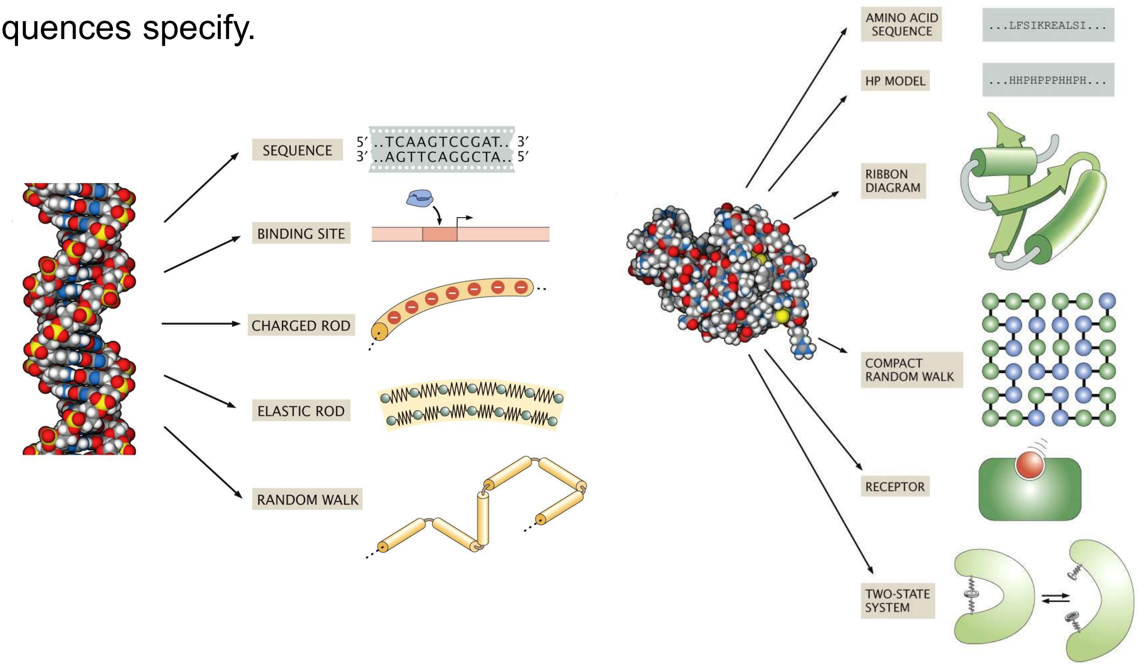
Quantitative principles in biological systems

8. Sequences and spin glass models

Spring 2026

Sequences specify.



Phillips et al

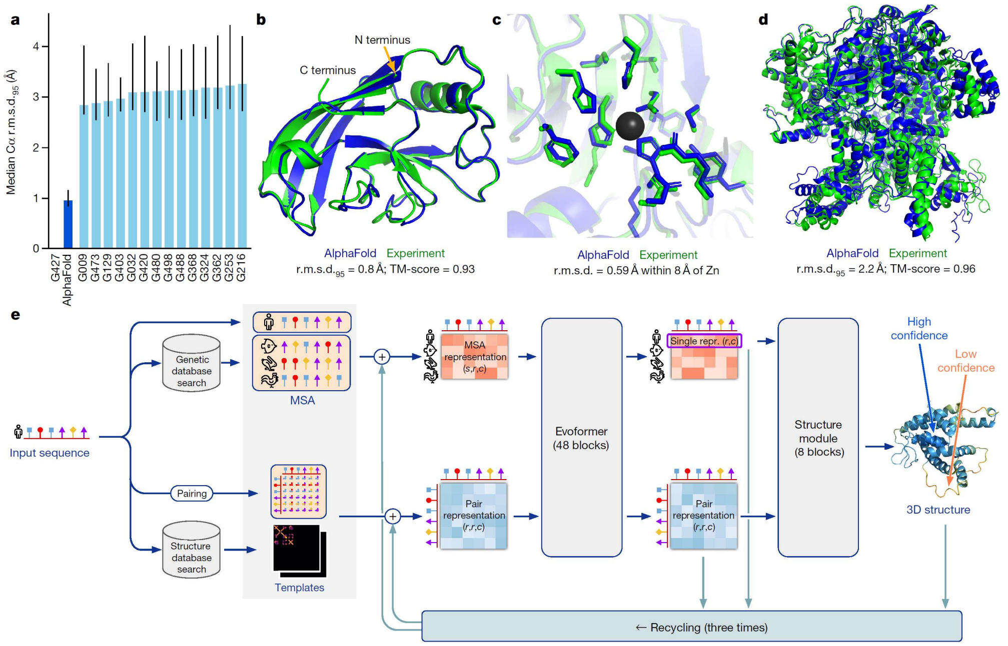
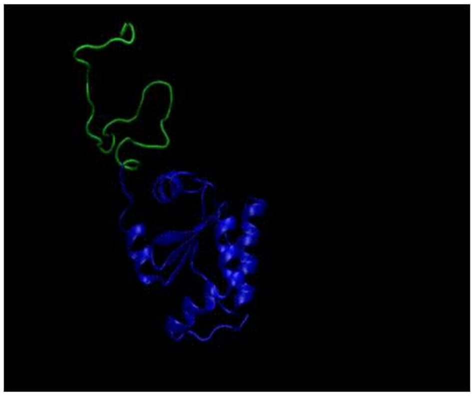
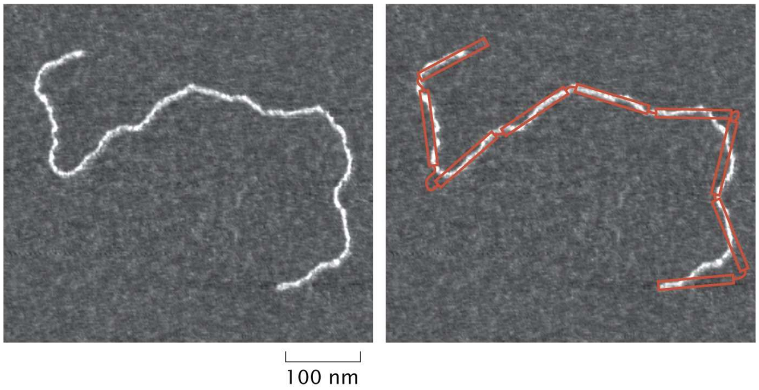


Fig. 1 | AlphaFold produces highly accurate structures. a, The performance   An example of a well-predicted zinc-binding site (AlphaFold has accurate side

Jumper et al. *Nature* (2021)

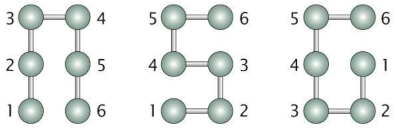
Random polymers almost never fold.





