Protein Simulation I

1. Objectives

- Understand proteins as polymers that transition between unfolded and folded states.
- Identify the structural hierarchy: sequence → secondary structure → tertiary structure.
- Analyze degrees of freedom and understand why conformational search is computationally intractable.
- Describe energy landscape theory and folding funnels.
- Explain the driving forces behind folding and unfolding.
- Quantify solvent exposure using rSASA.
- Understand backbone geometry including bond lengths, bond angles, and dihedral angles.

2. Key Concepts and Definitions

Protein Folding: The transition that a protein takes from a linear chain of polypeptides to its native folded state. A protein must be in its folded state to perform its biological function. Many proteins have their folded states solved in crystal structures. Unfolding can occur due to high temperature or denaturants in solution.

Amino Acids: The individual units that make up a peptide chain. They contain a carbonyl group and an amine group connected by a single alpha carbon. The alpha carbon can be attached to a side chain, which determines its identity, and there are 20 typical amino acids present in human proteins, characterized by polarity and charge. They can chain together by peptide bonds into longer polymers which form proteins. The chemical diversity of amino acid side chains directly shapes the energy landscape that governs protein folding

Levinthal's Paradox: A protein with N = 100 residues has roughly $3^{200} \approx 10^{95}$ possible backbone conformations. Random sampling would take longer than the age of the universe, yet proteins fold in 10^{-6} to 10^{-3} seconds. Folding must therefore follow guided pathways.

Energy Landscape: The free energy available associated with a given conformation of an amino acid chain. The landscape potential is determined by the forces acting on the atoms in the given conformation, and a minimum occurs when the gradient is zero in a convex area. The landscapes for wild-type proteins under evolutionary pressure are thought to be generally smooth with few local minima and a steep global minimum representing the folded state.

Driving Forces: Forces that promote folding include hydrophobicity, hydrogen bonding, and van der Waals interactions. Forces opposing folding include the increase in conformational entropy of the unfolded state and electrostatic repulsion.

rSASA: Relative solvent accessible surface area, defined as $\text{rSASA} = \text{SASA}_{\text{res}} / \text{SASA}_{\text{dip}} \in [0, 1]$, where SASA_{dip} is the surface area of the isolated dipeptide. Values near 0 indicate a buried core residue; values near 1 indicate full solvent exposure.

CASP: The Critical Assessment of Structure Prediction competition in which different models compete to most accurately predict the unreleased structure of a provided amino acid sequence.

Folding Driving Forces: The forces that act on a peptide chain that define the energy landscape and direct folding into the native state. Forces that contribute to folding include the hydrophobic effect, van der Waals interactions and hydrogen bonding.

Solvent Accessible Surface Area (SASA): The surface area of a given amino acid that could potentially be exposed to solvent, measured in square angstroms. The relative SASA (rSASA) normalizes this value relative to the total surface area of the amino acid to give a number in the range [0, 1]. The rSASA can be used to quantify if an amino acid is in the core; values below 10^{-2} to 10^{-3} usually indicate it is in the core, otherwise it is a surface residue.

Secondary Structure: The local structure of a protein, categorized into several different motifs. These include alpha helices and beta strands, both of which are heavily stabilized by hydrogen bonds. The amino acids in secondary structures often have strict bond angle requirements compared to more disordered regions; alpha helices often have phi and psi angles of -60 and -45 respectively while the same angles for a beta sheet are closer to -135, 135.

Bond Angles: The angle made by three adjacent covalently bonded atoms which can be plotted as a histogram or probability distribution. These distributions made for proteins and their side chains have relatively small variances (RMSD of 5-10 angstroms), since such angles are determined by fixed stereochemistry.

Dihedral Angles: The angle made by the two normal vectors of the adjacent planes that are made by four covalently bonded atoms in sequence. The most relevant ones in proteins are the phi angle, which rotates around the alpha carbon-amino nitrogen bond, and the psi angle, which rotates around the alpha carbon-carbonyl carbon bond. Because bond lengths and bond angles are relatively constrained, dihedral angles represent the primary degrees of freedom in protein structure

Ramachandran Plot: A graph showing the occurrence of pairwise phi and psi angles for a given protein or group of amino acids, typically with phi angles on the x-axis and psi angles on the y-axis. The distributions reflect steric constraints and typically reveals certain clusters associated with secondary structures such as alpha helices and beta sheets.

