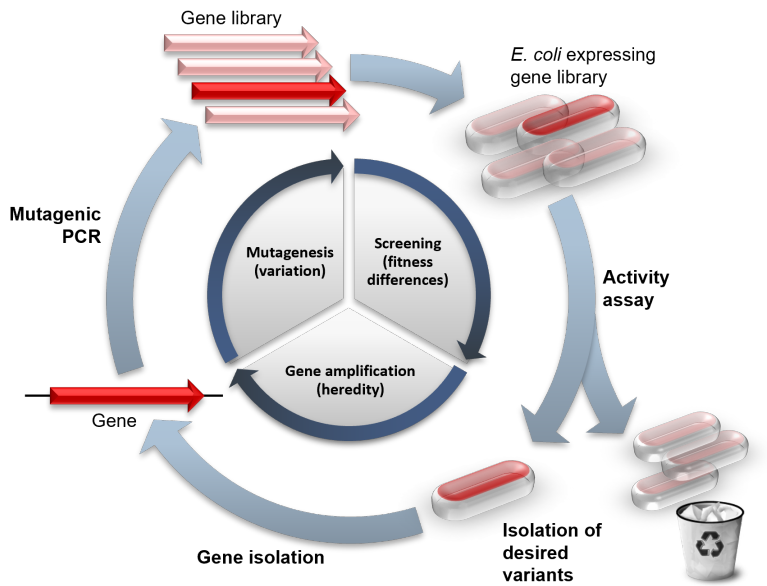
Engineer proteins for desirable properties

Experimental Approaches

- Techniques such as directed evolution, phage display, and yeast display.
- Iterative process of mutation and selection in directed evolution.
- Structural validation using X-ray crystallography and cryo-EM.
- Notable successes include enhanced green fluorescent protein (eGFP) and evolved enzymes.



(credits wiki directed evolution)

Computational Protein Design

- Tools like Rosetta and FoldX for structure prediction and design.
- Enables rapid exploration of vast sequence and conformational spaces.
- Design of novel enzymes, e.g., Kemp eliminase.
- Creation of protein switches and custom therapeutics.

Why Computational Design is Desirable

- Speed: Computational methods can quickly scan vast design spaces.
  - 100 amino acids proteins $\implies 20^{100} \approx 10^{91}$ |atoms in v. universe|
- Precision: Allows for targeted design based on specific criteria.
- Cost-Effective: Reduces the need for extensive lab work and resources.
- Predictive Power: Can forecast how changes in sequence affect structure and function.
- Synergy: Complements experimental methods by narrowing down candidates and guiding lab efforts.

The Problem: Physics-Based Approach

Definition

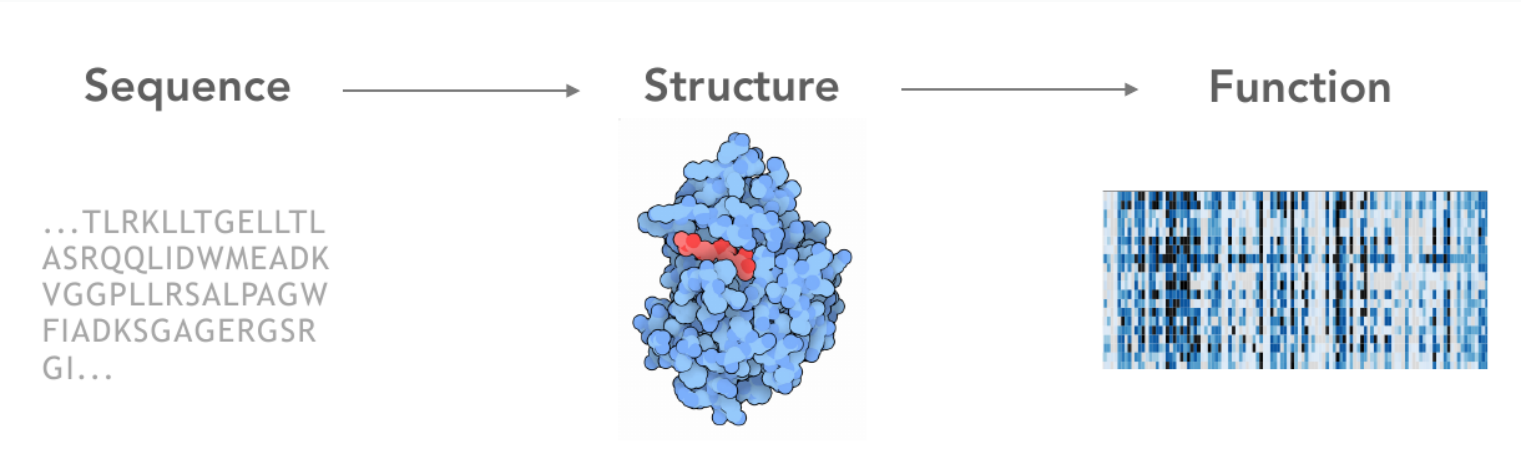
- Predict 3D protein structures from sequences.

