, predict its structure, and compare to the target structure
- Choose the best match

# Why doesn't the "direct" approach work?

- Computationally intractable
  - We'd need to use ab initio structure prediction
  - Ab initio structure prediction for even one sequence is computationally intensive
  - Huge number of sequences to consider
    - $20^N$  possible sequences with  $N$  residues
- May not be good enough!
  - Ab initio structure prediction is far from perfect, in part because energy functions are imperfect
  - Given an energy function, what we really want is to maximize the probability of the desired structure (compared to all other possible folded and unfolded structures)
  - We could do this by sampling the full Boltzmann distribution for each candidate sequence ... but that's even more computationally intensive!

Template based modeling won't work because we are designing proteins that are different from known proteins.

# We can dramatically simplify this problem by making a few assumptions

1. Assume the backbone geometry is fixed
2. Assume each amino acid can only take on a finite number of geometries (*rotamers*) Refers to side chain conformations
3. Assume that what we want to do is to maximize the energy drop from the completely unfolded state to the target geometry
  - In other words, simply ignore all the other possible folded structures that we want to avoid

We'll first address the problem under these assumptions, then consider relaxing them a bit

# The simplified problem

- At each position on the backbone, choose a rotamer (an amino acid type and a side-chain geometry) to minimize overall energy
  - Assume our energy function specifies a free energy. The Rosetta all-atom force field (physics-based/knowledge-based hybrid) is a common choice.
  - For each amino acid sequence, energy is measured relative to the unfolded state.
    - In practice a “reference energy” for each amino acid is subtracted off, corresponding roughly to how much that amino acid favors folded states
  - Assume that energy can be expressed as a sum of terms that depend on one or two rotamers each. This is the case for the Rosetta force fields (and for most molecular mechanics force fields as well).
- Thus, we wish to minimize total energy  $E_T$ , where

$$E_T = \sum_i \left[ E_i(r_i) + \sum_{i \neq j} E_{ij}(r_i, r_j) \right]$$

Note that  $r_i$  specifies both the amino acid residue at position  $i$  and that residue's side-chain geometry

# Protein design methodology

Protein design methodology

**Designing the backbone**

# Designing the backbone

- The first step of most protein design protocols is to select one or more target backbone structures.
  - This is as much art as science.
  - Often multiple target structures are selected, because some won't work. (Apparently proteins can only adopt a limited set of backbone structures, but we don't have a great description of what that set is.) [The goal is to get precise coordinates for backbone atoms](#)
- Methods to do this:
  - Use an experimentally determined backbone structure
  - Use a fragment assembly program like Rosetta, selecting fragment combinations that fit some approximate desired structure
  - Assemble secondary structure elements by hand
  - Current research direction: generating suitable backbone structures by machine learning

# Example of backbone design

“Top7” was the first designed protein without any known protein as a starting point!

- To design “Top7,” a protein with a novel fold, Kuhlman et al. started with a schematic, then used Rosetta fragment assembly to find 172 backbone models that fit it.