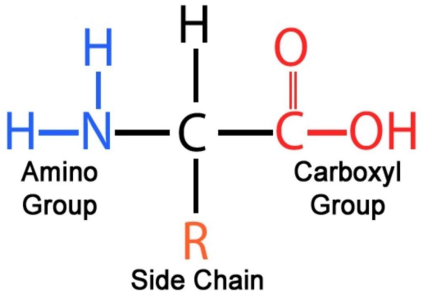
CASP: Critical Assessment of Structure Prediction, a community benchmark evaluating how well computational methods predict protein structure from sequence alone. Accuracy is measured with the GDT score.

3. Main Content

3.1 Proteins and Amino Acids

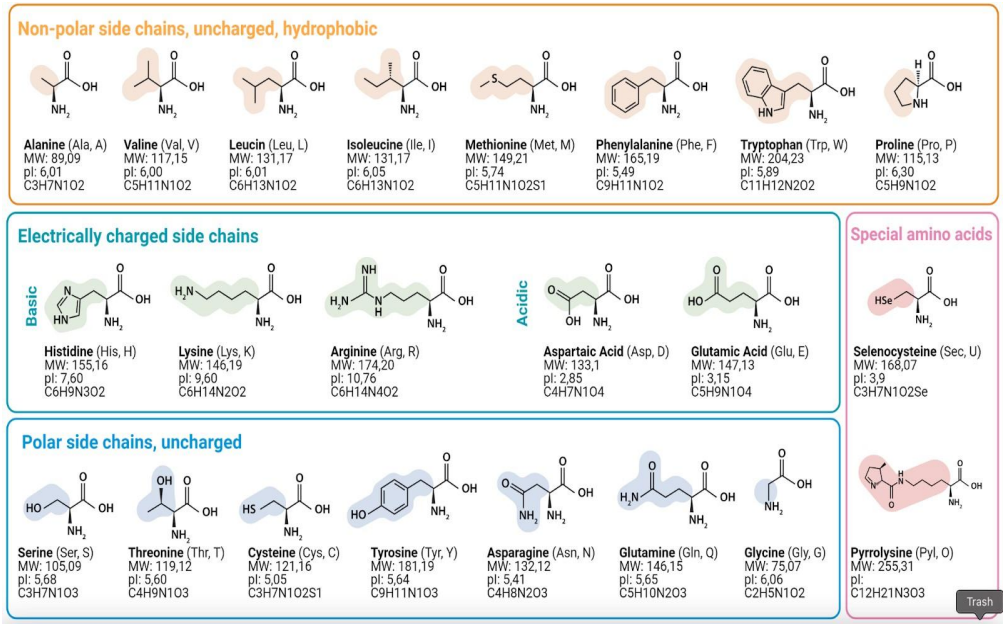
Proteins are linear chains of amino acids linked by peptide bonds. Different proteins have different amino acid sequences, and the sequence determines the 3D structure and biological function.

All amino acids share the same backbone: an amino group and a carboxyl group connected through a central alpha carbon (Cα). The side chain (R group) attached to Cα is what differentiates each amino acid.



There are 20 naturally occurring amino acids, grouped by side chain character:

- **Nonpolar / hydrophobic:** Gly, Ala, Val, Leu, Ile, Met, Phe, Trp, Pro. These tend to be buried in the protein core.
- **Polar uncharged:** Ser, Thr, Cys, Tyr, Asn, Gln. These can form hydrogen bonds and are often surface-exposed.
- **Electrically charged (acidic):** Asp, Glu. Negatively charged at physiological pH.
- **Electrically charged (basic):** Lys, Arg, His. Positively charged at physiological pH.



3.2 Levinthal's Paradox and Degrees of Freedom

Each amino acid backbone is described by dihedral angles (ϕ,ψ), giving ~ 2N degrees of freedom.