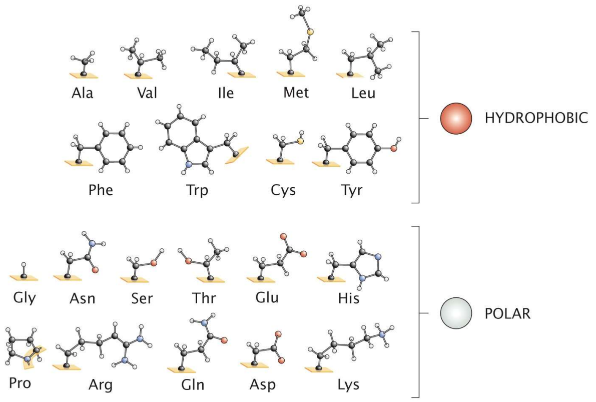
Wiggins et al. *Nat Nanotechnol* (2006)
Wikipedia

Proteins as compact polymers with interacting residues



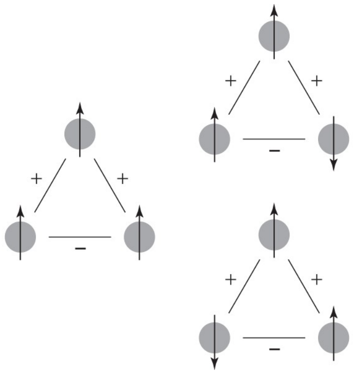
HPHPHP

PHPPHP



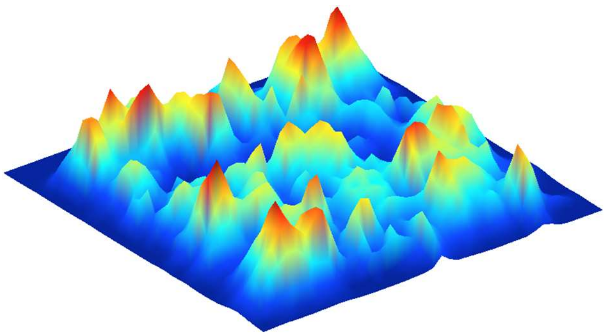
Phillips et al

Spin glass will be our model system for complex systems.



$$E = -\sum_{ij} J_{ij} s_i s_j$$

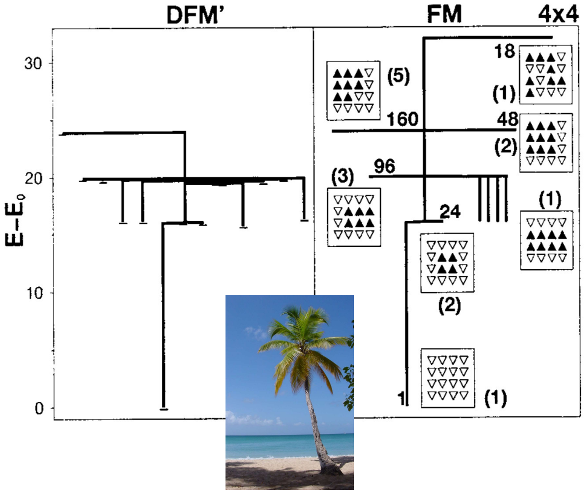
Frustration → Rugged landscapes



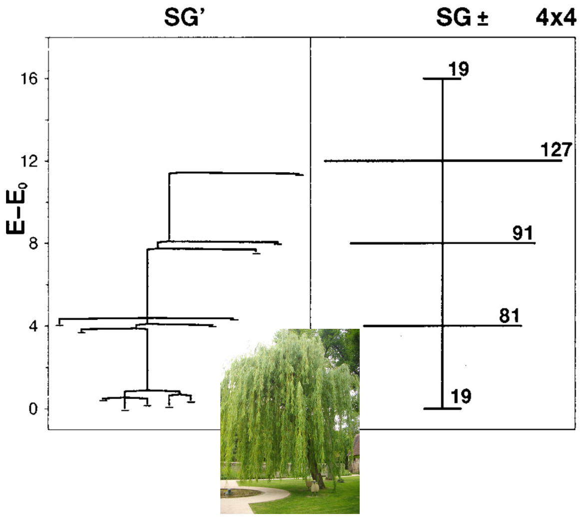
Bialek

Example landscapes for small systems:

Minimize frustration → Funnel landscape



Frustration → Rugged landscapes

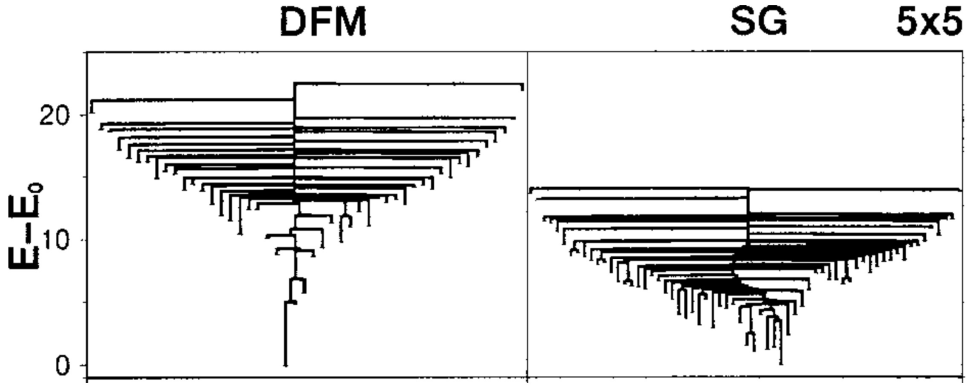


Garstecki et al. *Phys Rev E* (1999)

Example landscapes for small systems:

Minimize frustration → Funnel landscape

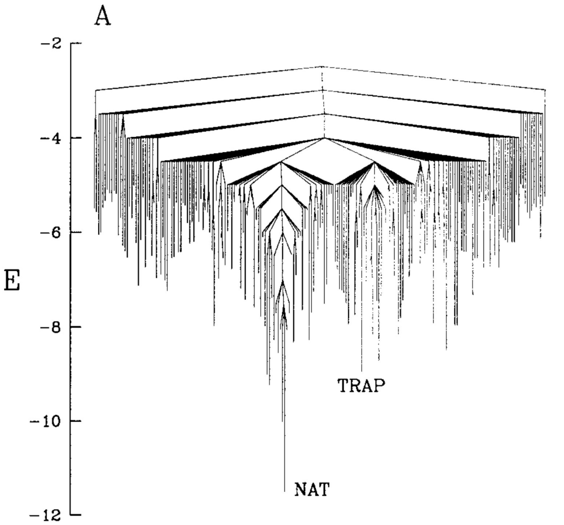
Frustration → Rugged landscapes



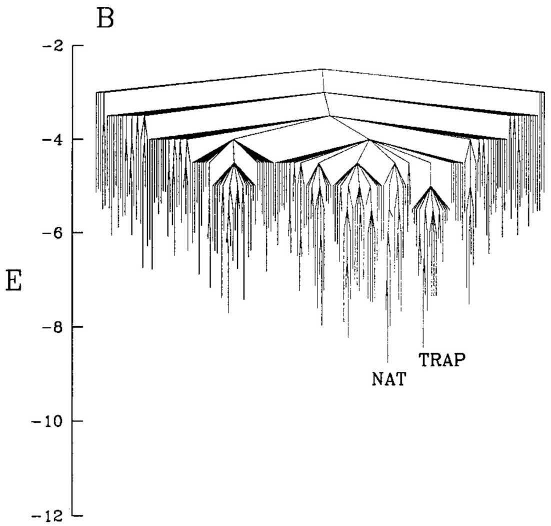
Garstecki et al. *Phys Rev E* (1999)

Example landscapes for small systems:

Minimize frustration → Funnel landscape

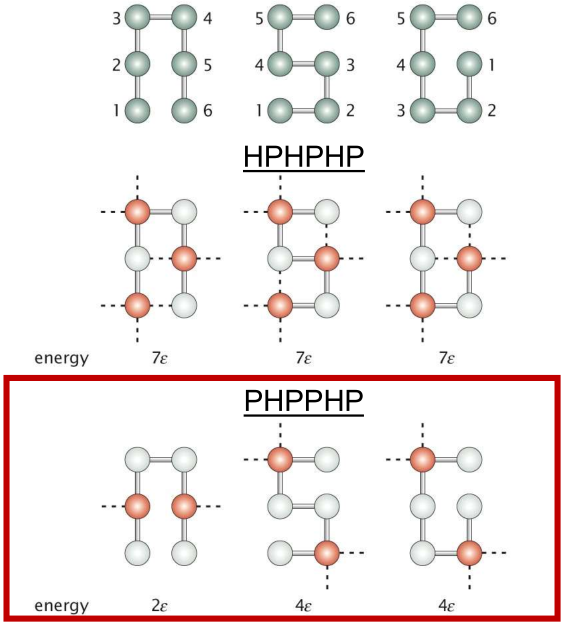


Frustration → Rugged landscapes

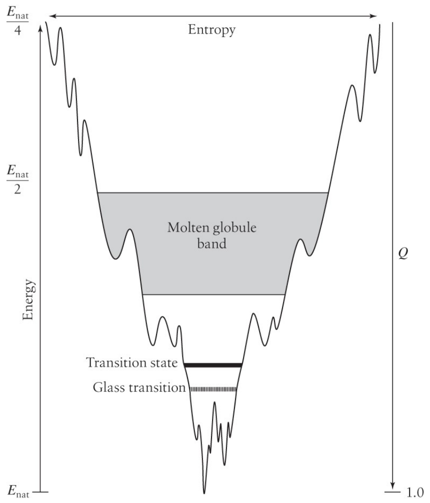


Garstecki et al. *Phys Rev E* (1999)

Proteins as compact polymers with interacting residues

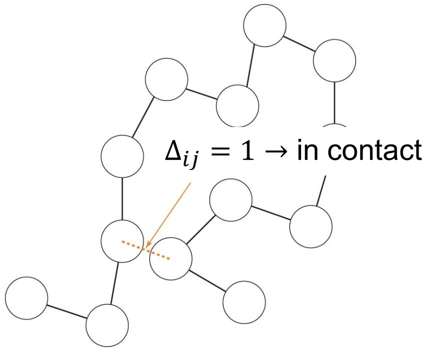


Minimize frustration → Funnel landscape



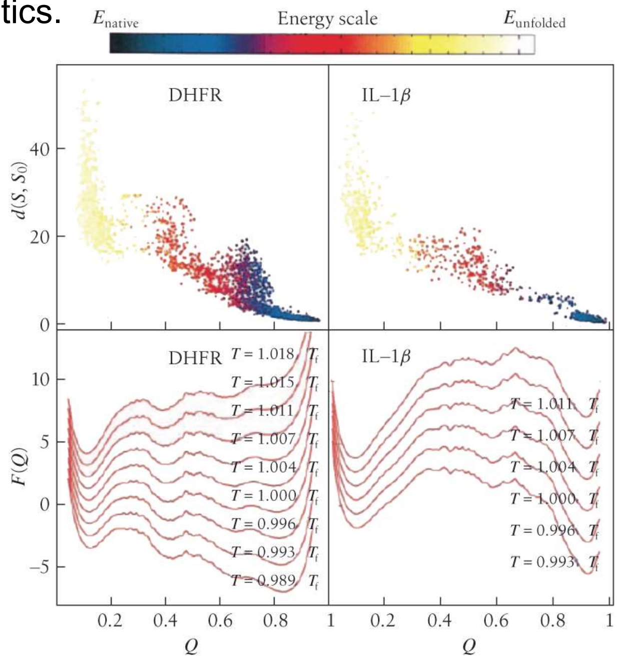
Onuchic et al. *PNAS* (1995)
Phillips et al

Native conformation predicts folding kinetics.