Historical Insight

- Structure dictates function.
- Nature's "folds" reveal protein functions.

Modern Bioinformatics Goal

- Design sequences for desired 3D structures and functions.



(credits E. Laine)

Protein Design Principles

The Core Challenge: Optimization

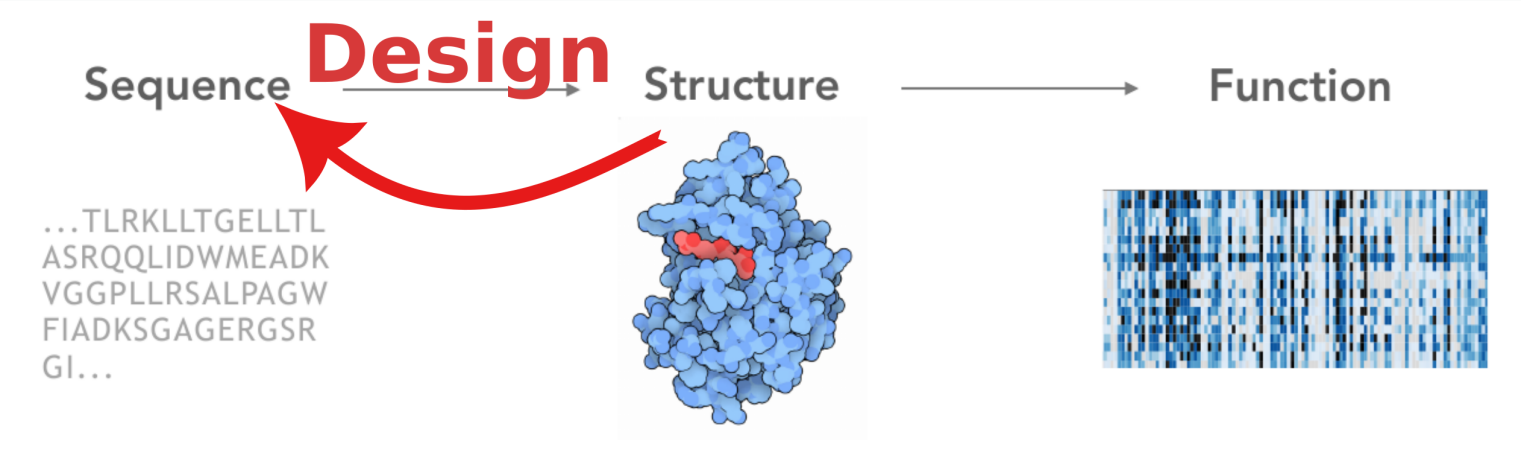
Objective: Design protein sequences to achieve specific structures and functions.

Starting Point: Known Structures

Utilize established protein folds with recognized functions as templates.