Deep Dive: Degrees of Freedom & Levinthal's Paradox

Key idea: Even coarse discretization yields exponential complexity. Assume μ allowed states per dihedral angle:

$$N_c \sim \mu^{2N}$$

For $\mu = 3$, $N = 100$:

$$N_c \approx 3^{200} \approx 10^{95}$$

Folding time estimate:

$$T_{fold} \sim N_c T_{sample} \gg T_{universe}$$

Intuition: Proteins cannot sample all states; folding must be guided.
References: Levinthal (1969), Dill & MacCallum (2012)

(from source 1)

For $\mu = 3$, $N = 100$: $Nc \approx 3^{200} \approx 10^{95}$. At 10^{-12} s per sample, folding time would exceed the age of the universe (10^{17} s). Yet real proteins fold in 10^{-6} to 10^{-3} s. This contradiction is Levinthal's Paradox.

The resolution is that proteins do not search randomly. The energy landscape is funneled, guiding the chain toward the native state without exhaustive sampling.

In terms of degrees of freedom, subtracting bond length constraints (N−1), bond angle constraints (N−2), and overall translation/rotation (6) from 3N total coordinates gives N−3 effective dihedral angle degrees of freedom.

3.3 Driving Forces of Folding

Hydrophobicity appears to drive folding, with hydrogen bonds, electrostatic forces, and van der waals forces between AAs being used to stabilize the interaction. Hydrophobicity can be measured as 0 - 1 where 0 is hydrophobic and 1 is hydrophilic. Hydrophobic AAs typically occur in the core of the protein, while hydrophilic ones arise on the surface. This leads to the idea of Solvent Accessible Surface Area, which is a measure of how much of the protein's surface area is accessible to the solvent.

3.4 Energy Landscape and Folding Funnel

The energy landscape is a mapping from every possible conformation (described by all dihedral angles) to a free energy value:

Deep Dive: Energy Landscape & Funnels

Energy is a function of all degrees of freedom:

$$E = E(\{\phi_i, \psi_i\}_{i=1}^{N})$$

Smooth funnel: guides folding efficiently.
Rough landscape: traps in local minima (misfolding).
Representative step (gradient descent intuition):

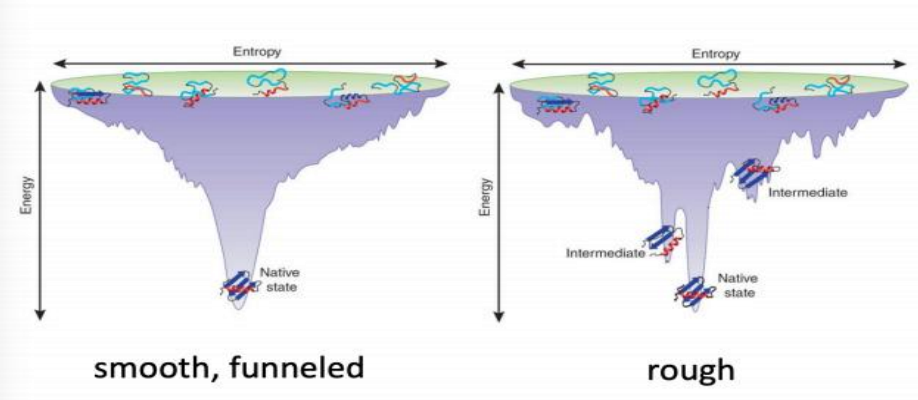
$$x_{t+1} = x_t - \eta \nabla E(x_t)$$

Insight: Evolution selects sequences with funneled landscapes.
References: Onuchic & Wolynes (2004), Wolynes et al. (1997)

(from source 1)

A smooth, funneled landscape guides folding efficiently toward the native state. A rough landscape has local minima that trap the protein in misfolded intermediates.

Evolution selects for sequences that minimize frustration (conflicting interactions), producing landscapes that are globally biased toward the native state even if locally rough.



