

MATHEMATICAL FOUNDATIONS OF BIOLOGICAL
ORGANISATION, OBERWOLFACH 2023

PROTEIN AS A MATHEMATICAL PROBLEM

J.-P. Eckmann & J. Rougemont (Geneva)

E. Moses & E. Weinreb (Weizmann)

G. Zocchi (UCLA)

A. Libchaber (Rockefeller)

M. Mitchell & S. Leibler (Rockefeller & IAS)

J. McBride, S. Dutta, M. Siek & TT (IBS)

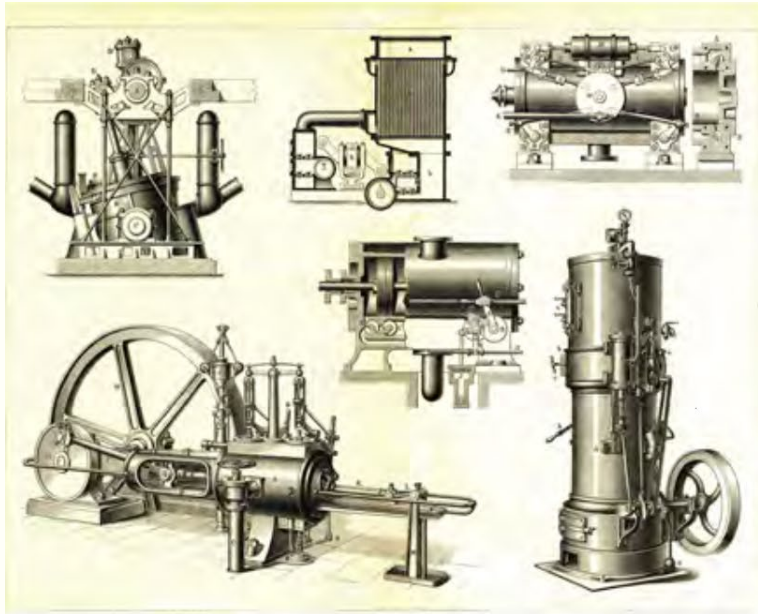


Please ask questions during the talk

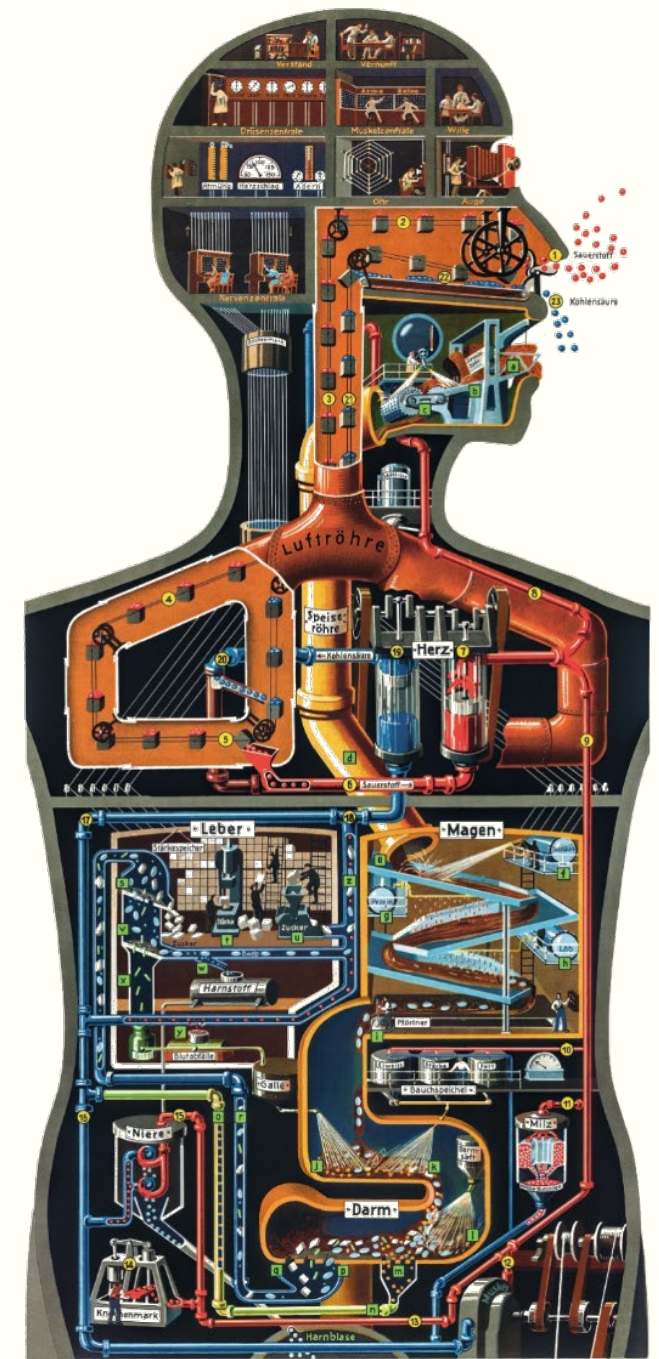
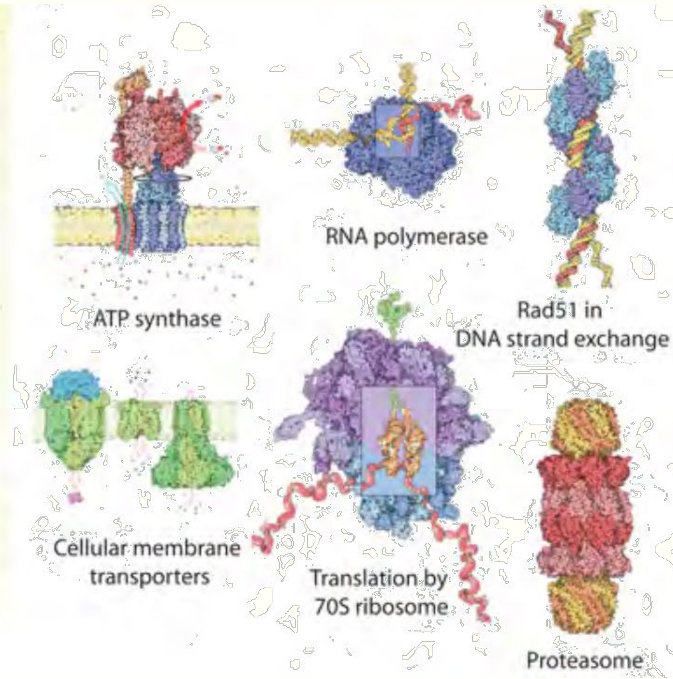
Proteins are the chief molecular machines of life.

As with any machine, one desires to understand the logic of design\evolution, production, and function

BLUEPRINT → STRUCTURE → DYNAMICS → FUNCTION



Michael Hons, 2015

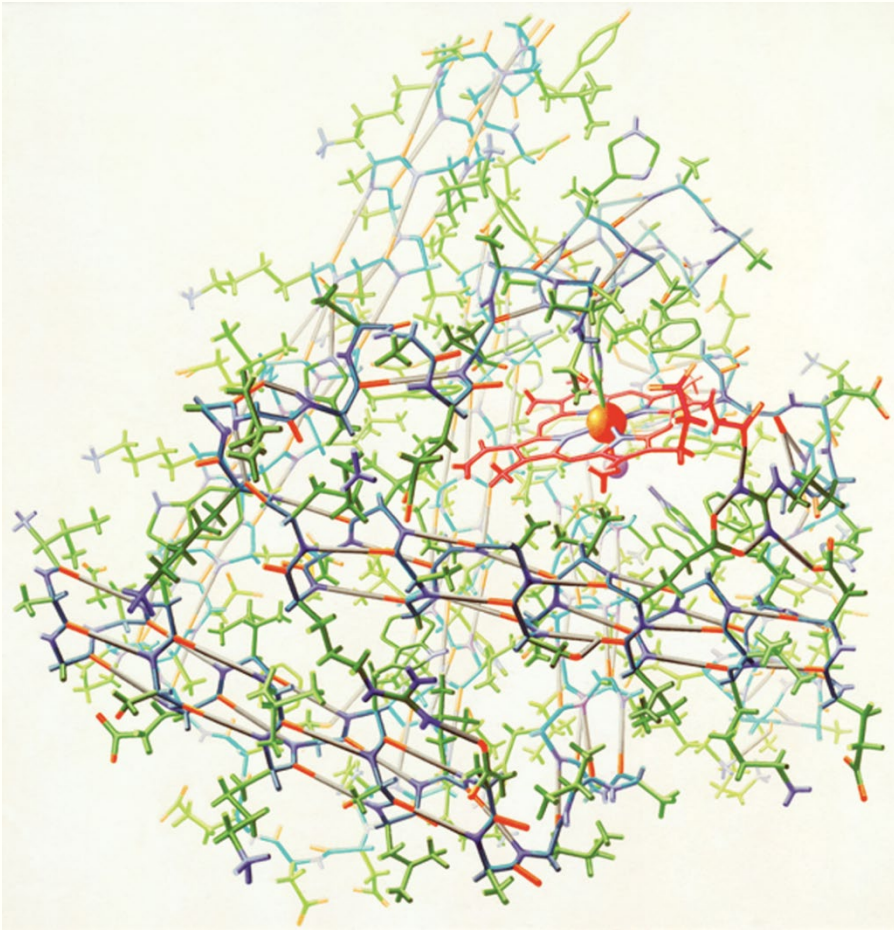


Fritz Kahn, 1927, der mensch als industriepalast

* **Machine:** an object that transforms a physical system non-randomly to perform a function

In the logic of protein machines, there has been a breakthrough in understanding folding.

But most fundamental problems remain open, despite the explosion of data and experiment



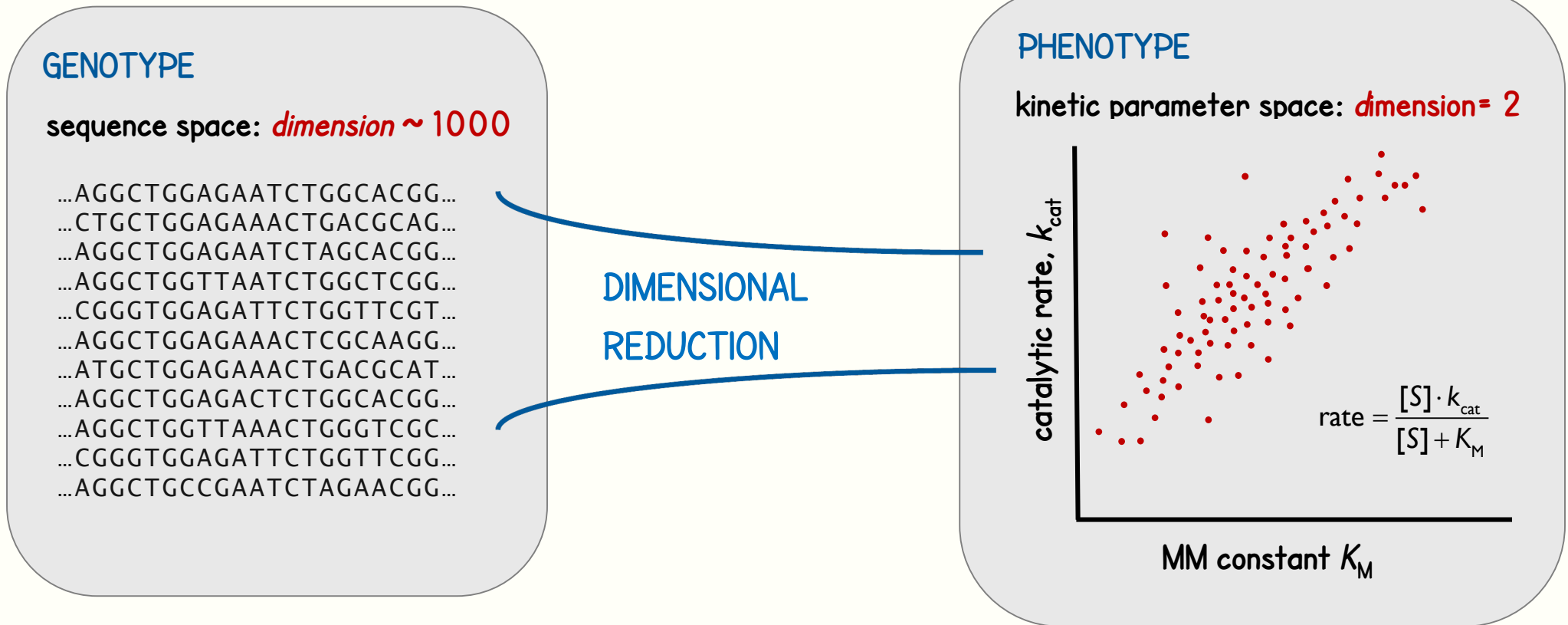
Kendrew (1961) – 3D structure
of myoglobin (art by I. Geis)

PROTEIN machines are made of dual matter

(= many-body physics + evolutionary dynamics)

- Non-random **information-rich** object.
- Long-term evolutionary **memory** (non-ergodic)
- **Small, heterogeneous** (no TD limit, Boundary is essential)
- **Sparsity of data** in high-D gene space (sampling)
- Collective **many-body** interactions.
- Function is part of **multiscale networks** of biochemistry, evolution, and ecology.

1. How the many-body physical interactions that determine protein function are encoded in the gene?



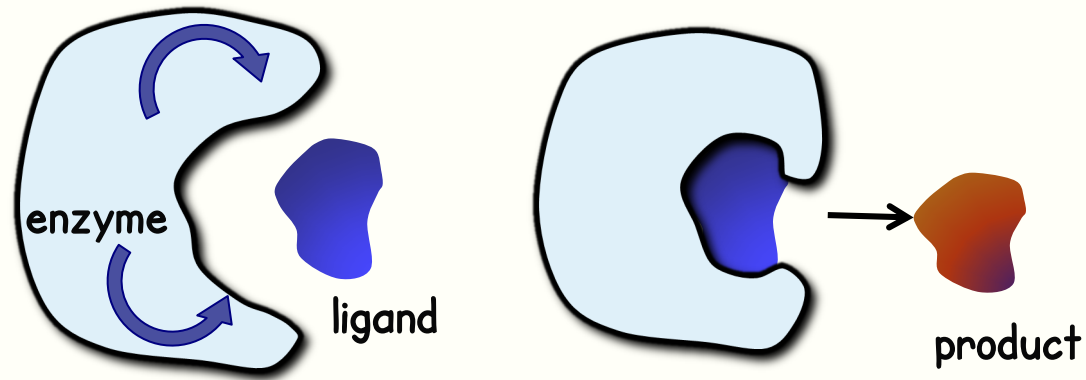
Genes and protein structures have *many* parameters, but protein function is often described by only *a few*

→ What is the origin of the *dimensional reduction* in the genotype-to-phenotype map?

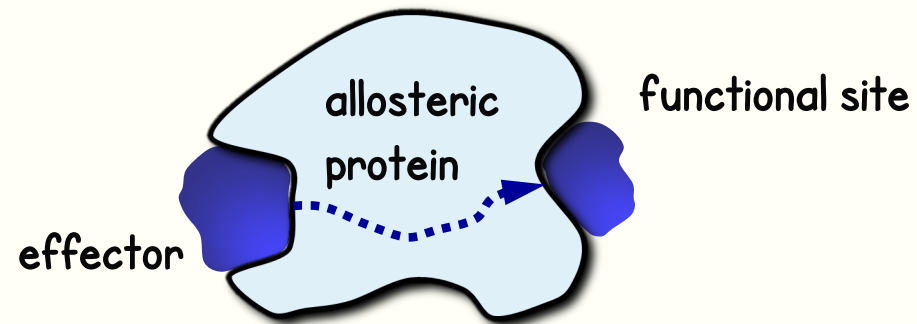
2. What is the role of dynamics in protein function?

Many proteins use large-scale collective motions to function

Induced fit



Allostery



** Localized chemical interactions induce global motions*

Outline: Questions about physics-evolution correspondence in protein

BLUEPRINT

folding

STRUCTURE

DYNAMICS

FUNCTION

1. How the **many-body** physical interactions that determine protein **function** are encoded in the gene?
 - the origin of **dimensional reduction** in genotype-to-phenotype?
2. What is the role of **dynamics** in function and evolution?

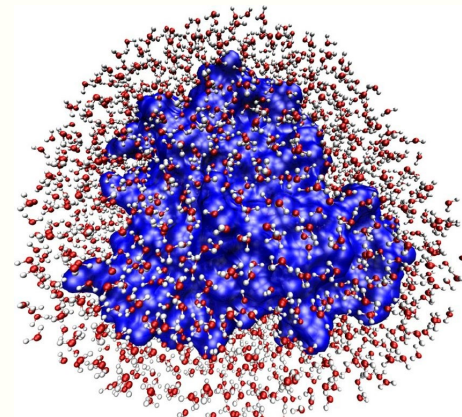
We will not discuss:

3. How do proteins achieve their phenomenal **specificity**?

(Eckmann, McBride et al., **General Theory of Specific Binding: Insights from a Genetic-Mechano-Chemical Protein Model**, Mol Bio Evol 2022)

4. What is the physics of **water** + ions + electrons?

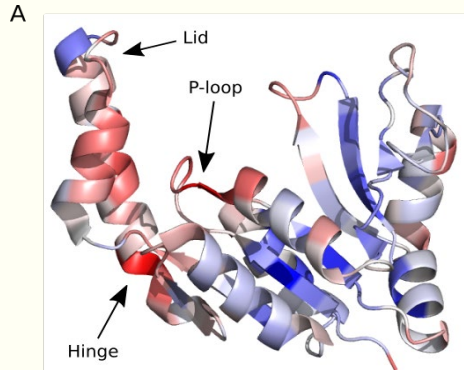
(Collective long-range modes, underscreening)



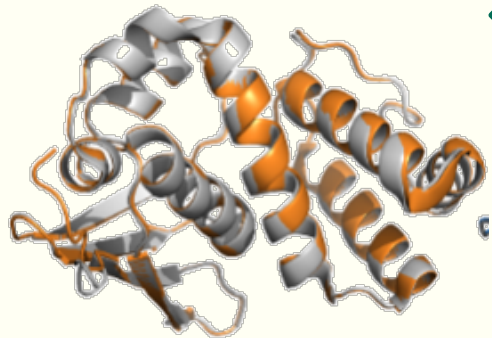
Impact of high-shear regions on protein dynamics and function

Outline: Methods and Results

Nanorheology: proteins are not simply elastic but **visco**-elastic



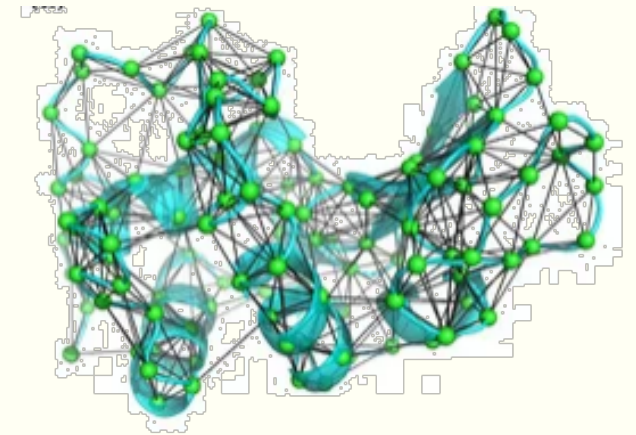
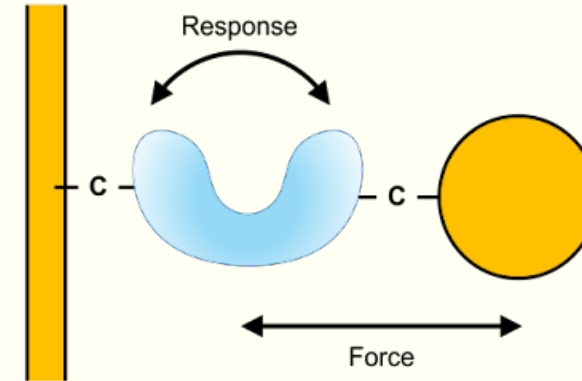
AlphaFold - how can we use it to probe evolution-physics correspondence



3.
Databases,
Experiments

2.
Machine
Learning

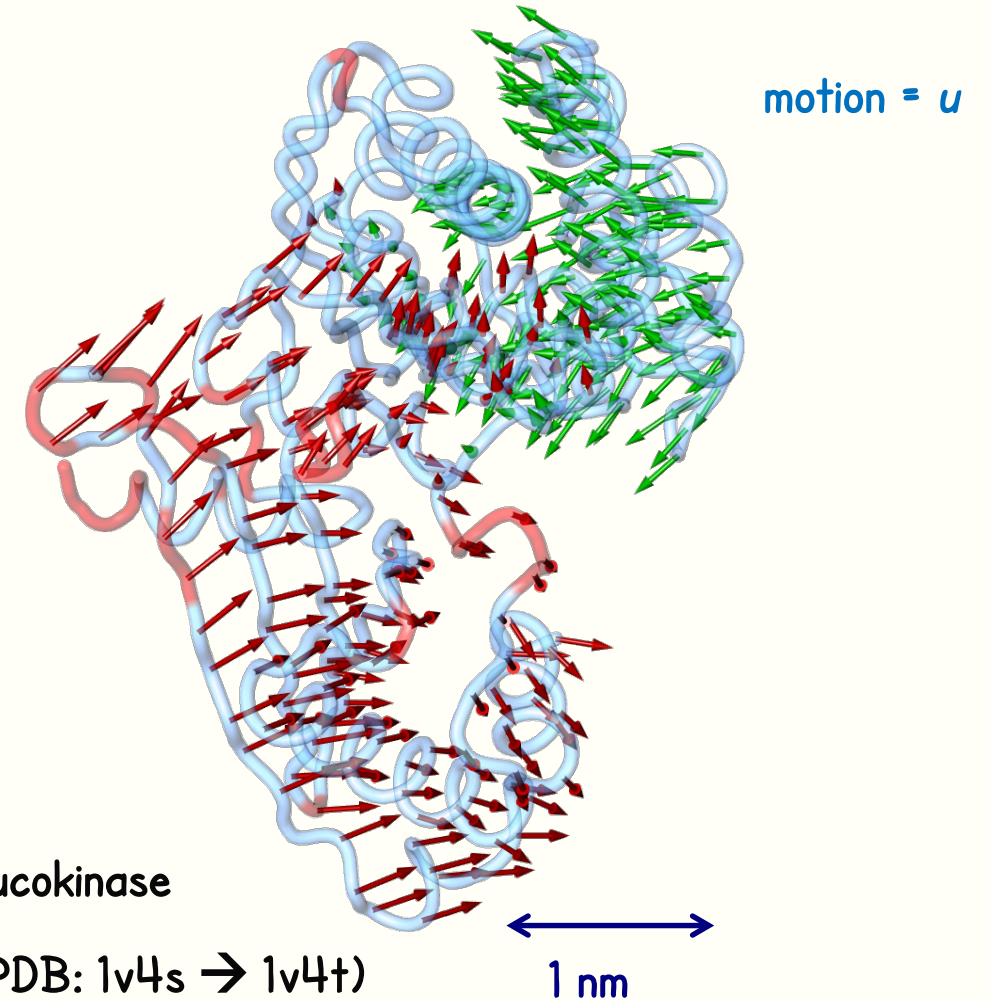
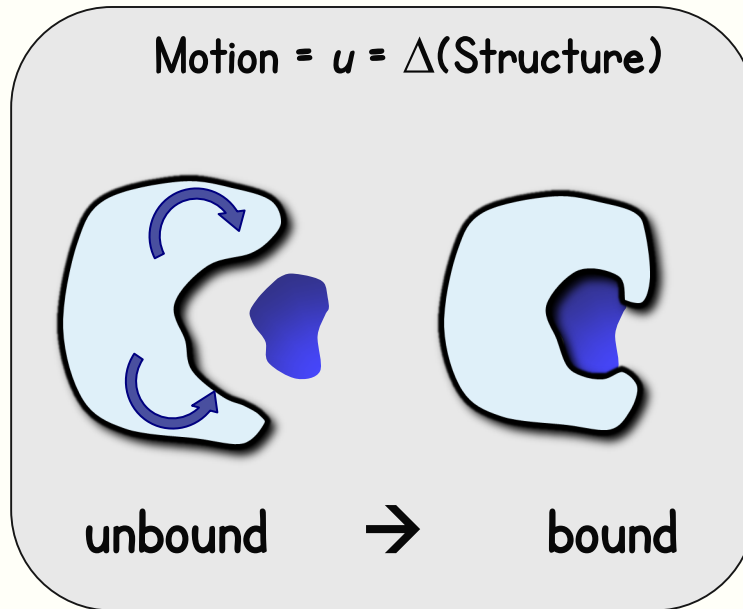
1.
Physical
Models