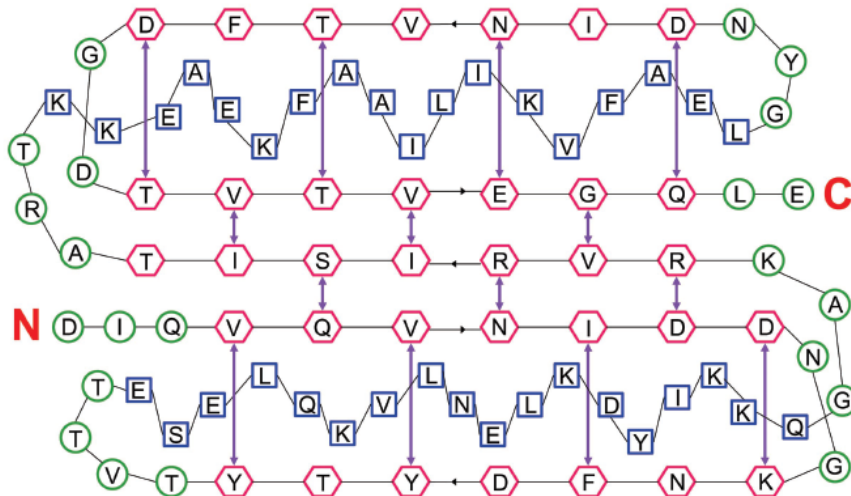
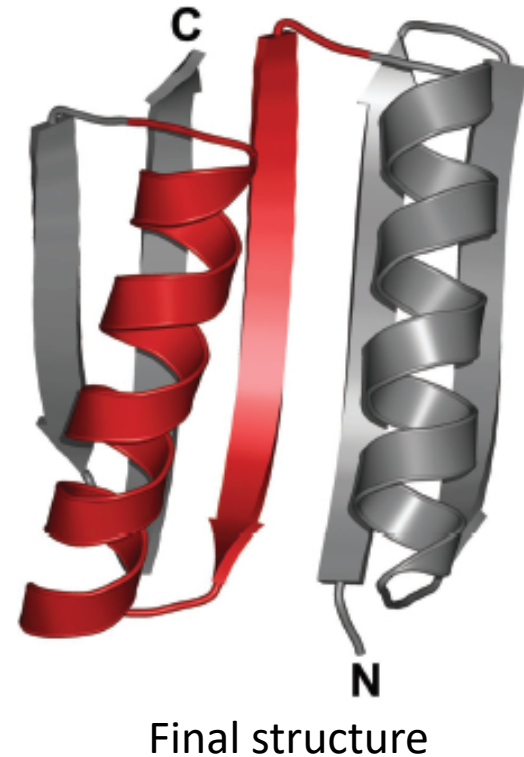


Fig. 1. A two-dimensional schematic of the target fold (hexagon, strand; square, helix; circle, other). Hydrogen bond partners are shown as purple arrows. The amino acids shown are those in the final designed (Top7) sequence.

Initial schematic of target fold. Hexagons =  $\beta$  sheet. Squares =  $\alpha$  helix. Arrows = hydrogen bonds. Letters indicate amino acids in final designed sequence (these were not determined until much later).



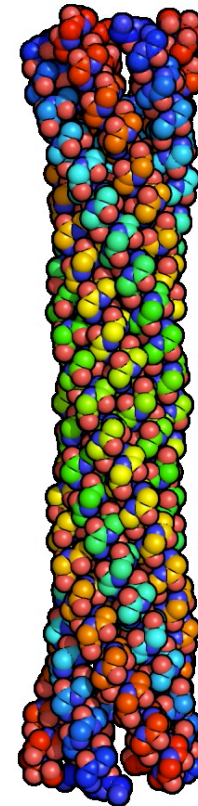
Final structure

# In some special cases, one can describe possible backbone structures by equations with a few tunable parameters

- Pro: Very elegant
- Con: Usually only applicable to long alpha helices (which adopt “coiled coil” structures)  
Called “coiled coil” because when zoomed out, an alpha helix itself tends to form a much larger helix

Early work: Pehr Harbury ...  
Peter Kim, *Science* 1998

You're not responsible for this



movie credit:  
Vikram Mulligan, Possu Huang

# Protein design methodology

**Select sidechain rotamers: the core optimization problem**

# The optimization problem

- Given a desired backbone geometry, we wish to select rotamers at each position to minimize total energy

$$E_T = \sum_i \left[ E_i(r_i) + \sum_{i \neq j} E_{ij}(r_i, r_j) \right]$$

where  $r_i$  specifies both the amino acid at position  $i$  and its side-chain geometry

# Optimization methods

- Heuristic methods
  - Not guaranteed to find optimal solution, but faster than exact methods
  - Used in great majority of protein design today
  - Most common is Metropolis Monte Carlo
    - Moves may be as simple as randomly choosing a position, then randomly choosing a new rotamer at that position
    - May decrease temperature over time (simulated annealing)
- Exact methods
  - Guaranteed to find optimal solution, but slow for larger proteins
  - Multiple proteins have been designed with the Dead-End Elimination method, which prunes branches of the exhaustive search tree by proving that certain rotamers are incompatible with the global optimum
  - An alternative: The A\* optimization algorithm (originally developed at Stanford, for robot path-finding)

These days, researchers are more focused on structures of larger proteins, thereby making exact methods less practical.

