

Dec 2021

PROTEIN:

AMORPHOUS EVOLVING MATTER

Jean-Pierre Eckmann & Jacques Rougemont (Geneva)

Albert Libchaber (Rockefeller)

Michael Mitchell & Stanislas Leibler (Rockefeller & IAS)

Sandipan Dutta (IBS)



Please ask questions during the talk

Proteins are complex 3D arrangements of hundreds to thousands amino acids

PROTEIN

- Non-random **information-rich matter**.
- Long-term **evolutionary memory**.
- **Small, heterogeneous**.
- **Sparsity of data** in high-D gene space.
- **Collective many-body interactions**

*Q: how are the collective
interactions that determine protein
function are encoded in the gene?*

Kendrew (1961) – 3D
structure of myoglobin (art
by I. Geis)

Genes and protein structures have **many** parameters
but protein function is often described by only **a few**

Genotype

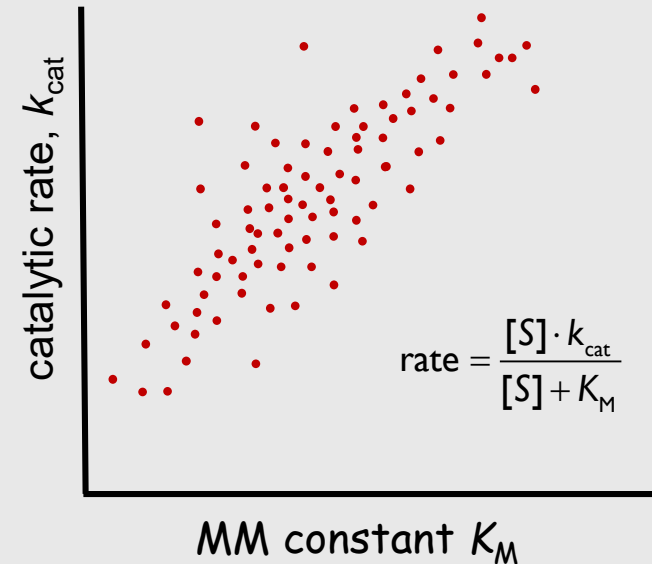
sequence space: *dimension* ~ 1000

...AGGCTGGAGAATCTGGCACGG...
...CTGCTGGAGAACTGACGCAG...
...AGGCTGGAGAATCTAGCACGG...
...AGGCTGGTTAATCTGGCTCGG...
...CGGGTGGAGATTCTGGTTCGT...
...AGGCTGGAGAACTCGCAAGG...
...ATGCTGGAGAACTGACGCAT...
...AGGCTGGAGACTCTGGCACGG...
...AGGCTGGTTAAACTGGGTTCGC...
...CGGGTGGAGATTCTGGTTCGG...
...AGGCTGCCGAATCTAGAACGG...

**dimensional
reduction**

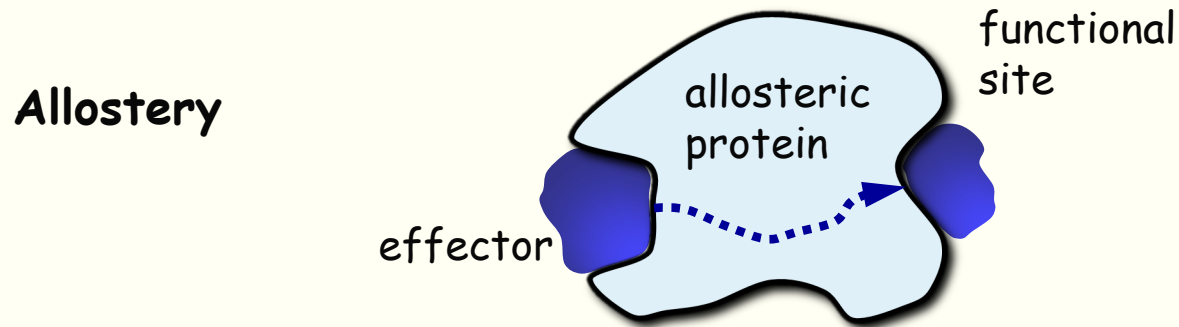
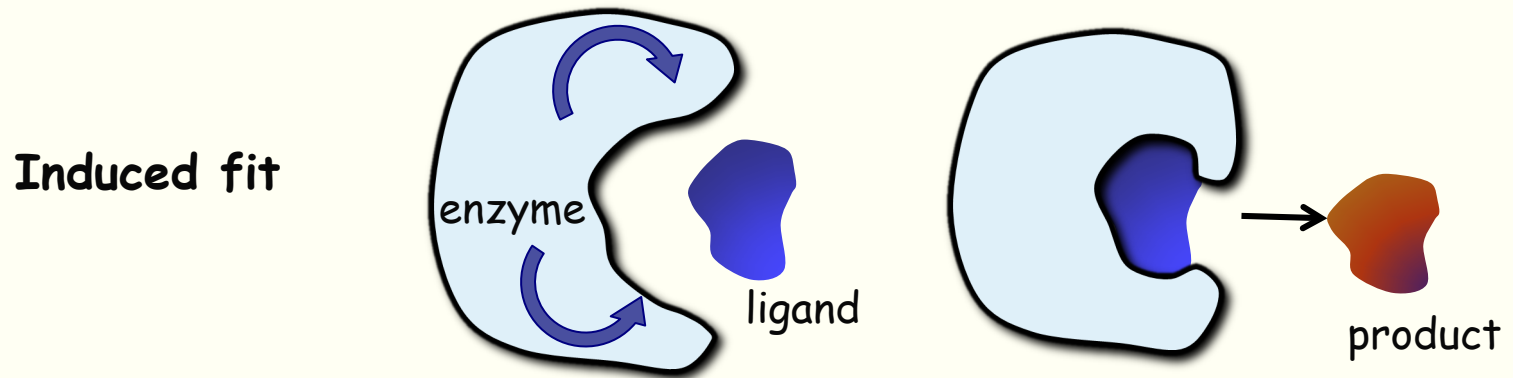
Phenotype

kinetic parameter space: *dimension*= 2



Information ~ gene length $\xrightarrow{\text{compression}}$ Information ~ a few bits

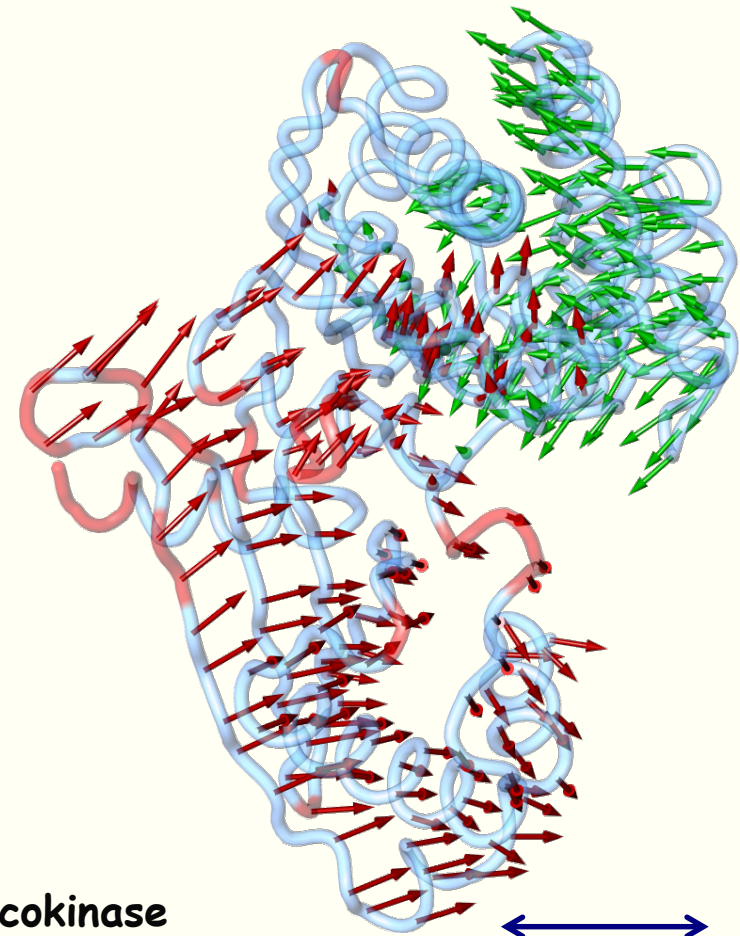
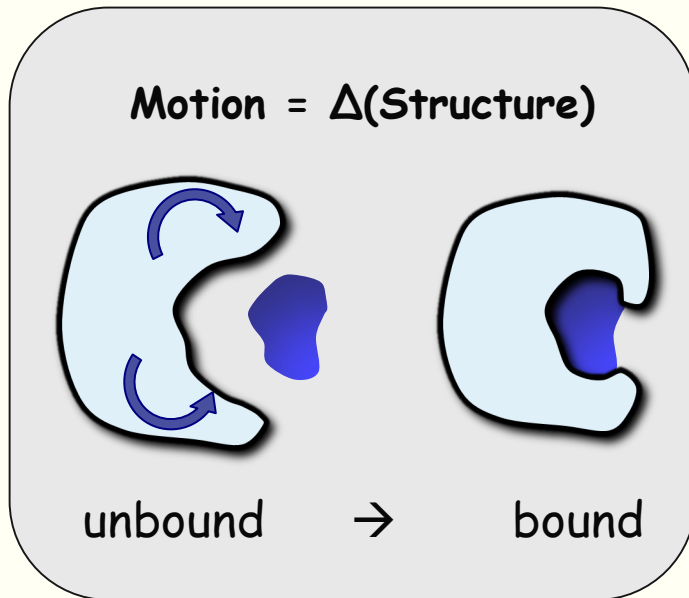
The **dimension** question is linked to dynamics and mechanics of protein:
Many proteins use **large-scale collective motions** to function



- **Slow global motions** in enzymes, ion-channels, ribosome...

Motivation:

Looking at dynamics suggests a mechanical basis for mapping genetic information to collective function



Glucokinase

(PDB: 1v4s → 1v4t)

1 nm

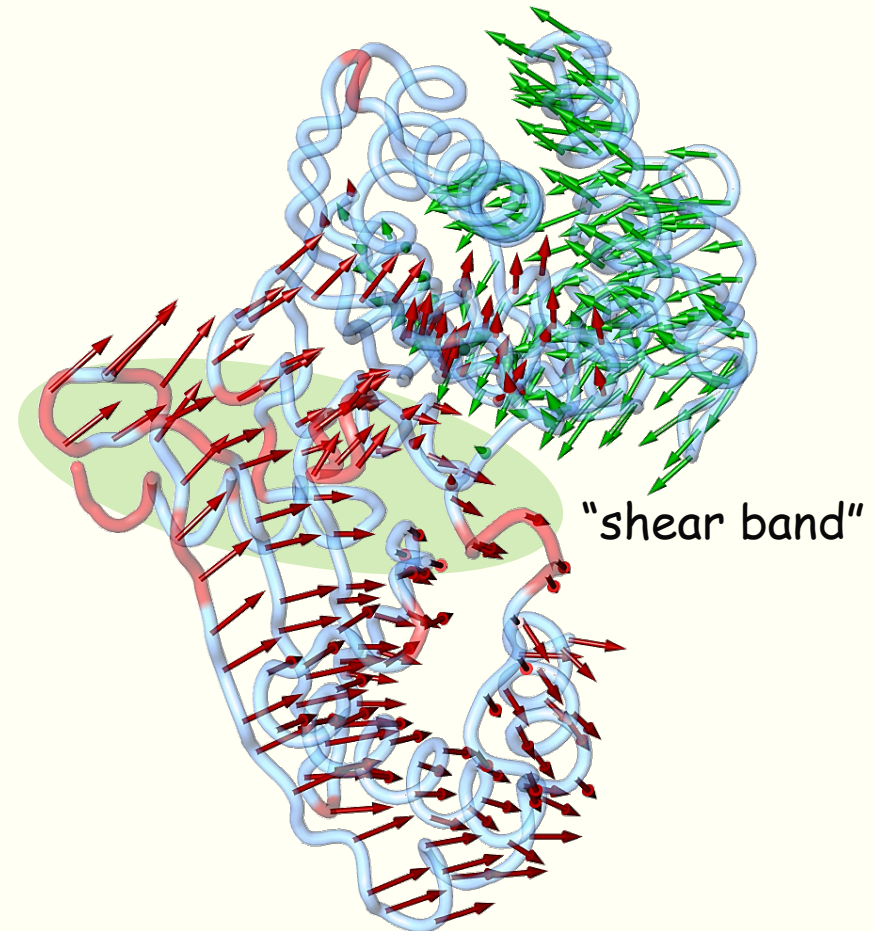
Motivation:

The shear is localized in “hinges” or “bands”
separating rigid protein domains

Shear = Δ motion of neighboring a.a.
= $\Delta\Delta$ Structure

$$\text{shear} = \frac{1}{2} \left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i} + \sum_k \frac{\partial u_k}{\partial x_i} \frac{\partial u_k}{\partial x_j} \right)$$

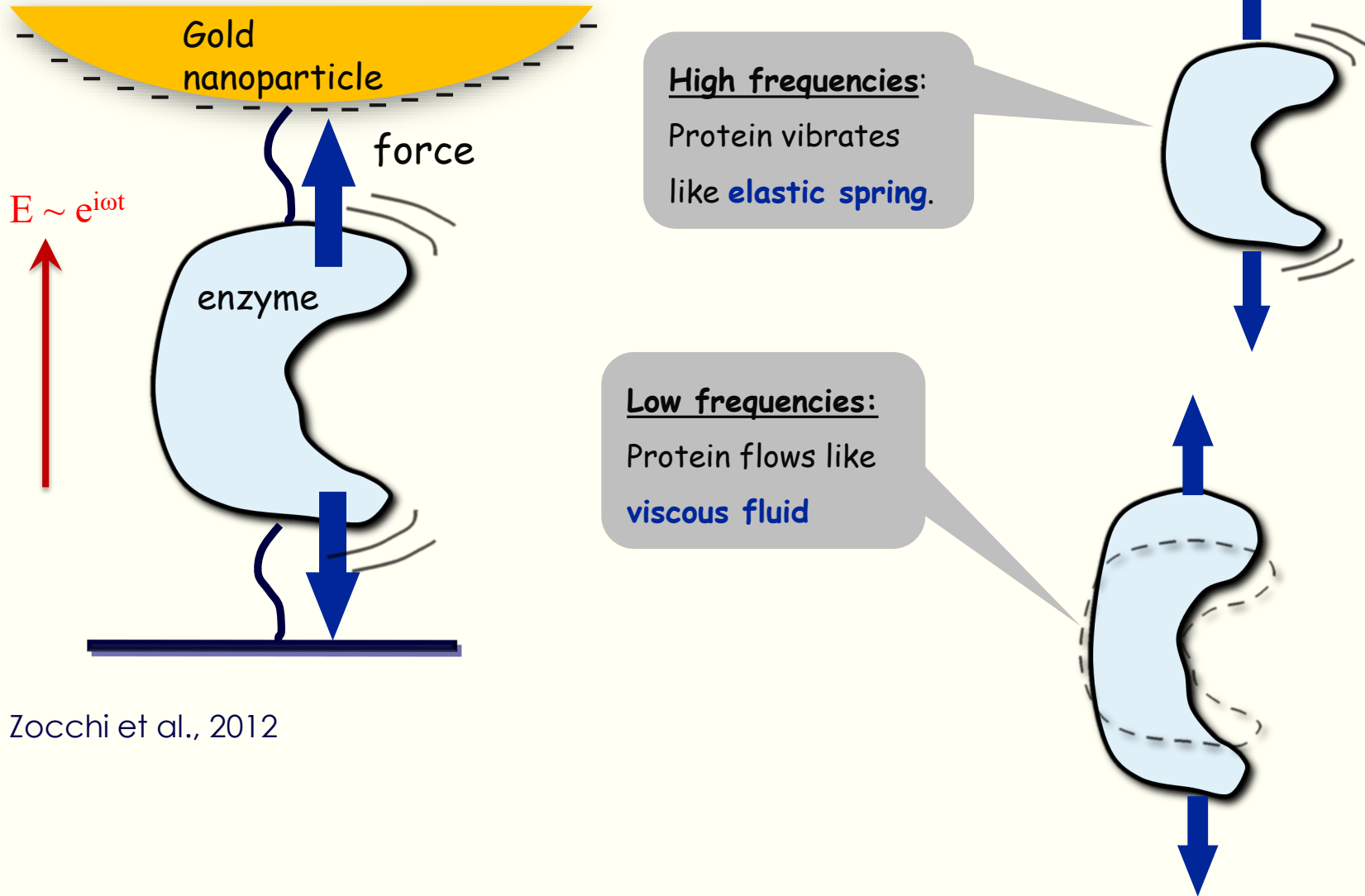
- The shear highlights the internal motion of amino acids in the protein



Glucokinase (PDB: 1v4s $\rightarrow$ 1v4t)

Motivation:

Proteins are soft and viscous-like at low frequencies

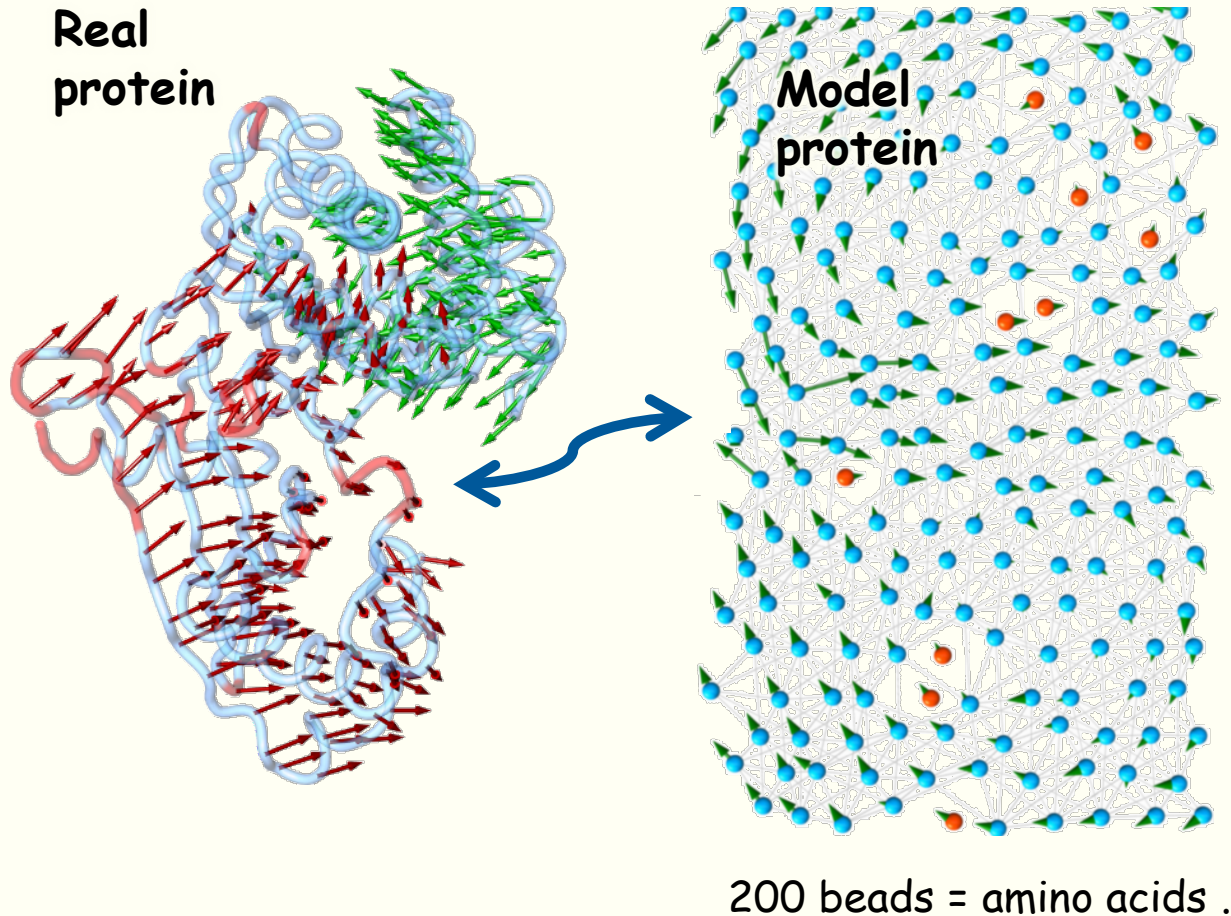


Zocchi et al., 2012



We also use this technique, see later

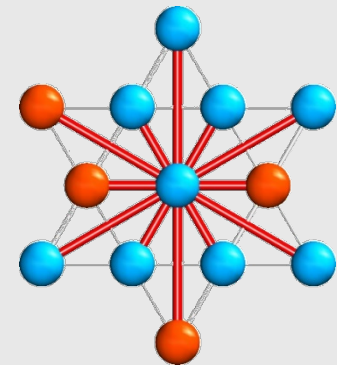
Minimal coarse-grained model of protein as adaptive matter gives evolution a mechanical interpretation



Two types of amino acids
connected by "springs"

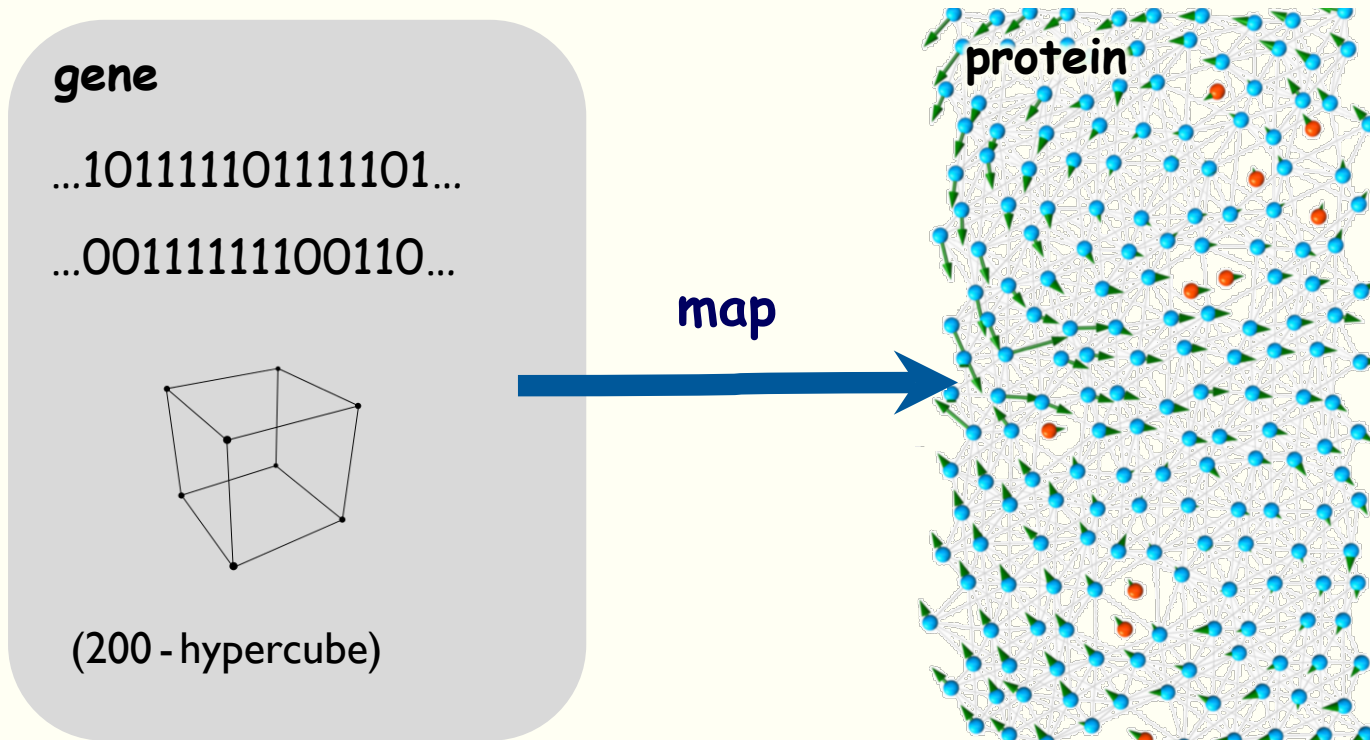
amino acids	H	P
H	strong	weak
P	weak	weak