(from source 3)

3.5 Solvent Accessible Surface Area (rSASA)

Solvent accessible surface area (SASA) measures how much of a residue's surface can be reached by a solvent probe (rolling sphere model). The relative version normalizes this:

Deep Dive: Solvent Accessible Surface Area

Relative exposure:

$$\mathrm{rSASA} = \frac{\mathrm{SASA}_{\mathrm{res}}}{\mathrm{SASA}_{\mathrm{max}}} \in [0, 1]$$

Interpretation:

- rSASA ≈ 0: buried (core)
- rSASA ≈ 1: exposed (surface)

Key idea: Surface definition depends on probe size (rolling sphere model).
Insight: Predicting core vs surface residues strongly correlates with structure accuracy.
References: Lee & Richards (1971), Chothia (1976)

(from source 1)

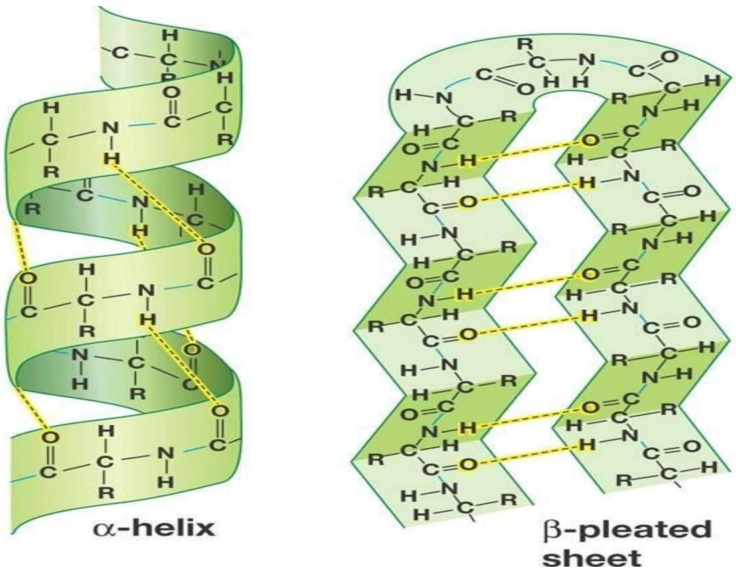
A value of rSASA ≈ 0 means the residue is fully buried (core); rSASA ≈ 1 means fully exposed (surface). Core residues are typically defined by rSASA ≤ 10^{-2} to 10^{-3}.

In practice, hydrophobic residues concentrate in the core (low rSASA) while hydrophilic and charged residues appear on the surface (high rSASA). Predicting which residues are core-buried is strongly correlated with structure prediction accuracy.

3.6 Secondary Structure: α-Helices and β-Sheets

α-Helix: right handed, three turns. There are ~3.6 AA/turn, which corresponds to roughly 100 degrees per AA. They're stabilized by hydrogen bonding. Side chains are found on the outside of the helix, typically pointing toward the N-terminus. Commonly formed by Met, Ala, Leu, and Glu. The phi, psi angles for this structure are about (-60, -45).

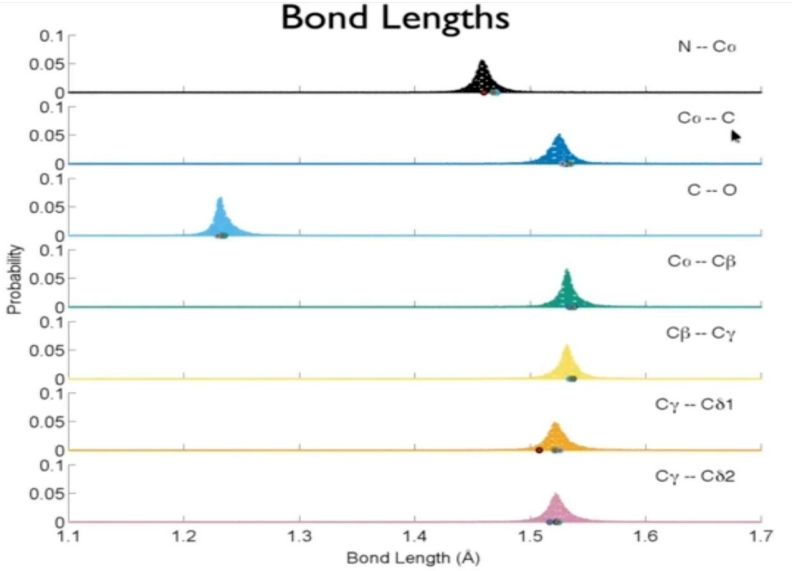
β-Sheet: peptide backbone is fully extended, can run parallel or antiparallel where one strand is hydrogen bonded to the next. Commonly formed by Val, The, Tyr, Trp, Phe, Ile. The phi, psi angles for this structure are about (-135, 135).