# Protein design methodology

**Optional: giving the backbone wiggle room**

# “Flexible backbone” design

- One of our key simplifying assumptions was that of a fixed backbone geometry.
- For many applications, protein design works better if you give the backbone some limited “wiggle room.”
- This requires optimizing simultaneously over rotamers and backbone geometry.
  - Often addressed through a Monte Carlo search procedure that alternates between local tweaks to backbone dihedrals and changes to side-chain rotamers
  - One can also refine a designed structure by local energy minimization, then re-optimize the side chains

Protein design methodology

**Optional: negative design**

# Negative design

- Another simplifying assumption was that we simply minimize the energy of the desired structure
  - We do not consider all other possible structures. It's possible that their energy ends up even lower.
- In negative design, we identify a few structures that we want to *avoid*, and we try to keep their energies high during the design process.
  - This can help, but we cannot explicitly avoid all possible incorrect structures without making the problem much more complicated. So the overall approach is still heuristic.

# Protein design methodology

**Optional: complementary experimental  
methods**

# Complementary experimental methods

- Computational protein design is often combined with experimental protein engineering methods
- For example, computational designs can often be improved by directed evolution
  - Directed evolution involves introducing random mutations to proteins and picking out the best ones
  - Usually this is done in living cells, with the fittest cells (i.e., those containing the “best” version of the protein) selected by some measure

“Fittest” means best able to survive in the given conditions



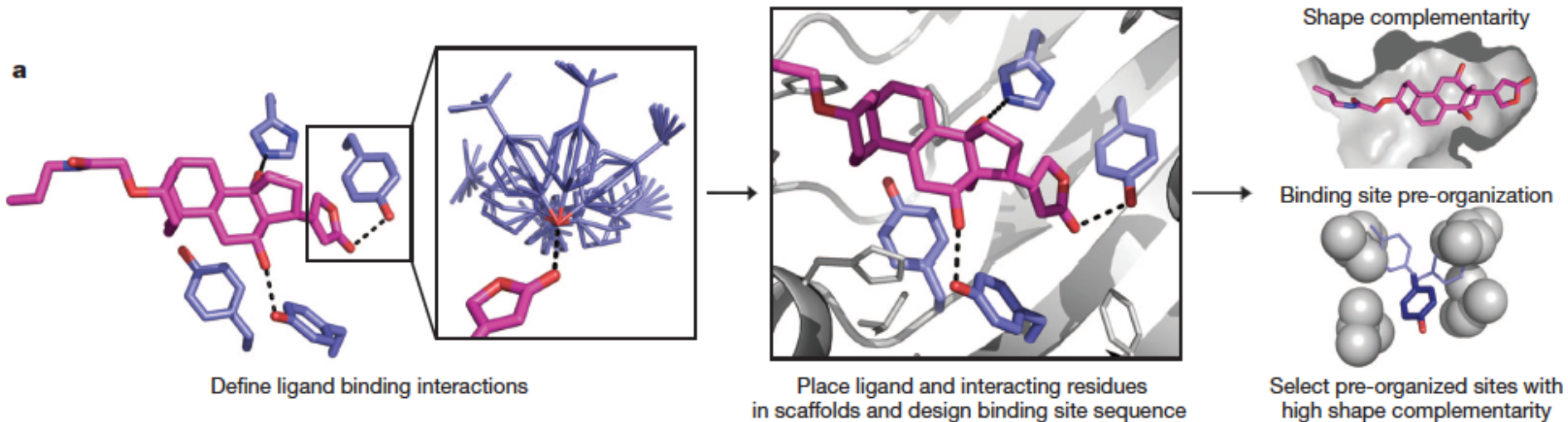
Frances Arnold

2018 Nobel Prize “for the directed evolution of enzymes”

# Examples of recent designs

# Designing proteins that bind specific ligands

- The example below required specification of the position of certain side chains that will form favorable interactions with the ligand