12 neighbors



Main goal:

Understanding the emergence and evolution of mechanical function in the genotype-to-phenotype map



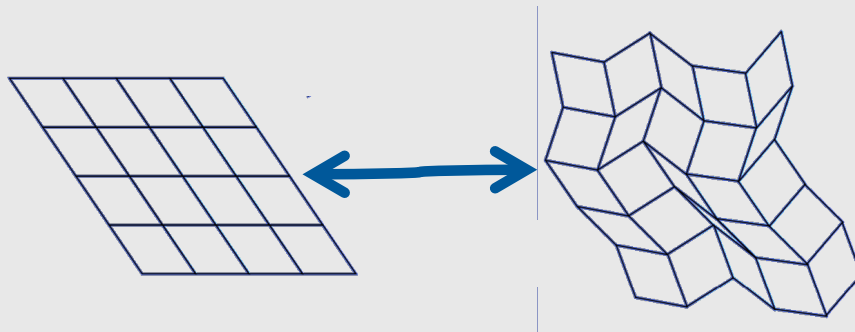
Mutations are random walks among the $2^{200} = 10^{60}$ genes:

- *How many of them are functional?*
- *How they are arranged in the high-dimensional genotype space ?*

Our coarse-grained approach models protein as **amorphous evolving matter**

- The emergence of “**floppy modes**”

[Shlomo Alexander 1998: Physics of Amorphous solids]



- **Constraint theory** and **rigidity transitions**

[Maxwell 1864: On reciprocal figures and diagrams of forces]

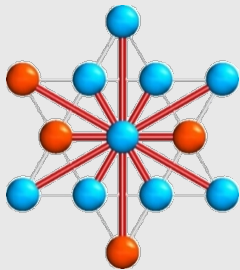


*We will focus on the mechanical basis of the **genotype-to-phenotype map**.*

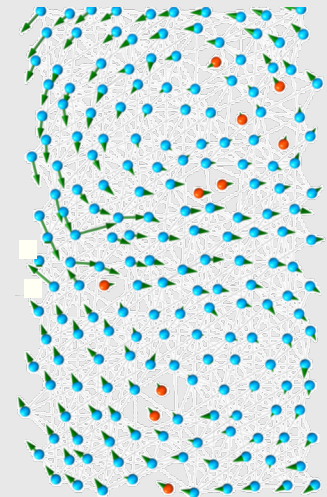
The mechanical response of the protein is determined by the elastic network

The elasticity tensor $\mathbf{H}$ of the protein is a function of its gene:

$$\mathbf{H}(\text{gene}) = \text{Connectivity} \times \text{spring strength} \times \text{geometry}$$



amino acids	H	P
H	strong	weak
P	weak	weak

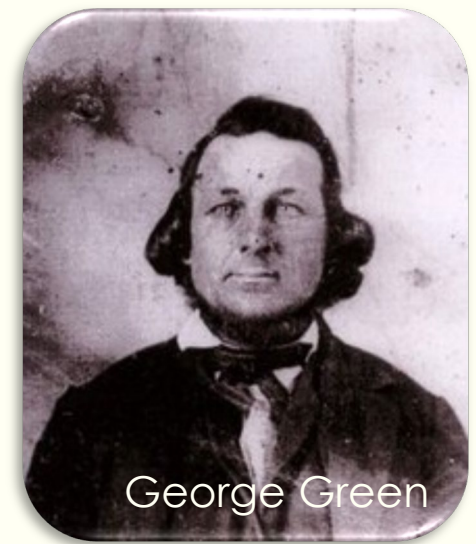


$$\mathbf{H}_{ij}(g_i, g_j) = \mathbf{A}_{ij} \cdot k_{ij}(g_i, g_j) \cdot \mathbf{n}_{ij} \times \mathbf{n}_{ij}$$



Please ask questions

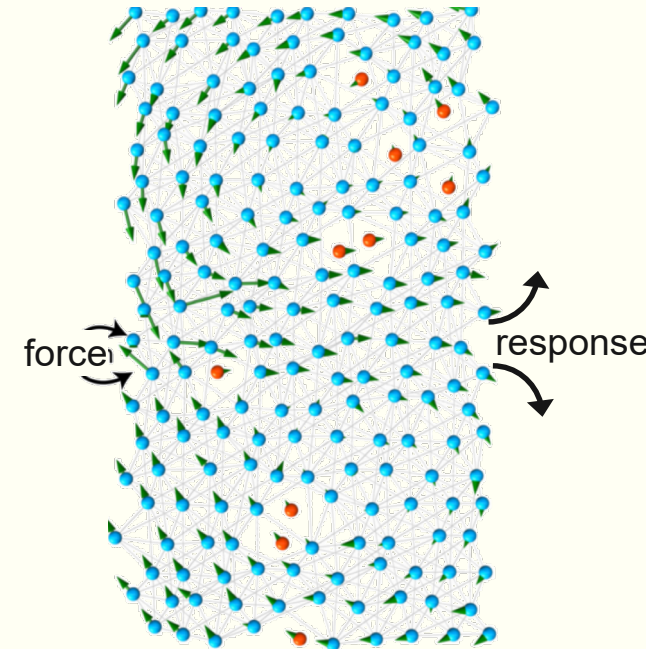
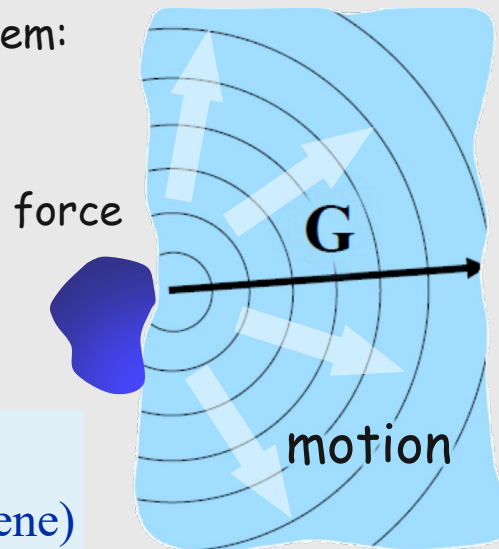
In the protein: Green's function measures the motion induced by a local force, such as ligand binding



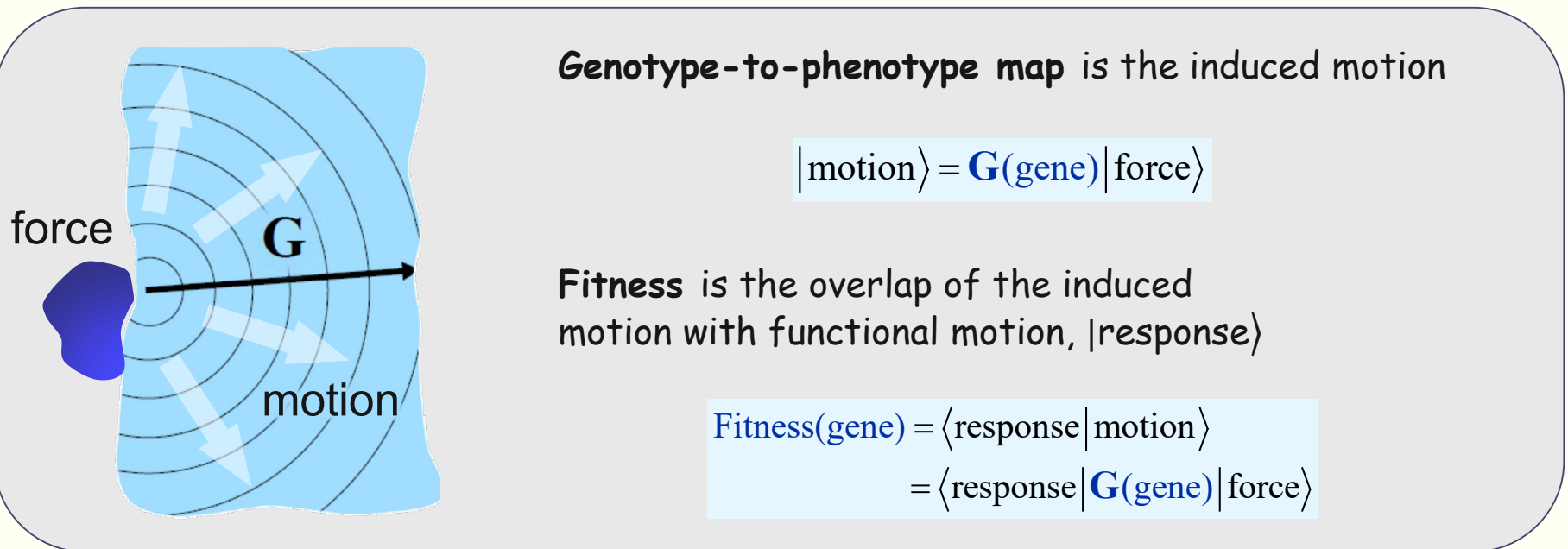
Green's function G solves
the inverse elastic problem:

$$|\text{motion}\rangle = G |\text{force}\rangle$$

with $G(\text{gene}) = H^{-1}(\text{gene})$

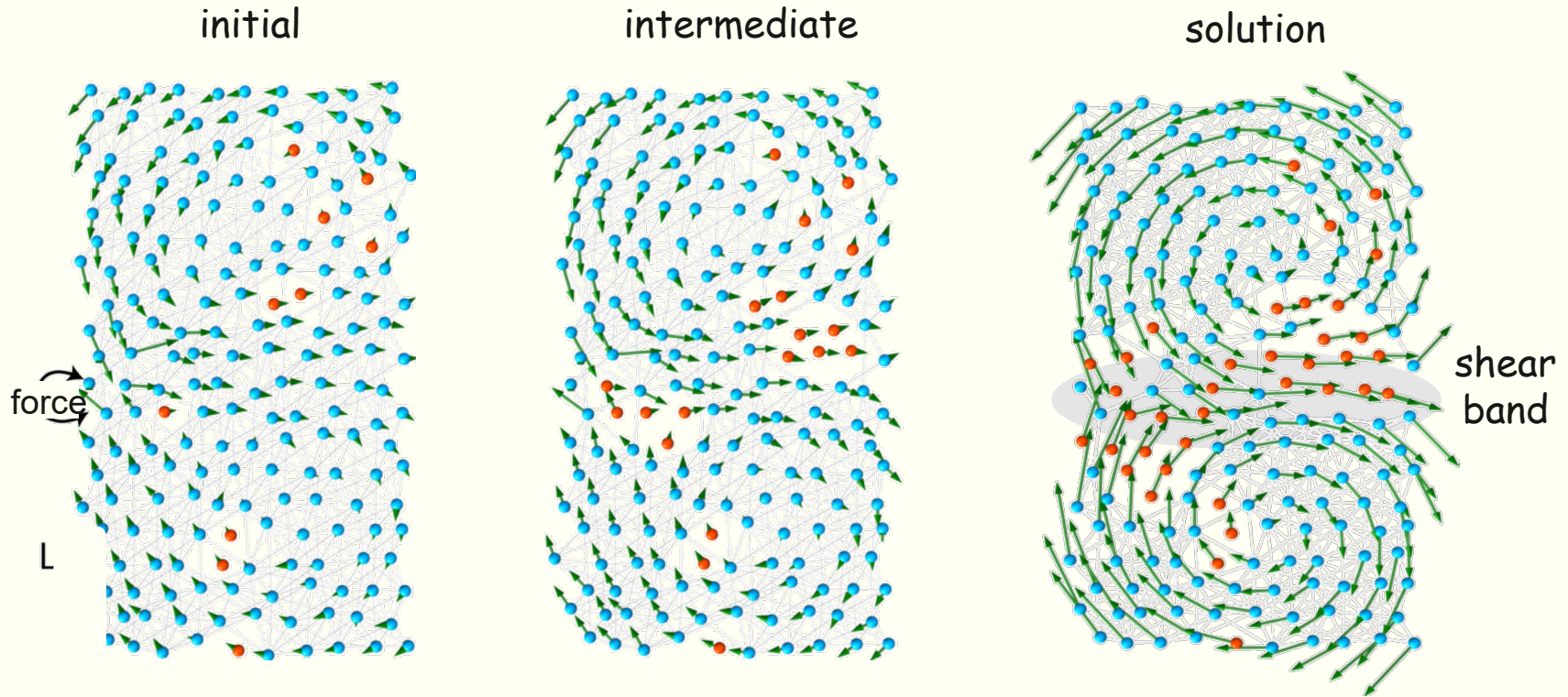


Green's function gives explicit gene-to-function map and fitness landscape



G: *gene* → *amino acid network* → *protein dynamics* → *function* → *fitness*

Evolution searches in the mechanical fitness landscape for a solution with shear band and soft mode



$$\text{Fitness} = \text{overlap of } \langle \text{response} || \text{motion} \rangle$$



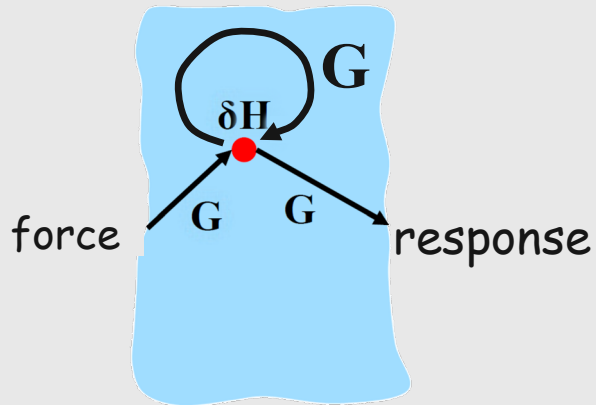
One can look at more general forces and responses: for example, specific recognition

Mutations locally perturb the amino acid network, scatter the force, and thereby change the fitness

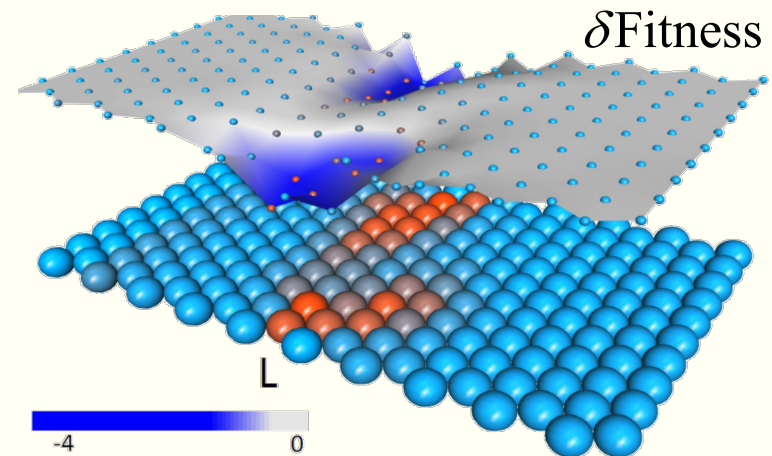
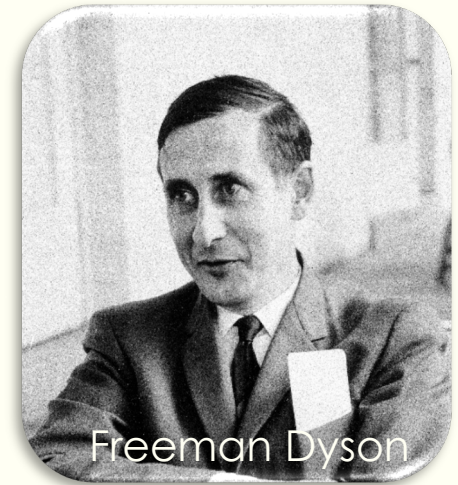
Effect of mutations on elastic network

$$\delta \mathbf{H} = \mathbf{H}(\text{gene}') - \mathbf{H}(\text{gene})$$

Effect on the mechanical response = sum over all **scattering paths** (Dyson's series)

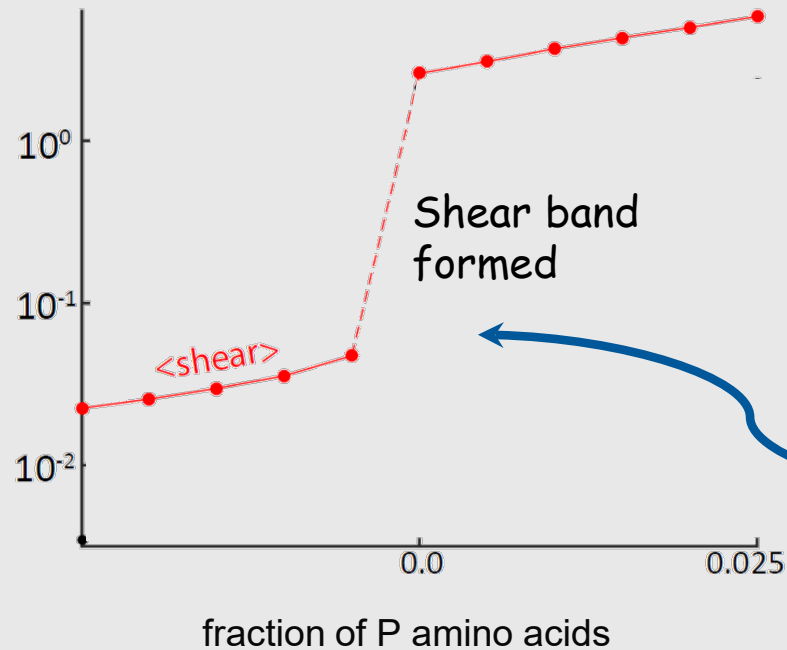


$$\delta \mathbf{G} = -\mathbf{G} \delta \mathbf{H} \mathbf{G} + \mathbf{G} \delta \mathbf{H} \mathbf{G} \delta \mathbf{H} \mathbf{G} - \dots$$



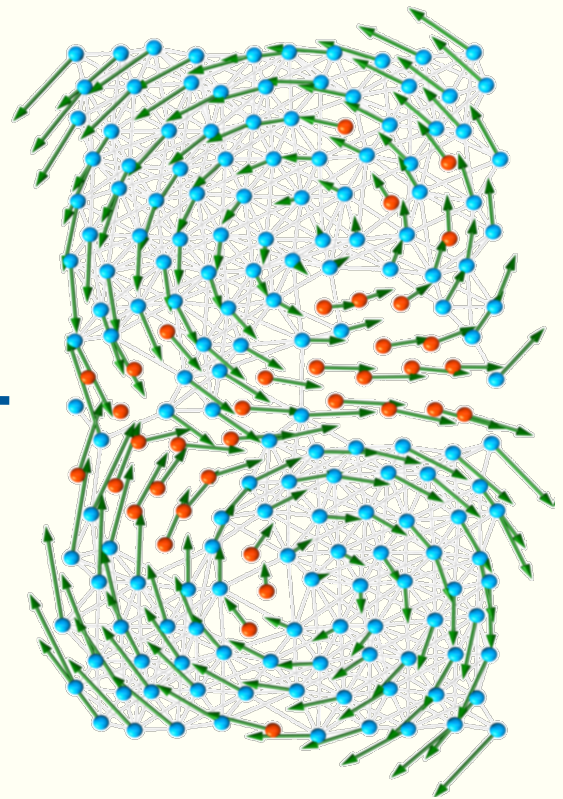
scattering is strongest
in the shear band

Protein function emerges at a topological transition when
a shear band divides the protein and a soft mode appears



soft mode energy $\rightarrow 0$

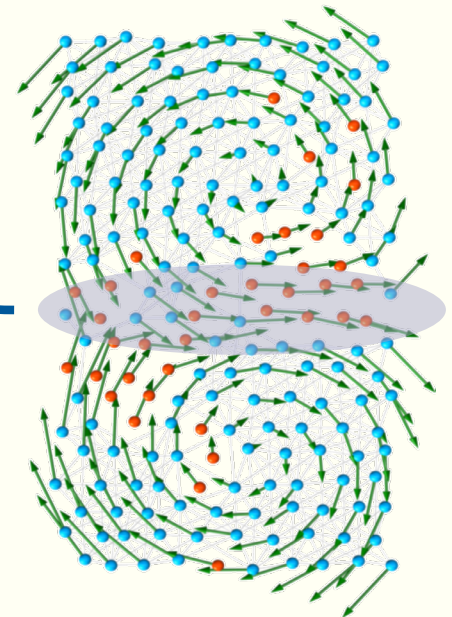
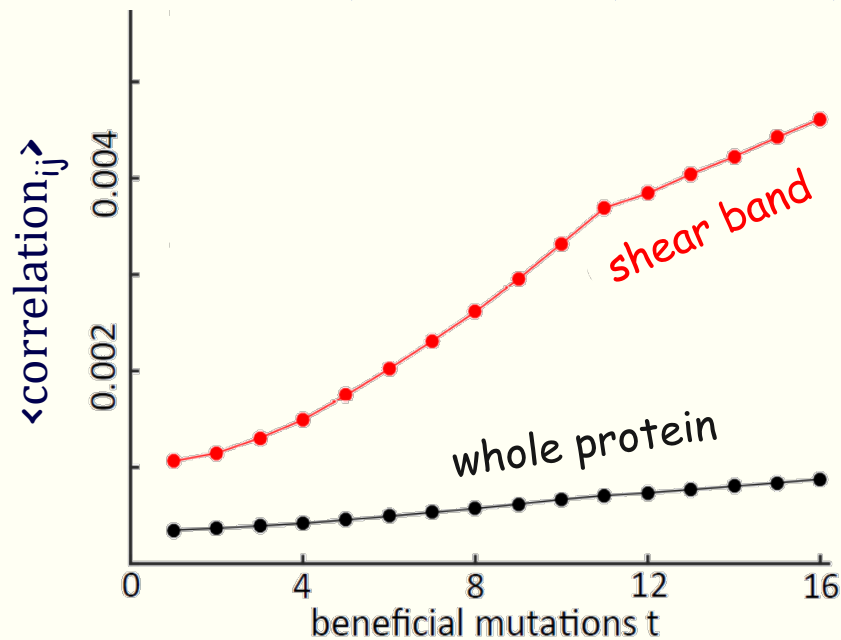
$$\text{Fitness} = \langle \text{response} | G | \text{force} \rangle \sim \frac{1}{\text{soft mode energy}} \rightarrow \infty$$



The genetic correlation among solutions increases towards the emergence of mechanical function

10^6 solutions $\left\{ \begin{array}{l} 11100101011111000010101... \\ 10010001000101011101110... \\ 00111110111110100111010... \\ 10011111110011001111000... \\ \dots \end{array} \right.$