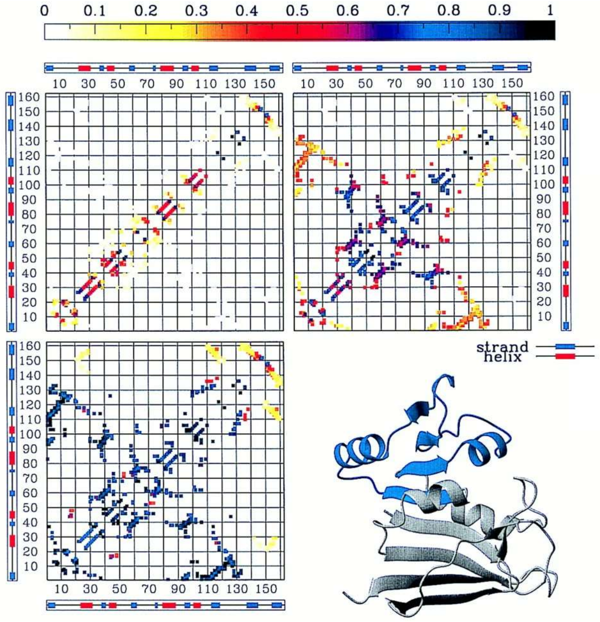
$\Delta_{ij} = 1 \rightarrow$ in contact

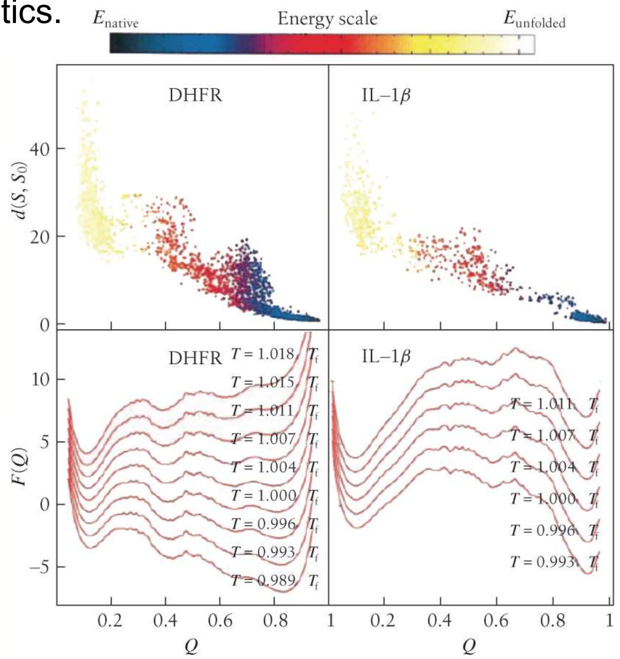
$$E = -\sum_{ij} C_{ij}^{\text{native}} \Delta_{ij} + \cdots$$



Clementi et al. *PNAS* (2000)

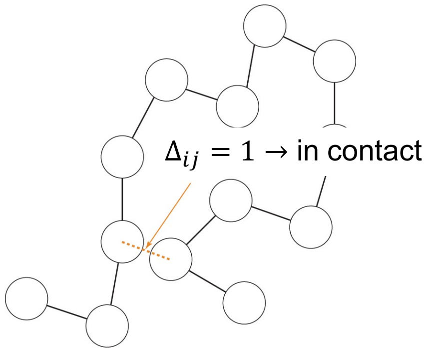
Native conformation predicts folding kinetics.





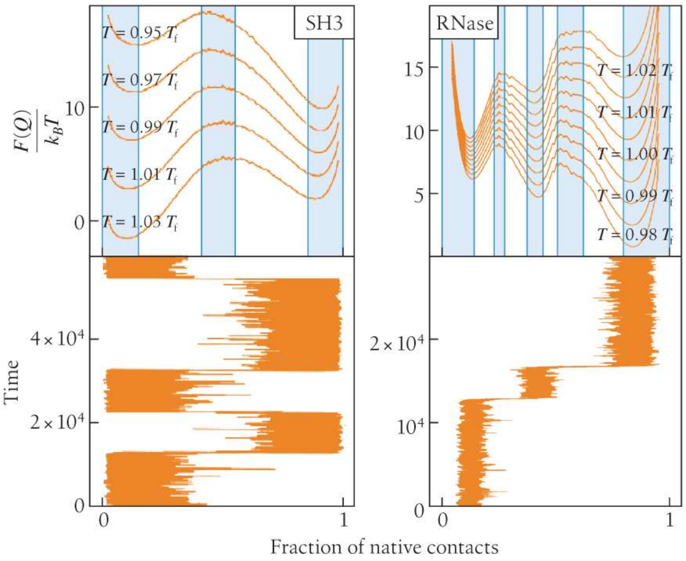
Clementi et al. *PNAS* (2000)

Native conformation predicts folding kinetics.



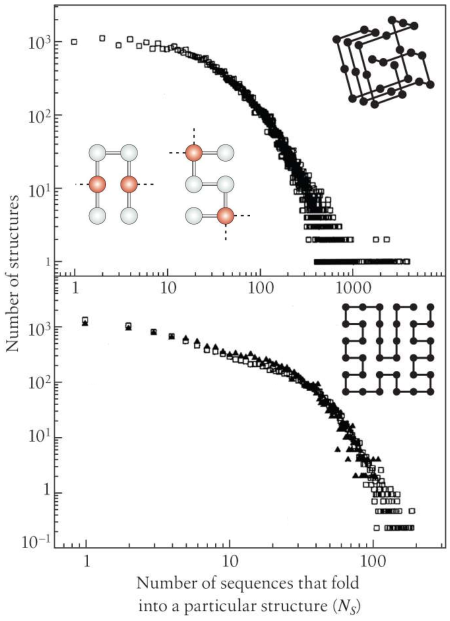
$\Delta_{ij} = 1 \rightarrow$ in contact

$$E = -\sum_{ij} C_{ij}^{\text{native}} \Delta_{ij} + \cdots$$



Clementi et al. *J Mol Biol* (2000)

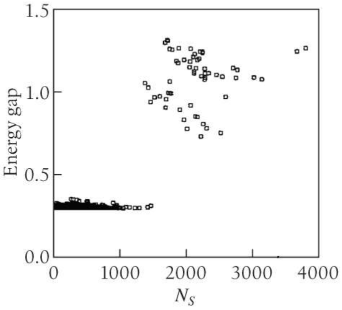
The problem of protein design

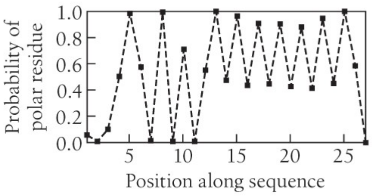


Emergence of Preferred Structures in a Simple Model of Protein Folding

Hao Li, Robert Helling,* Chao Tang,† Ned Wingreen

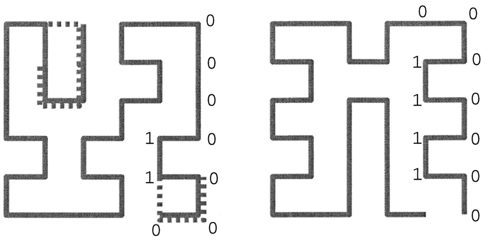
Protein structures in nature often exhibit a high degree of regularity (for example, secondary structure and tertiary symmetries) that is absent from random compact conformations. With the use of a simple lattice model of protein folding, it was demonstrated that structural regularities are related to high "designability" and evolutionary stability.