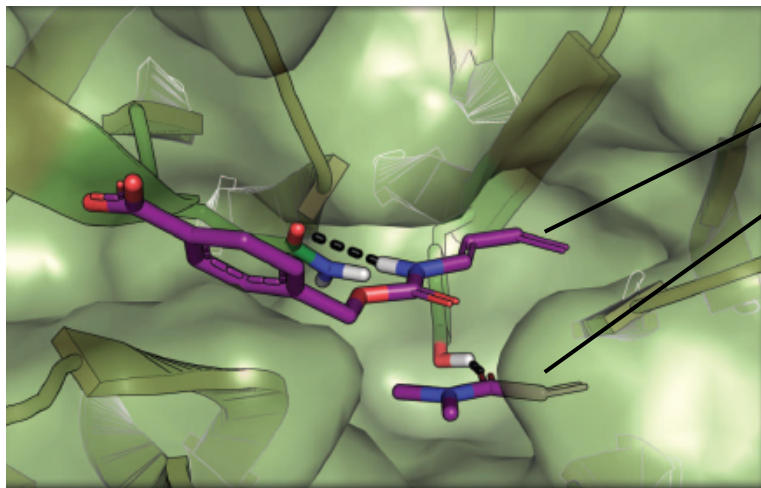
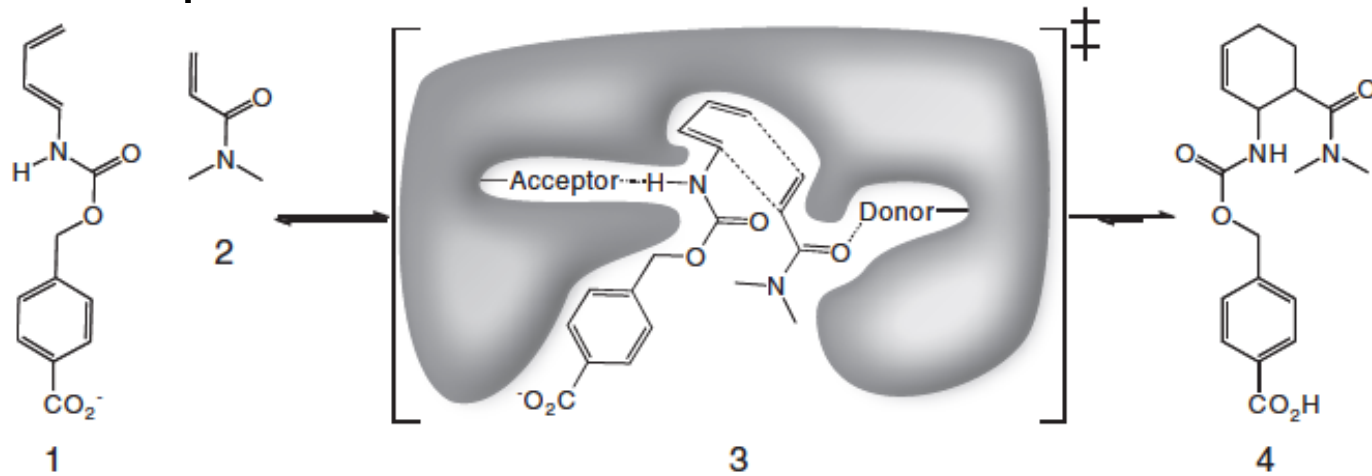
Protein designed to bind tightly to a specific steroid, but not to related molecules

Steroids refer to organic compounds with a 4 ring structure. Used in cellular membranes and signaling Tinberg et al., Nature 501:212-6 (2013)

# Designing enzymes

Enzymes are proteins that speed up chemical reactions

- In the example below, the protein holds two molecules in just the right relative positions for them to react. This speeds up the reaction.