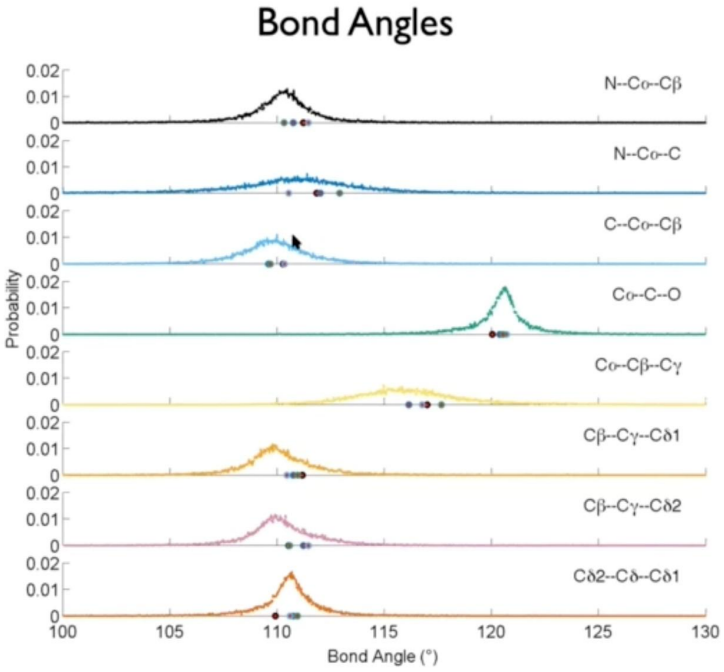
3.7 Bond Geometry: Lengths and Angles

From a dataset of ~62,000 monomeric crystal structures in the PDB, we can extract statistical distributions of bond lengths and bond angles.

Bond lengths: Most backbone bonds (N–Cα, Cα–C', Cα–Cβ, side chain C–C) cluster around 1.5 Å. The C=O double bond is shorter at about 1.23 Å. These distributions are very narrow, meaning bond lengths are essentially fixed.

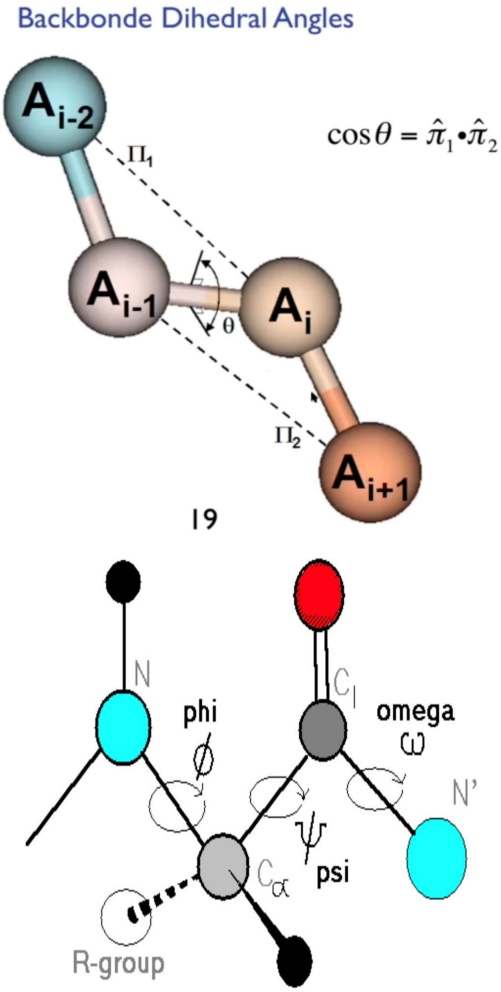


Bond angles: Backbone and side chain angles are centered around 110° for tetrahedral carbons and around 120° for the carbonyl carbon (Cα–C'–O). Distributions are narrow with an RMSD of approximately 5°. Note these are measured in degrees, not angstroms.