Emergence of **allostery** and **specific recognition** in simple mechanical modes.
(and link to **dimensional reduction**)

Motivation:

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



Eckmann TT, Bioessays 2021
Rougemont Eckmann TT, *Rev Mod Phys* 2019
Eckmann Libchaber TT, *PRX* 2018
Dutta Eckmann Libchaber TT, *PNAS* 2017
Mitchell Leibler TT, *PNAS* 2016

The shear measures the internal motion of amino acids in the protein

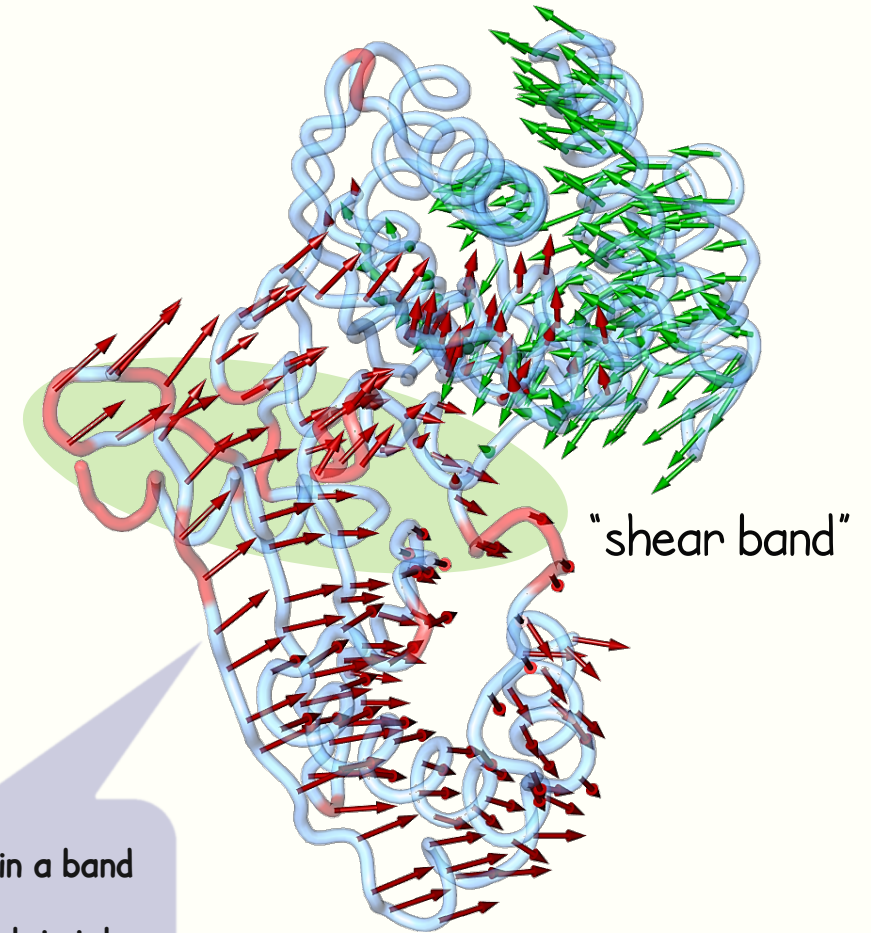
Shear = how much amino acids move with respect to their neighbors = $\Delta\Delta$ Structure

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

Geometric meaning: **metric** of deformation $\varepsilon = \frac{1}{2}(\mathbf{g} - \mathbf{1})$

Discrete version by tracing motion of neighboring atoms (or a.a.)

$$\mathbf{g} = (\nabla \mathbf{u} + \mathbf{1})(\nabla \mathbf{u}^T + \mathbf{1}) \rightarrow \text{discrete approx.}$$



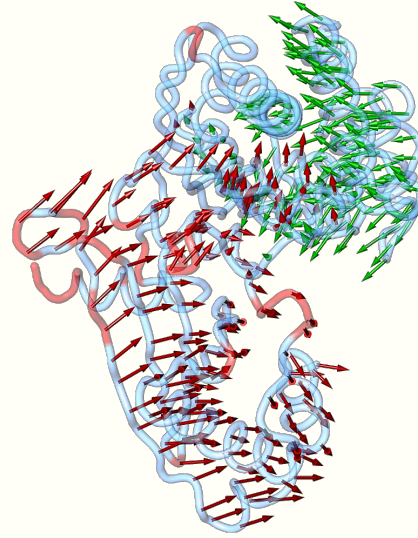
Shear is localized in a band separating the protein into rigid domains

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

We will examine deformations induced by physics and evolution

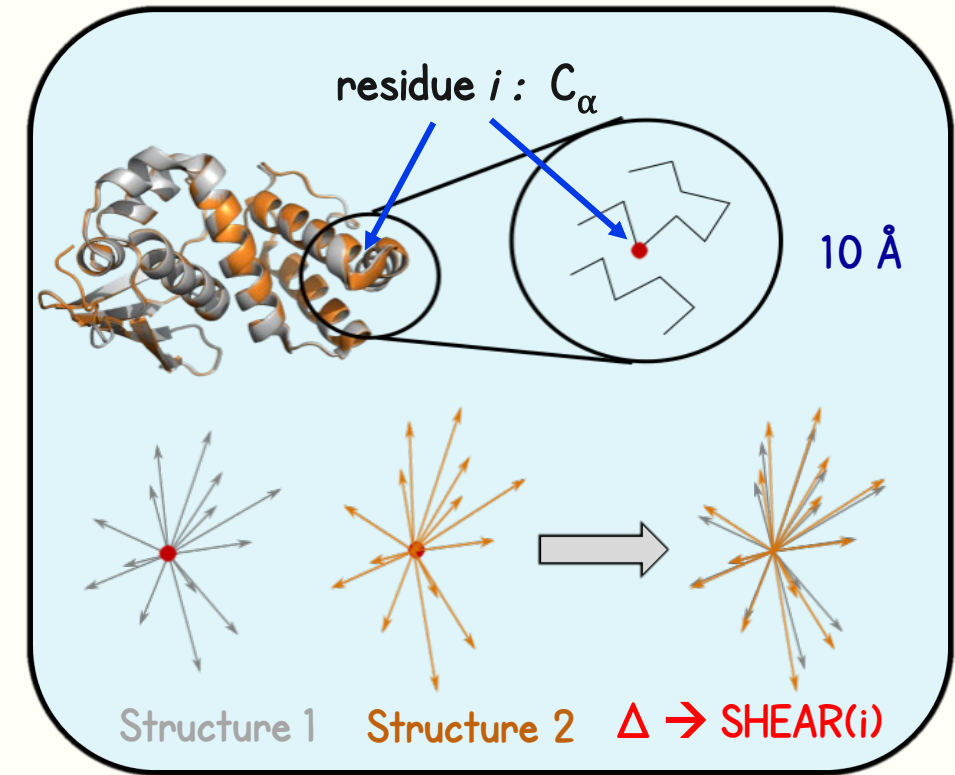
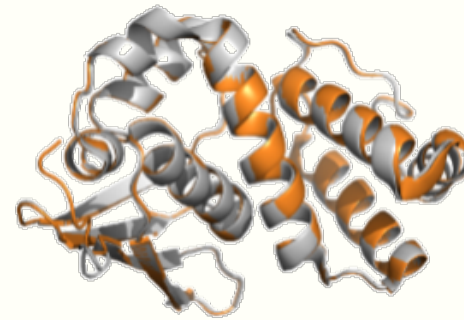
PHYSICAL

- Deformation during binding ligands
 - Example: induced fit, allostery
- Deformation induced by forces acting on **elastic** (or **viscoelastic**) networks.



EVOLUTIONARY

- Mutations induce structural changes.

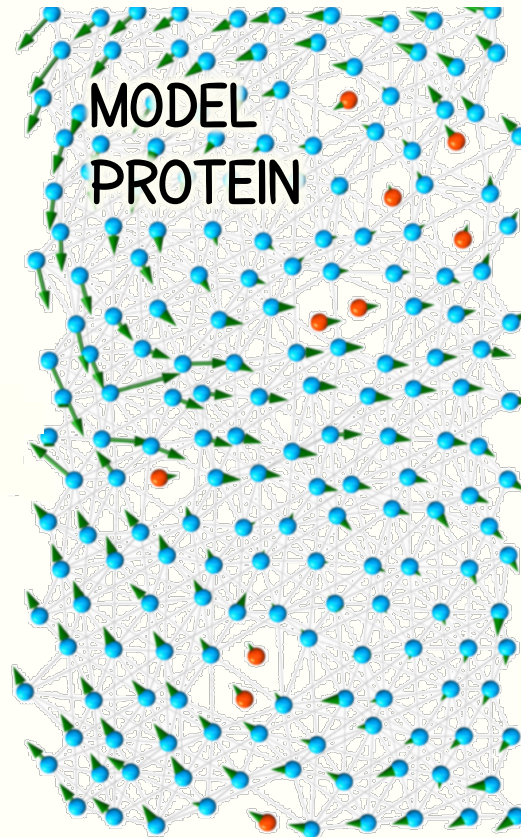
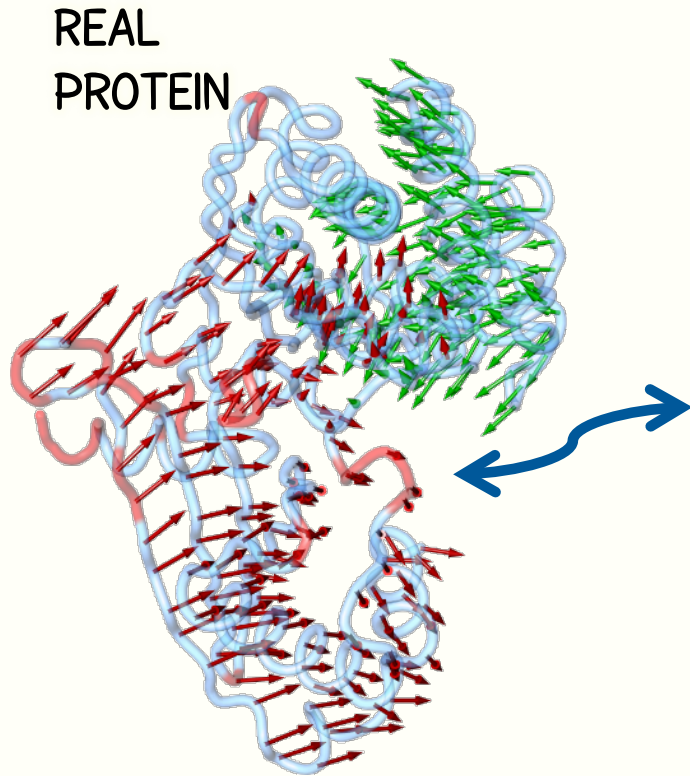


In all cases, the deformation is measured by the shear tensor

1. SIMPLE THEORETICAL IDEAS

- Mechanical deformations reveal the **correspondence** between physics and evolution.
- **Green functions** are the natural theoretical object to this end.
- Mutations correspond to localized **mechanical\structural perturbations** and evolutionary interactions (epistasis) correspond to interactions of "defects".
- **Dimensional reduction** originates from physics-evolution correspondence.

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

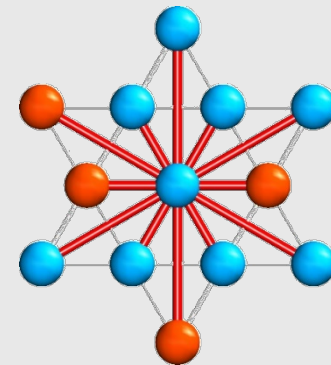


200 beads = amino acids .

Two types of amino acids
connected by "springs"

amino acids	H	P
H	strong	weak
P	weak	weak

12 neighbors



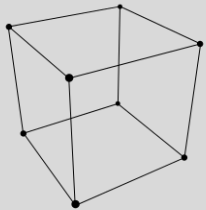
Main goal:

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

GENE

...101111101111101...

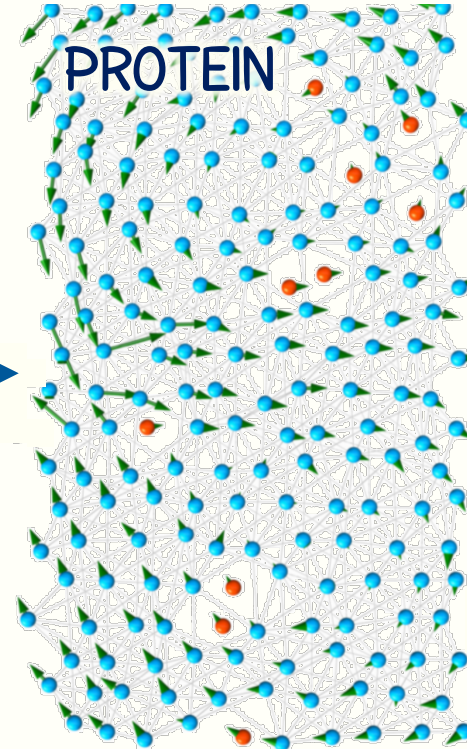
...00111111100110...



MAP Γ

$$\Gamma : \{0,1\}^{200} \rightarrow \text{proteins}$$

PROTEIN



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

- How many proteins are functional?
- How are they distributed in the high-dimensional genotype and phenotype spaces?

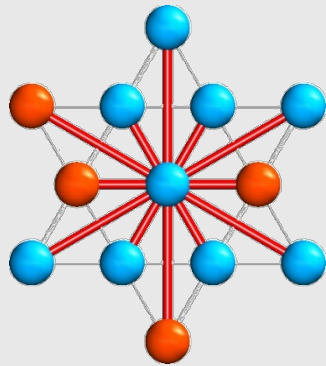
The mechanical response of the protein is determined by the elastic network

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

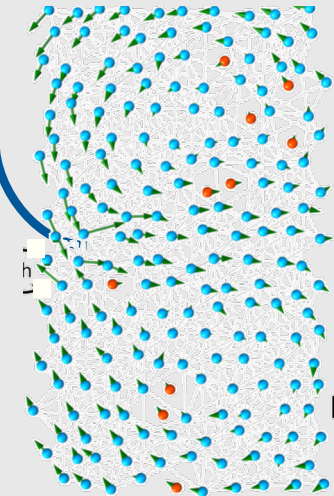
$$\mathbf{H}_{ij}(g_i, g_j) = \mathbf{A}_{ij} \cdot k_{ij}(g_i, g_j) \cdot \mathbf{n}_{ij} \times \mathbf{n}_{ij}$$

$$|\text{force}\rangle = \mathbf{H} |\text{motion}\rangle$$

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



amino acids	H	P
	strong	weak
H	strong	weak
P	weak	weak



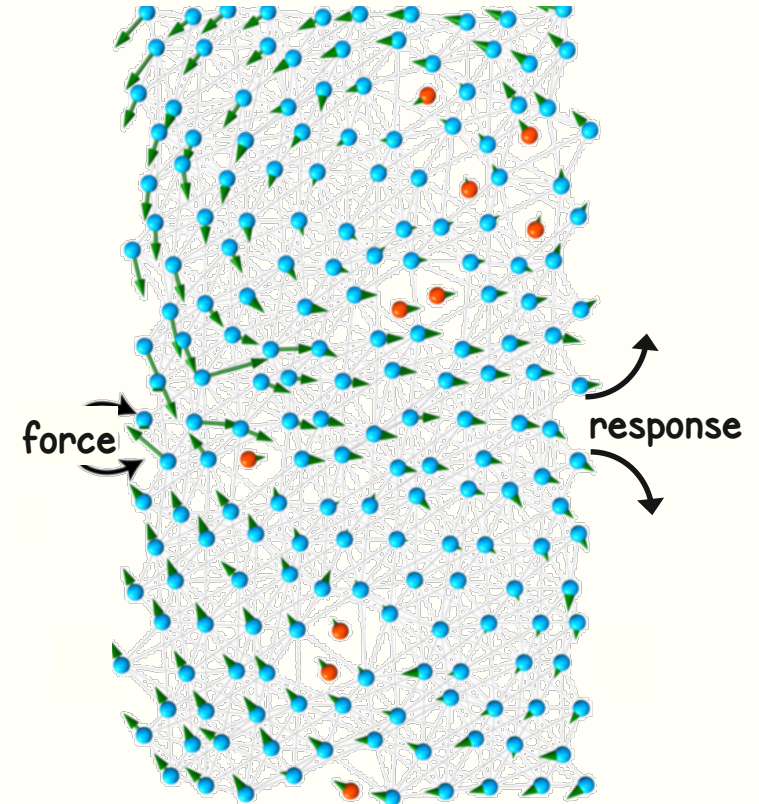
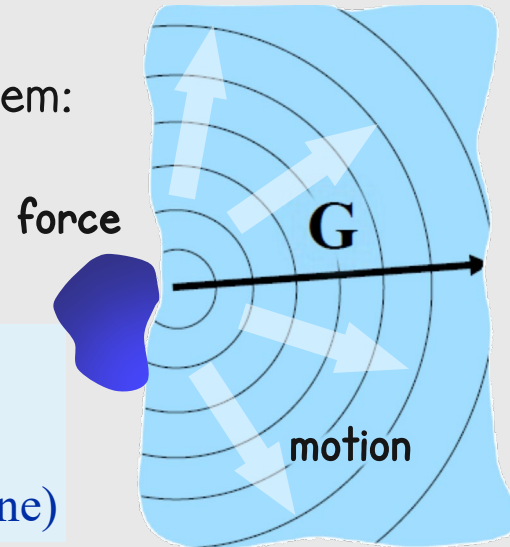
Green's function measures the protein motion induced by a local force, such as ligand binding

Green's function G

solves the inverse problem:

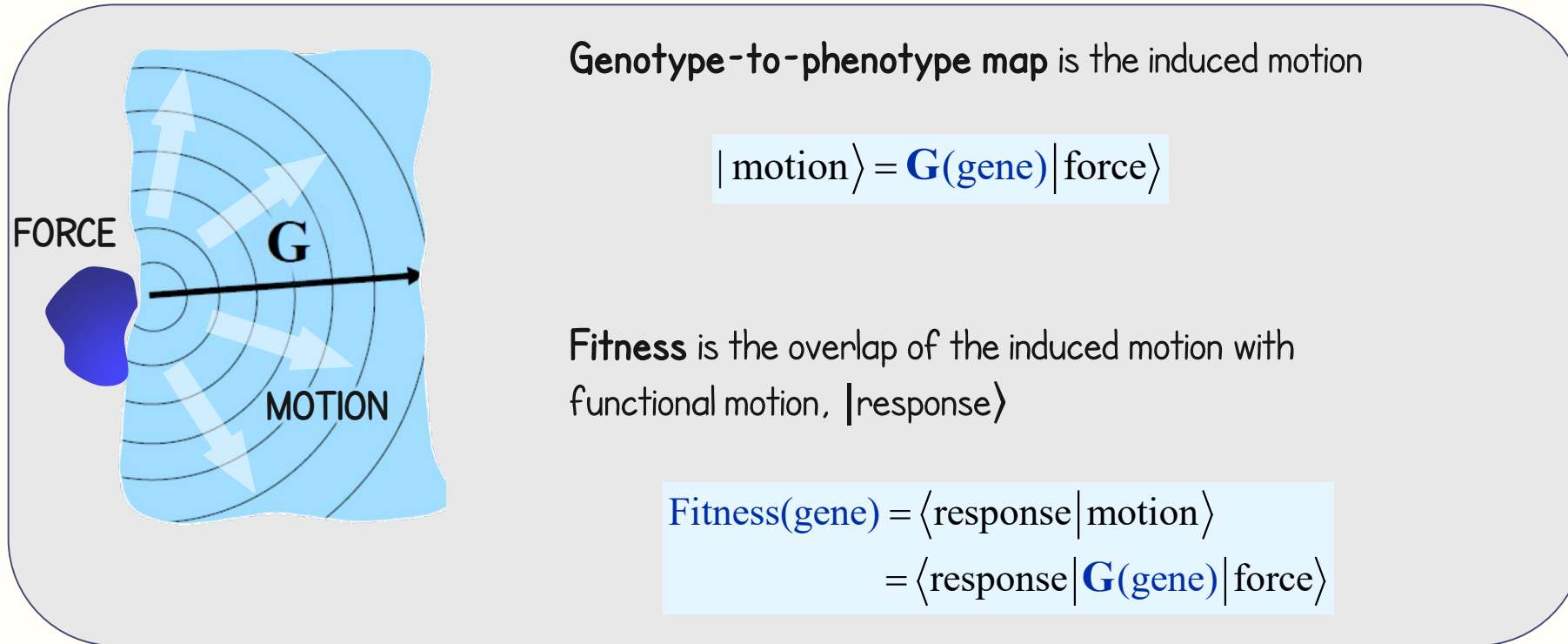
$$|\text{motion}\rangle = \mathbf{G} |\text{force}\rangle$$

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



* G is the pseudoinverse: $G = H^*$

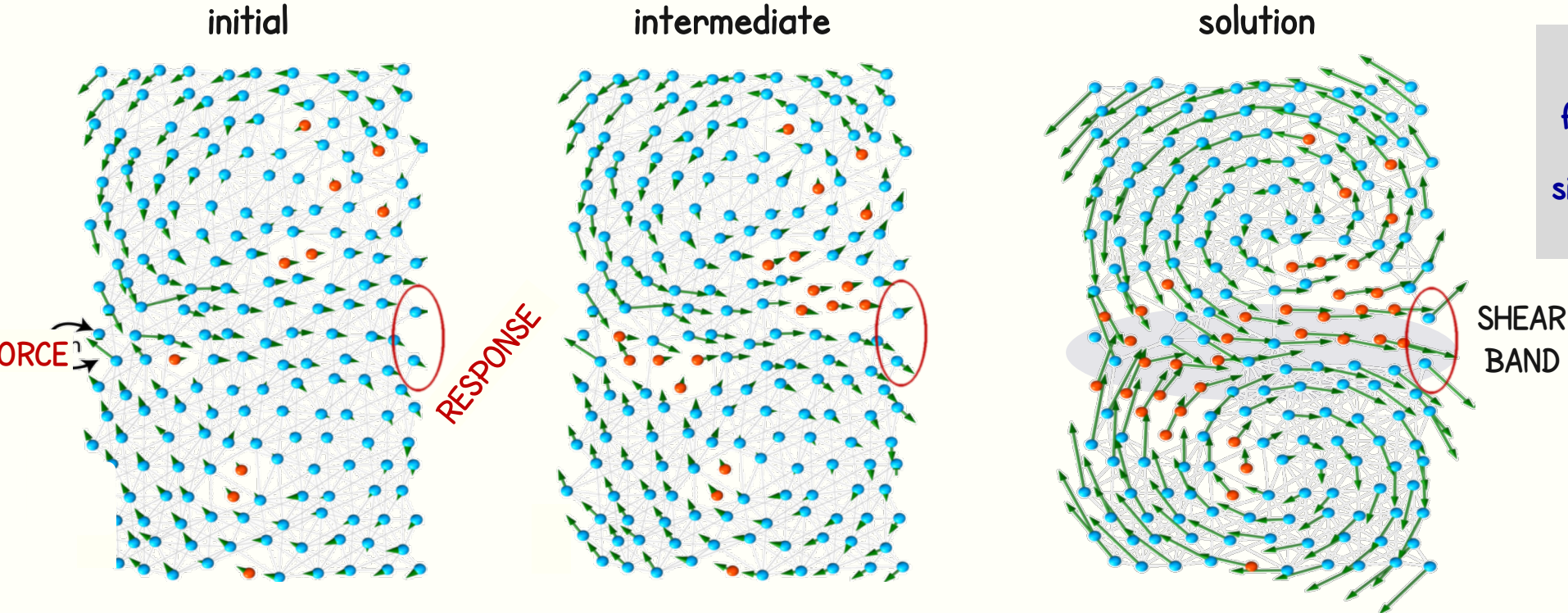
Green's function provides explicit gene-to-function map and fitness landscape



G: *gene* $\rightarrow$ *amino acid network* $\rightarrow$ *protein dynamics* $\rightarrow$ *function* $\rightarrow$ *fitness*

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

$$\text{Fitness} = \text{overlap of } \langle \text{response} | \mathbf{G} | \text{force} \rangle$$



fitness cares only about output site, not about internal structure

$$\mathbf{G} | \text{force} \rangle \approx \nabla \mathbf{G} | \text{force dipole} \rangle$$



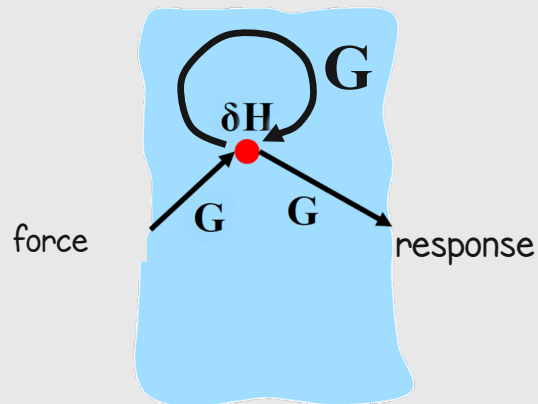
One can look at more general forces and responses: for example, specific recognition