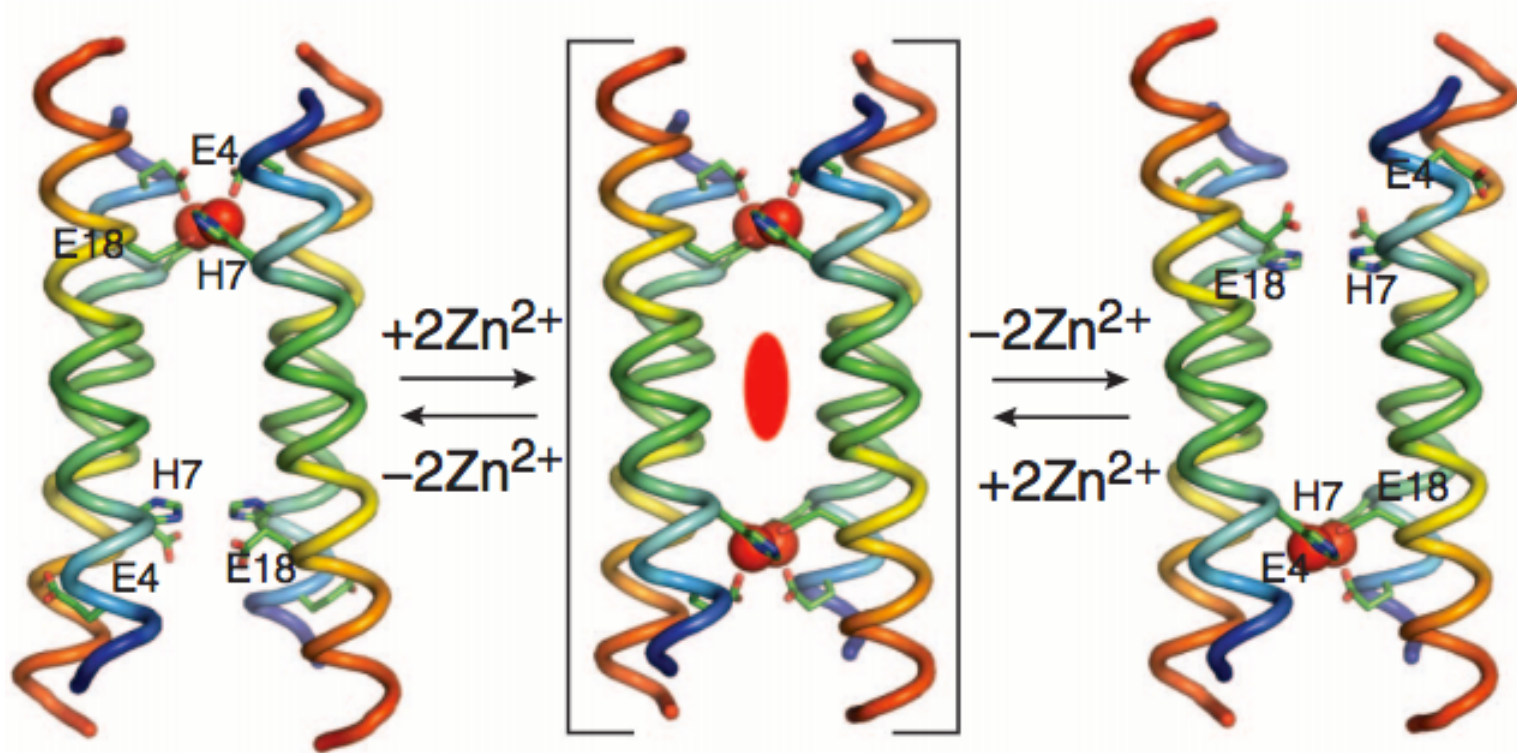


Molecule 1

Molecule 2

# Design of a transporter

Transporters are proteins that carry specific molecules into/out of the cell



- De novo design of a protein that transports zinc ions ( $Zn^{2+}$ ), but not calcium ions ( $Ca^{2+}$ ), across a cell membrane

How well does protein design work?

# How well does protein design work?

- Very impressive recent successes!
- However, one should keep in mind that:
  - Successful protein design projects often involve making and experimentally testing dozens of candidate proteins to find a good one
  - Projects and design strategies that fail generally aren't published
  - Protein design is not yet a matter of simply “turning the crank”