The Dual Nature of the Problem

Traditional Folding: Given a sequence, predict its structure. Inverse Folding (Design): For a desired structure, determine or design the optimal sequence.



(credits E. Laine)

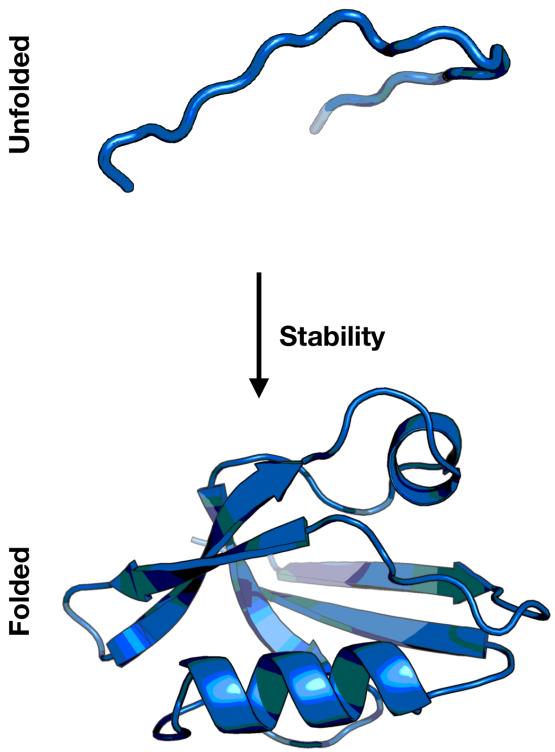
Design = inverse protein design

The Essence of Folding: Free Energy

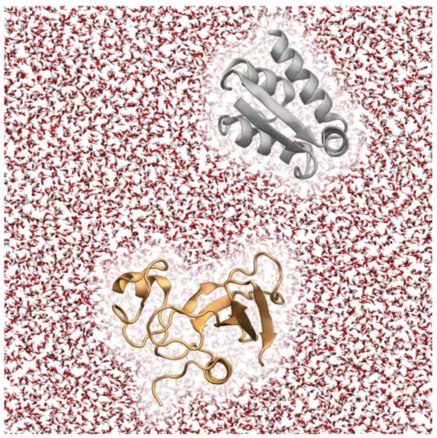
Protein folding is driven by changes in free energy. Equation: $\Delta G = \Delta H - T \Delta S$

Visualizing the Folding Process

Proteins transition from a less ordered (unfolded) state to a highly ordered (folded) structure. Representation: "Unfolded $\rightarrow$ Folded"