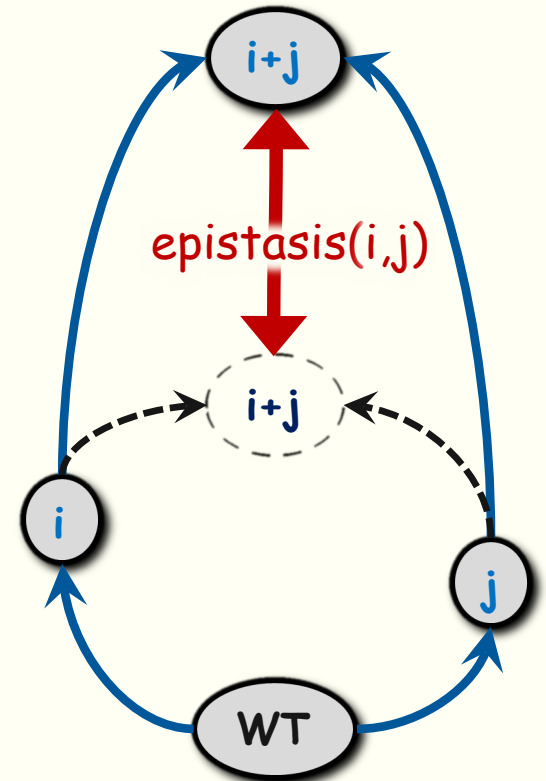
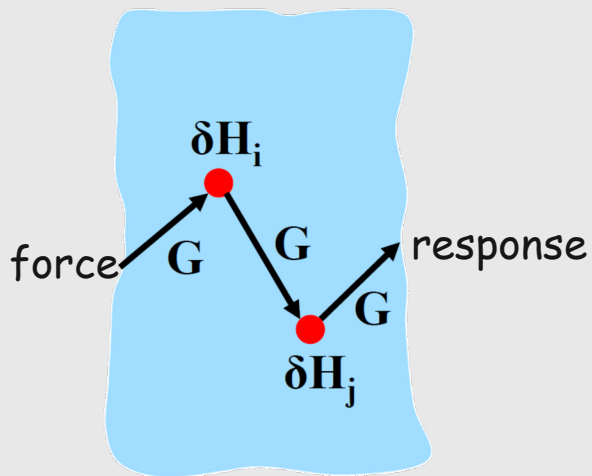
$$\text{correlation}_{ij} = \langle \text{codon}_i \text{codon}_j \rangle - \langle \text{codon}_i \rangle \langle \text{codon}_j \rangle$$



Epistasis is collective interaction of mutations
resulting in nonlinearity of the fitness landscape

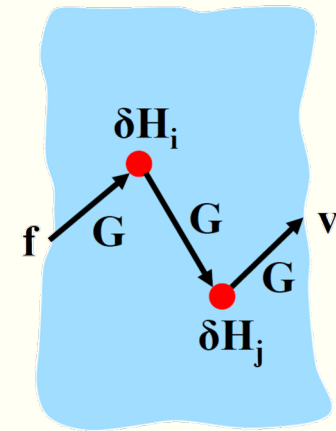
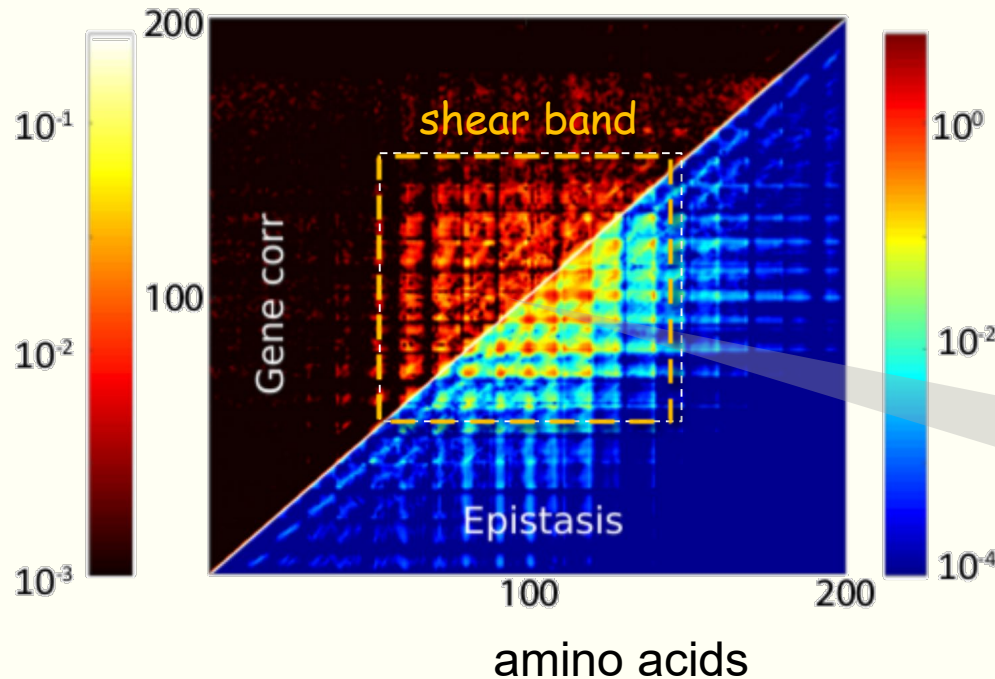
$$\text{epistasis}(i,j) = \Delta\text{fitness}(i+j) - [\Delta\text{fitness}(i) + \Delta\text{fitness}(j)]$$

Mechanical interpretation: sum (Dyson series)
over all scattering paths passing through i and j



Mechanical interactions (= epistasis) are origin of genetic correlations, and both exhibit similar patterns

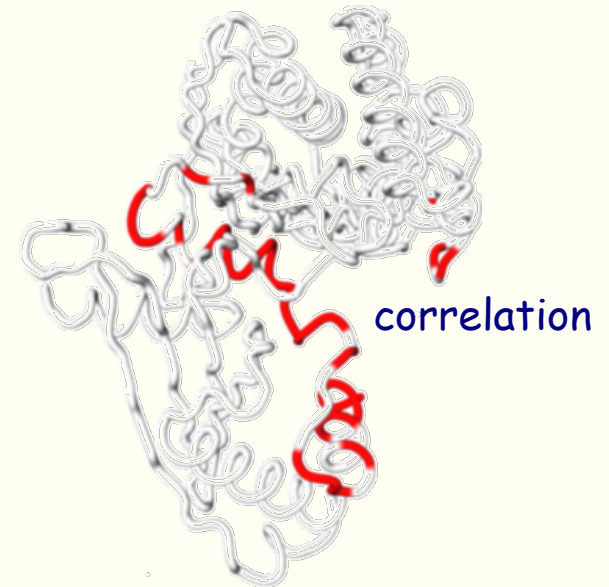
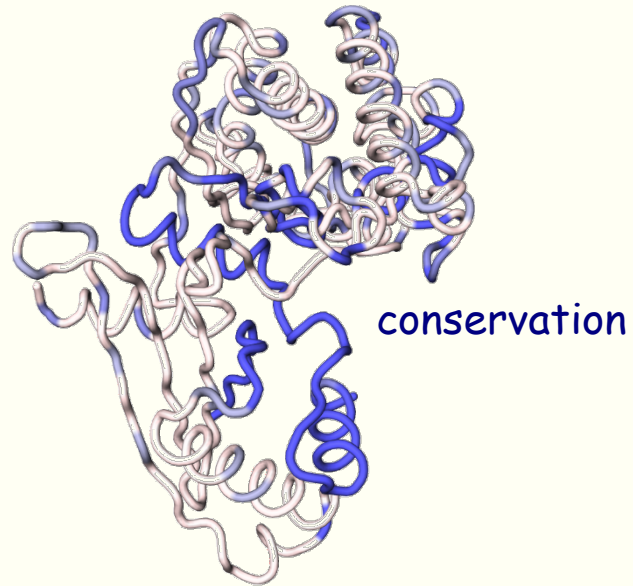
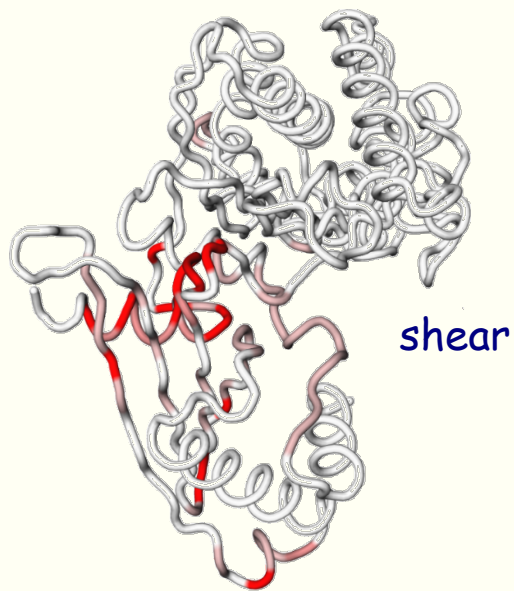
$$\text{correlation}_{ij} = \langle \text{codon}_i \text{codon}_j \rangle - \langle \text{codon}_i \rangle \langle \text{codon}_j \rangle$$



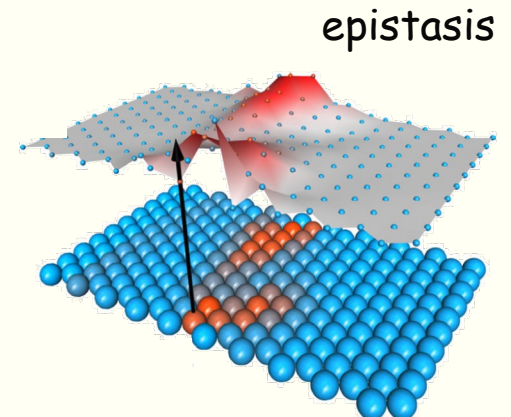
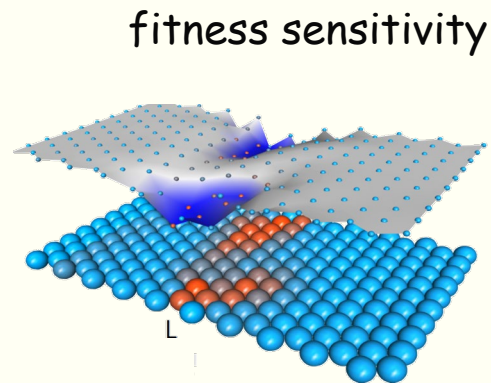
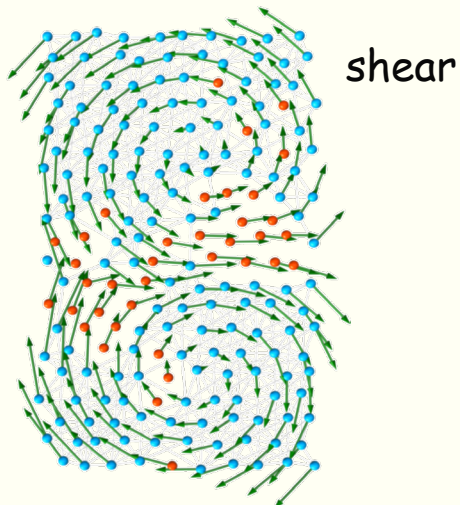
Epistasis and sequence correlation are strongest in high shear region separating the rigid subdomains