Li et al. *Science* (1996)

Designable structures are atypical.



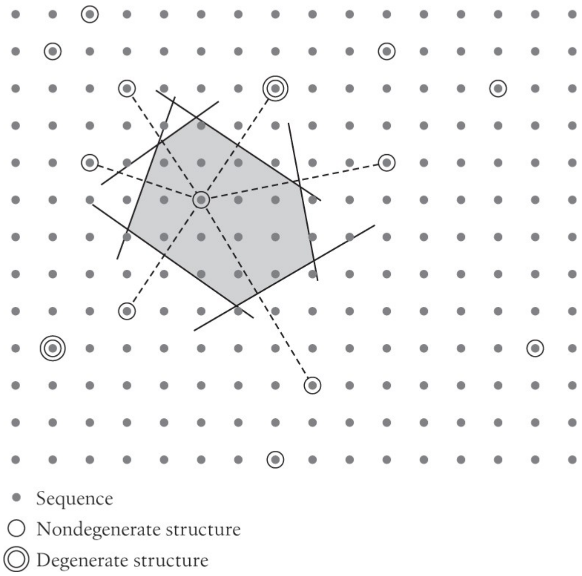
$S_a = 000110000001111111100000110000011110$

$S_b = 001100110000110000110011000011111100$

$$E = \sum_i (s_i - \sigma_i)^2$$

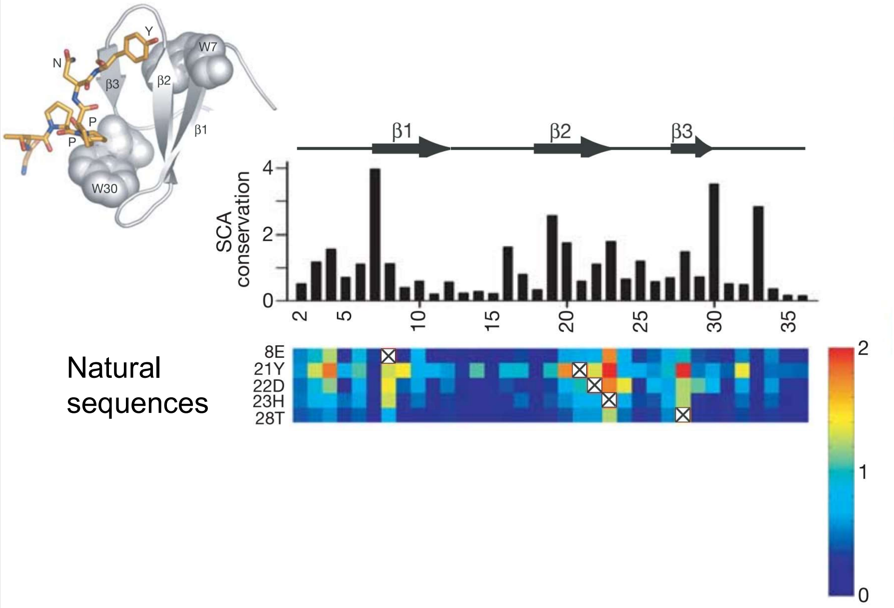
Hydrophobic/ Polar

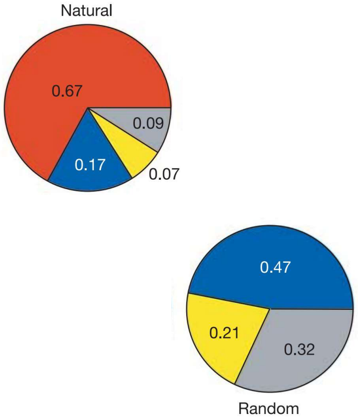
Core/ Surface



Li et al. *PNAS* (1998)

Pairwise correlations are required for natural-like functions.

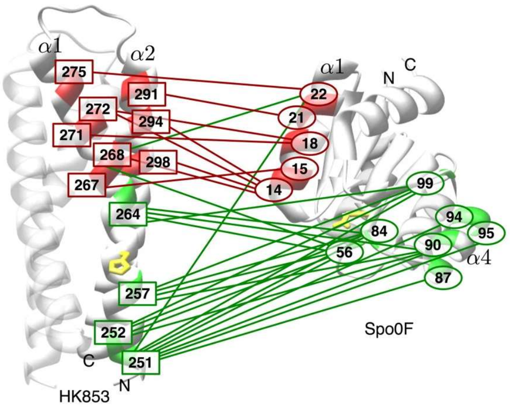




Folded
Unfolded
Insoluble
Unexpressed

Socolich et al. *Nature* (2005)

Pairwise correlations predict residue contacts.



$$E = -\sum_{ij} J_{ij} s_i s_j$$

Infer

Detour: Computing marginals via message passing

Calculating marginals require $O(20^N)$:

$$P(A_i) = \sum_{\{A_j\}\backslash A_i} P(A_i, \{A_j\}\backslash A_i)$$

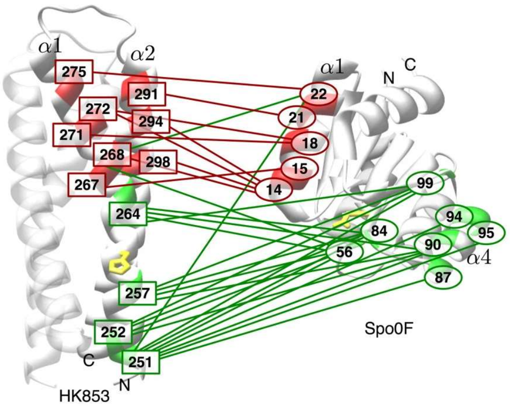
What if the joint probability factors?

$$P(A_1, \ldots, A_N) = \prod_{\{j\} \in \text{Ne}(i)} F_i(A_i, \{A_j\})$$

Then the sum-product rule can speed up calculations to $O(N)$.