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

Effect of mutations on elastic network

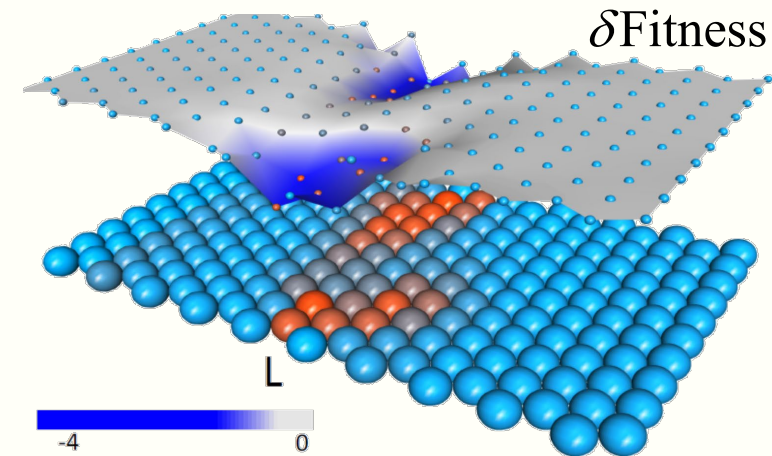
$$\delta \mathbf{H} = \mathbf{H}(\text{gene}') - \mathbf{H}(\text{gene})$$

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

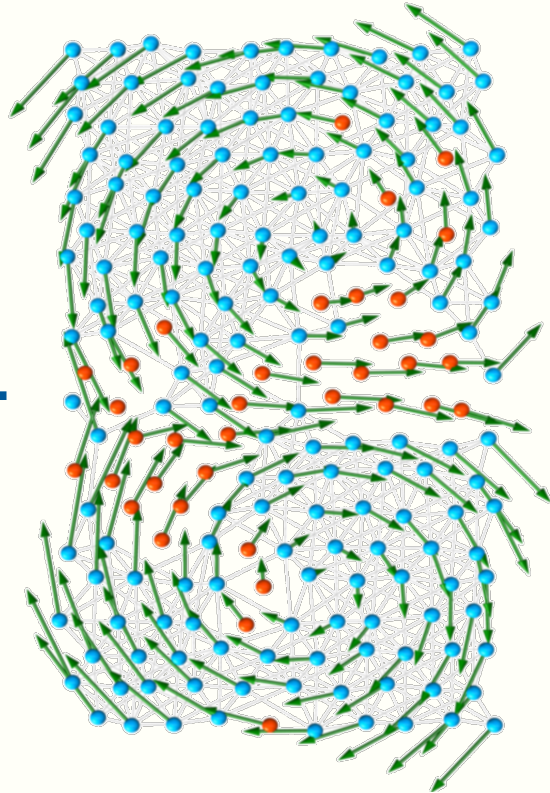
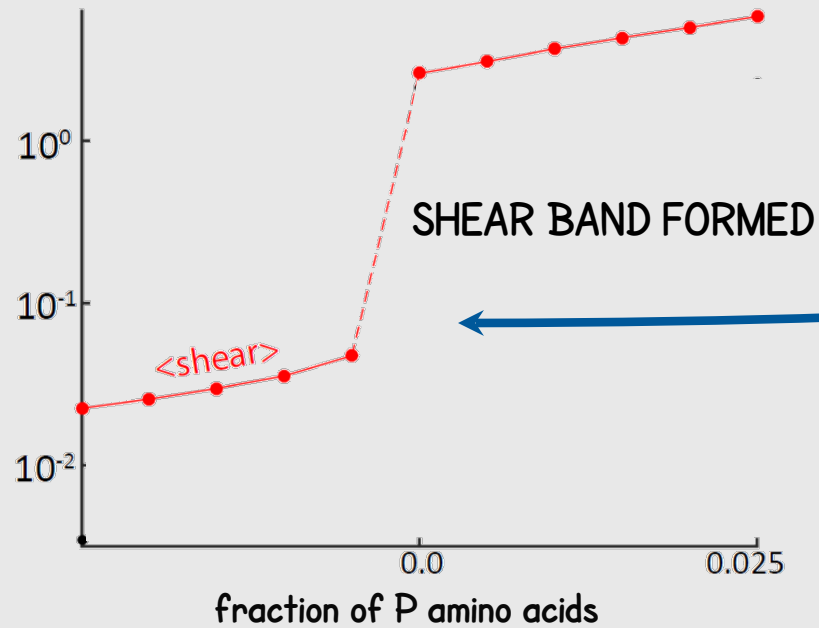


$$\delta \mathbf{G} = -\mathbf{G} \delta \mathbf{H} \mathbf{G} + \mathbf{G} \delta \mathbf{H} \mathbf{G} \delta \mathbf{H} \mathbf{G} - \dots$$

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



$\lambda_* = \text{soft mode} \rightarrow 0$

$$\mathbf{H} = \sum_i \lambda_i |\mathbf{v}_i\rangle\langle\mathbf{v}_i| \Rightarrow \mathbf{G} = \sum_i \frac{|\mathbf{v}_i\rangle\langle\mathbf{v}_i|}{\lambda_i}$$

$$\text{Fitness} = \langle \text{response} | \mathbf{G}(\text{gene}) | \text{force} \rangle \sim \frac{1}{\lambda_*} \rightarrow \infty$$

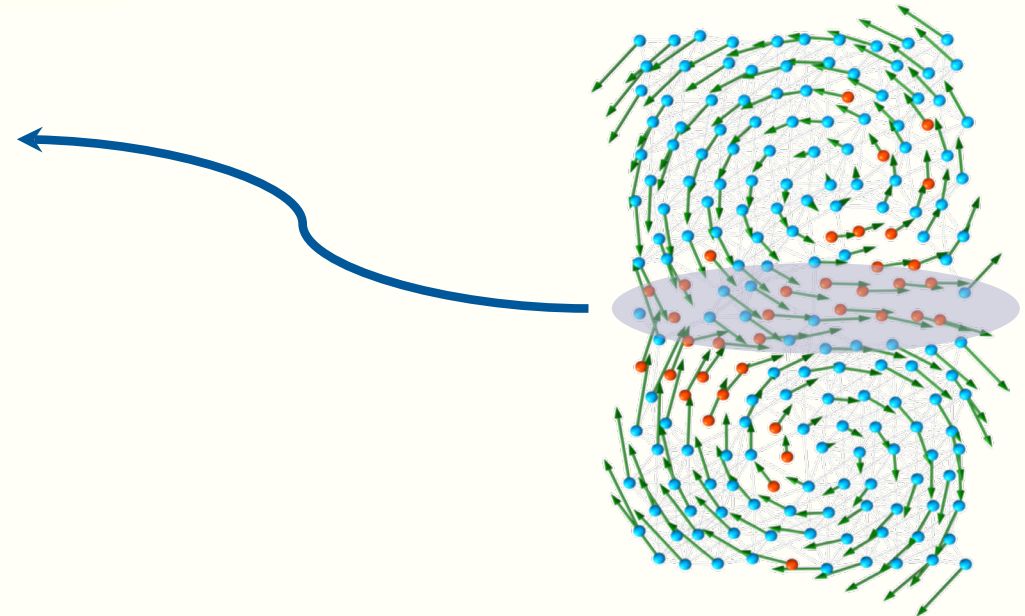
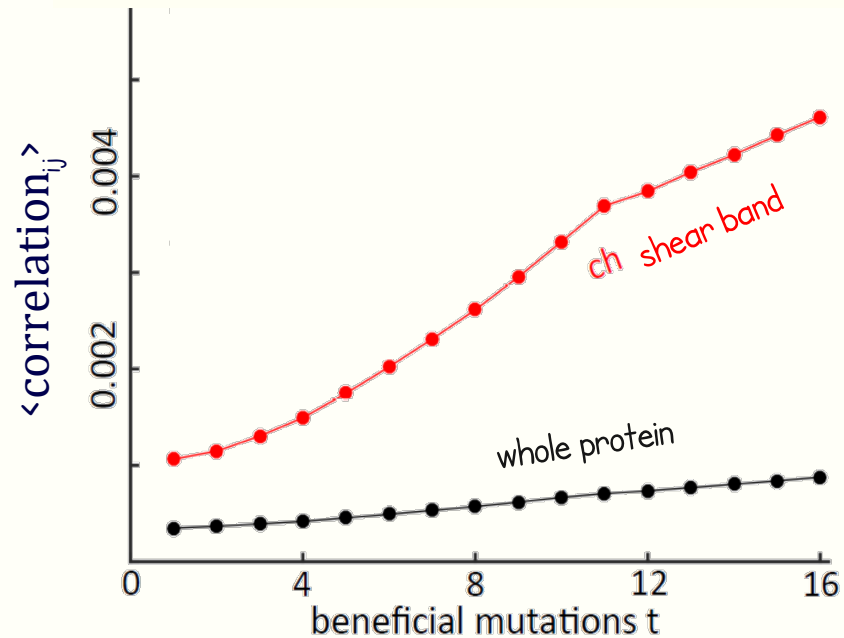
• Matrix-tree theorem

$$\# \text{ spanning trees} = \frac{1}{n} \lambda_* \lambda_2 \dots \lambda_{n-1}$$

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

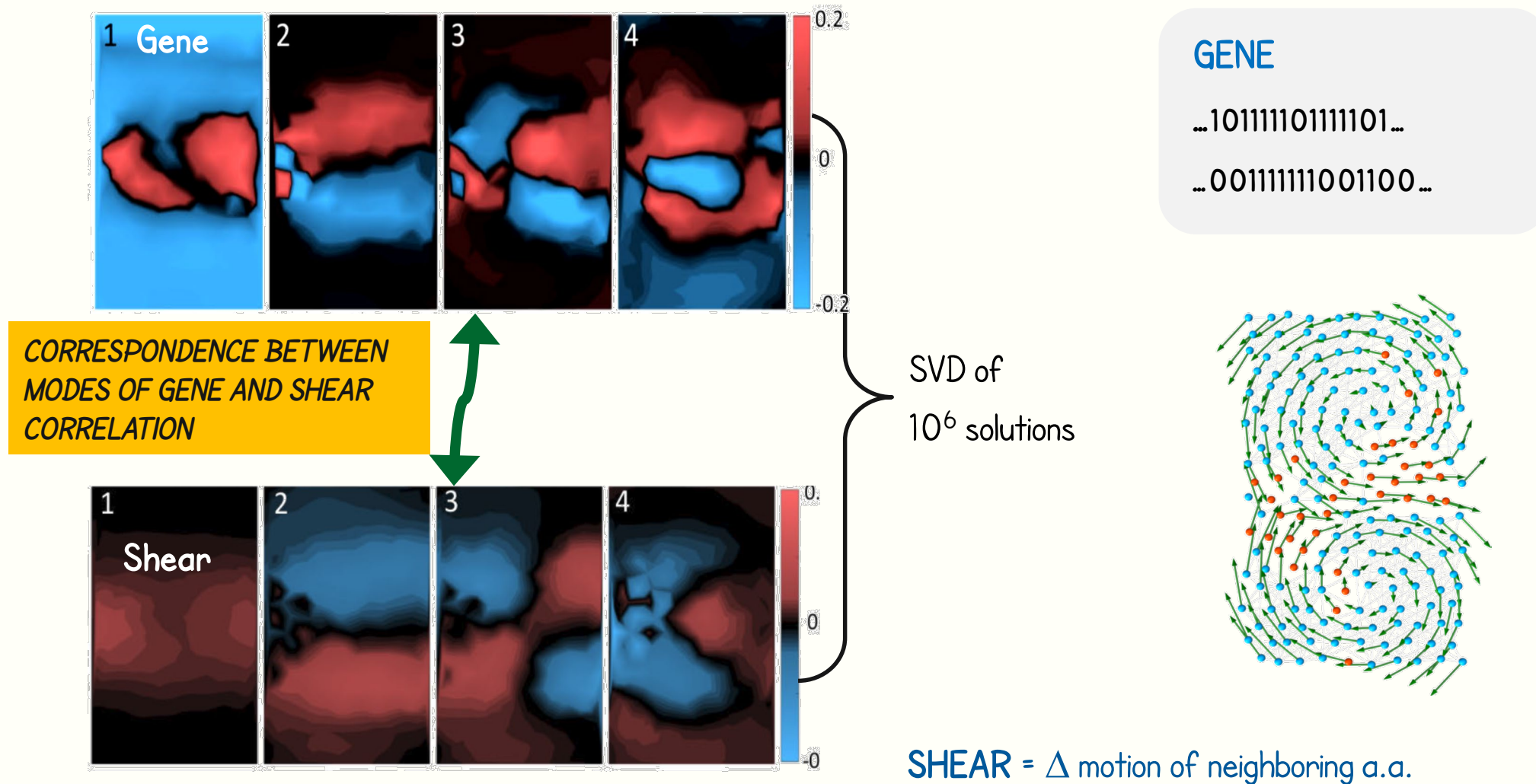
10^6 solutions {
11100101011111000010101...
10010001000101011101110...
00111110111110100111010...
10011111110011001111000...
...

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

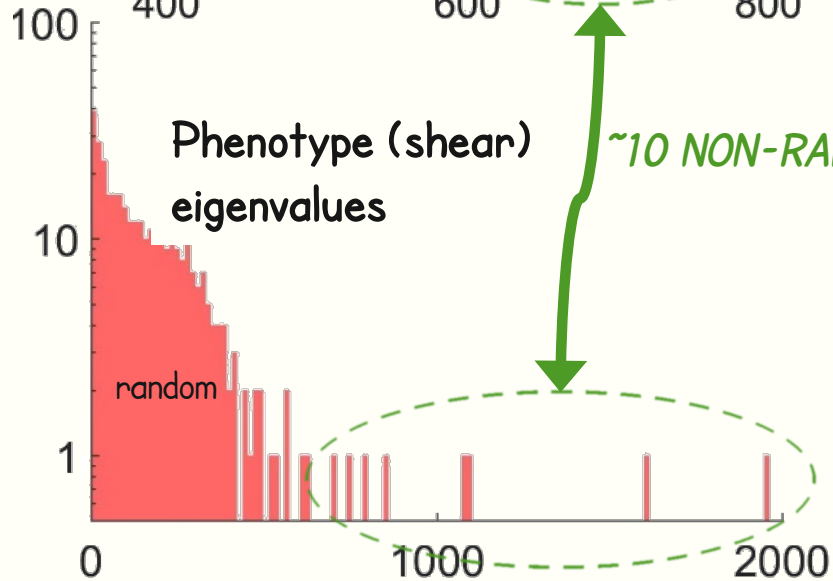
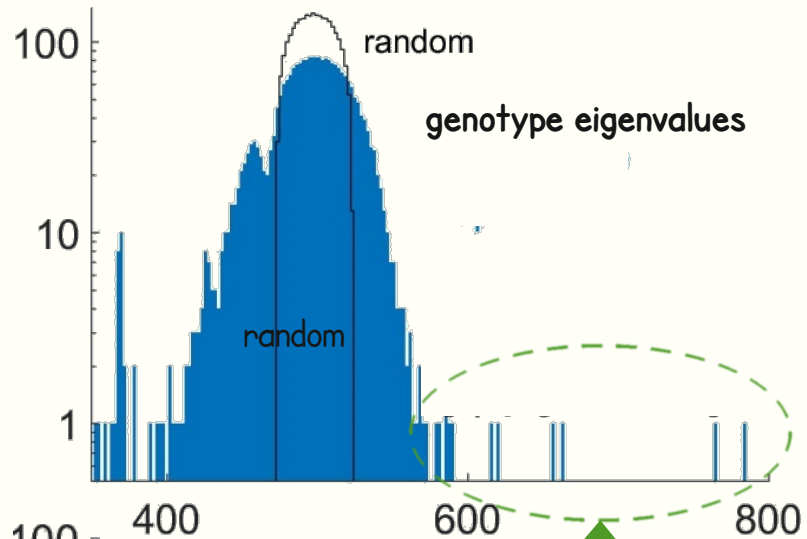


Spectral analysis:

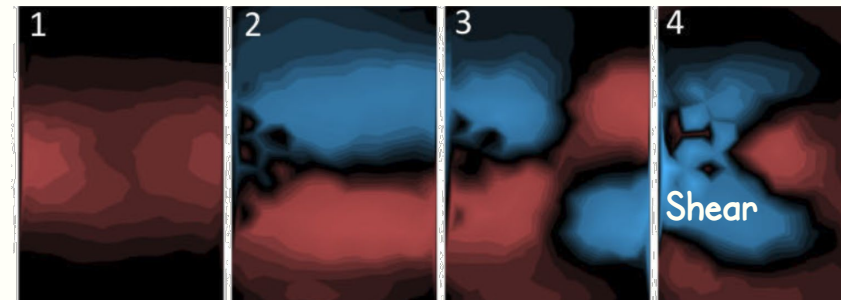
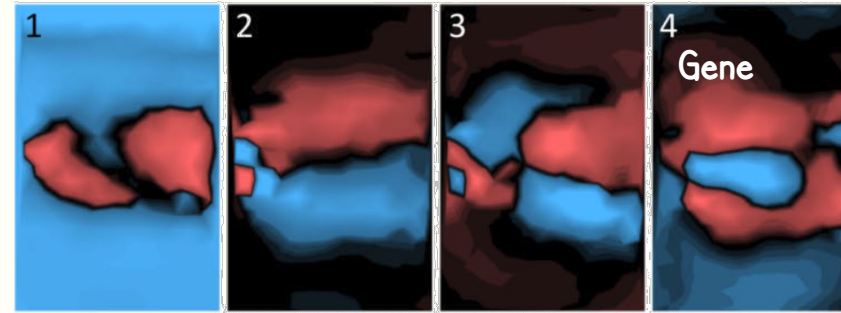
Correlation spectra of phenotypes and genotypes capture the relevant modes of mechanical function and evolution



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



~10 NON-RANDOM MODES

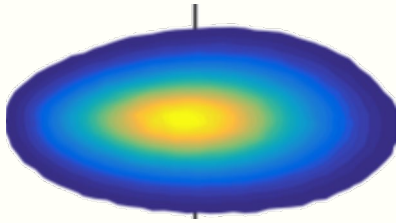


correlation modes are not the mechanical normal modes

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

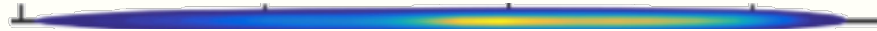
In genotype space

200-dimensional sphere



In phenotype space

~10-dimensional "disc"



Genotype

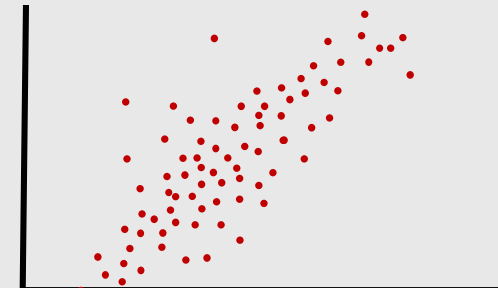
sequence space: *dimension ~ 1000*

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

dimensional
reduction

Phenotype

kinetic parameters *dimension= 2*

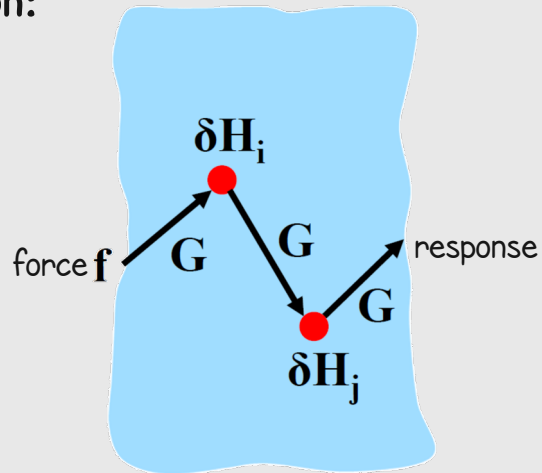


* Epistasis is collective interaction of mutations resulting in curvature of the fitness landscape

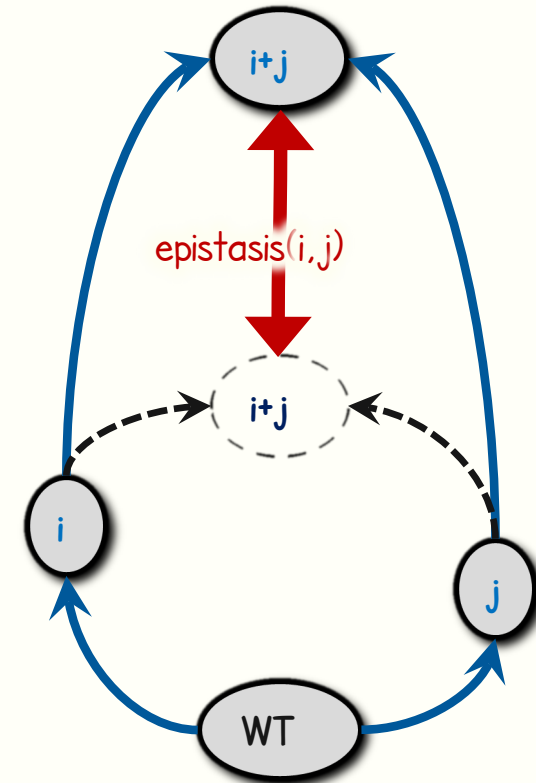
$$\text{epistasis}(i,j) = \Delta\text{fitness}(i+j) - [\Delta\text{fitness}(i) + \Delta\text{fitness}(j)]$$

Mechanical interpretation:

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

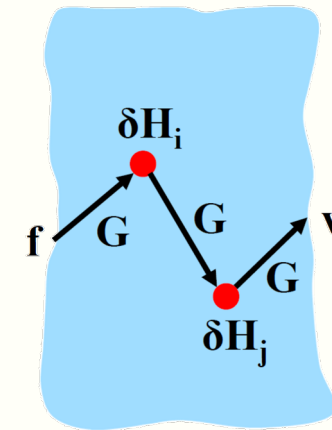
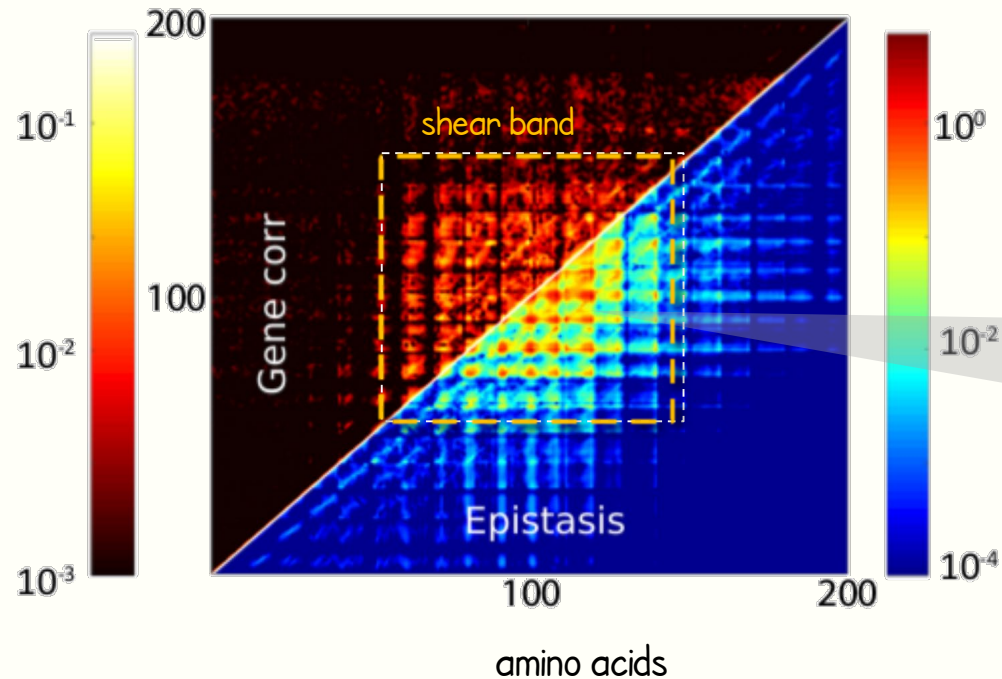


$$\text{epistasis}_{i,j} = \langle \text{force} | \mathbf{G} \delta H_i \mathbf{G} \delta H_j \mathbf{G} + \mathbf{G} \delta H_j \mathbf{G} \delta H_i \mathbf{G} \dots | \text{response} \rangle$$