Predictions of the model can be compared qualitatively to real proteins

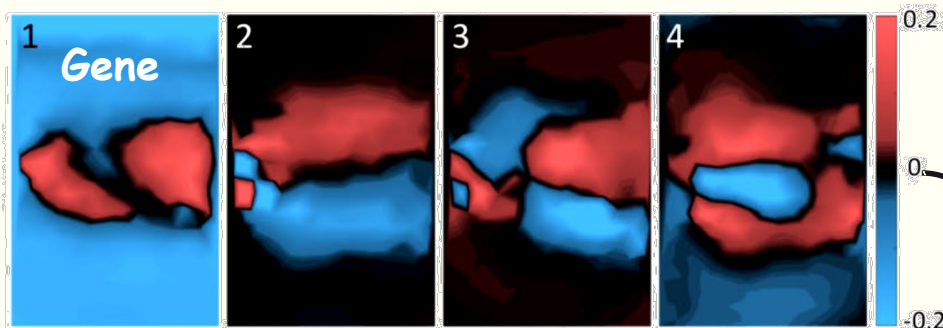
Glucokinase



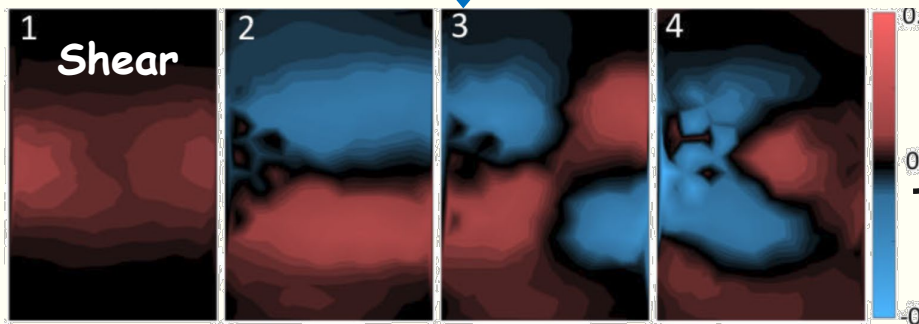
Model



Correlation spectra of phenotypes and genotypes capture the relevant modes of mechanical function and evolution



*correspondence
between modes of
gene and shear*

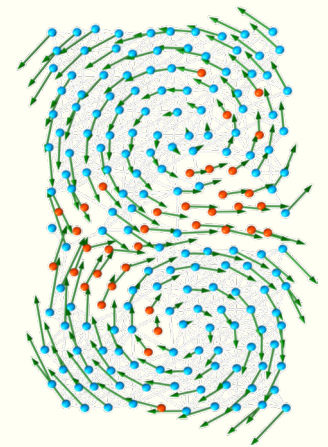


gene

...101111101111101...

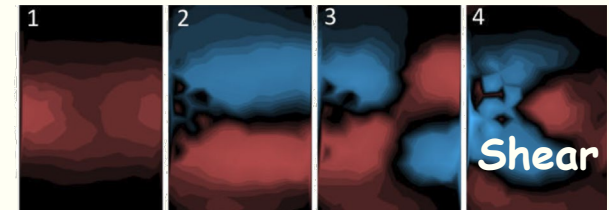
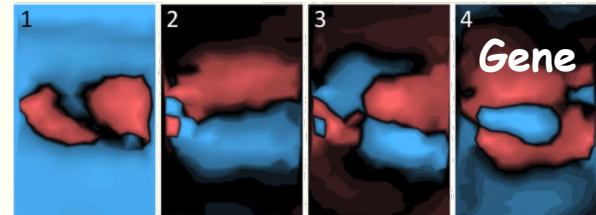
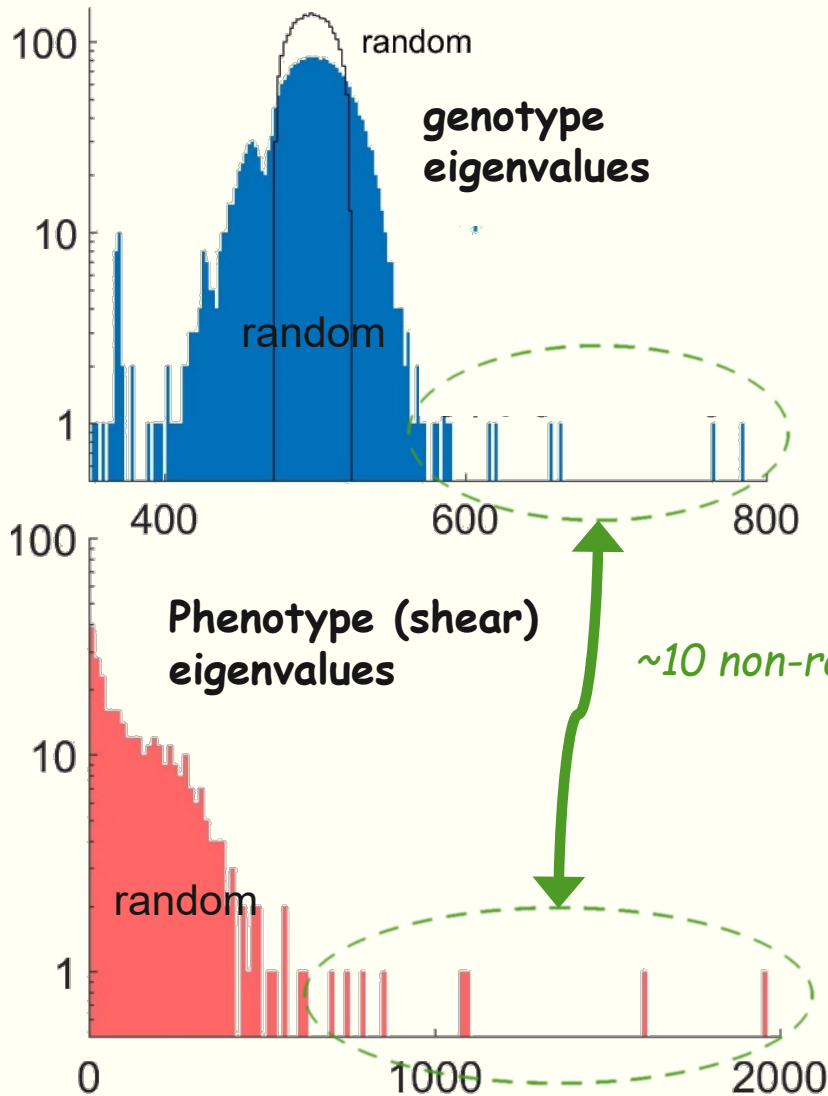
...001111111001100...

PCA of
 10^6 solutions



Shear = Δ motion of neighboring a.a.

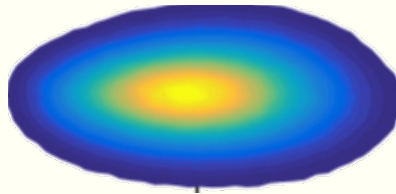
A few non-random modes capture the functionally-relevant bits of information in the gene



 correlation modes are not the mechanical normal modes

The genotype-to-phenotype map involves huge dimensional reduction whose origin is the mechanical interaction

In genotype space
200-dimensional sphere



In phenotype space
~10-dimensional "disc"



Genotype

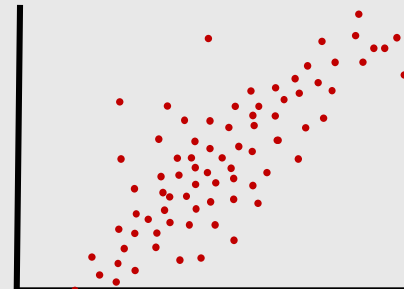
sequence space: dimension ~ 1000

...AGGCTGGAGAATCTGGCACGG...
...CTGCTGGAGAACTGACGCAG...
...AGGCTGGAGAATCTAGCACGG...
...AGGCTGGTTAATCTGGCTCGG...
...CGGGTGGAGATTCTGGTTCGT...
...AGGCTGGAGAACTCGCAAGG...
...ATGCTGGAGAACTGACGCAT...
... ..

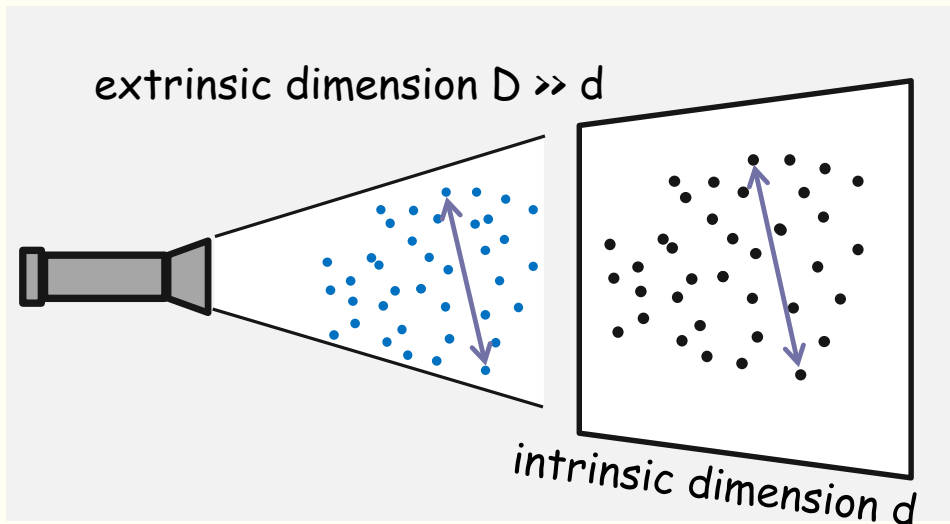
dimensional
reduction

Phenotype

kinetic parameters dimension= 2



Is there something special about dimensional reduction in biology?



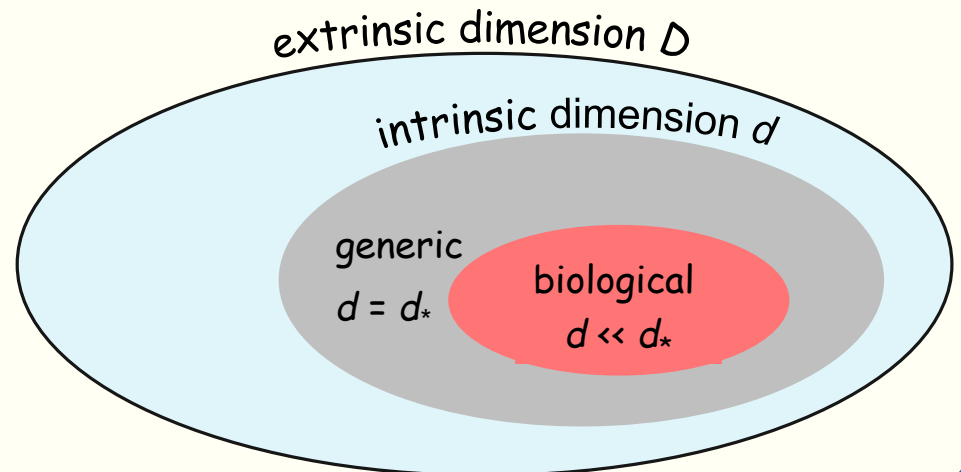
Intrinsic dimension: minimal dimension needed to project data with little deformation.

