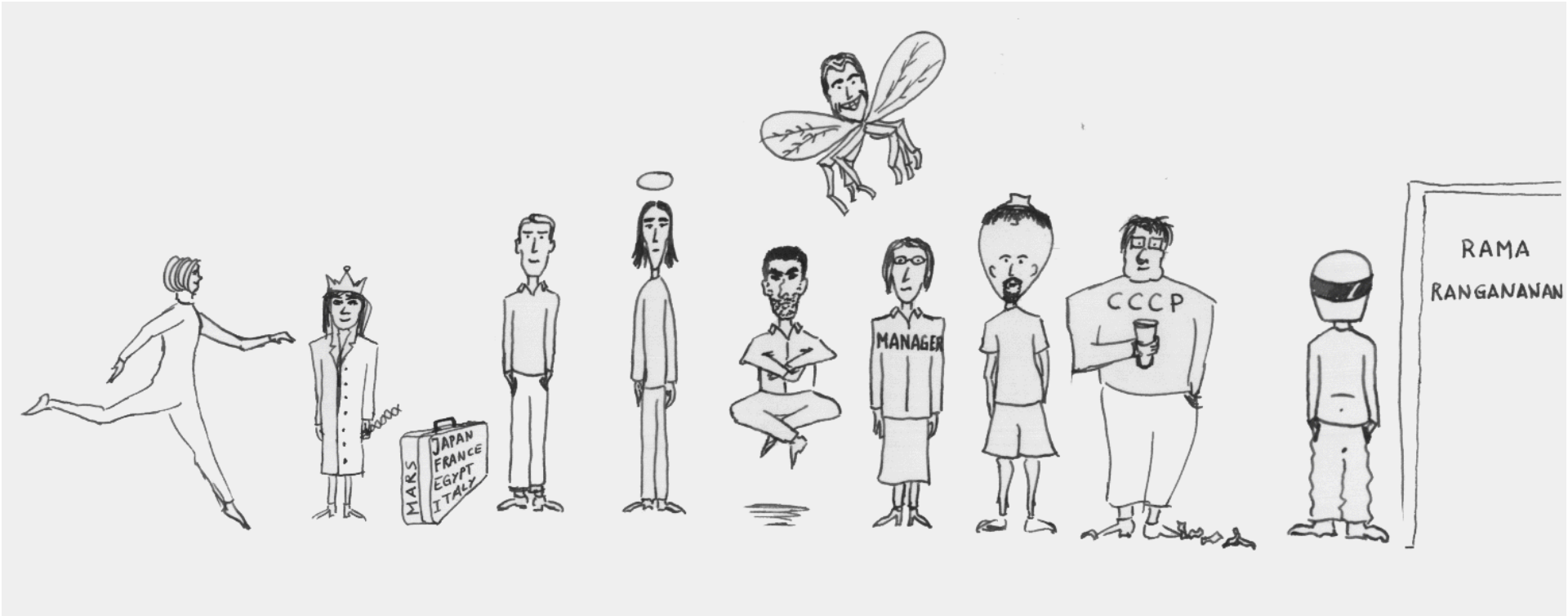
Lecture 13: Large-scale non-linear problems - part 1

R. Ranganathan
Green Center for Systems Biology, ND11.120E

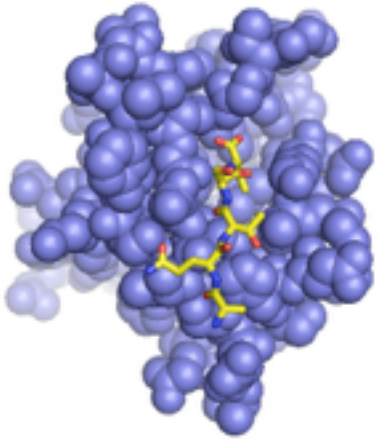
What are proteins? How do they fold, function, and evolve?