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

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

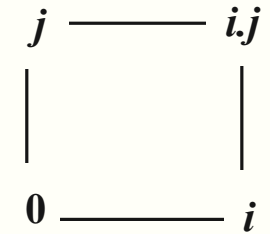


Epistasis and sequence correlation are strongest in high-shear region separating the rigid subdomains

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

* Higher-order epistatic interactions are also significant

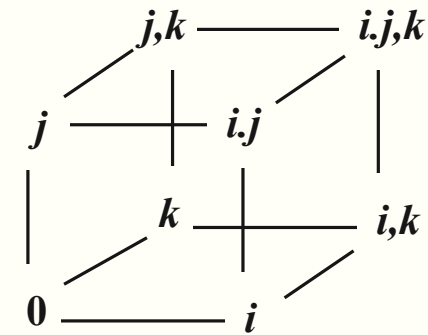
$$2^{\text{nd}} \text{ - order : } e_{i,j} = \delta F_{i,j} - (\delta F_i + \delta F_j)$$



$$3^{\text{rd}} \text{ - order : } e_{i,j,k} = \delta F_{i,j,k} - (\delta F_{i,j} + \delta F_{j,k} + \delta F_{i,k}) + \delta F_i + \delta F_j + \delta F_k$$

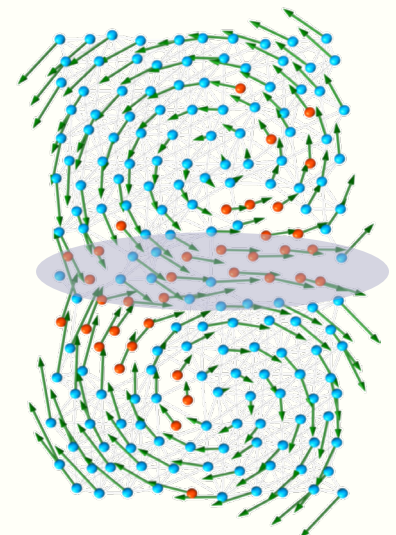
⋮

$$N^{\text{th}} \text{ - order : } e_{i_1, i_2, \dots, i_N} = \sum_{q=1}^N (-1)^{N-q} \sum_{i_1 < \dots < i_q} \delta F_{i_1 \dots i_q},$$

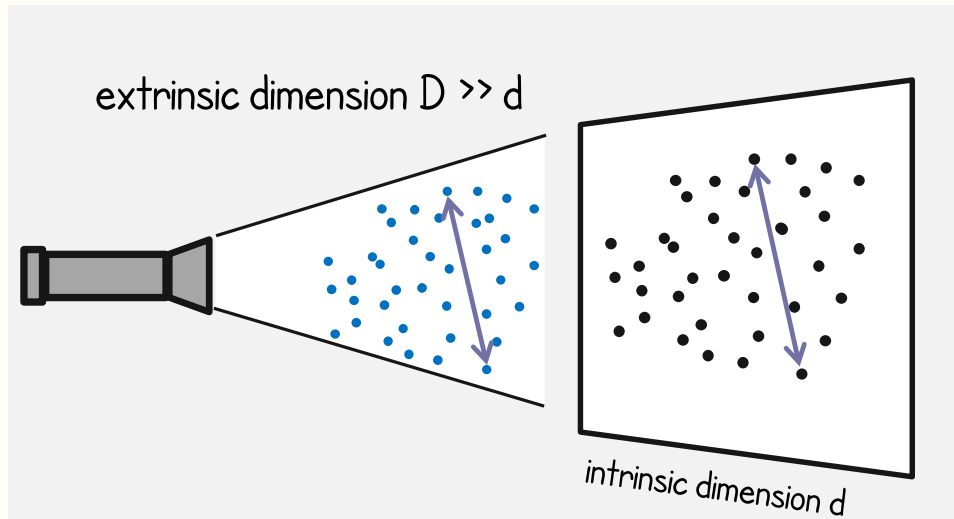


→ **curvatures** of Green functions on the hypercube that shape the genotype-to-phenotype map

All orders are significant: **Mechanically critical** regions are strongly coupled with **many-body epistatic interactions**.



Even more generally: Is there something special about dimensional reduction in biology?

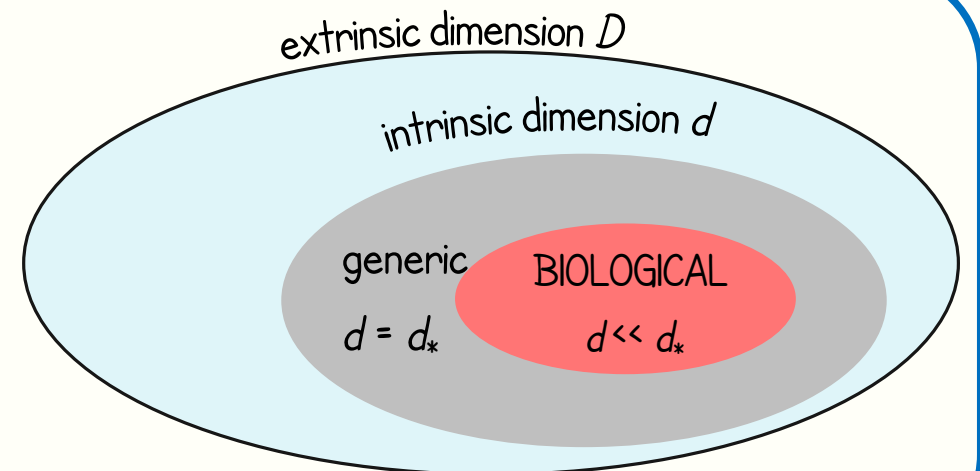


Intrinsic dimension: minimal dimension needed to project data with little deformation.

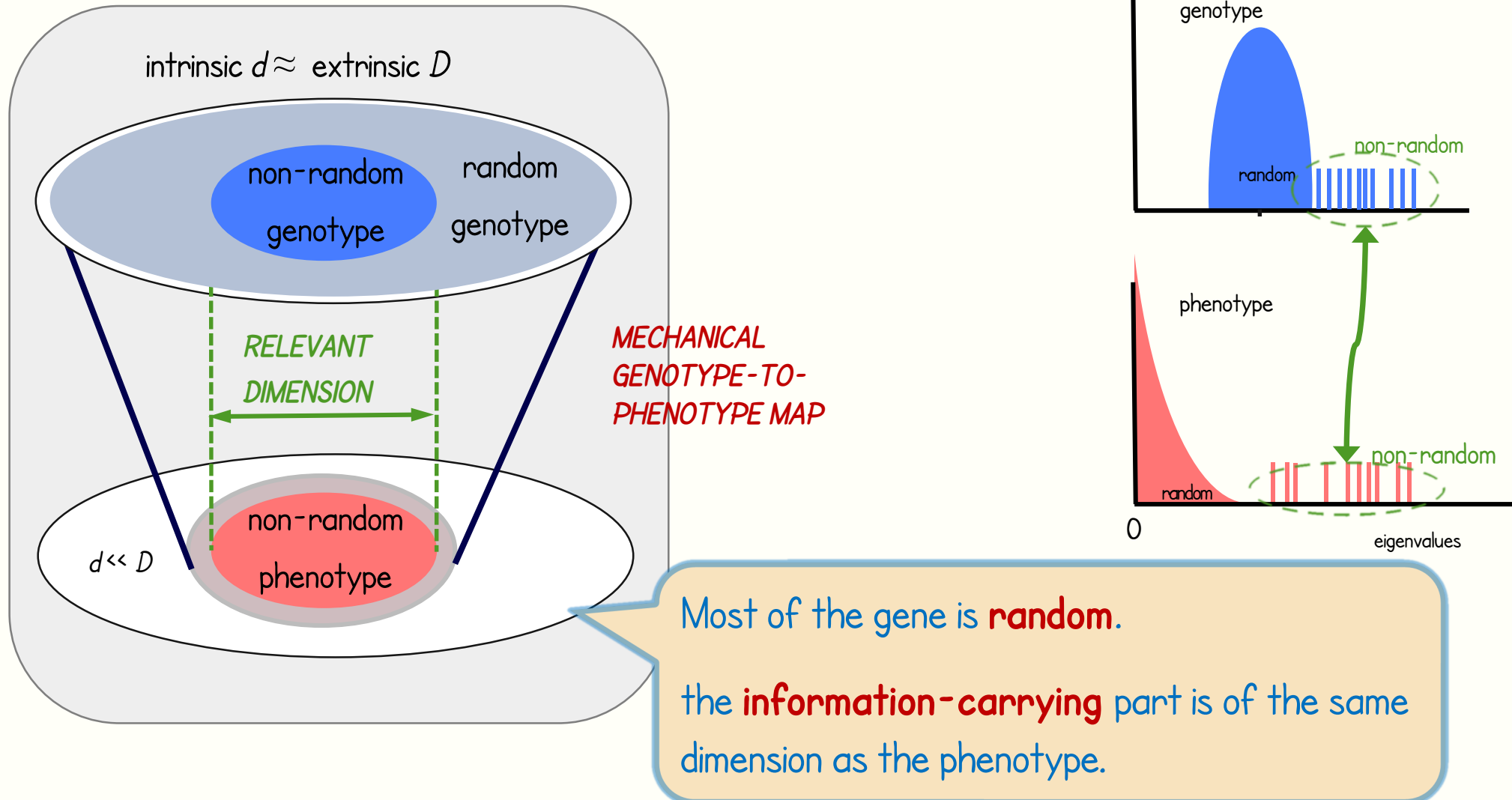
Johnson-Lindenstrauss compression: Any **generic data** can be reasonably compressed into a space of dimension

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

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

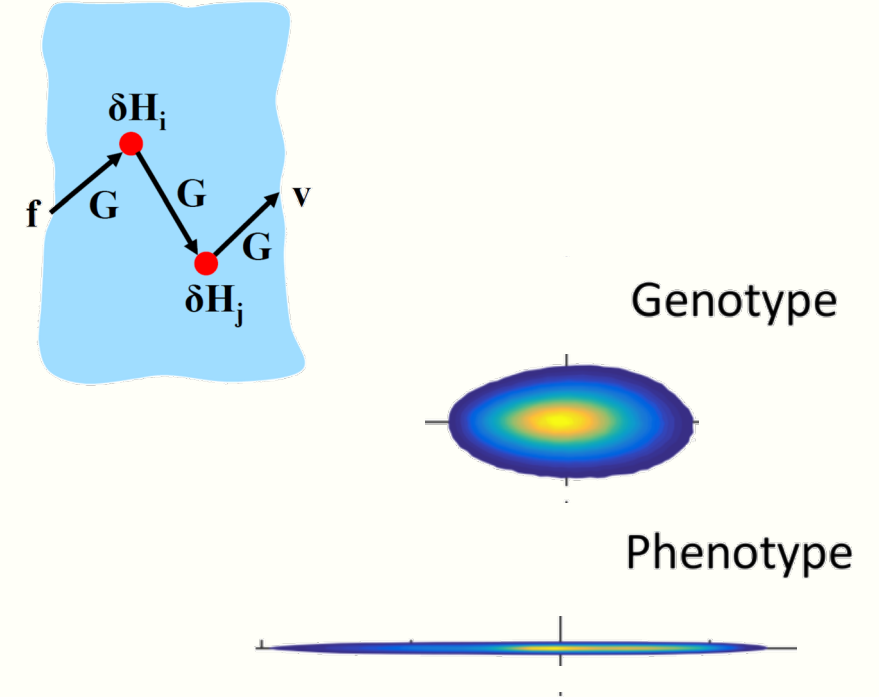


Hypothesis: Biological systems exhibit extreme dimensional reduction, far beyond the generic limit



SIMPLE THEORETICAL IDEAS – Mechanical view

- Mechanical perturbations reveal the **correspondence** between physics and evolution.
- **Green functions** are the main theoretical object.
- Mutations are localized **mechanical\structural defects**.
- scattering among mutations shapes
genotype-to-phenotype map and fitness landscape.
- **Dimensional reduction** originates
from physics-evolution correspondence.
- Hypothesis: dimensional reduction = general hallmark of biology
(applicable to other strongly-coupled systems?)



Dutta Eckmann Libchaber TT, PNAS 2018
Eckmann Libchaber TT, PRX 2017
Eckmann Rougemont TT, Rev Mod Phys 2019
Eckmann TT , Bioessays, 2021

2. USING DEEP LEARNING TO UNDERSTAND THE LOGIC OF PROTEIN MACHINES

(looking at Green functions of real proteins)

Can AlphaFold (AF) predict single mutation effects?

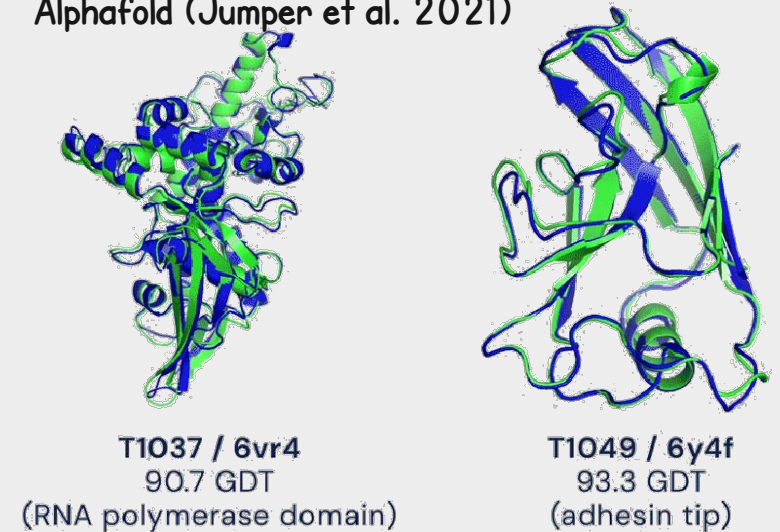
PROBLEM

- ❑ AF not trained for single mutations.
- ❑ AF not evaluated on predicting motion u , not deformation ∇u

APPROACH

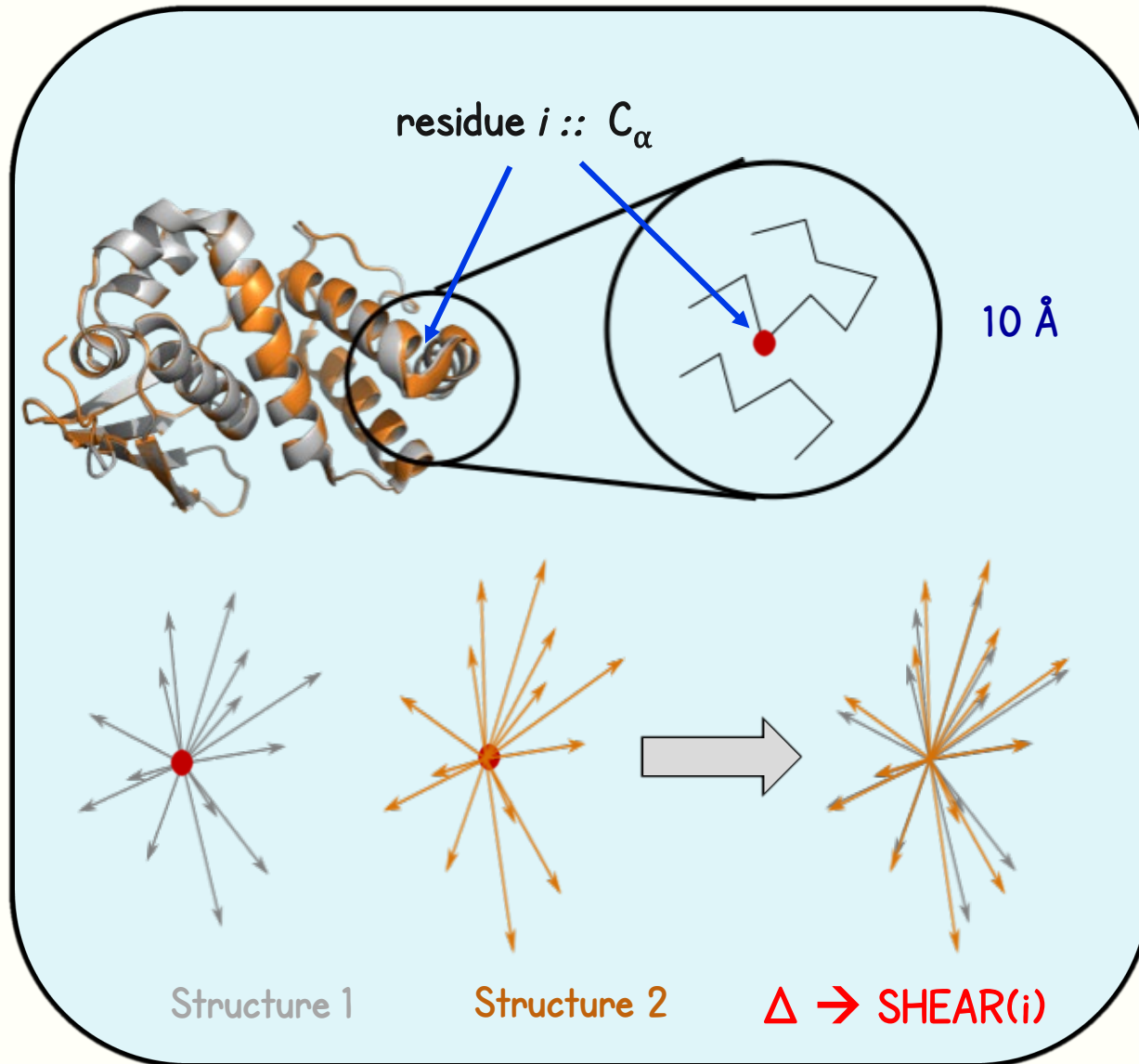
- ❑ **Direct:** structural comparison AF vs PDB.
- ❑ **Indirect:** predicting high-through **phenotype** measurements.
- ❑ Structure-phenotype correspondence?

AlphaFold (Jumper et al. 2021)



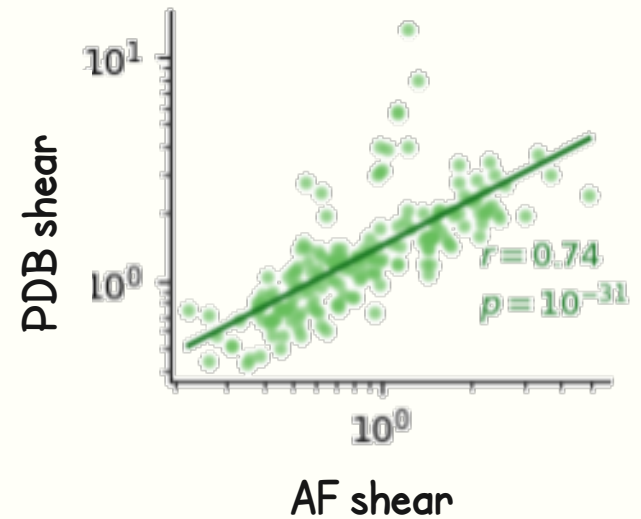
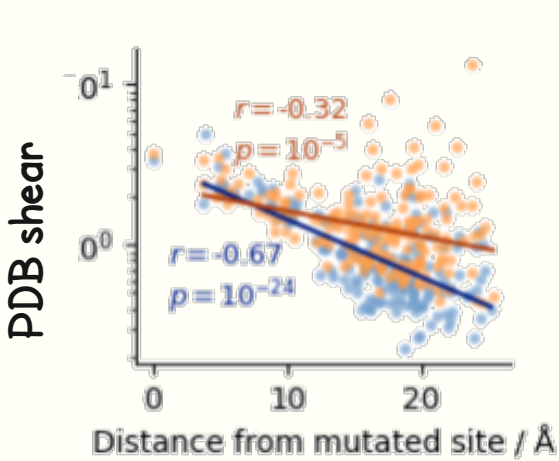
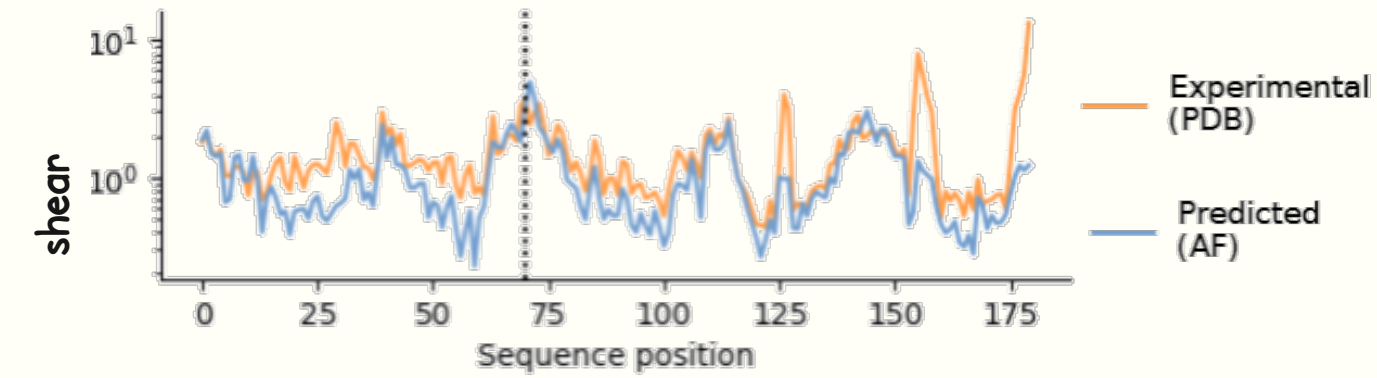
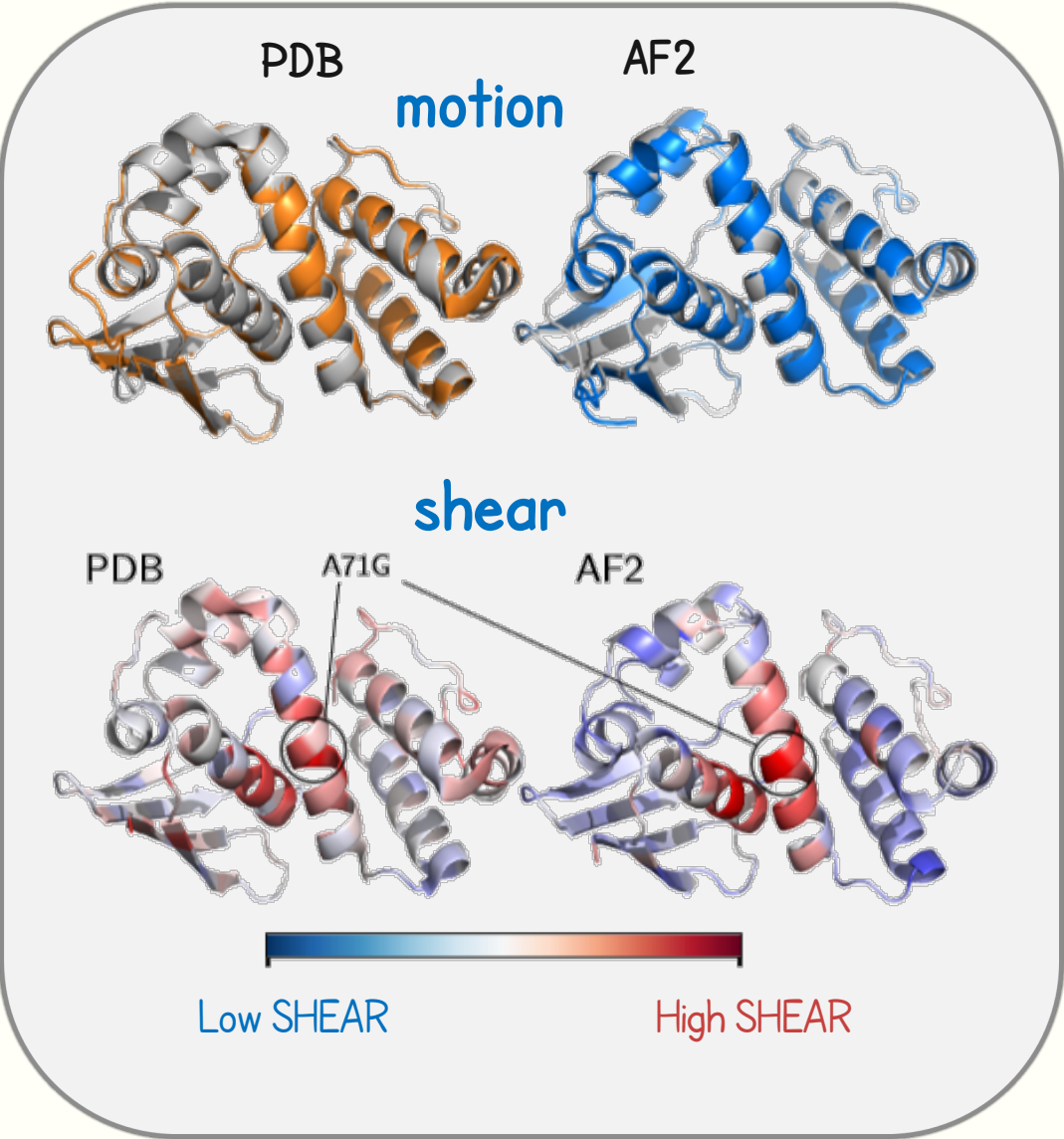
● Experimental result
● Computational prediction