Usually not realistic



very long time

a lot of energy

PhD

1 non-transferable model

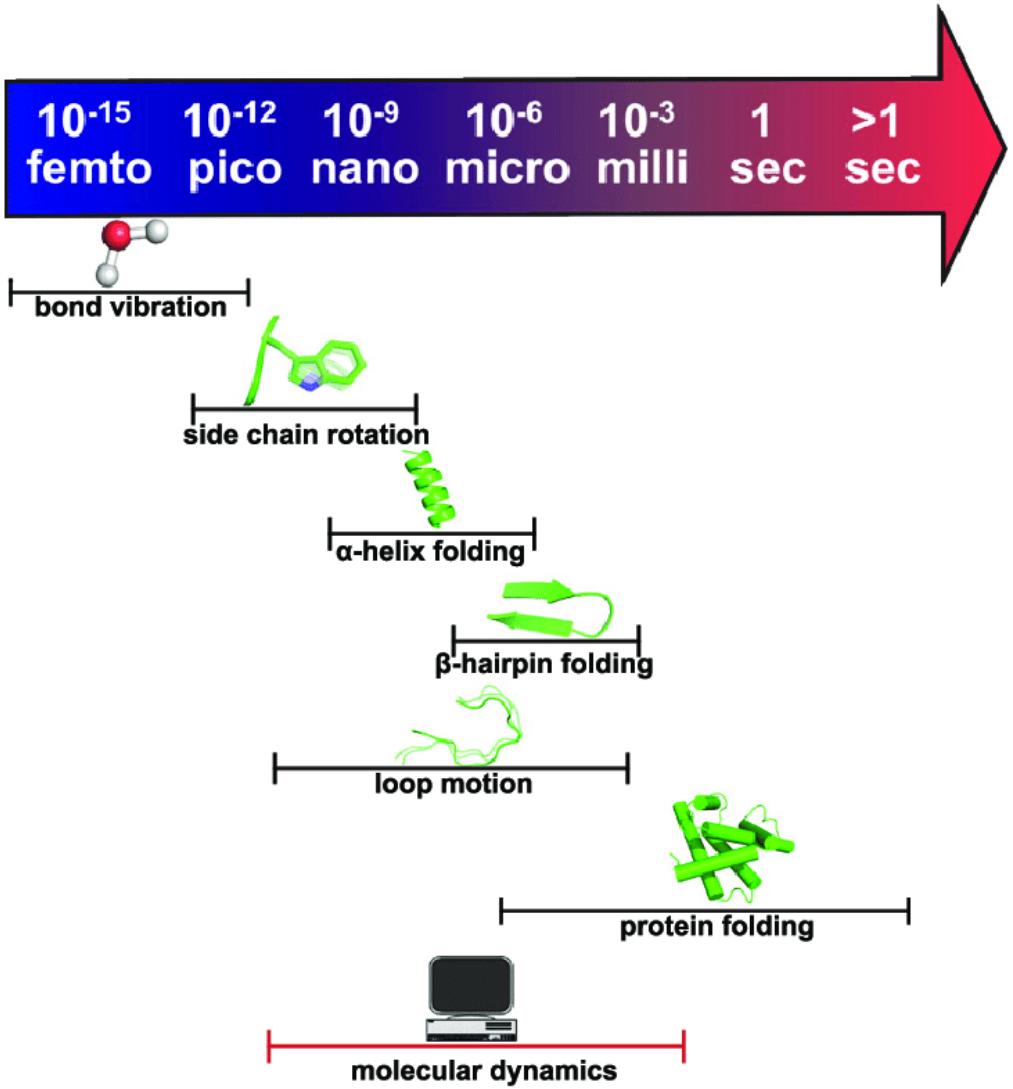
2 ms MD = 20,000 GPU days
500 gigajoule

Burn a Saturn V rocket and deliver
50 ton payload to lunar orbit

1500 gigajoule

(credits Pr. F. Noe, F. U. Berlin)

Usual timescale for the folding process



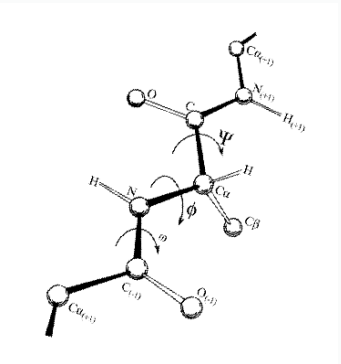
Interesting phenomena take time (credits Werner, Tim, et al. Advanced drug delivery reviews (2012))

Exploring Protein Conformations

The Challenge of Flexibility

Proteins are inherently flexible, leading to a vast conformational space. Combinatorial explosion example: For a rotation () of 30°:

- N=2: 144 conformations
- N=3: 21,000 conformations
- N=5: 430,000,000 conformations



Simplifying with Approximations

Fixed Backbone: Retain the protein's main structure, only vary side chains.
Use of rotamer libraries: Discretized conformations for amino acid side chains.

Modeling the Unfolded State

- Primary approach: The solvated short peptide model.
- Alternative: Statistical models derived from observational data.

Navigating the Protein Sequence Space

Probabilistic Approaches

Delve into the vast sequence space using stochastic methods.

- Markov Chain Monte Carlo (MCMC): A random sampling method to explore possible sequences.
- Genetic Algorithm: Mimics natural selection to optimize sequences.