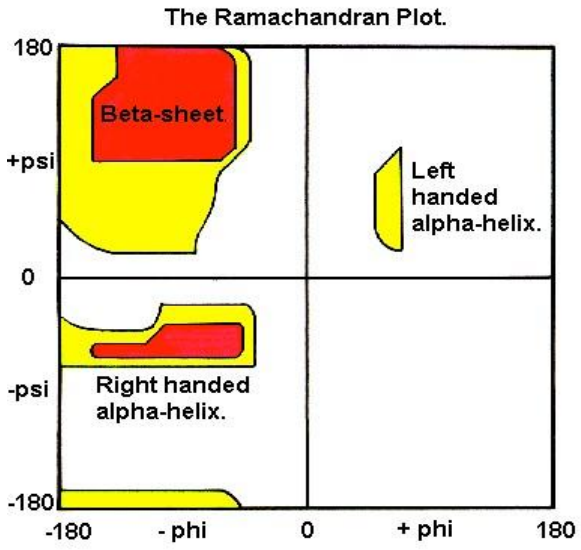
3.8 Backbone Dihedral Angles

Whereas the bond angle is defined by two atoms and the bond length three, these are defined by the position of four atoms. Three atoms can define a plane, two groups of three atoms defines two planes, and the two planes generate an angle between them, which is defined as a dihedral angle. They can be calculated with backbone or side chain atoms. The *phi* angle is the angle around the -N-CA- bond (where 'CA' is the alpha-carbon). The *psi* angle is the angle around the -CA-C- bond. The *omega* angle is the angle around the -C-N- bond (i.e. the peptide bond).



3.9 Ramachandran Plot

These graphs plot the phi and psi angles of atoms. Invented by Ramachandra through his use of atom modeling, they are useful in determining the structure of a protein and estimating how much it is able to move. The yellow regions correspond to greater movement (generated by using smaller atoms) while the red regions correspond to lesser movement (generated by using larger atoms). Beta sheets and alpha helices have been found to correspond to specific areas of the plot.



3.10 CASP: Benchmarking Structure Prediction

CASP evaluates computational structure prediction methods using proteins whose structures have been solved experimentally but not yet released. Prediction quality is measured using GDT (Global Distance Test), which counts the fraction of Cα atoms within a distance cutoff of the true structure.

Prediction accuracy generally decreases with protein length and with decreasing sequence similarity to known structures. Even strong methods can fail on specific targets, showing the problem is not fully solved.

4. Discussion/Comments

Protein Folding Problem (Levinthal's Paradox): How do proteins spontaneously fold from a linear chain of amino acids into their correct structures? Theoretically, if a protein tried every possible conformation to find its most stable state, it would take longer than the age of the universe. However, in reality, proteins fold within seconds. This contradiction is known as Levinthal's Paradox. Rather than relying on random sampling, it is now believed that proteins follow multiple guided folding pathways (rough landscapes), forming intermediate structures within energy minimums that help lead to the final folded form.

Quite astonishingly, Google's AlphaFold has managed to tackle this problem quite successfully, being able to accurately predict protein folding and structure based off of amino acid sequence alone by utilizing deep learning methods, using established protein databases to recognize protein patterns and to conducting multiple sequence alignments (MSA) to better hypothesize structure.

Degrees of Freedom (DoF): The concept of DoF in molecules has many layers that work as limitations governing molecular behavior. When you start, theoretically you could have 3N possible movements (N is the number of atoms and this describes x,y,z coordinates), molecules then lose degrees of freedom through various constraints. First, bond lengths are fixed, removing N-1 degrees of freedom. Then, bond angles introduce further restrictions, reducing mobility by another N-2 degrees. Finally, dihedral angles - the rotational possibilities around bonds - further constrain the molecule, ultimately leaving just N-3 degrees of freedom. 3N−6 of those account for overall translation and rotation and would be the result of

the summation of all the previous limitations. Therefore, the number of DoF would be 3N-6 and would account for all the constraints.