Examination of localized structural deformation in experiment (PDB) and AI-prediction (AF)



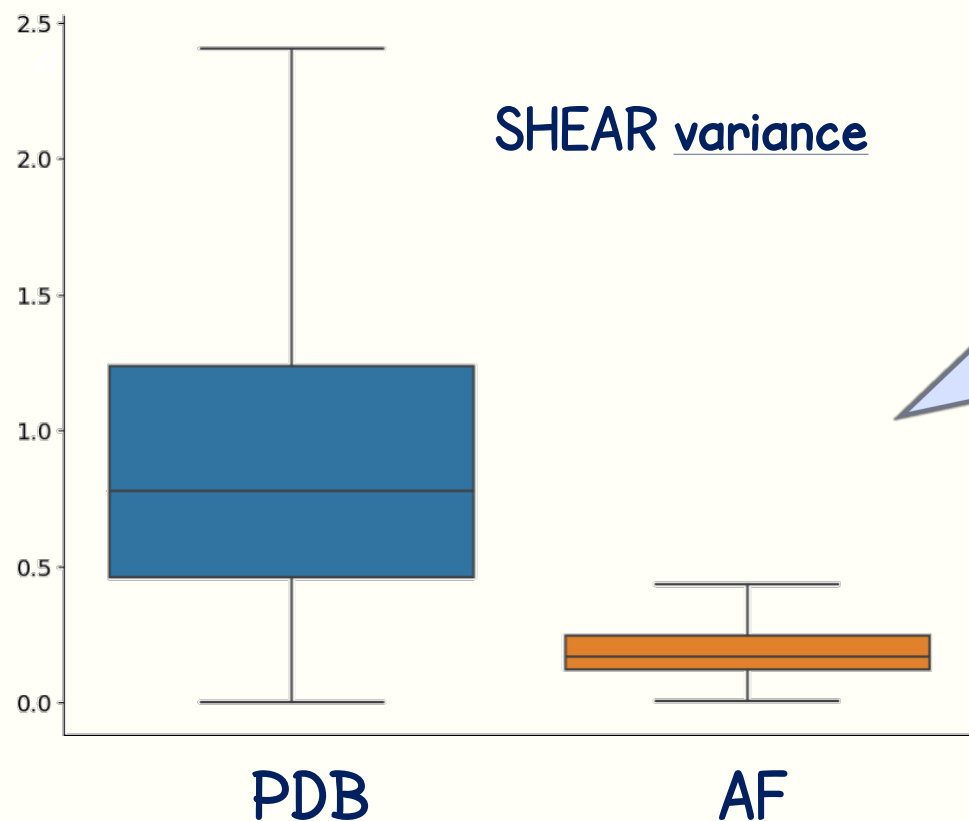
- About 5,000 pairs of sequences that differ in 1-3 positions.
- Compare SHEAR in PDB vs. AF

Comparing local deformations (shear) in a case study: H-NOX

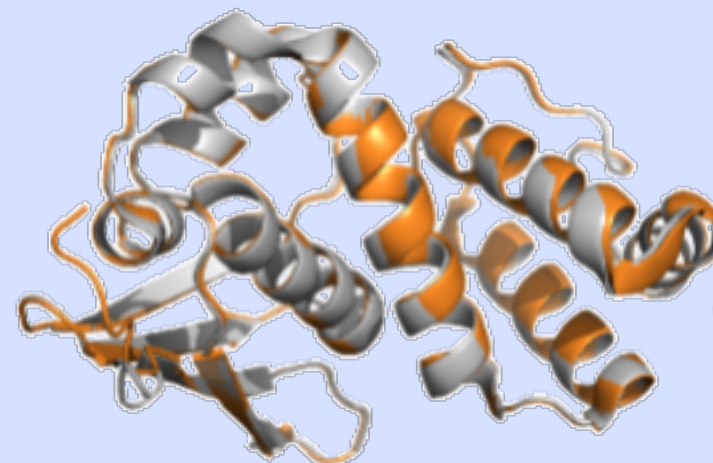


H-NOX = HEME-NITRIC OXIDE/OXYGEN BINDING

PDB repeat measurements are more variable than AlphaFold's



Find sequences with more than one PDB structure

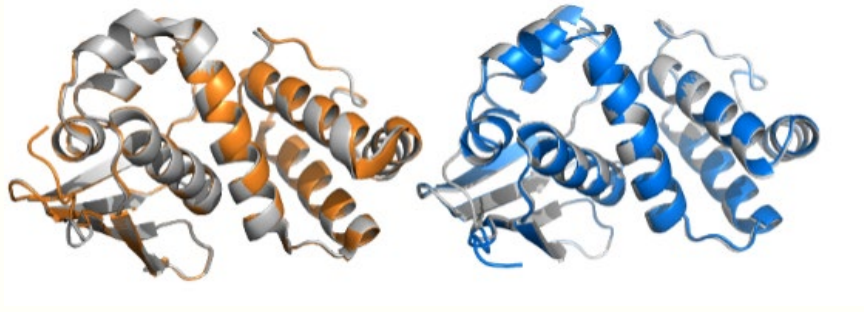


Calculate SHEAR for all residues
for 6,000 pairs of structures

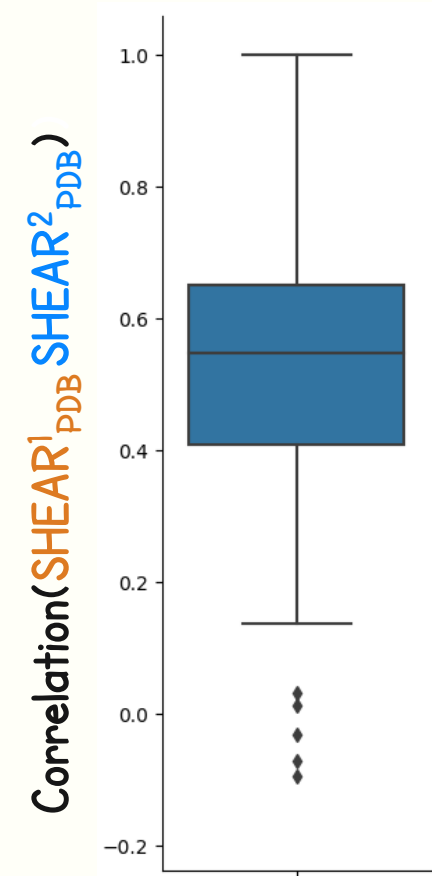
What is the (physical) origin of fluctuation? Is the comparison valid?

Fluctuations are correlated across repeat PDB measurements

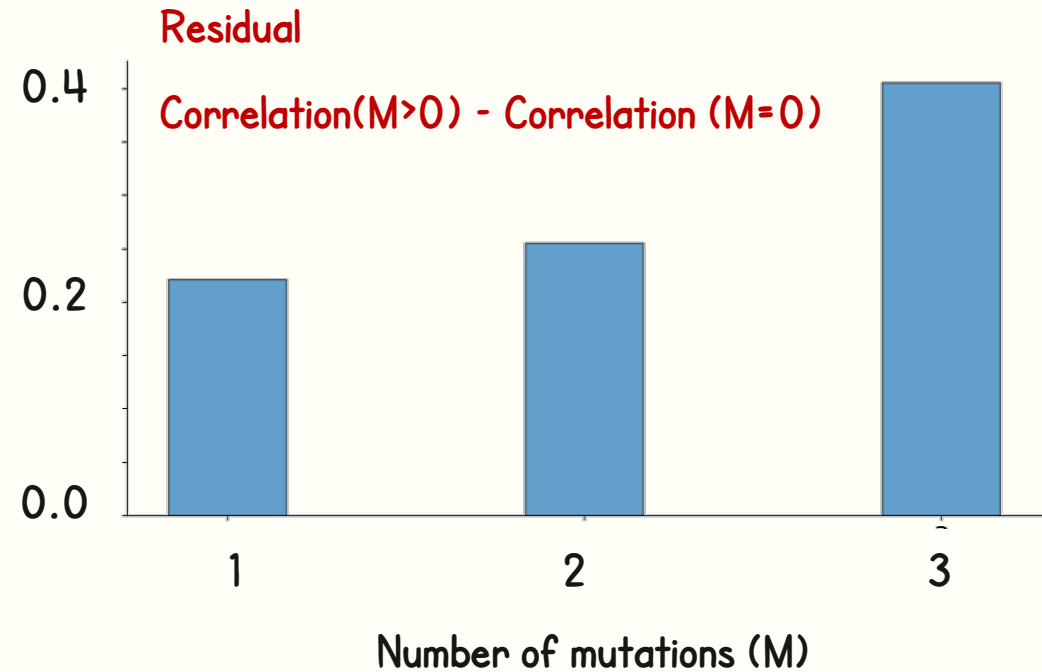
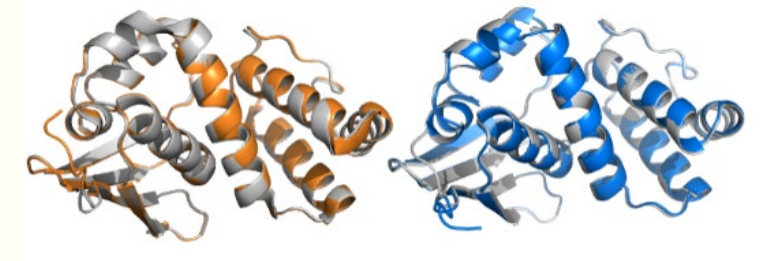
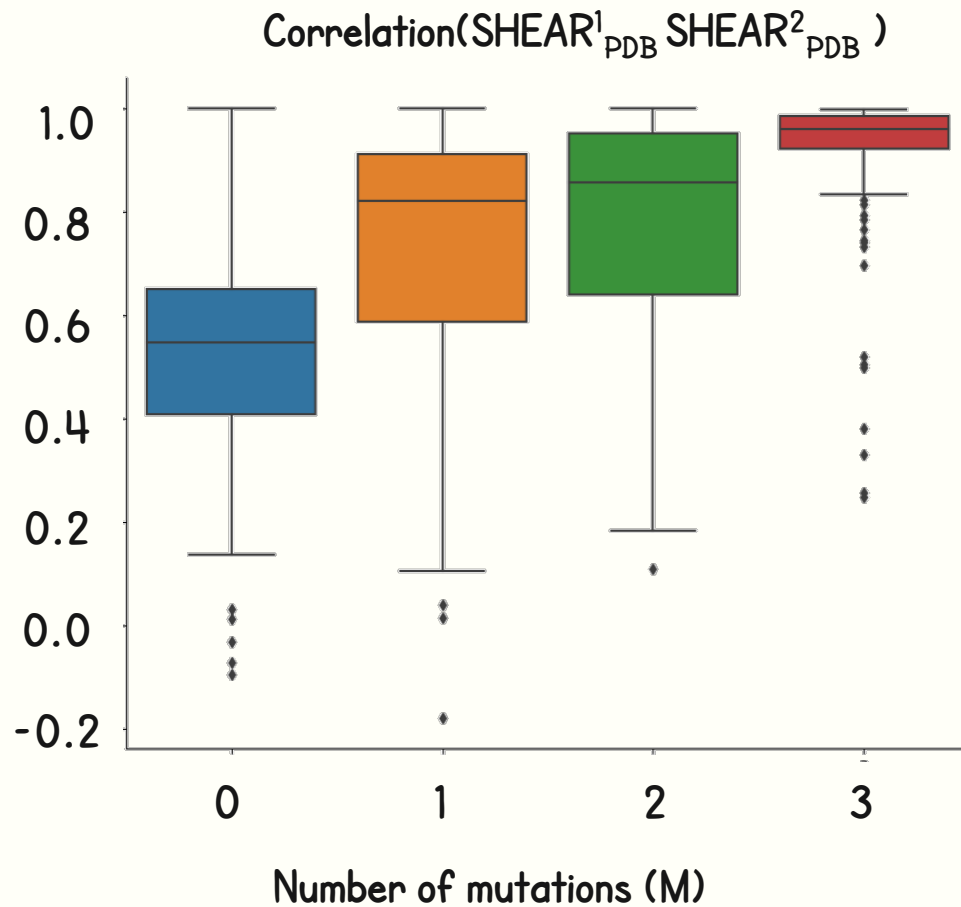
Find sequences with more than two PDB structures



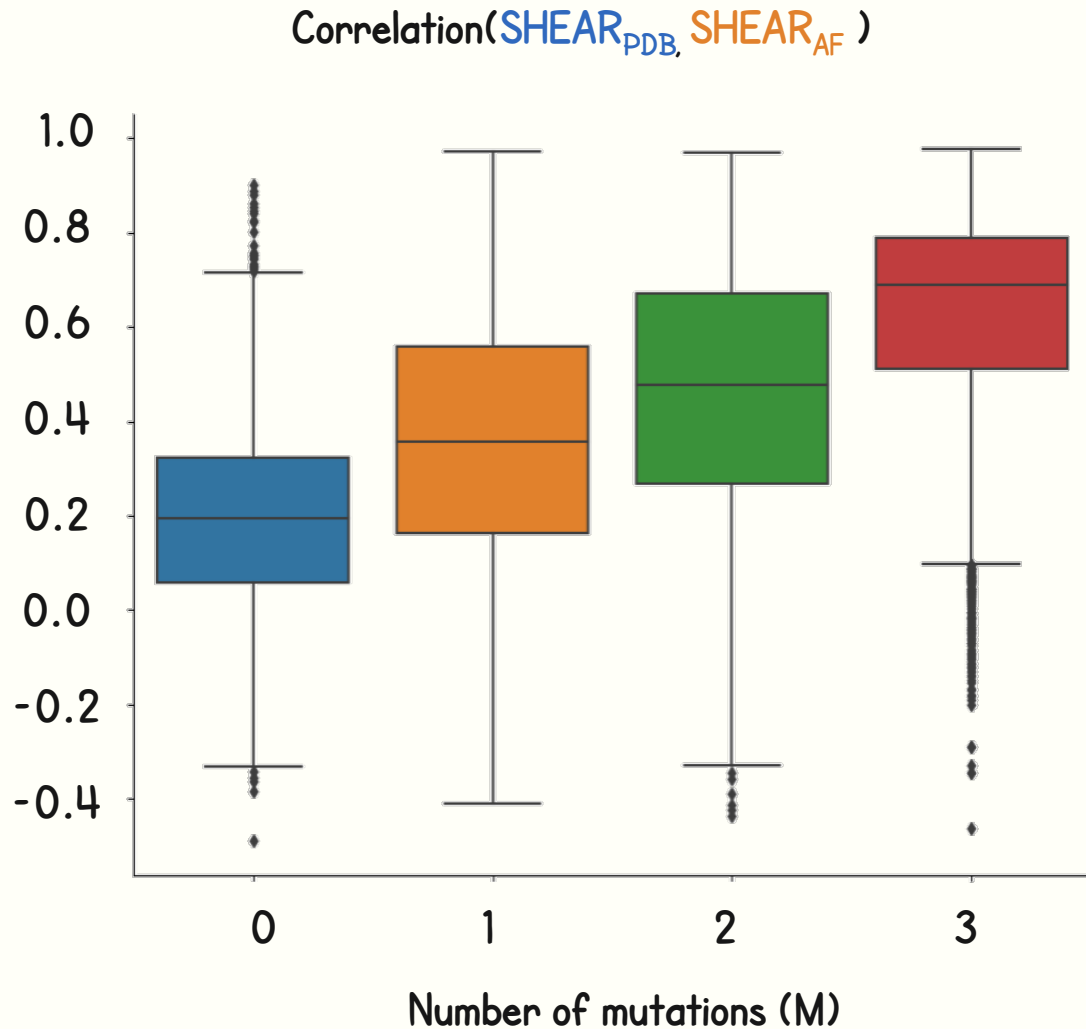
Calculate $\text{Correlation}(\text{SHEAR}^1_{\text{PDB}}, \text{SHEAR}^2_{\text{PDB}})$
for 2,000 groups of structures



* Mutation-induced SHEAR can be measured from PDB despite fluctuations

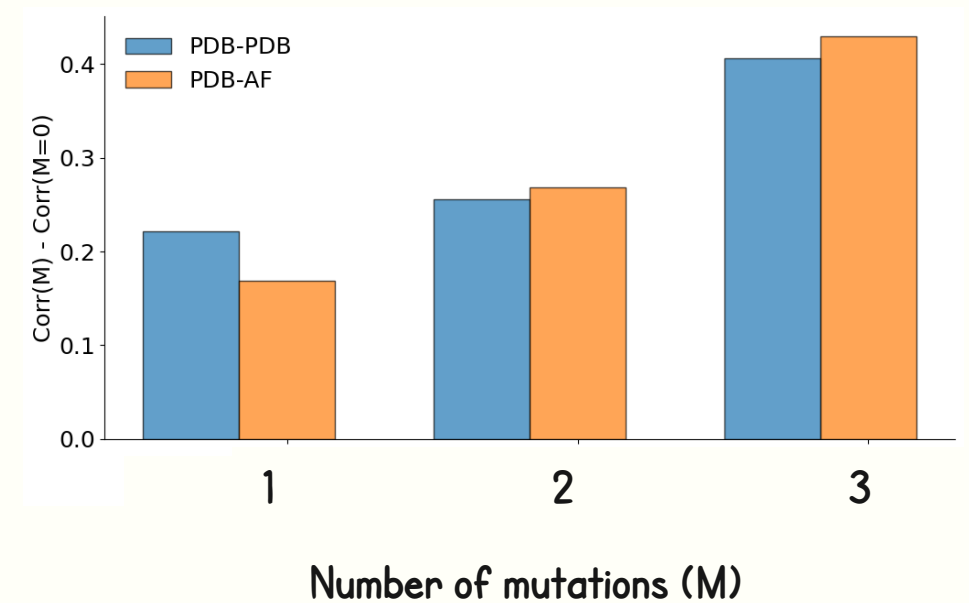


Alphafold captures ~whole experimental signal above noise



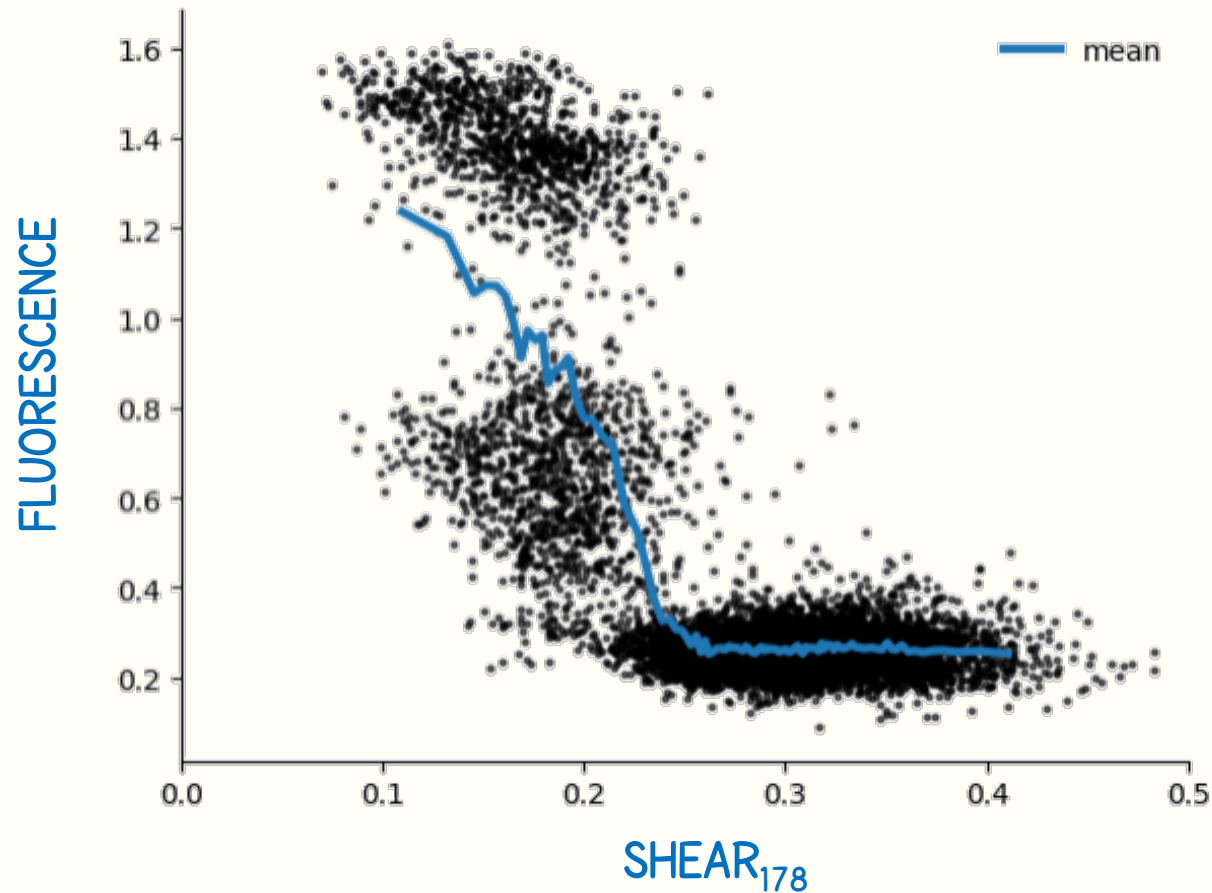
Fluctuations in experimental measurements

hinder evaluation of mutation effects

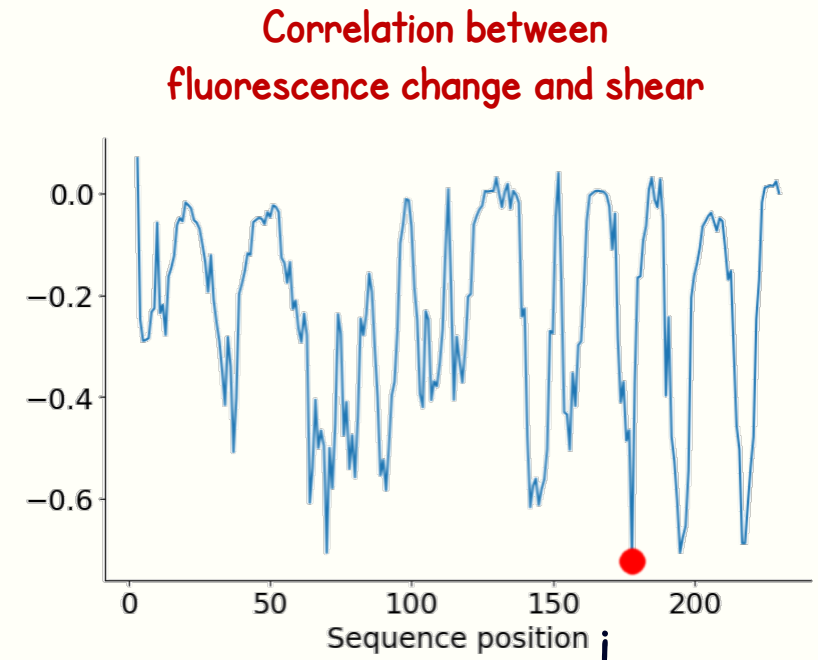


But AF predictions of mutation effects are (on average)
as reliable as repeat exp. measurements

Shear is a binary predictor of fluorescence level in BFP



...on average, AF can predict mutation effects



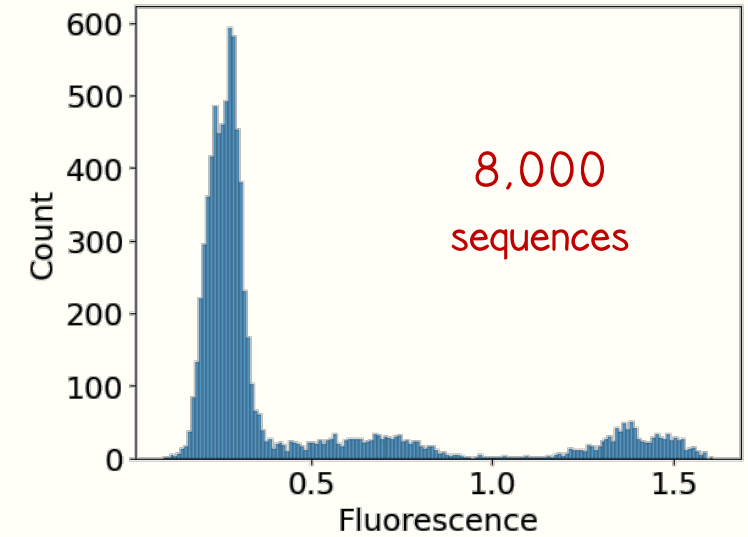
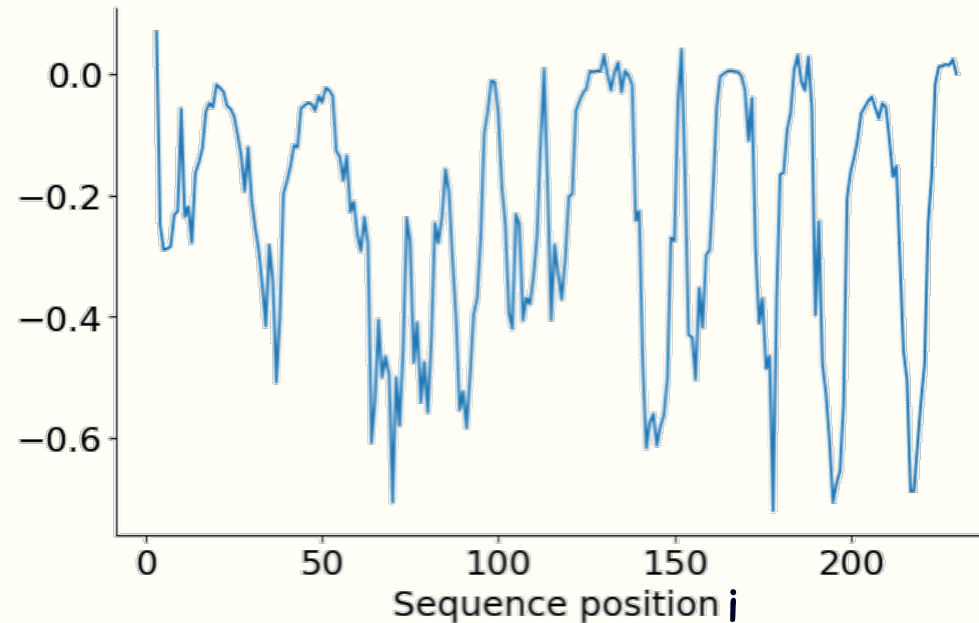
Also in other high-throughput
datasets: GFP, PafA (kinase)

Function: Structural change correlates with phenotypic changes in BFP

Blue fluorescent protein (BFP)



Correlation between fluorescence change and $SHEAR_i$



2.