Deterministic Search for Optimal Conformations

- Aim: Identify the sequence with the lowest possible energy.
- Dead End Elimination (DEE): Systematically eliminates unlikely sequences to pinpoint the Global Minimum Energy Conformation (GMEC).

Understanding the Unfolded State

Modeling Exposed Residues

- In the unfolded state, residues are fully exposed $\implies$ no interactions.
- Energy Model:
  - $E^{uf}(S) = \sum_i E^{uf}(S_i)$
  - Calculates the energy based on individual amino acids.

Refining the Model

- Parameterization for accuracy.
  - Use experimental data and computational methods.

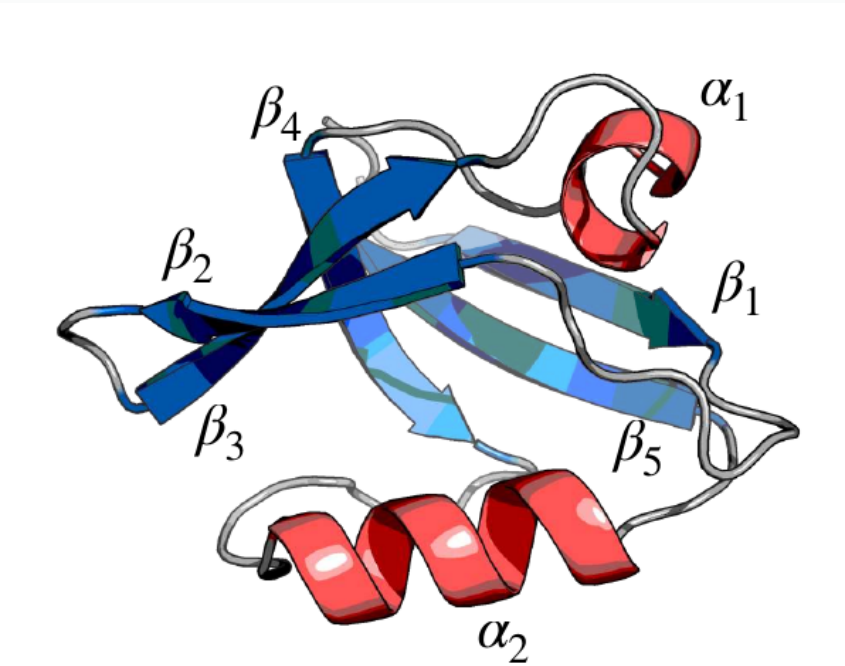
Test case: redesigning the PDZ fold

The target fold: PDZ

- PDZ domains are protein structural domains that are often found in signaling proteins. They play a role in anchoring receptor proteins in the membrane and recruiting specific proteins to specific sites in the cell.

What it does?

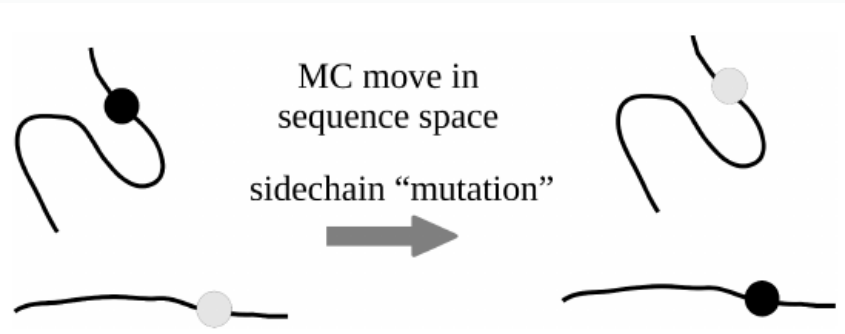
- PDZ domains typically bind to the C-terminus of other proteins, facilitating protein-protein interactions.