Mezard and Mora. *arXiv* (2008)
Weigt et al. *PNAS* (2009)

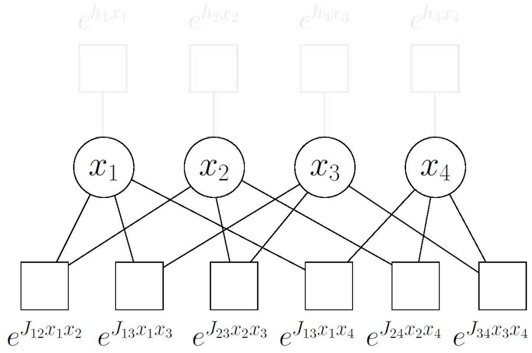
Pairwise correlations predict residue contacts.



$$E = -\sum_{ij} J_{ij} s_i s_j$$

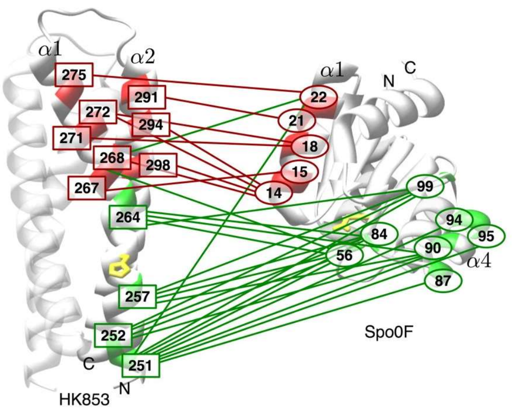
Infer

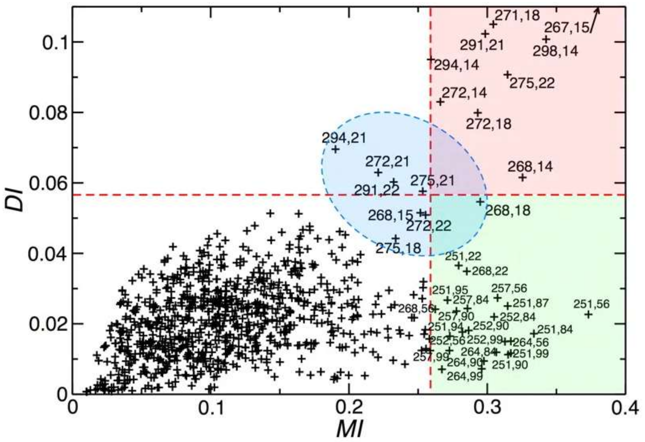
Detour: Computing marginals via message passing



Mezard and Mora. *arXiv* (2008)
Weigt et al. *PNAS* (2009)

Pairwise correlations predict residue contacts.

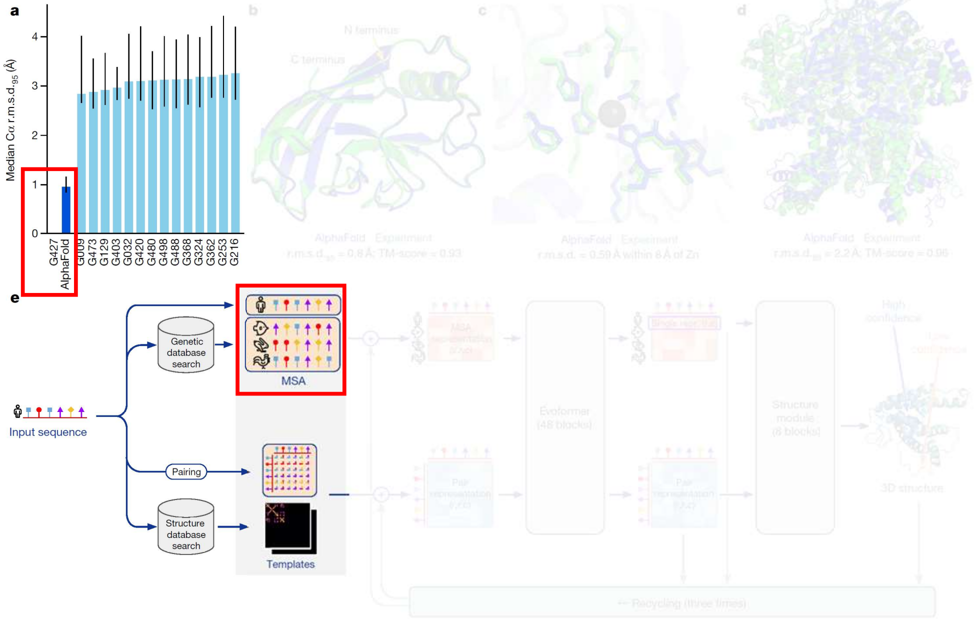




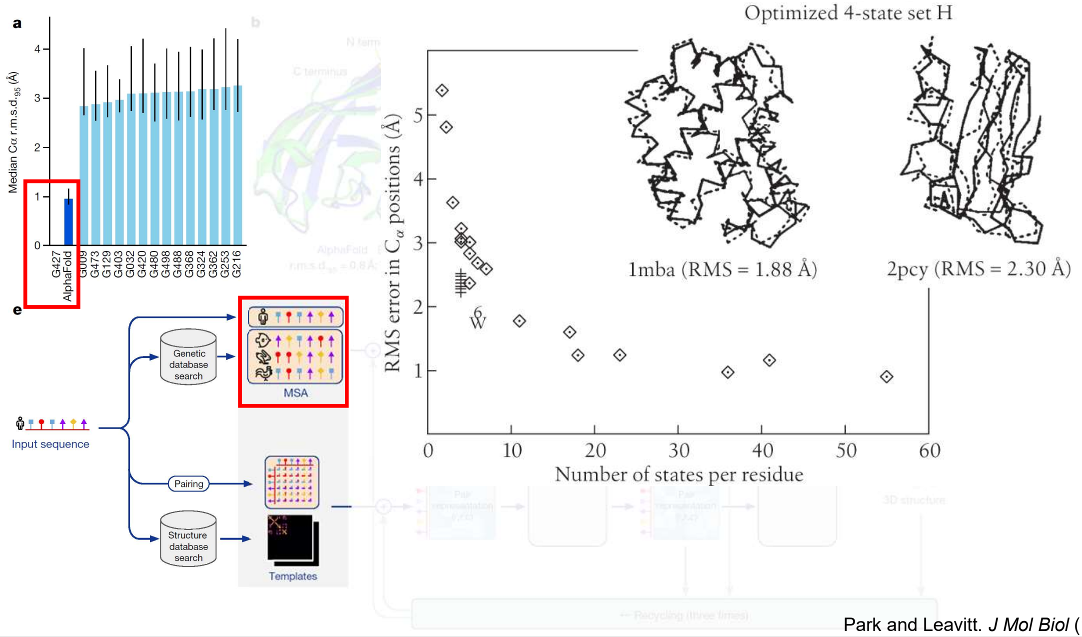
$$E = -\sum_{ij} \underset{\overset{\uparrow}{\text{Infer}}}{J_{ij}}\, s_i s_j$$

Weigt et al. *PNAS* (2009)

How much information is required to specify structure?