Johnson-Lindenstrauss compression:

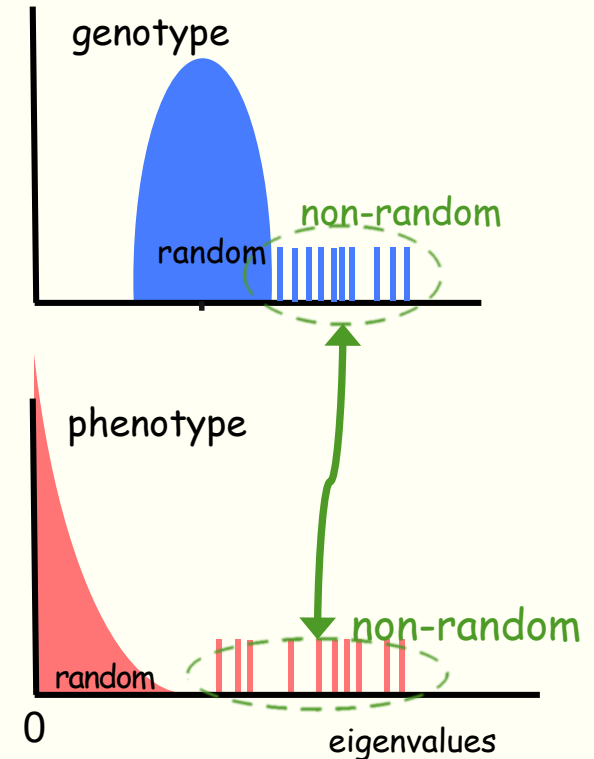
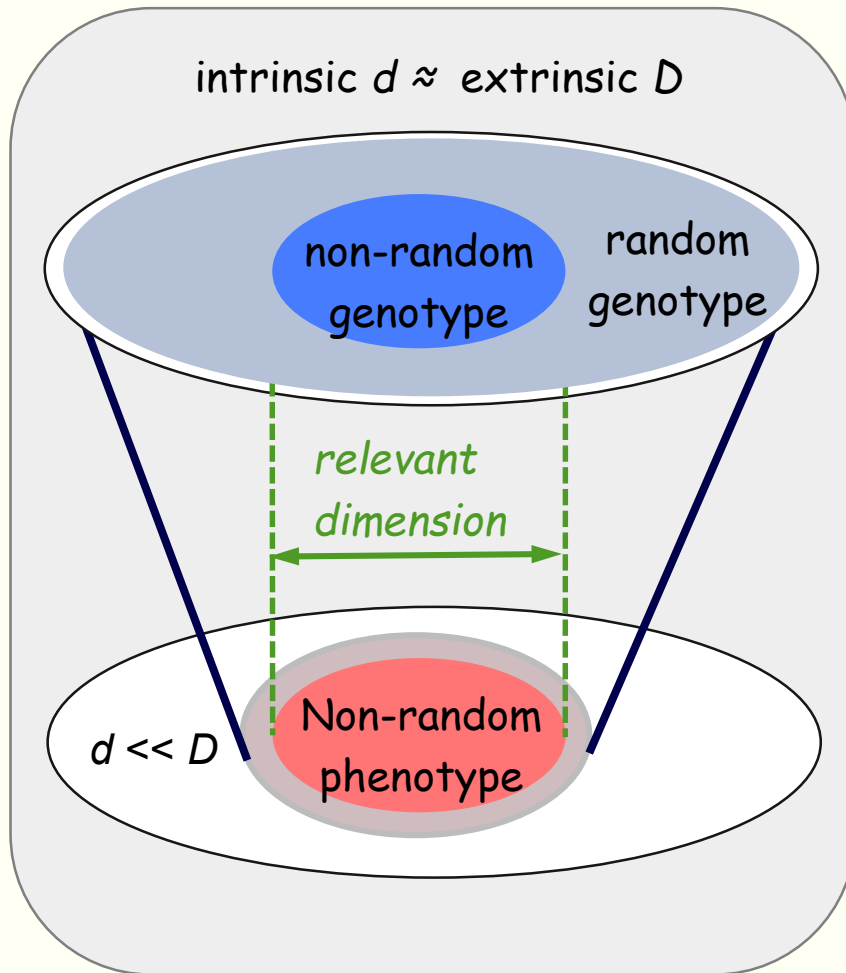
Any generic data can be reasonably compressed into a space of dimension

$$d_* \sim \frac{18}{\varepsilon^2} \log_{10} m$$

m = sample size $\sim 10^6$, ε = tolerance



Hypothesis. Biological systems exhibit extreme dimensional reduction, far beyond the generic limit

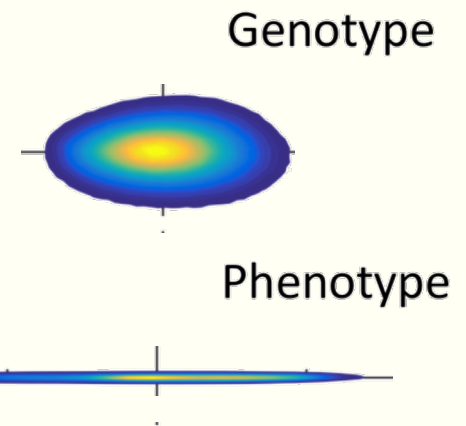
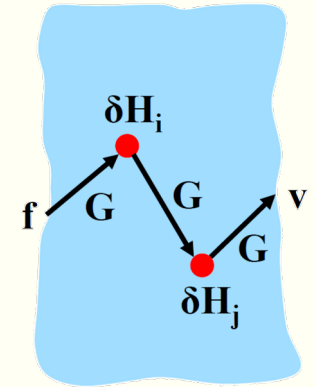


Most of the gene is **random**.

the **information carrying** part is of the same dimension as the phenotype.

Looking at protein dynamics suggests an intuitive picture of proteins as amorphous evolving matter

- **Mutations:** mechanical perturbations.
- **Epistasis:** scattering among mutations shapes *genotype-to-phenotype map* and *fitness landscape*.
- **Soft modes** emerge as function evolves.
- Mechanics explains **dimensional reduction** and **correspondence** function $\leftrightarrow$ non-random genetic info.

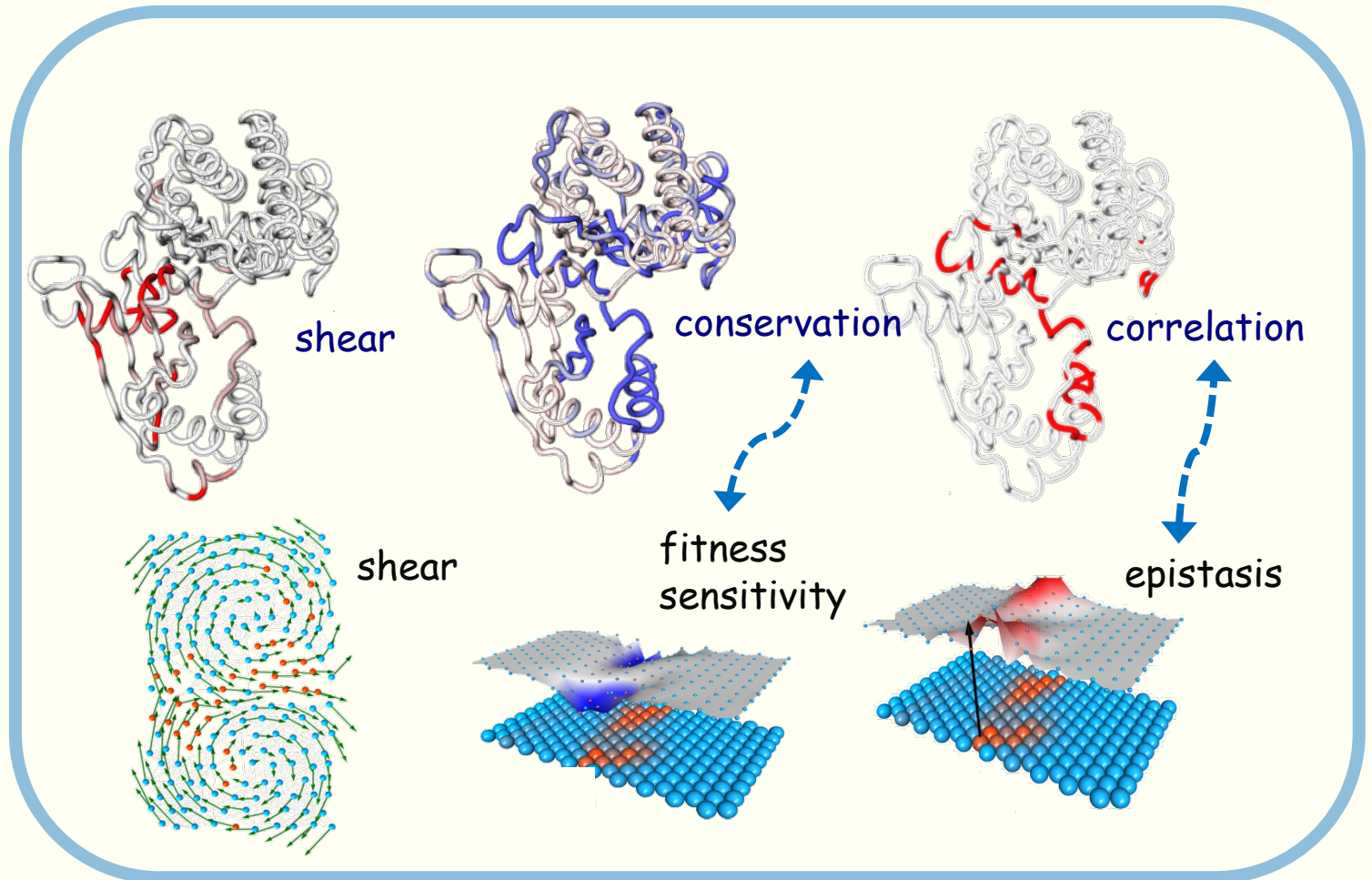


- **Applicable to other strongly-coupled living systems?**

Future directions for the amorphous evolvable matter approach

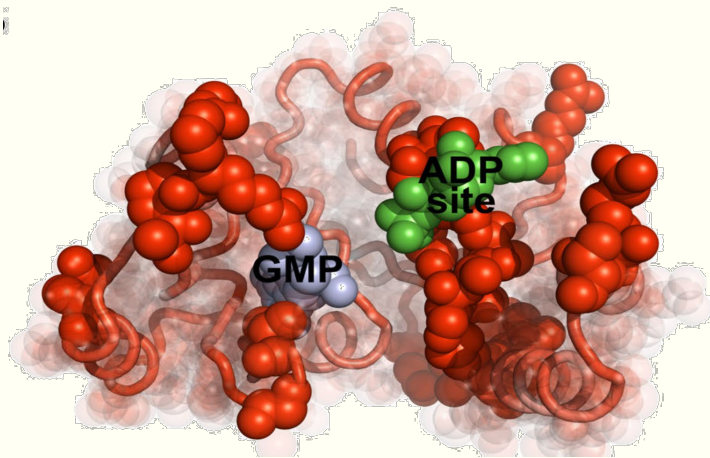
The physical picture suggests simple experimental tests

But so far, direct structural and mechanical data was unavailable

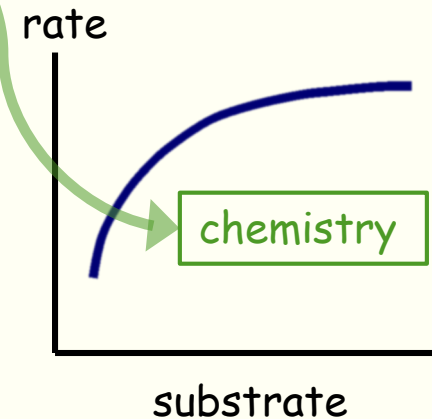
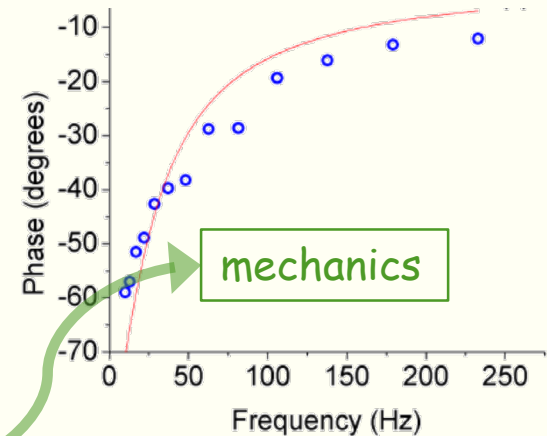
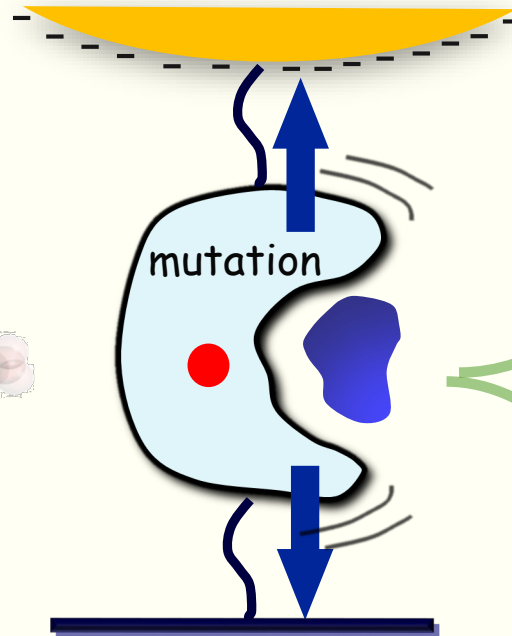