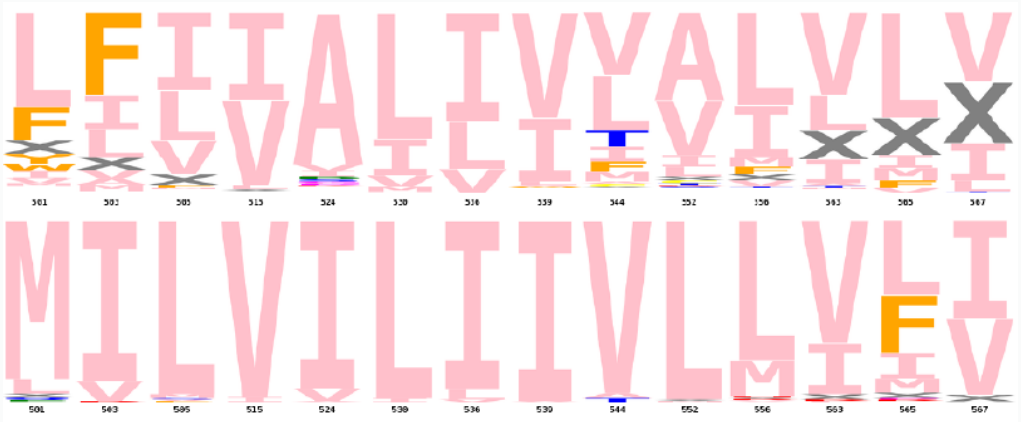
Sampling sequences

The model

- MM + Implicit model + Empirical unfolded model
- 10^8 MCMC steps

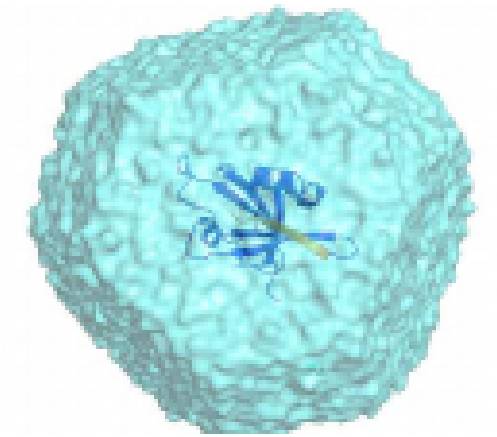


MCMC sampling



MD simulations

You actually don't fold them as the folding is too expensive



For each selected sequence: 1) Solvate then 2) simulate for stability.

MD settings

- $0.2\mu s \leftrightarrow$ 2 weeks on 200 CPUs

(Opuu *et al* (2020) Sci. Rep.)

Design for the binding or catalysis

The problem definition

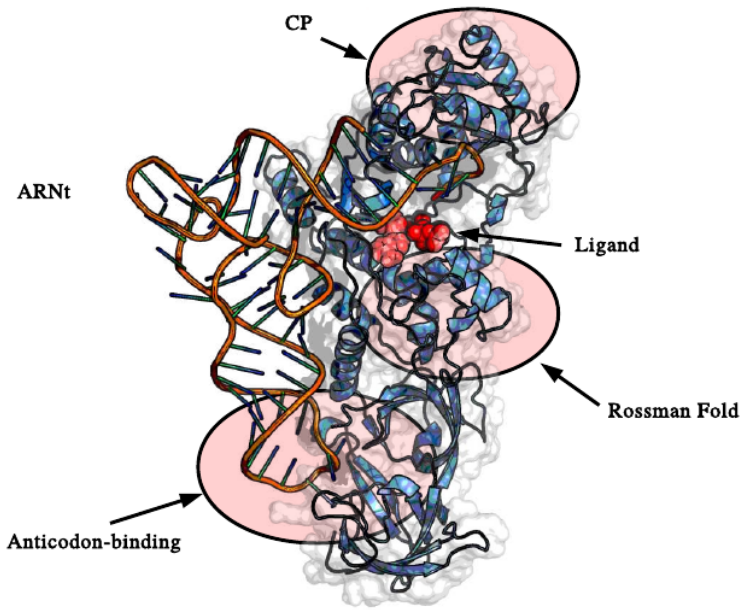
Definition of the problem

Find a sequence that can bind a specific substrat | catalyze a specific reaction
Shift of paradigm!