- Alphafold is on average good enough to predict shear
- Alphafold's shear is linked to function (at least in some cases)

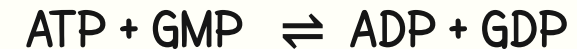
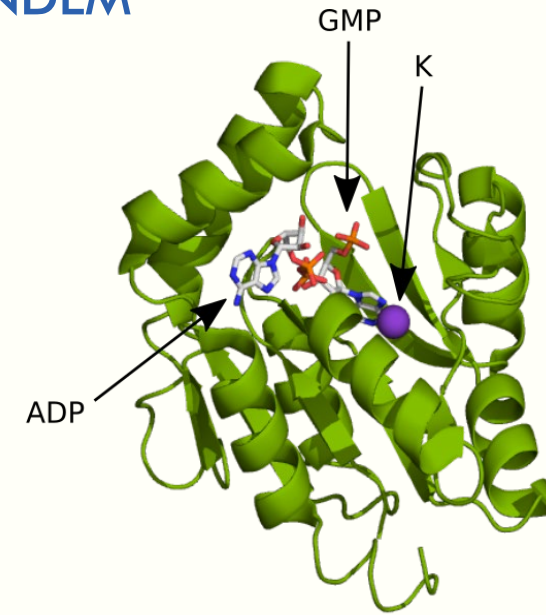
3. THEORY PREDICTIONS TESTED BY MEASURING IN TANDEM

SHEAR(PDB & AF) +

NANORHEOLOGY +

BIOCHEMICAL FUNCTION

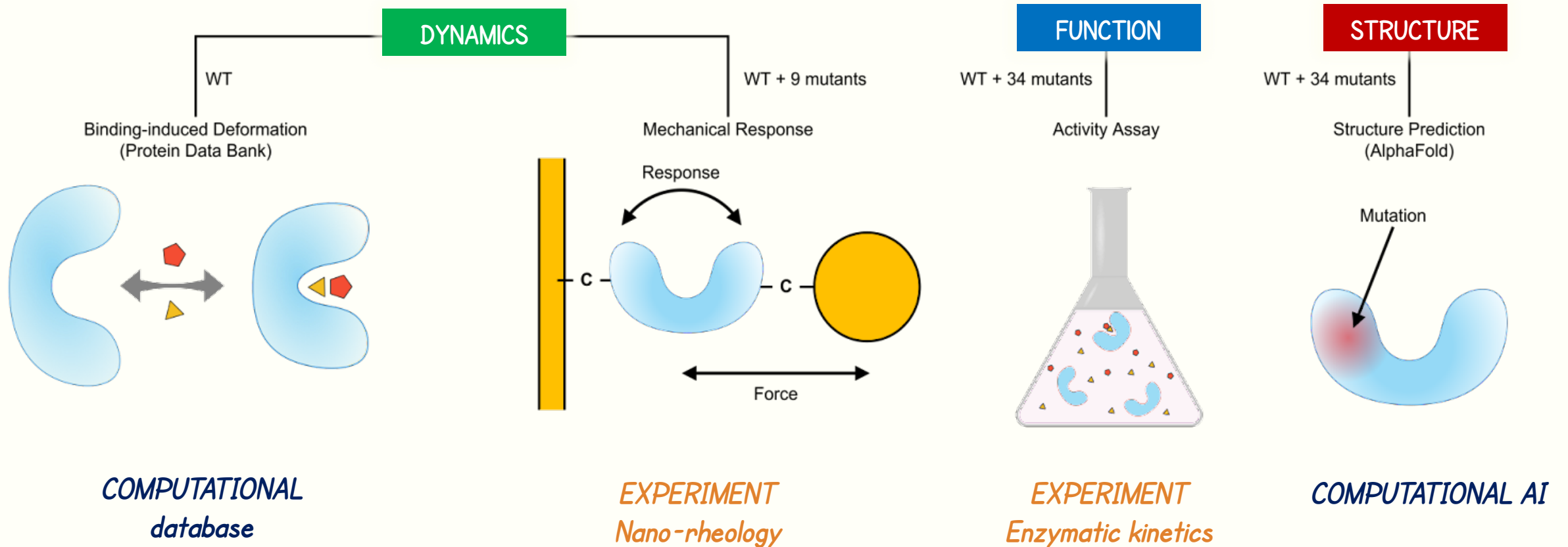
IN GUNAYLATE KINASE



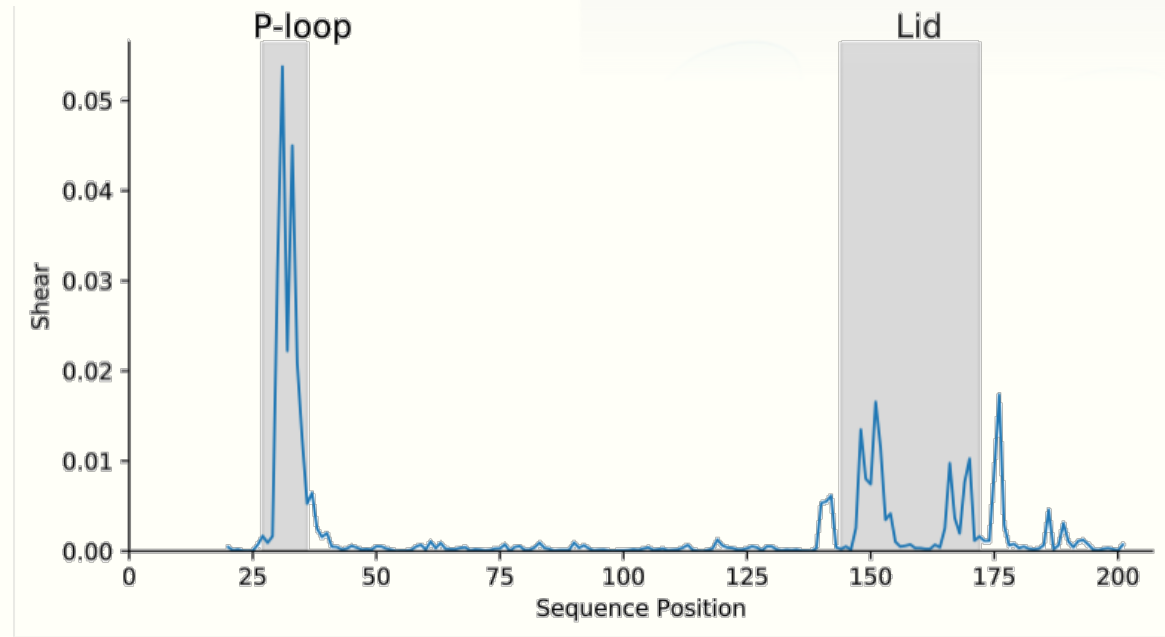
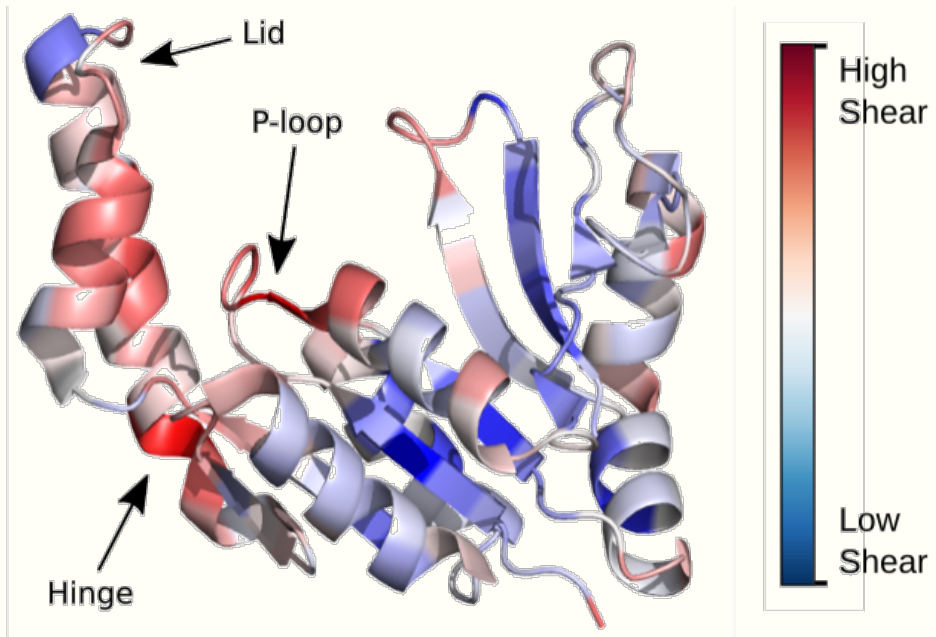
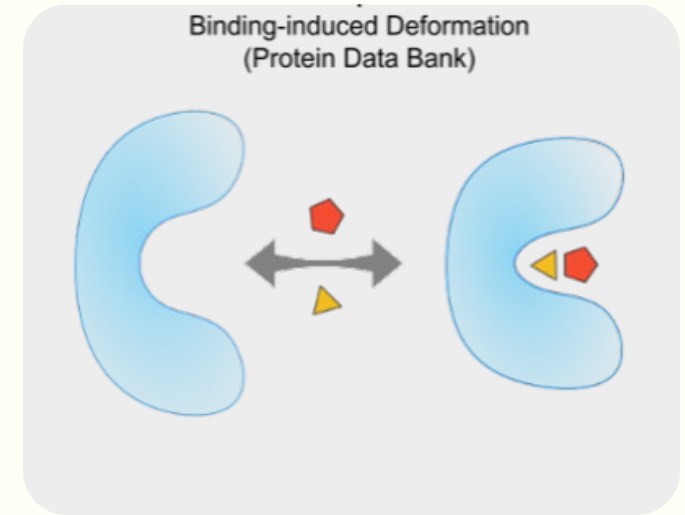
High-shear regions in guanylate kinase impact viscoelastic mechanics and enzymatic function

Eyal Weinreb, John M. McBride, Marta Siek, Jacques Rougemont, Renaud Renault, Yoav Peleg, Tamar Unger, Shira Albeck, Joel L. Sussman, Bartosz A. Grzybowski, Giovanni Zocchi, Jean-Pierre Eckmann, Elisha Moses, TT (submitted)

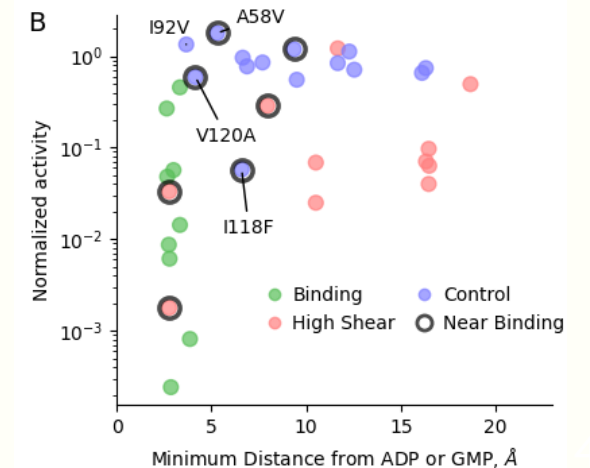
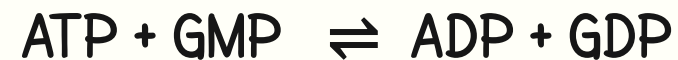
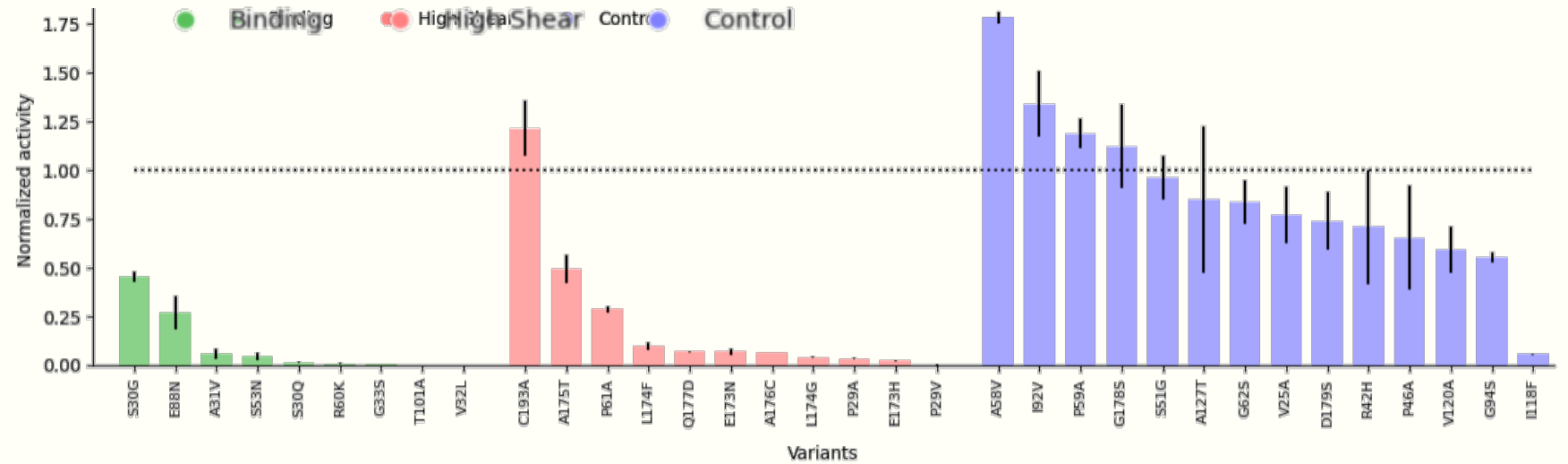
Measurements of Guanylate Kinase mutants allow correlating dynamics – function – structure



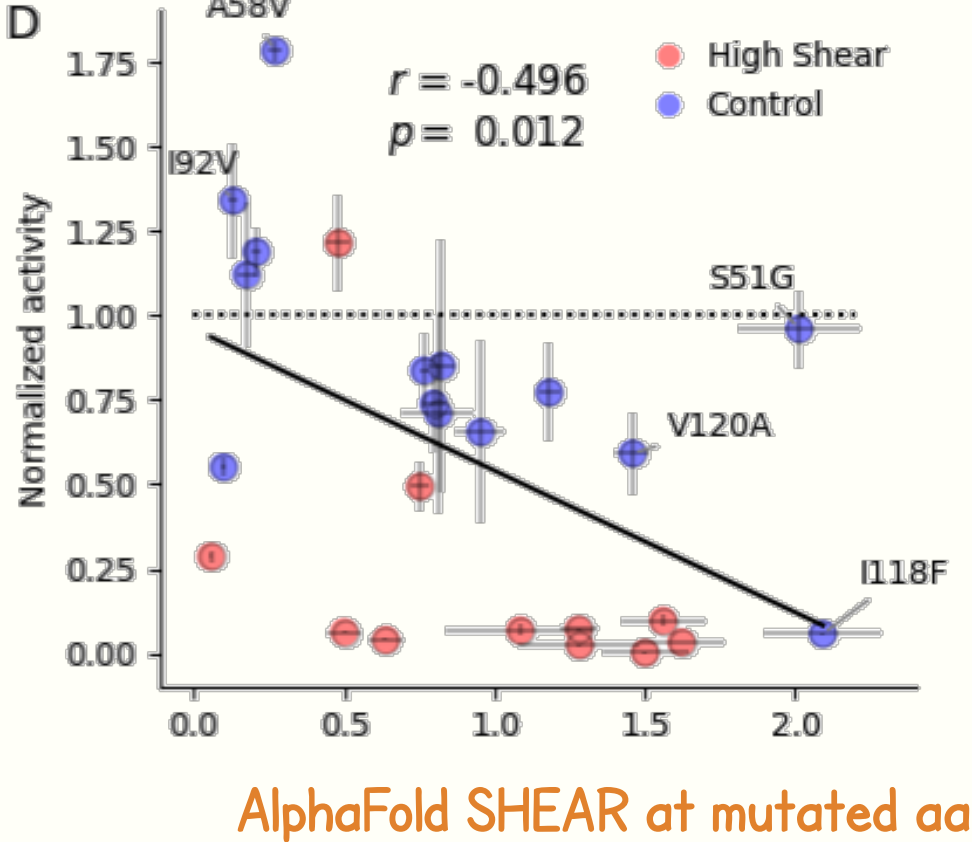
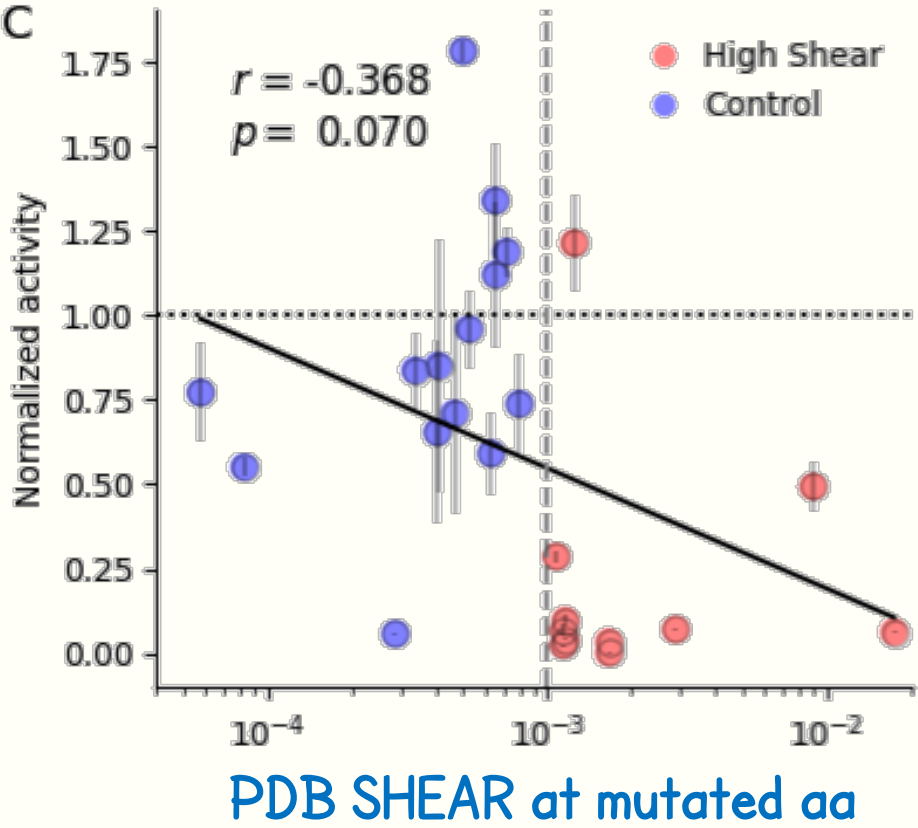
We calculate the shear and introduce 34 mutations in high-shear, binding, and control regions



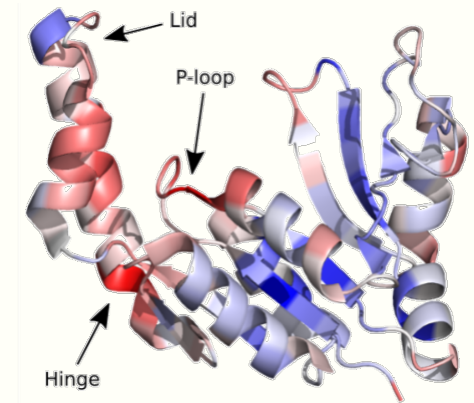
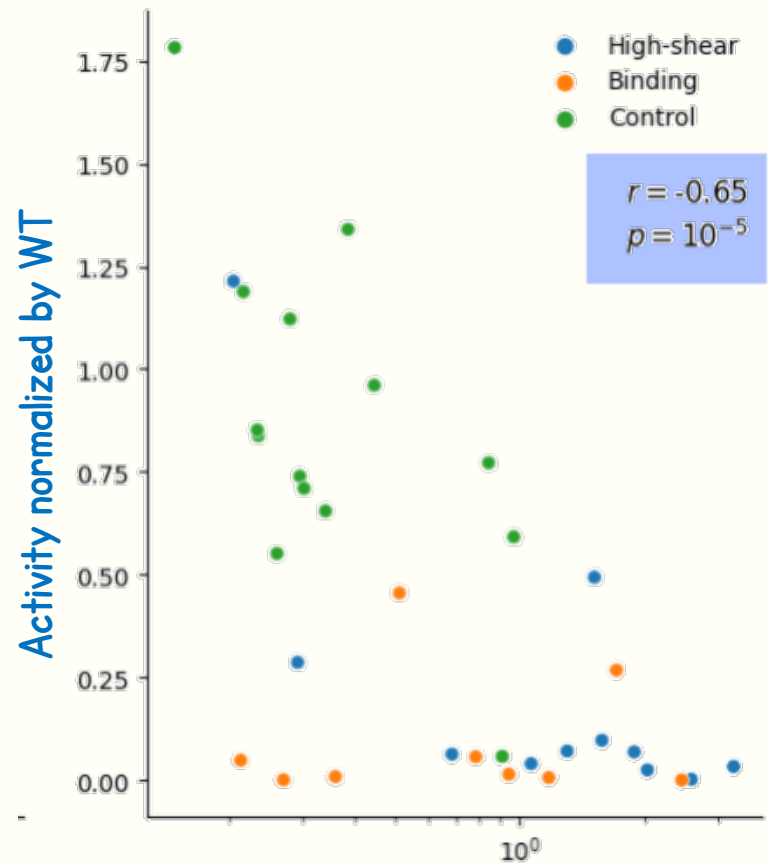
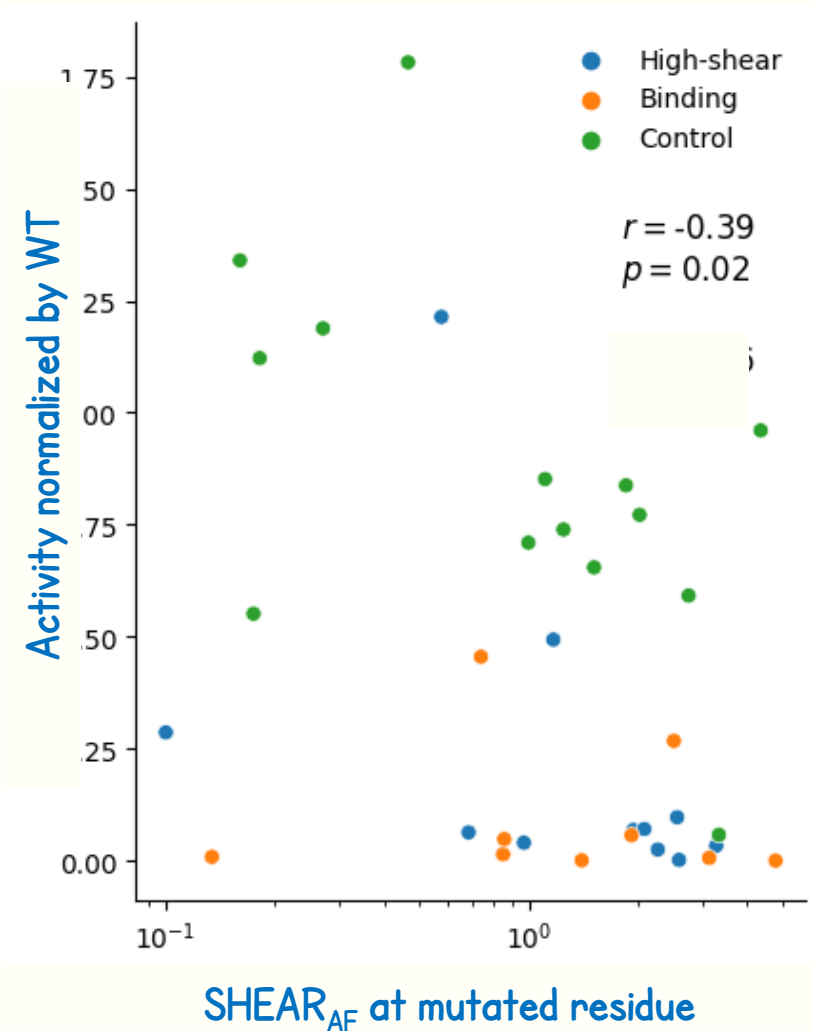
The effect of mutation on enzymatic activity correlates with region