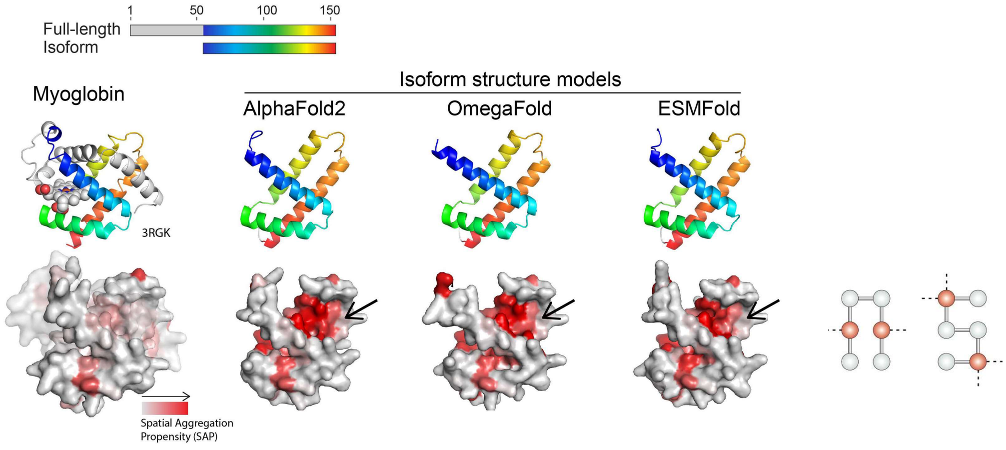


How much information is required to specify structure?



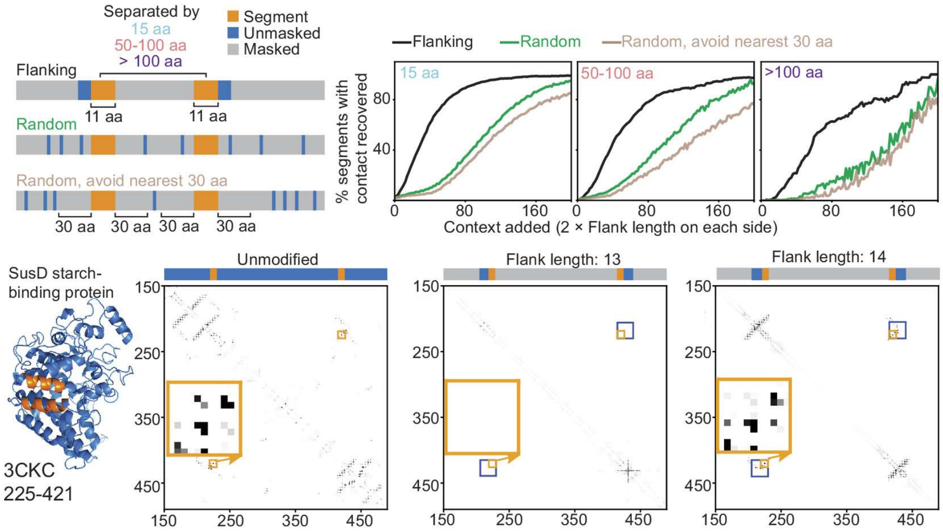
Park and Leavitt. *J Mol Biol* (1995)

What information is contained in large models?



Zhang et al. *PNAS* (2024)

What information is contained in large models?



Zhang et al. *PNAS* (2024)

Native protein sequences are close to optimal for their structures

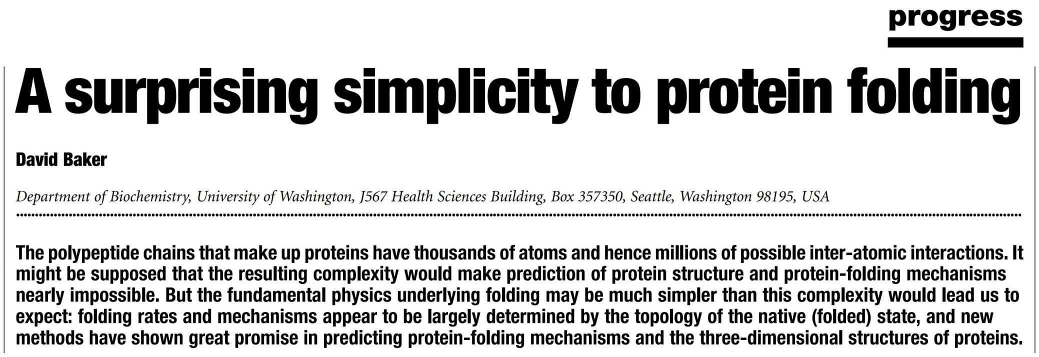
Brian Kuhlman and David Baker*

Department of Biochemistry and Howard Hughes Medical Institute, University of Washington School of Medicine, Seattle, WA 98195

Edited by William F. DeGrado, University of Pennsylvania School of Medicine, Philadelphia, PA, and approved July 11, 2000 (received for review March 20, 2000)

How large is the volume of sequence space that is compatible with a given protein structure? Starting from random sequences, low

many common features with the energy functions used for protein design. Therefore, it is possible that refinement builds a



progress

A surprising simplicity to protein folding

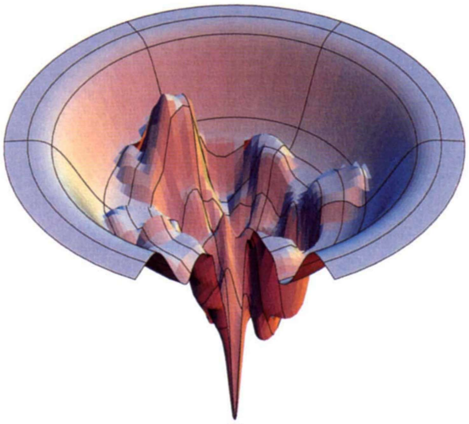
David Baker

Department of Biochemistry, University of Washington, J567 Health Sciences Building, Box 357350, Seattle, Washington 98195, USA

The polypeptide chains that make up proteins have thousands of atoms and hence millions of possible inter-atomic interactions. It might be supposed that the resulting complexity would make prediction of protein structure and protein-folding mechanisms nearly impossible. But the fundamental physics underlying folding may be much simpler than this complexity would lead us to expect: folding rates and mechanisms appear to be largely determined by the topology of the native (folded) state, and new methods have shown great promise in predicting protein-folding mechanisms and the three-dimensional structures of proteins.