What for?

Develop enzymes:

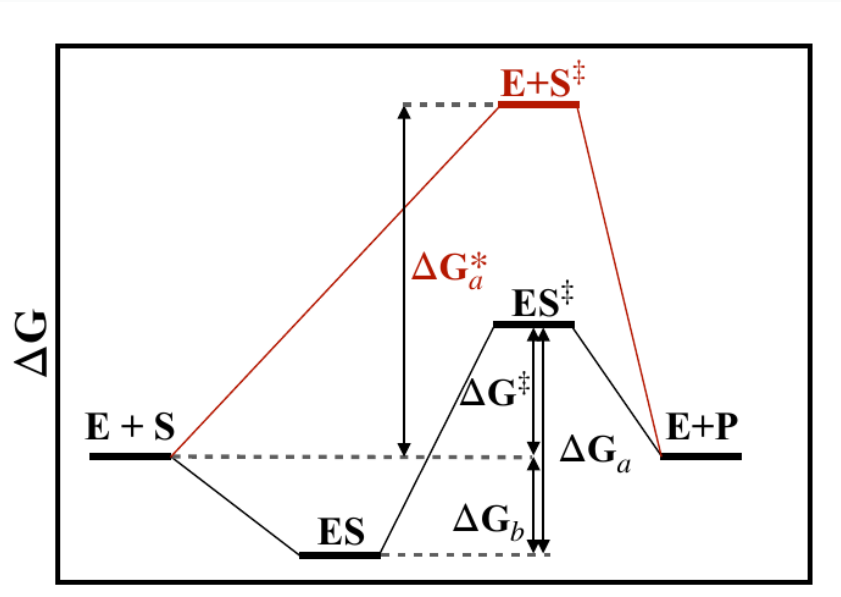
- that degrade plastic
- that catalyze the synthesis of bio-fuel
- that allow the incorporation of new chemistry in proteins



The energy bottleneck

The reaction pathway

- The series of molecular events during a reaction.



Define the optimization problem

- Binder: find mutations that improve the binding ΔG_b
- Catalyzer: find mutations that improve the catalysis $\Delta G^{\ddagger}$