Effect on enzymatic function correlates with SHEAR – of both PDB and AF

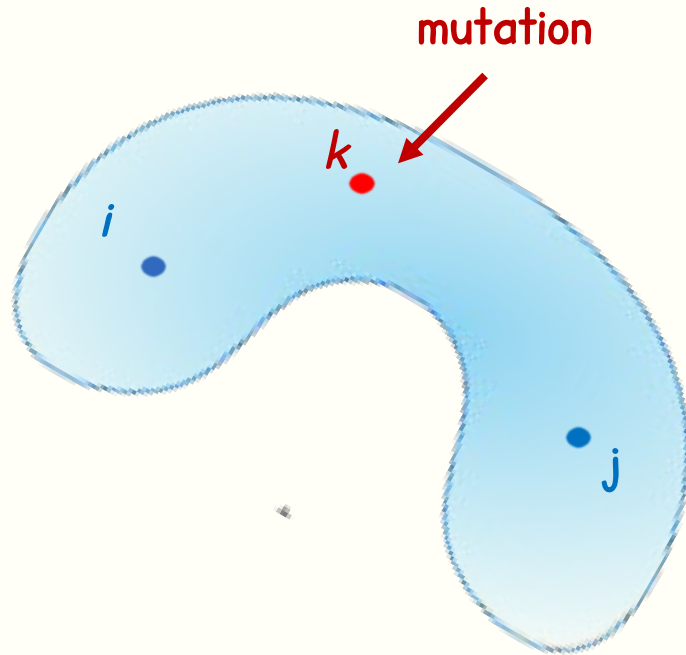


Effect on activity shows strong allosteric effect

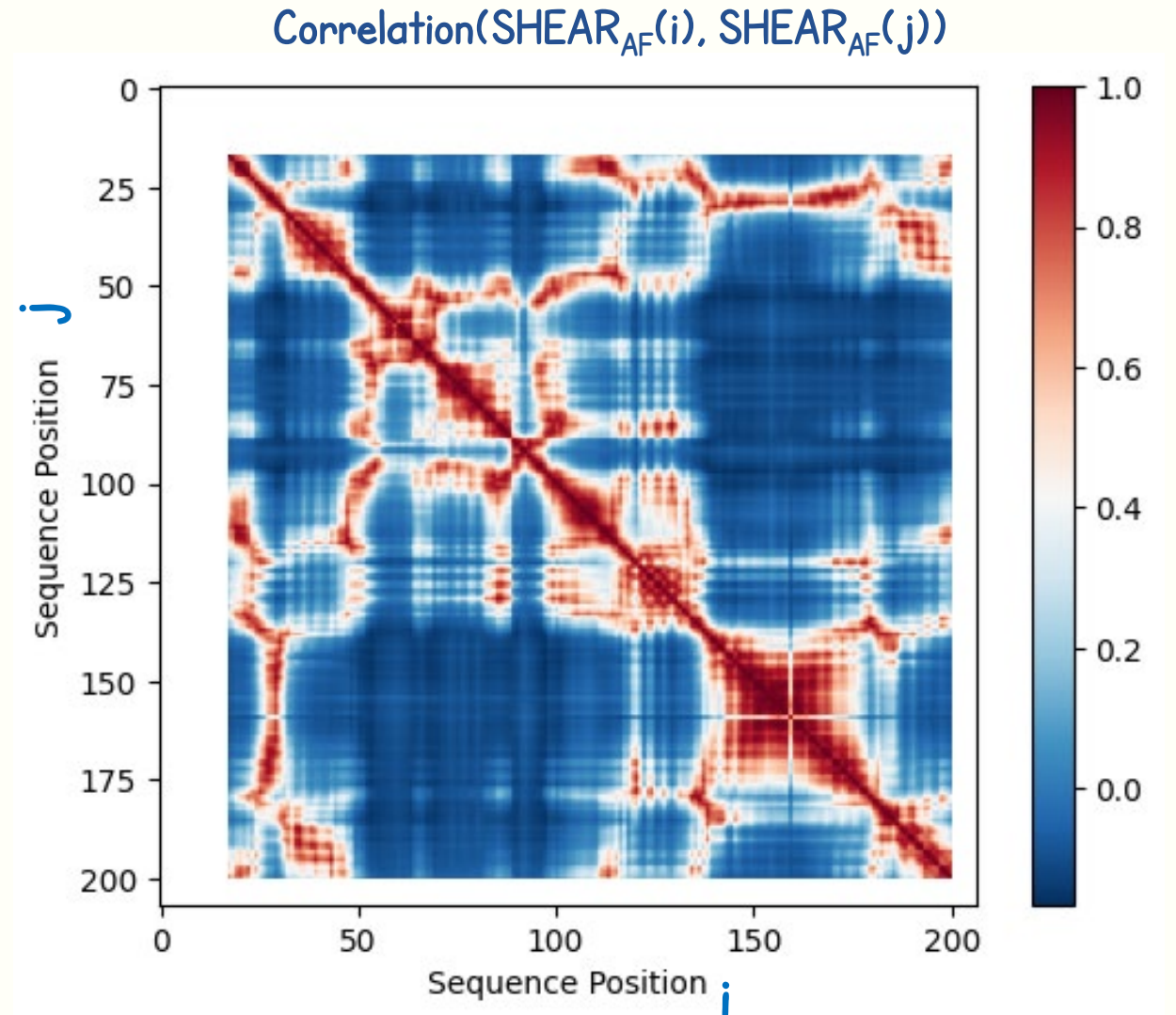


SHEAR_{AF} induced by mutation at $i = 29$ (binding)

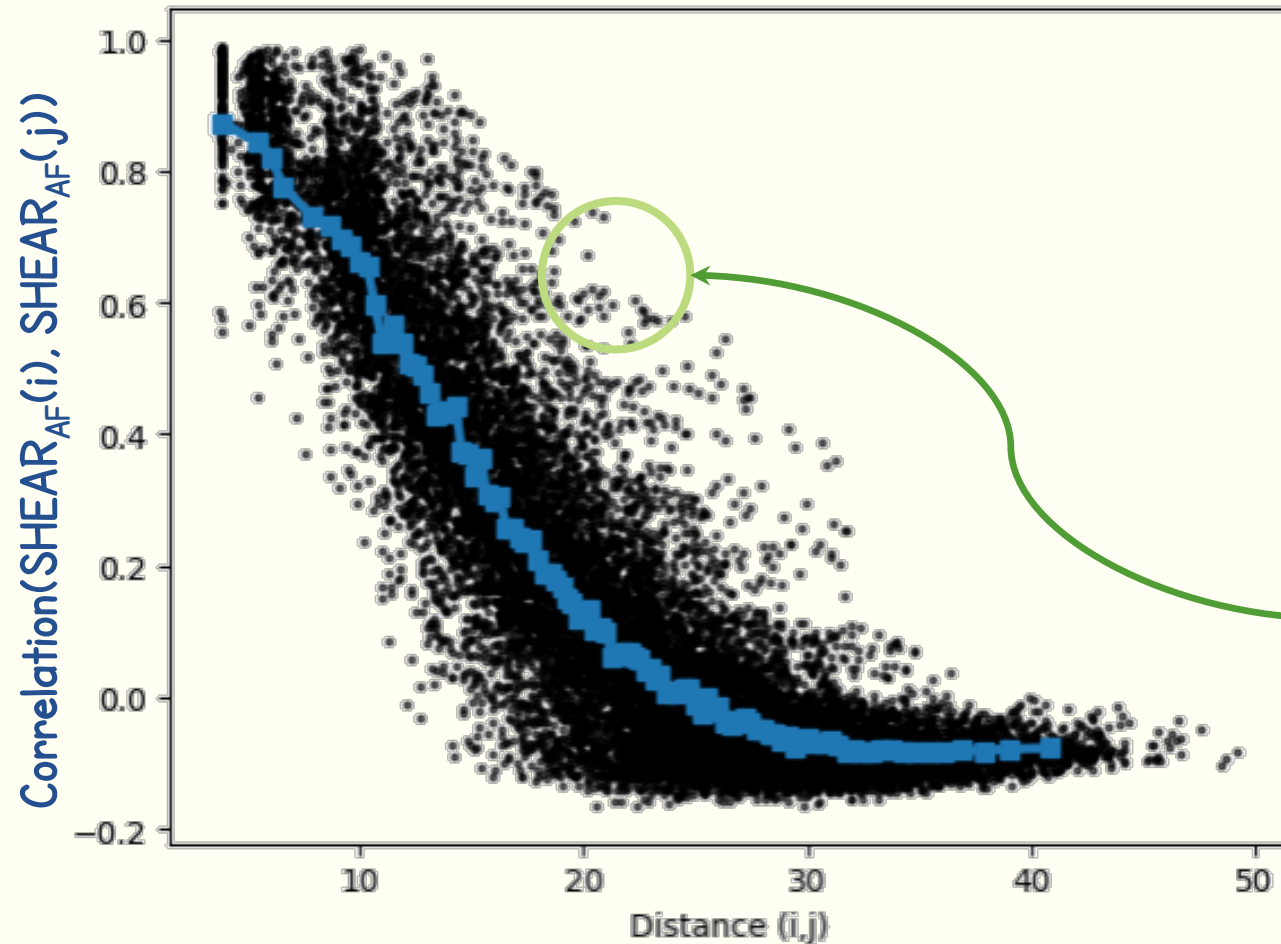
AF predicts non-trivial, **long-range** allosteric interaction



- Sample many mutants (k)
- Get $\text{Correlation}(\text{SHEAR}_{\text{AF}}(i), \text{SHEAR}_{\text{AF}}(j))$



Alphafold predicts non-trivial long-range allostery



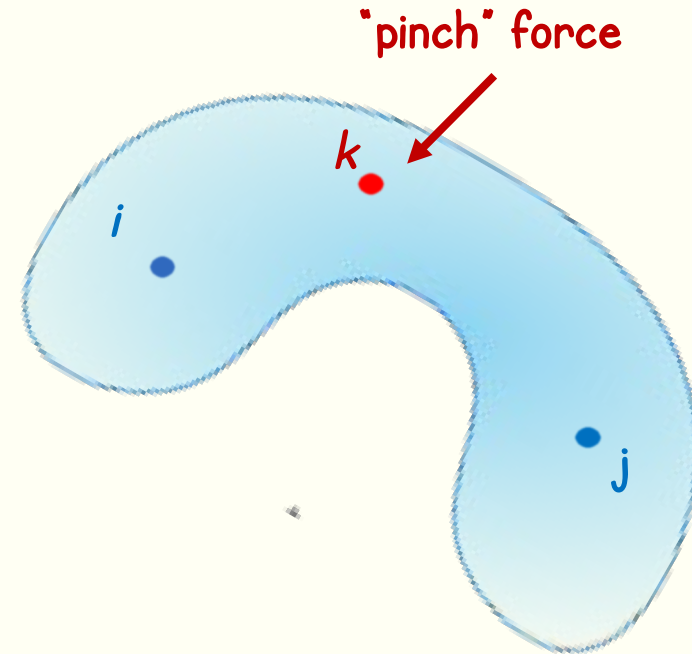
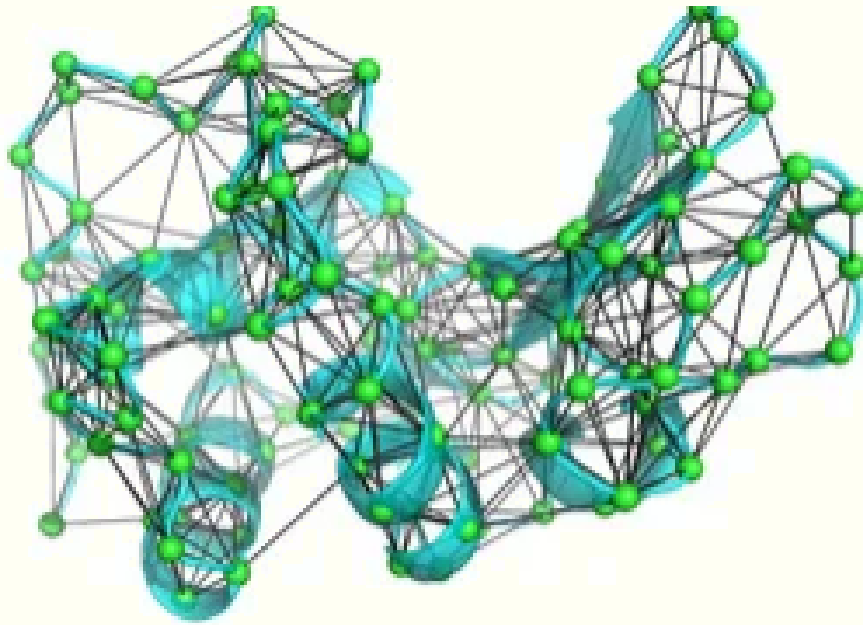
On average: Allosteric coupling decays with distance

But also: non-trivial long-range allosteric coupling

AlphaFold predicts non-trivial long-range allostery in guanylate kinase

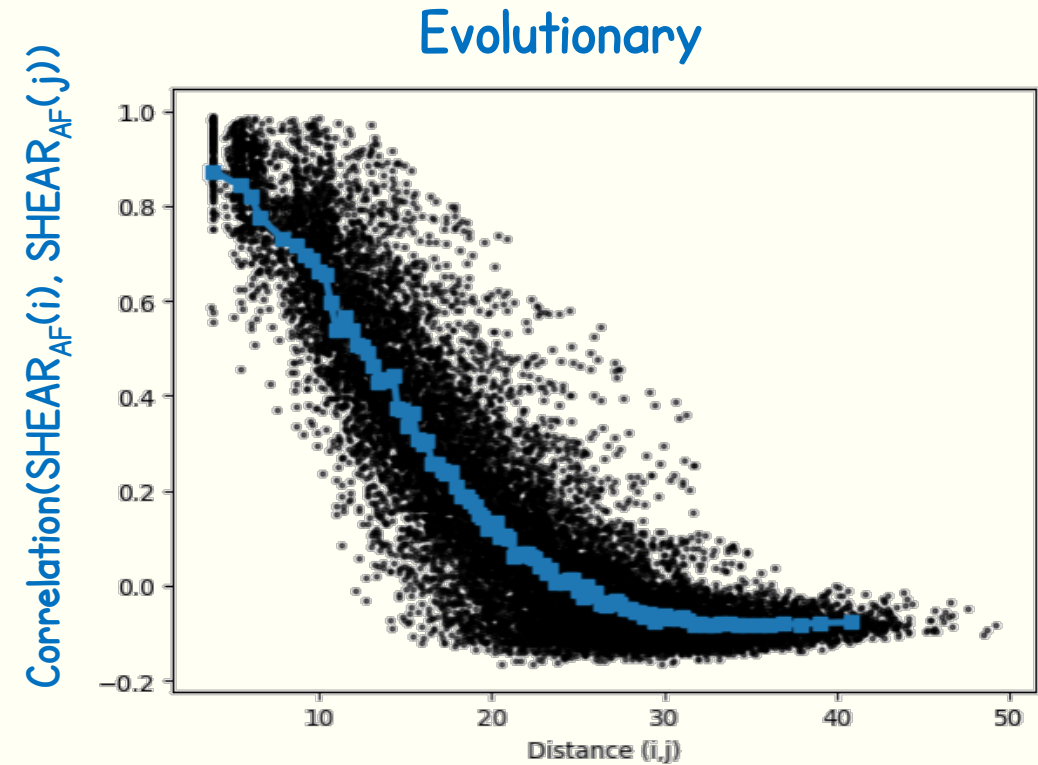
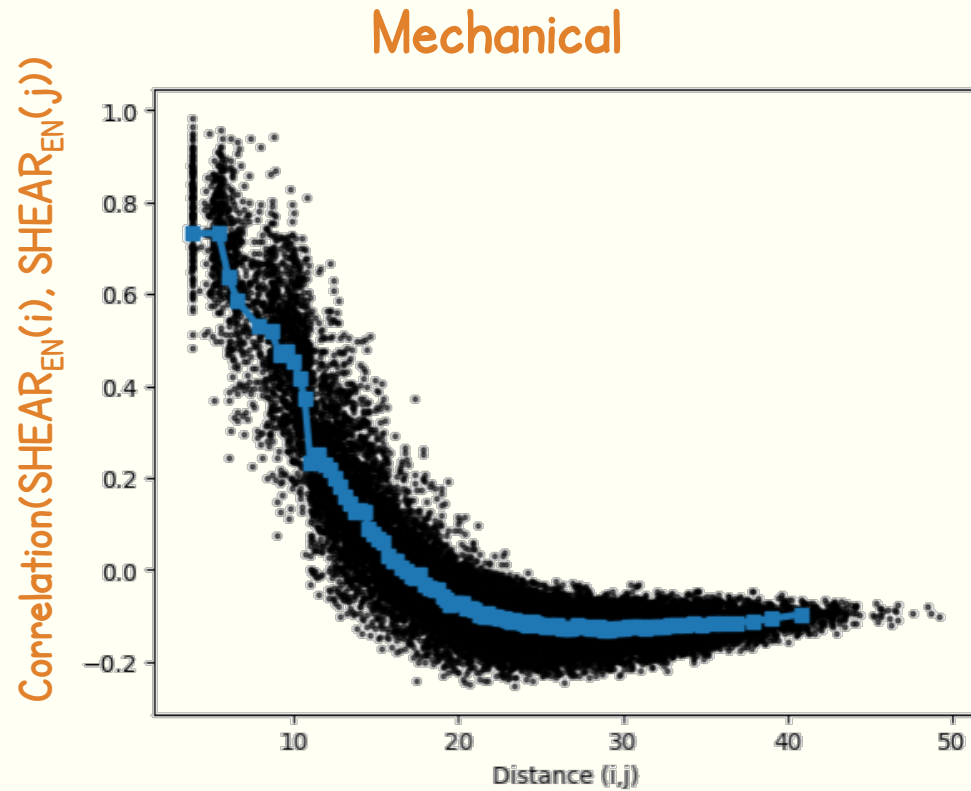


Mechanical allostery can be calculated from elastic network models



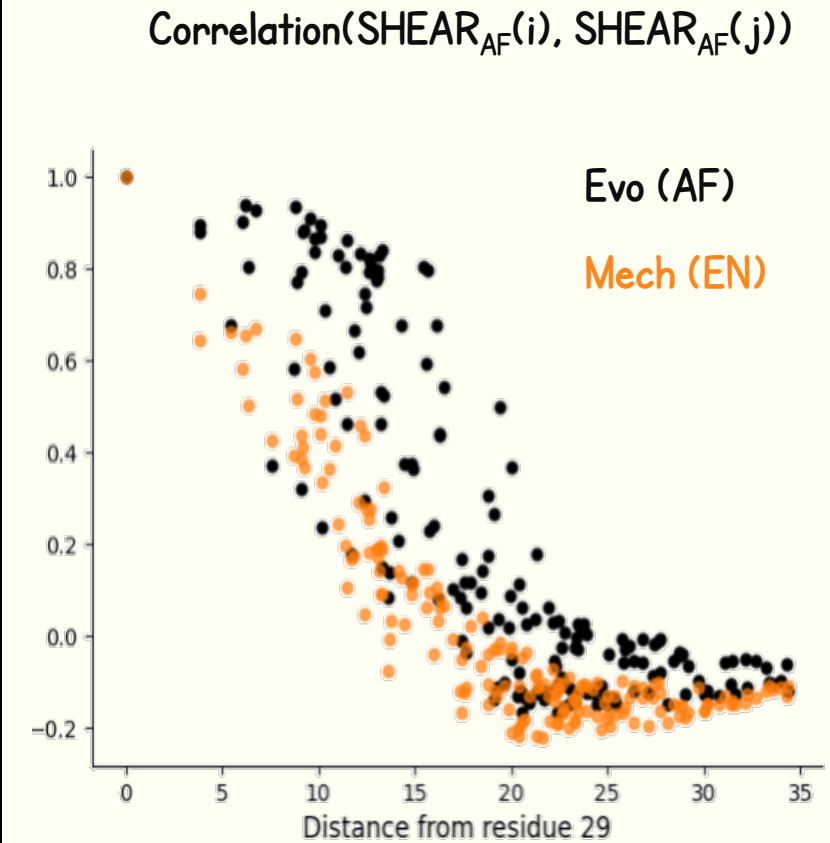
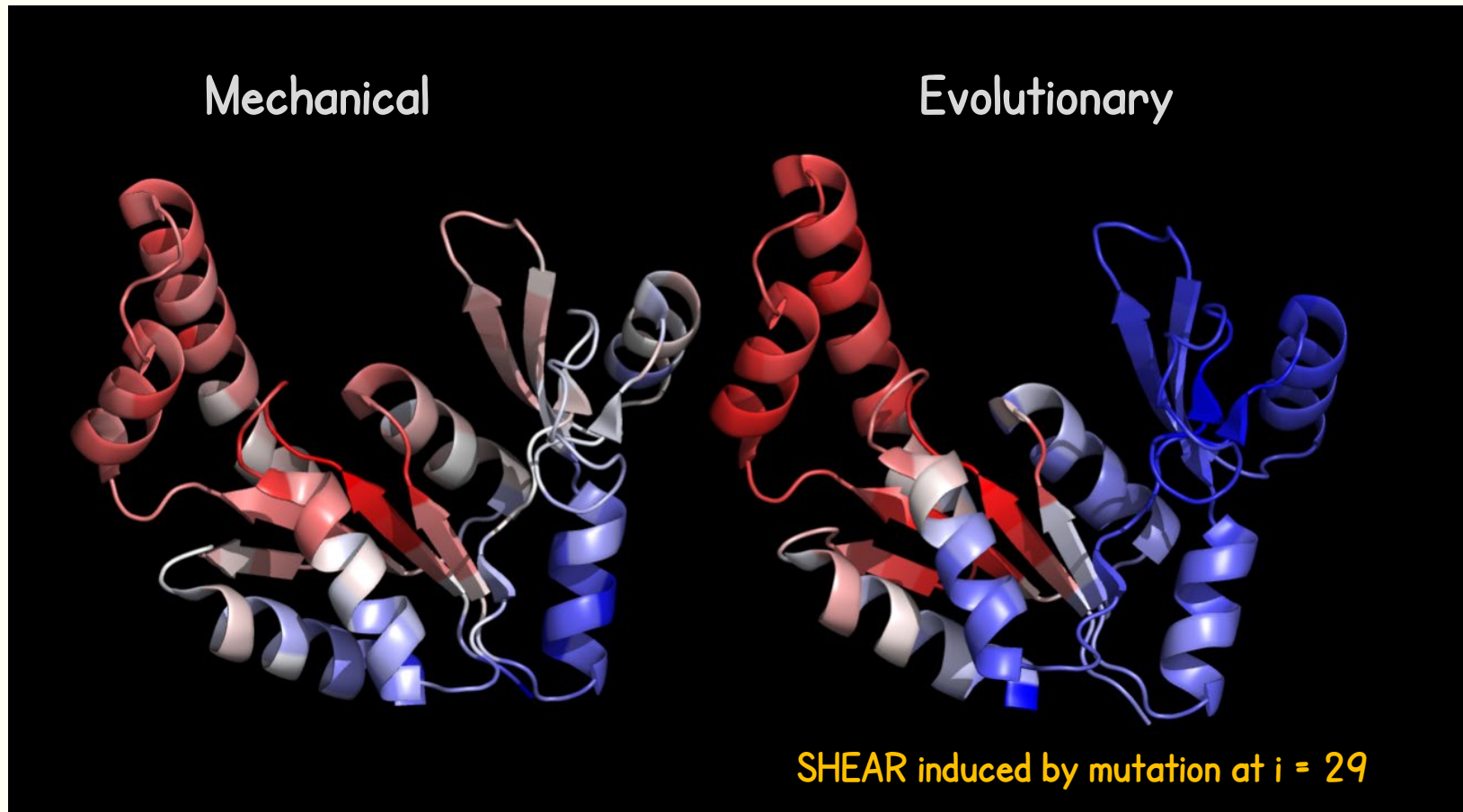
- Apply force to all residues (k)
- Get Correlation($\text{SHEAR}_{\text{EN}}(i), \text{SHEAR}_{\text{EN}}(j)$)