Computational Modeling of Protein Folding

- Proteins can be modeled using **potential energy (force field) functions** that approximate atomic interactions
- Total energy is decomposed into:
  - **Bonded terms**: bond lengths, bond angles, dihedral angles
  - **Non-bonded terms**: van der Waals interactions (Lennard-Jones potential), electrostatics (Coulomb's law
- **Molecular Dynamics (MD):** Simulates protein motion over time using **Newton's laws of motion (F = ma)**. Generates trajectories showing how a protein explores conformational space
- **Monte Carlo Sampling:** Randomly samples conformations and accepts/rejects changes based on energy

5. Suggested readings:

Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2 .

Baldwin, R. L. Energetics of Protein Folding. *Journal of Molecular Biology* **2007** , *371* (2), 283–301. https://doi.org/10.1016/j.jmb.2007.05.078 .

Relevant Papers

Paper 1: Evolution, Energy Landscapes, and the Paradoxes of Protein Folding – Peter G. Wolynes (2015)

This paper develops the modern energy landscape theory of protein folding, providing an explanation for how proteins can fold so quickly despite the enormous number of possible conformations described by Levinthal's paradox. Wolynes argues that proteins don't explore conformational space randomly, but instead fold on a "funneled" energy landscape, where many different unfolded states all converge toward a single low-energy native protein structure. A key idea that is introduced is "minimal" frustration, suggesting that through evolution, protein sequences have been optimized so that most interactions are energetically compatible, so folding pathway are biased and efficient rather than random and slow. The paper also connects this to thermodynamics and kinetics by showing that protein folding is not only governed by the final stability of the folded state, but also by the shape of the energy landscape that determines how the protein gets there. The energy landscape contains hills and valleys, and as a protein folds, it moves downhill in free energy, showing how folding intermediates and transition states can arise naturally from the shape of this energy landscape. Overall, the summary's discussion on protein folding energy landscapes and Levinthal's paradox tie closely into the discussion of this paper. The summary's

mention of folding driving forces like hydrophobicity and var der Waals also fits into this framework as these interactions shape the landscape that Wolynes describes. Notably, this paper builds upon the core theory of protein folding by unifying multiple ideas into one framework and emphasizing evolution's role in optimizing proteins to fold efficiently.

Paper 2: Solution of Levinthal's Paradox and a Physical Theory of Protein Folding Times – Dmitry N. Ivankov et al. (2020)

This is a review paper that focuses on resolving Levinthal's paradox and explaining how proteins fold efficiently from a kinetic viewpoint. It emphasizes that protein folding is not random, but a stochastic process guided by small energetic differences between conformations. The paper talks about how even slight energy preferences can significantly reduce the number of accessible states, allowing proteins to follow preferred pathways toward their native structure. The paper also emphasizes the importance of folding intermediates and transition states, which act as steppingstones across the energy landscape, explaining how folding rates are determined not only by product stability but also the height of energy barriers between states. This paper directly ties into our discussion on Levinthal's paradox and folding dynamics. It also connects to our discussion of degrees of freedom, showing how constraints on bond angles and rotations can reduce the search space.

Paper 3: **Highly accurate protein structure prediction for the human proteome – Tunyasuvunakool et al.**

This paper expands on AlphaFold by applying it at scale to predict the structures of entire proteomes, including nearly all human proteins. The study demonstrates how deep learning models can generate high confidence structural predictions for millions of proteins. The system uses the same core ideas as AlphaFold2, and leverages multiple sequence alignments, evolutionary conservation, and attention-based neural networks to infer spatial relationships between amino acids and build 3D structures. A key contribution of this paper is showing that protein structure prediction is now a practical tool for biology. The paper also introduces confidence metrics like per-residue confidence scores that allow researchers to assess which parts of a predicted structure are reliable.