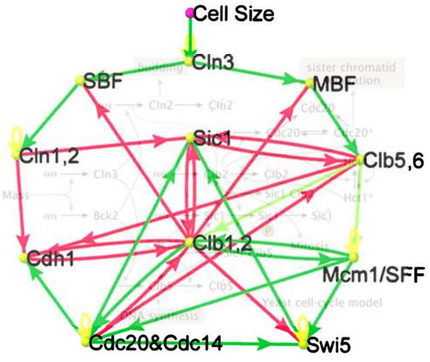
Summary

- Proteins are designable (in theory and in practice).
- Spin glasses are everywhere and frustration is a key feature of complex systems.

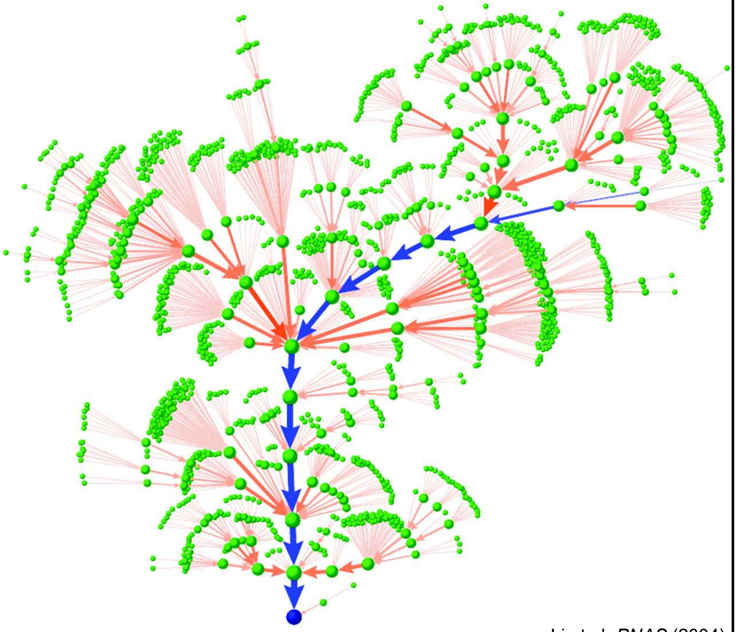


Dill and Chan. *Nat Struct Mol Biol* (1997)

Spin glass for cell cycle

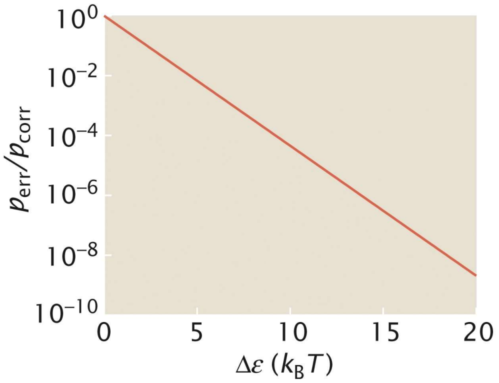


$$s_i(t+1) = \theta\left[\sum_j J_{ij} s_j(t)\right]$$

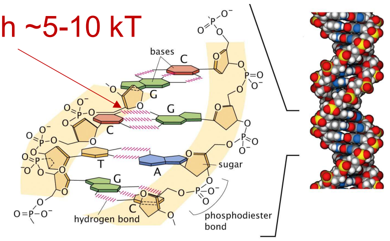


Li et al. *PNAS* (2004)
Phillips et al

Detour: How specific?



Mismatch ~5-10 kT



No error correction

A = “correct” substrate
B = “incorrect” substrate

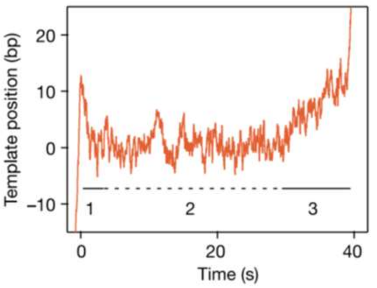
$$\begin{tikzcd}
 & EA \arrow[dl, shift left, "k_-"] \arrow[r, "V_{\max}"] & E + \text{correct product} \\
E \arrow[ur, shift left, "k_+[A]"] \arrow[dr, shift right, "k'_+[B]"'] & & \\
 & EB \arrow[ul, shift right, "k'_-"'] \arrow[r, "V'_{\max}"] & E + \text{incorrect product}
\end{tikzcd}$$

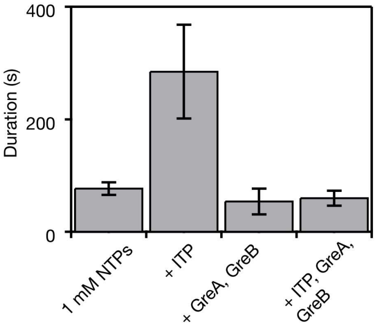
Kinetic proofreading

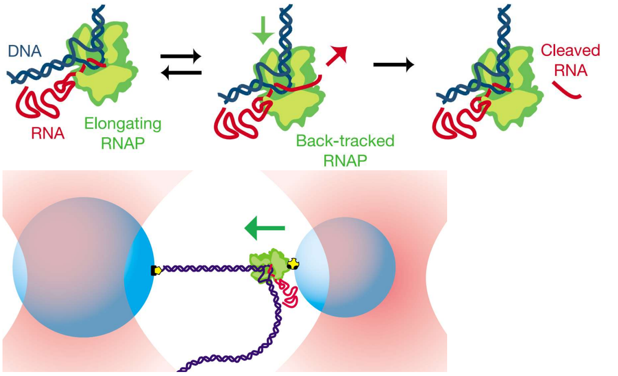
$$\begin{tikzcd}
E \arrow[r, shift left, "k_+[A]"] & EA \arrow[l, shift left, "k_-"] \arrow[r, "r"] & E^*A \arrow[r, "V_{\max}"] \arrow[d, shift left, "k_-"] & E + \text{product} \\
 & & E \arrow[u, shift left, "k_+[A]"] &
\end{tikzcd}$$

Phillips et al
Bialek

Detour: How specific?







Kinetic proofreading

Replication of DNA, or transcription of RNA, a = DNA template

$$\begin{tikzcd}
 & & \mathrm{PP_i} & & \\
\mathrm{dATP} + \mathrm{a} \arrow[r, shift left] & \mathrm{a}\cdot\mathrm{dATP} \arrow[l, shift left] \arrow[r, shift left] & & \mathrm{a}\cdot\mathrm{dAMP} \arrow[ll, shift left] \arrow[r, shift left] \arrow[d, shift left] & \text{Product (incorporation)} \arrow[l, shift left] \\
 & & & \mathrm{a} + \mathrm{dAMP} \arrow[u, shift left] &
\end{tikzcd}$$

Shaevitz et al. *Nature* (2003)
Bialek