Correspondence between mechanical and evolutionary allostery

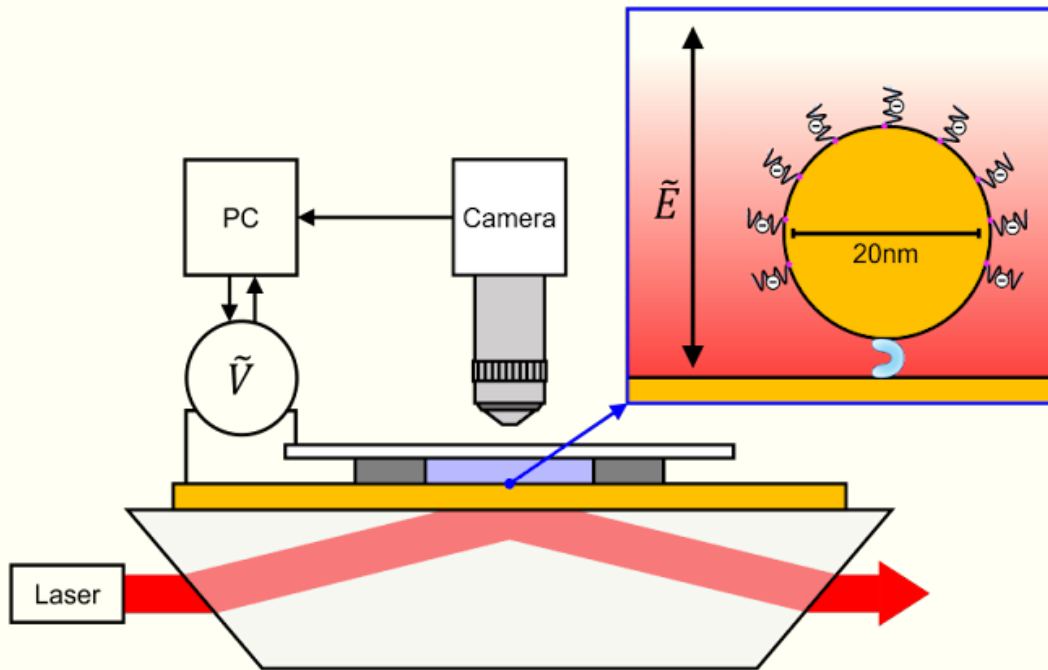


Both predict correlations at ranges of up to 2 nanometers

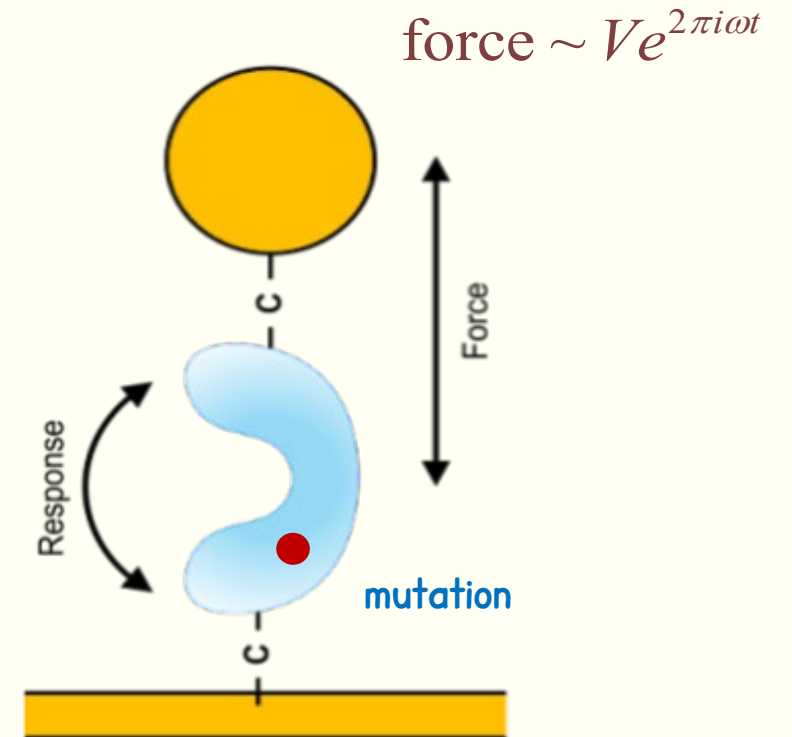
Allostery exhibits correspondence of Evolution and Mechanics



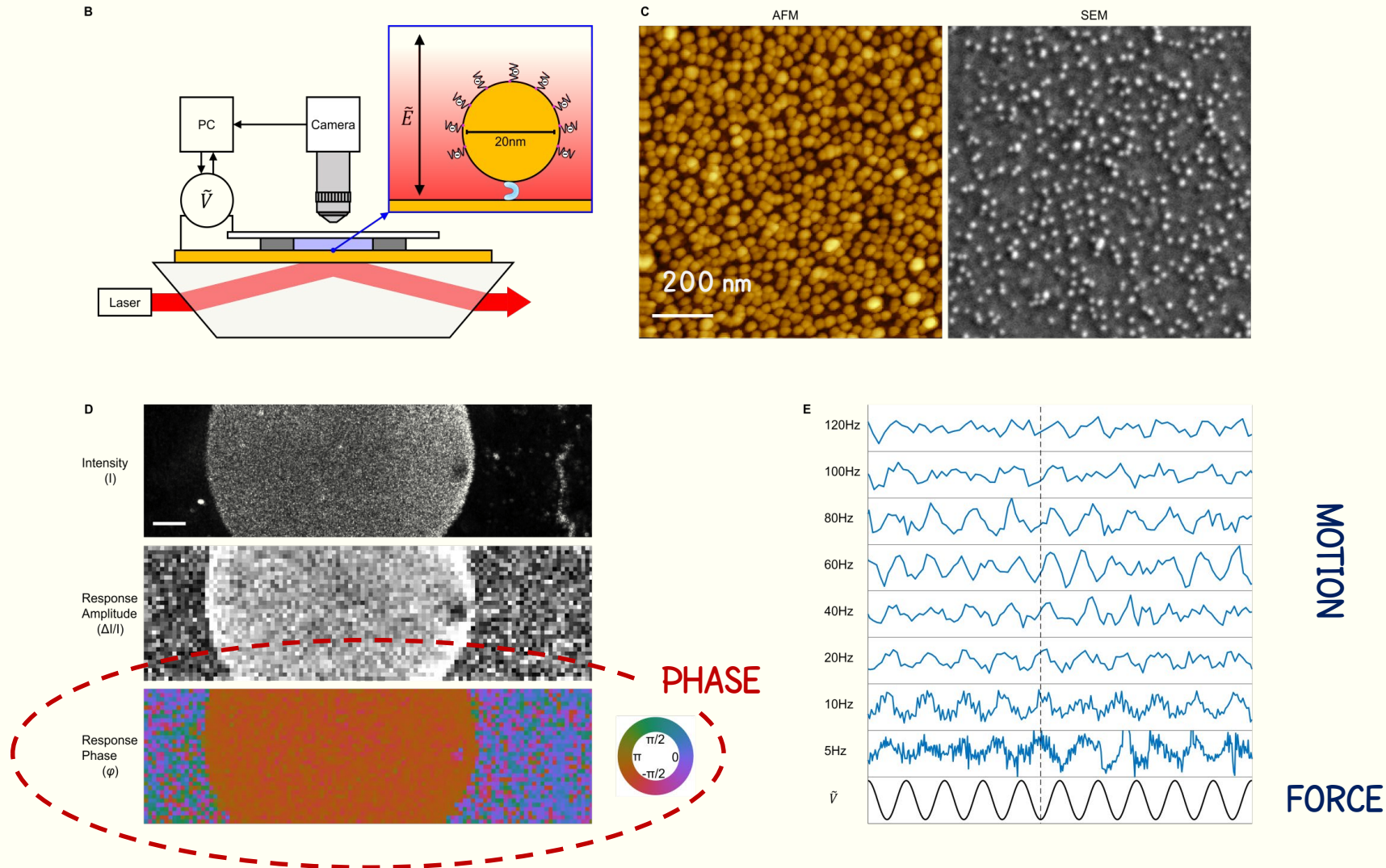
Nano-rheology of guanylate kinase reveals viscoelastic dynamics



Averaging over 10^6 nanoparticles allows
a resolution of 0.1 angstrom (Zocchi et al.)



Nano-rheology exhibits phase lag between applied electric force and internal motion

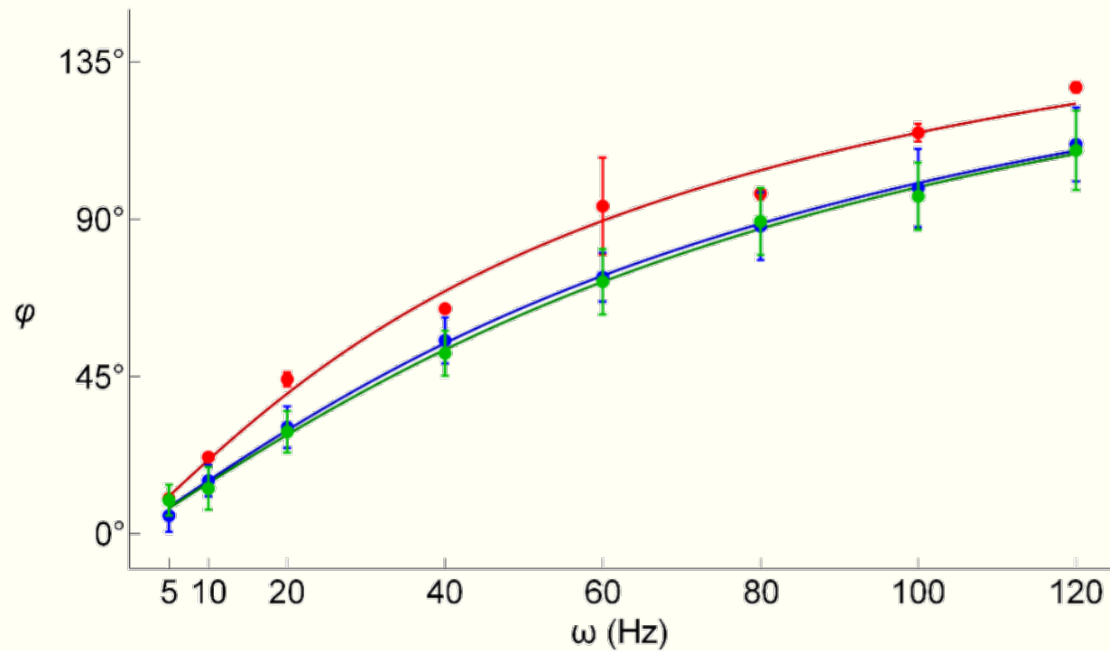


$$\text{force} \sim e^{2\pi i \omega t}$$

$$\text{motion} \sim e^{2\pi i (\omega t + \varphi)}$$

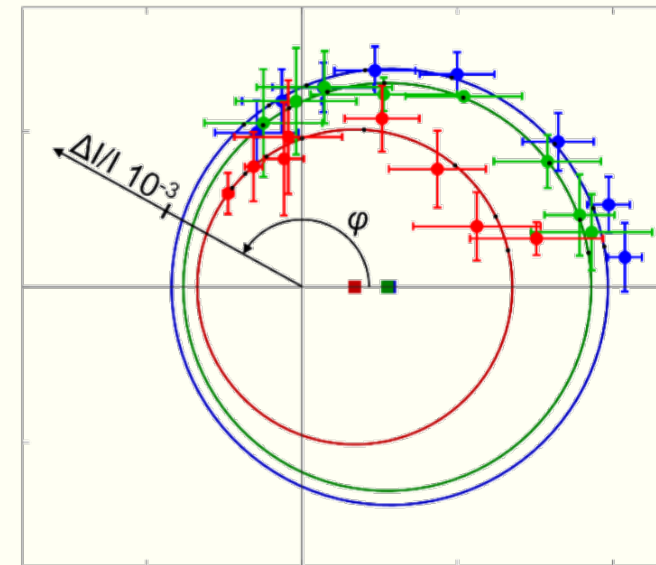
Viscoelastic behavior is affected by high-shear mutations

■ High shear ■ Control ■ Binding

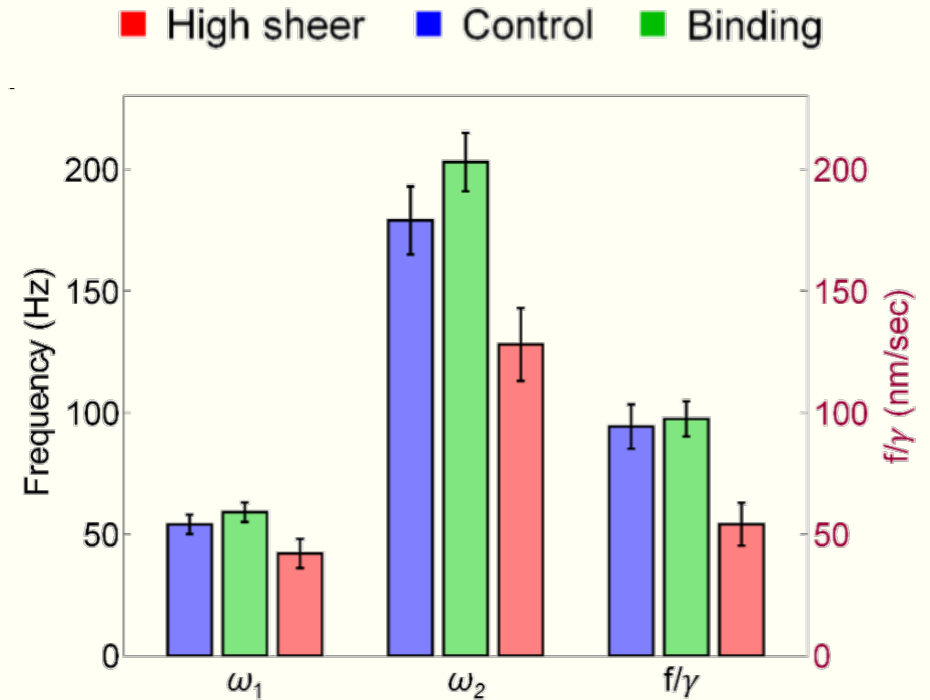
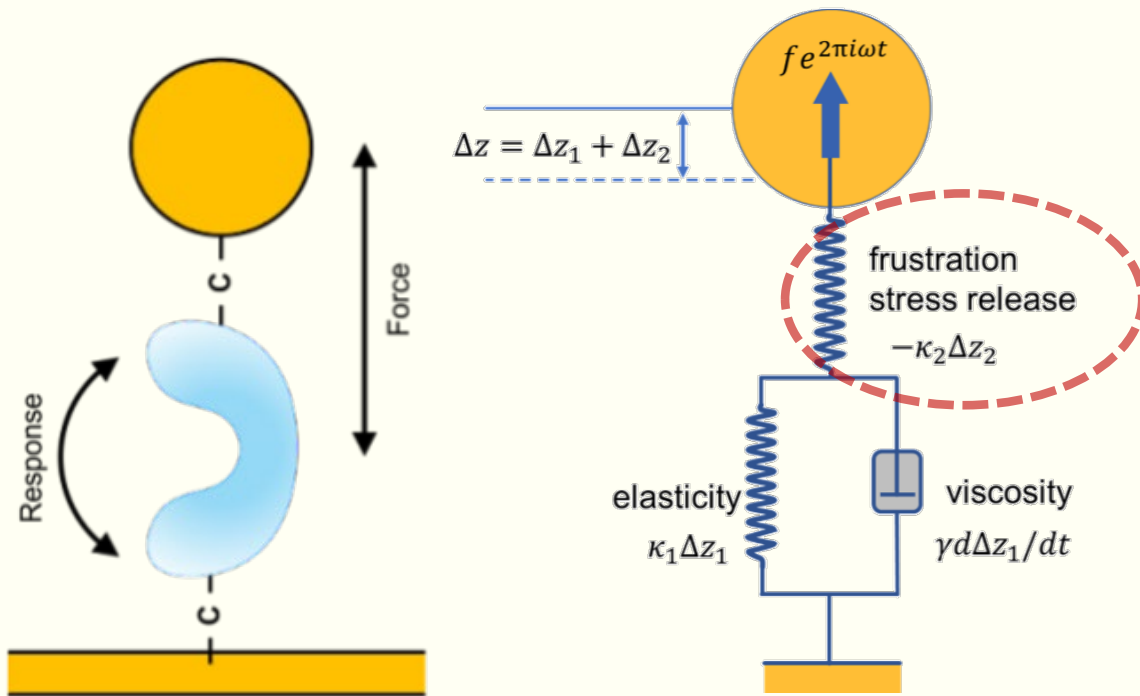


$$\text{force} \sim e^{2\pi i \omega t}$$

$$\text{motion} \sim e^{2\pi i (\omega t + \varphi)}$$



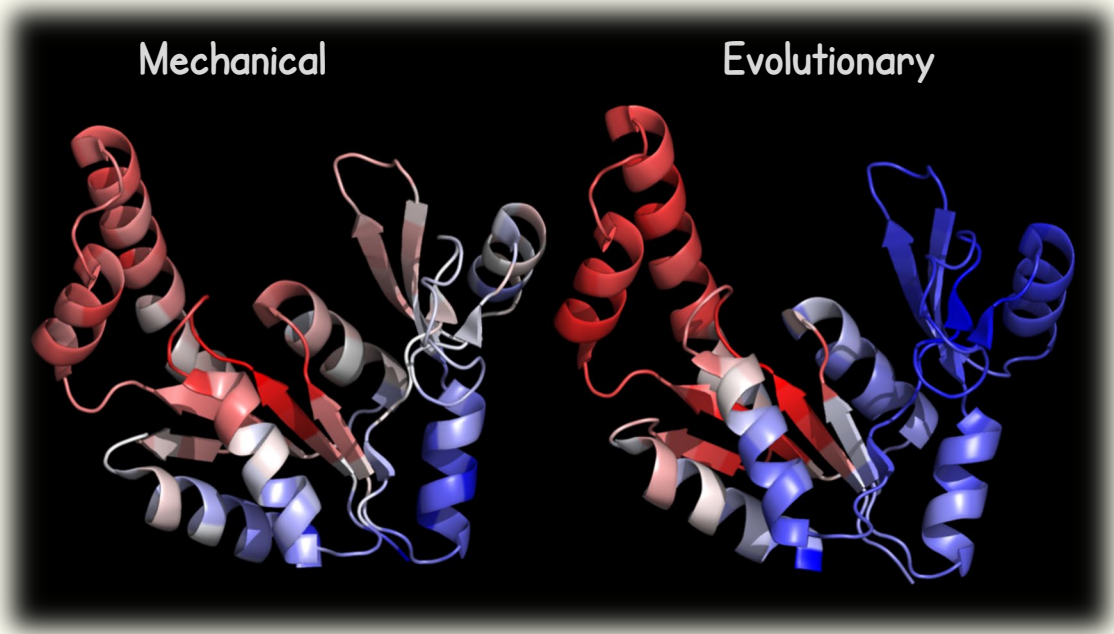
Protein viscoelasticity captures stress released from frustrated bonds



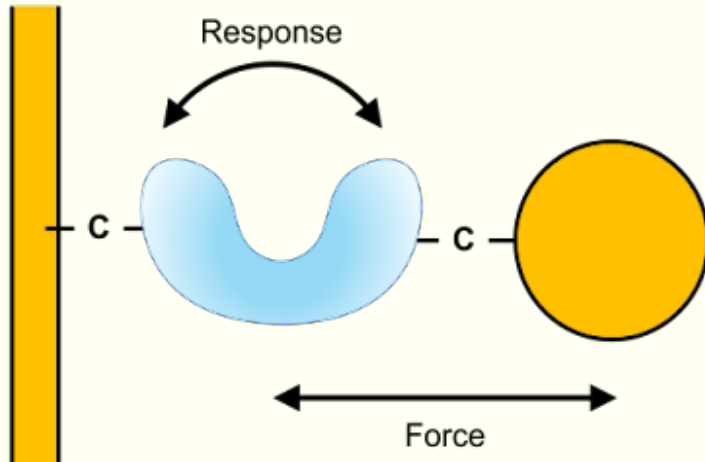
$$\Delta z(\omega) = \Delta z_1 + \Delta z_2 = \frac{f}{\gamma} \left(\frac{1}{\omega_1 + i\omega} - \frac{1}{\omega_2} \right)$$

$\omega_1 = \frac{\kappa_1}{\gamma}$ viscoelastic relaxation frequency
 $\omega_2 = \frac{\kappa_2}{\gamma}$ stress release frequency (frustration)

Logic of proteins as sequence encoded viscoelastic machines



- **Correspondence** of physical interactions, function, and evolution in patterns of **high-shear regions**
- **Evolutionary correlation** mirrors **mechanical** propagation of force (allostery)
- **Physical models** provide **insight** into basic principles.
- **AI (AlphaFold)** is sufficiently accurate to **predict** evolution of this correspondence and its effect on phenotype.
- **Experiment:** function of proteins is governed by **viscoelastic** specifications (e.g. turnover rate of guanylate kinase 24 Hz close to characteristic frequencies of the **machine**).



Jean-Pierre Eckmann & Jacques Rougemont (Geneva)

A. Libchaber (Rockefeller)

M. Mitchell & S. Leibler (Rockefeller & IAS)

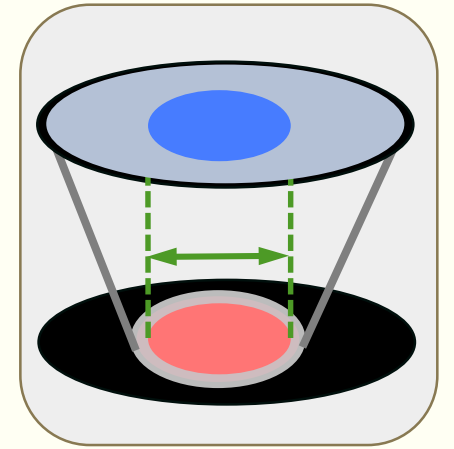
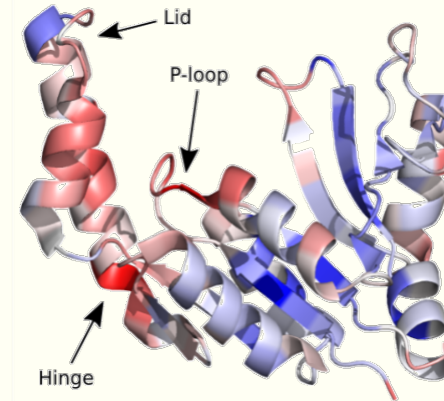
E. Moses & E. Weinreb (Weizmann)

G. Zocchi (UCLA)

J. McBride, S. Dutta, M. Seik & TT (IBS)



WE ARE RECRUITING



- McBride, Plev, Reinharz, Grzybowski, Tlusty, AlphaFold2 can predict structural and phenotypic effects of single mutations, *arxiv* 2022
- McBride, Eckmann, Tlusty, General theory of specific binding: insights from a genetic mechano-chemical protein model, *Mol. Biol. Evo* 2022
- Eckmann TT, Bioessays 2021
- Rougemont Eckmann TT, *Rev Mod Phys* 2019
- Eckmann Libchaber TT, *PRX* 2018
- Dutta Eckmann Libchaber TT, *PNAS* 2017
- Mitchell Leibler TT, *PNAS* 2016

- High-shear regions in guanylate kinase impact viscoelastic mechanics and enzymatic function

Weinreb, McBride, Siek, Rougemont, Renault, Peleg, Unger, Albeck, Sussman, Grzybowski, Zocchi, Eckmann, Moses, TT