Protein *nano-rheology* directly probes the mapping:

gene $\rightarrow$ mechanics $\rightarrow$ function



Guanylate kinase



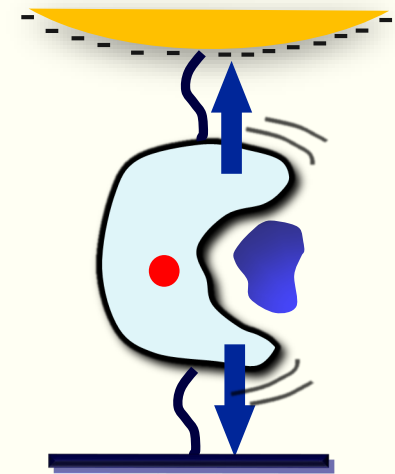
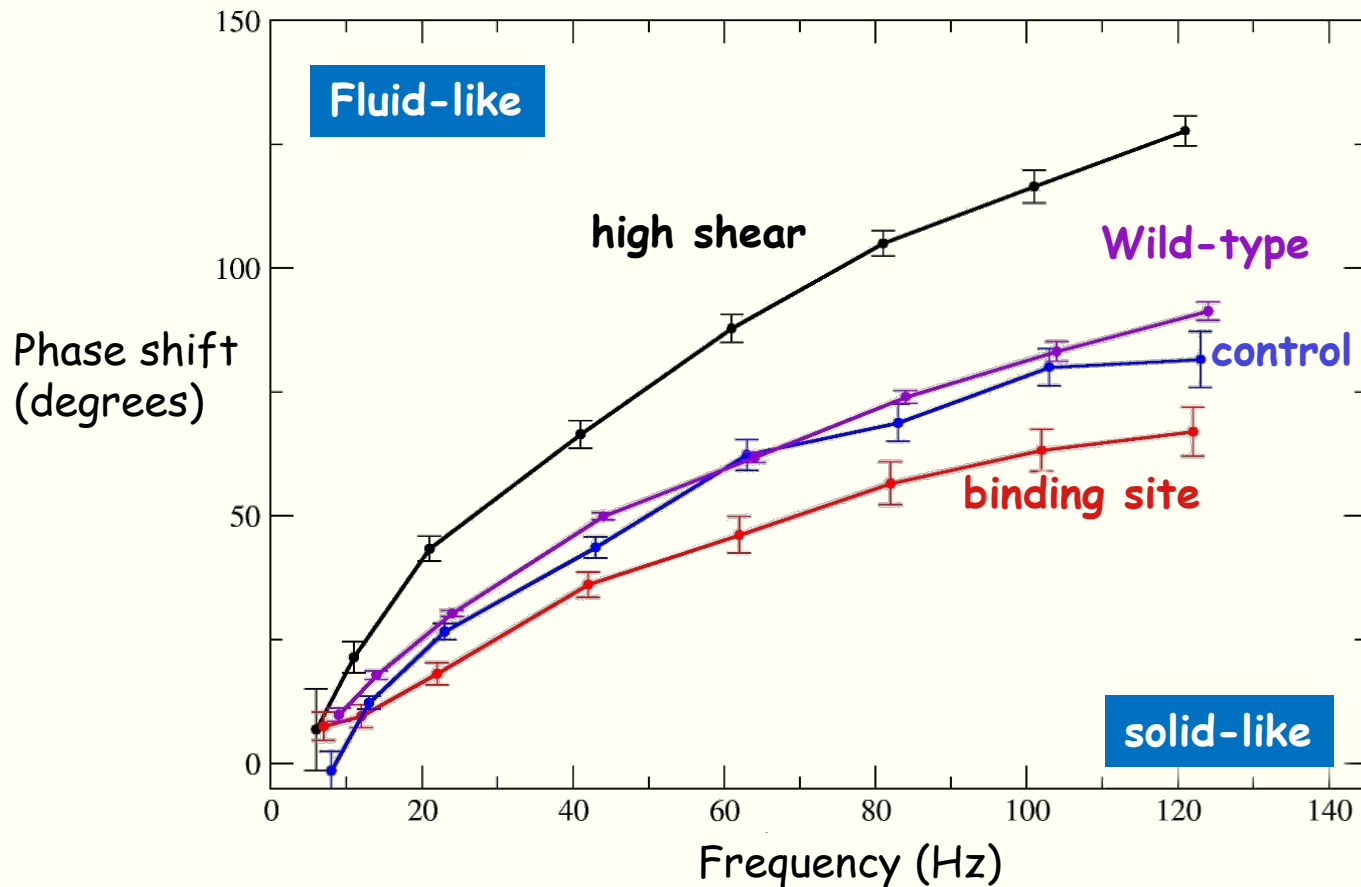
Measurement: E. Moses, E. Weinreb (WIS),

G. Zocchi (UCLA) M. Siek (IBS)

Analysis: J-P Eckmann (Geneva)

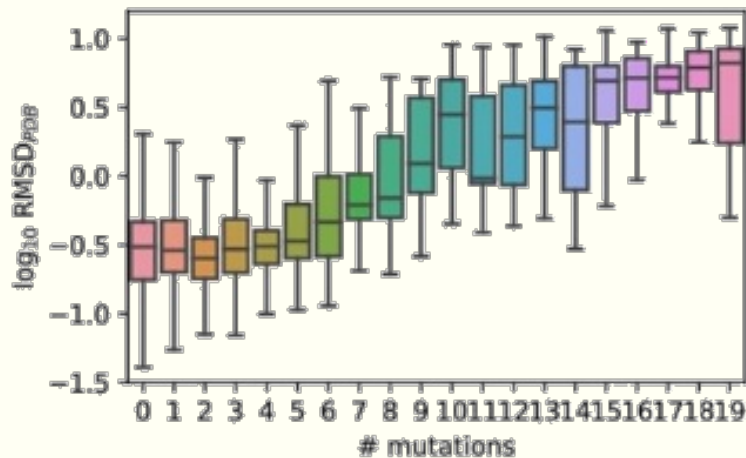
Mutant synthesis: Y Peleg, T Unger, J Sussman (WIS)

Preliminary results show strong effect of mutations in mechanically sensitive regions

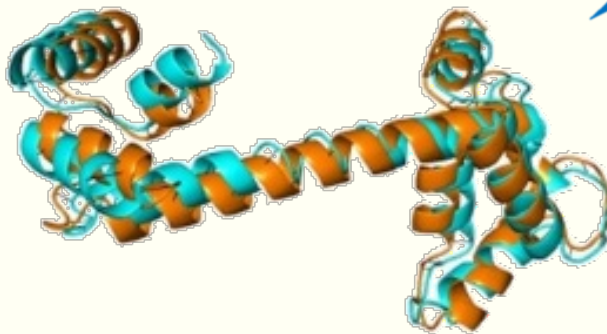


The protein folding algorithm AlphaFold2 can chart regions of the gene-to-structure map

John M. McBride
(IBS)



In principle, allows to calculate changes in *mechanical response* and thereby changes in *fitness* and *epistasis*

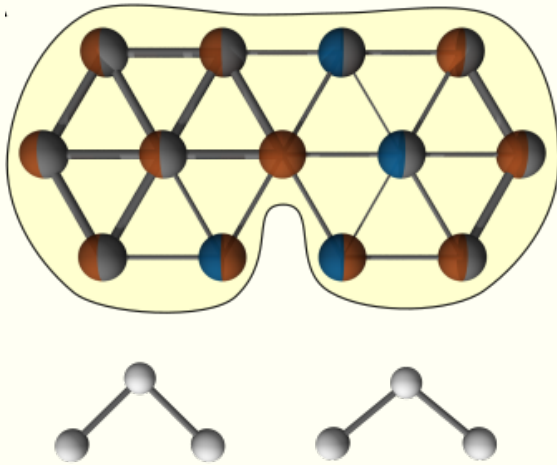


Example: predicted effect of a point mutation in troponin C

Mechanical approach to the problem of recognition:

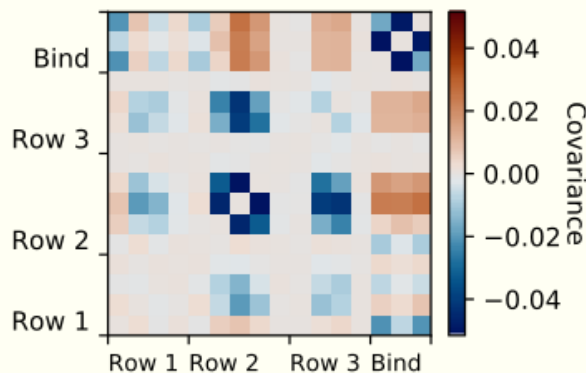
John M. McBride
(IBS)

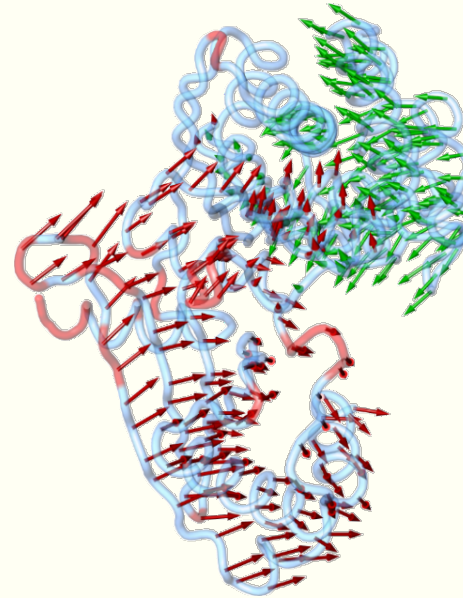
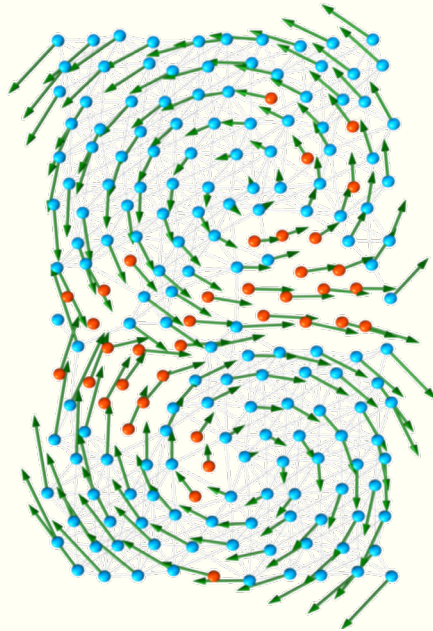
Why proteins are so big?



- Protein = spring network with elasticity and chemical interaction encoded in a gene
- Protein needs to discern a set of similar ligands.
- Sequence covariance among functional proteins shows long-range correlation deep into the protein.

→ Proteins are big to allow fine-tuned discrimination?





Thanks

Sandipan Dutta (IBS)

Jean-Pierre Eckmann & Jacques Rougemont (Geneva)

Albert Libchaber (Rockefeller)

Michael Mitchell & Stanislas Leibler (Rockefeller & IAS)



Center for Soft
& Living matter