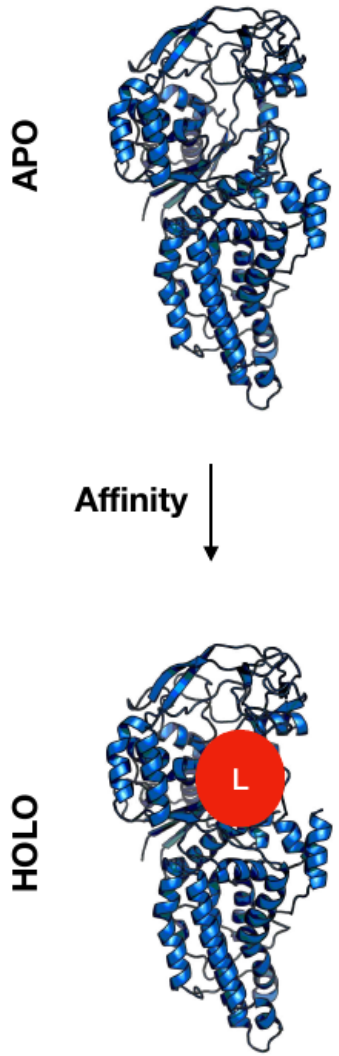
Designing a strong binder

MetRS: binding unatural ligand

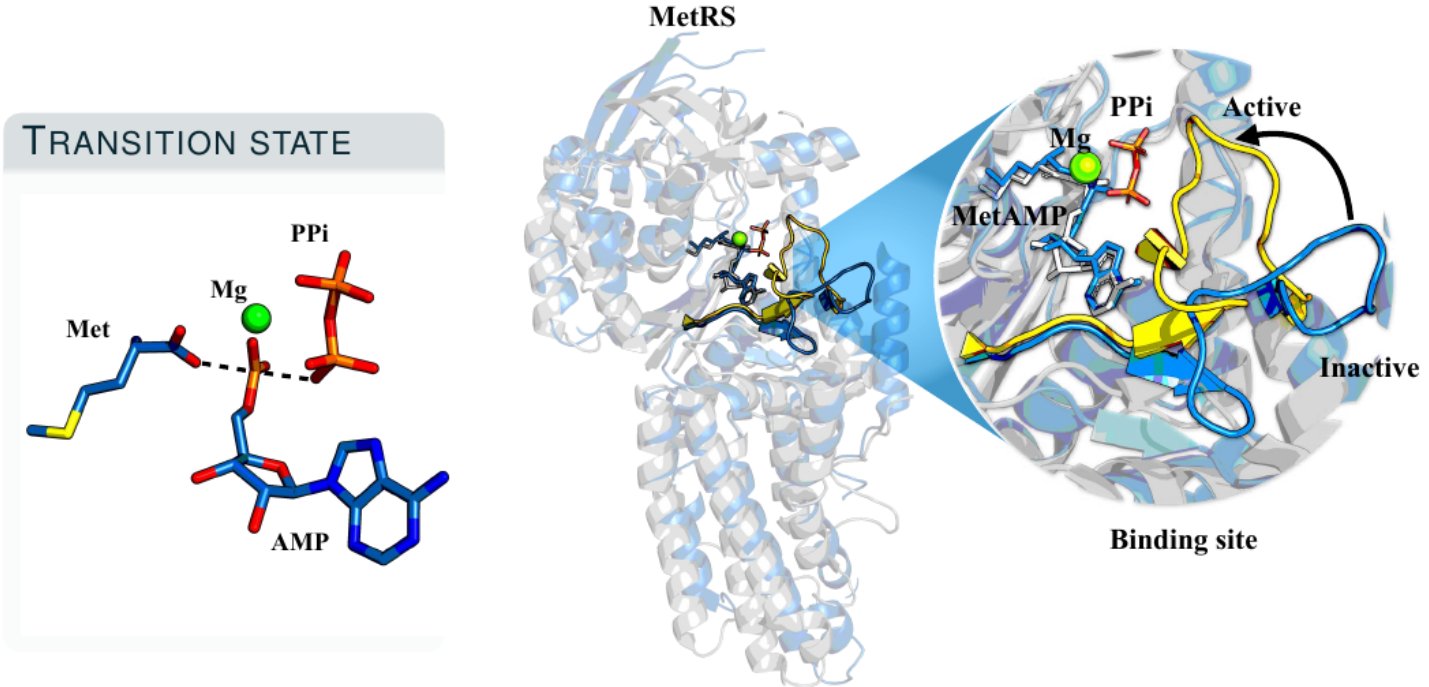
- ΔG(Unbound $\rightarrow$ Bound)

Sampling technique

- Adaptive landscape flattening:
  - MCMC to create a surrogate model of the unbound state ΔG_u
  - Combine with the bound state $\rightarrow$ $\Delta G_b - \Delta G_u$



Designing catalysts: the transition state



Optimize the binding with the transition state to optimize the catalysis.

Conclusion: Computational Protein Design - A Paradigm Shift

Key Takeaways

- Computational vs. Experimental: Computational methods offer speed, precision, and cost-effectiveness, complementing experimental techniques like directed evolution.
- Challenges: Navigating vast conformational and sequence spaces, handling computational intensity, and ensuring accurate approximations.
- Applications: From designing enzymes for bio-fuels and plastic degradation to predicting protein structures and functions.

Future Directions

- Integration: Combine computational and experimental methods for holistic protein design.
- Technological Advancements: Leverage AI and machine learning for enhanced design efficiency.