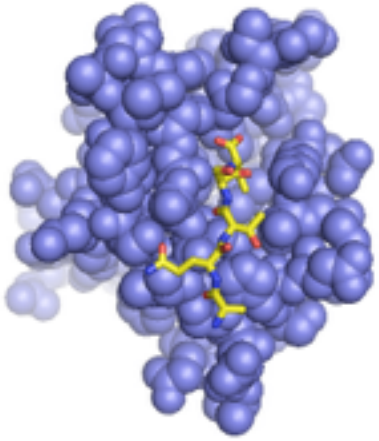
So, large-scale **non-linear dynamical systems**….how can we approach these problems?

	n = 1	n = 2 or 3	n >> 1	continuum
Linear	exponential growth and decay single step conformational change fluorescence emission pseudo first order kinetics	second order reaction kinetics linear harmonic oscillators simple feedback control sequences of conformational change	electrical circuits molecular dynamics systems of coupled harmonic oscillators equilibrium thermodynamics diffraction, Fourier transforms	Diffusion Wave propagation quantum mechanics viscoelastic systems
Nonlinear	fixed points bifurcations, multi stability irreversible hysteresis overdamped oscillators	anharmomic oscillators relaxation oscillations predator-prey models van der Pol systems Chaotic systems	systems of non-linear oscillators non-equilibrium thermodynamics protein structure/ function neural networks the cell ecosystems	Nonlinear wave propagation Reaction-diffusion in dissipative systems Turbulent/chaotic flows

adapted from S. Strogatz



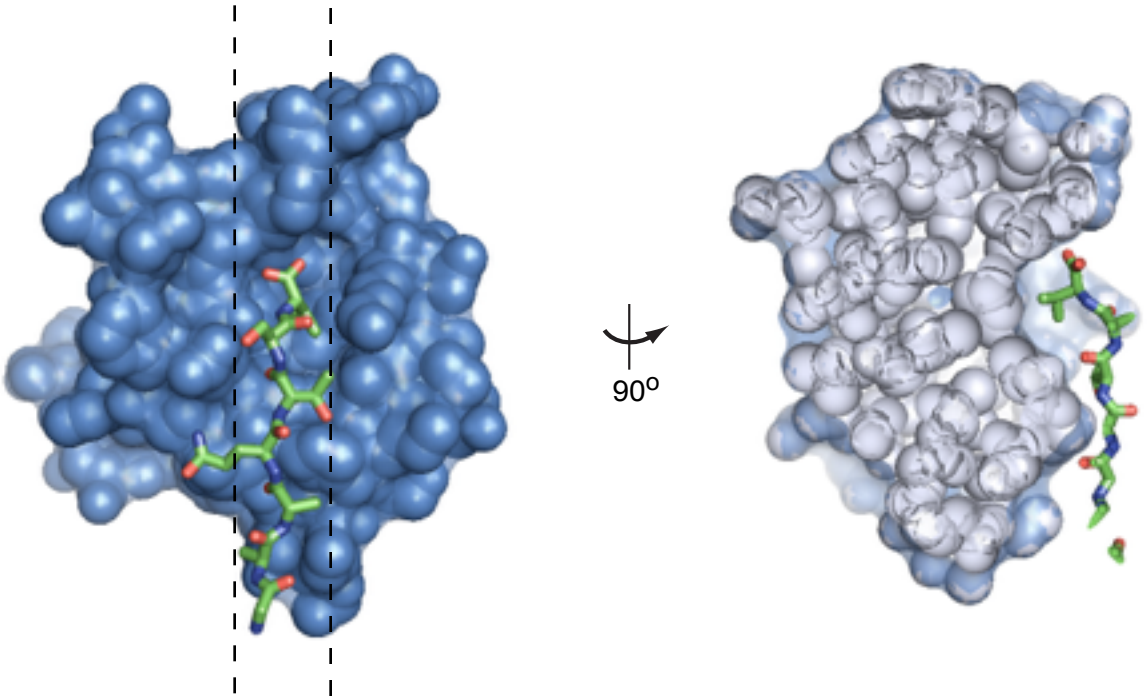
What is the “design” (in evolution) of proteins? The basic characteristics are **folding**, **function** (binding, catalysis, and allostery), and **evolvability**. We want to understand how these arise from the information stored in the gene sequence.



~ 100 - 1000 aa

So, here we have a system comprised of a lot of elementary parts (the amino acids). There are clearly cooperative (**non-linear, epistatic**) interactions between the residues…though the pattern of interactions is not obvious. It is hard to have intuition about this issue….

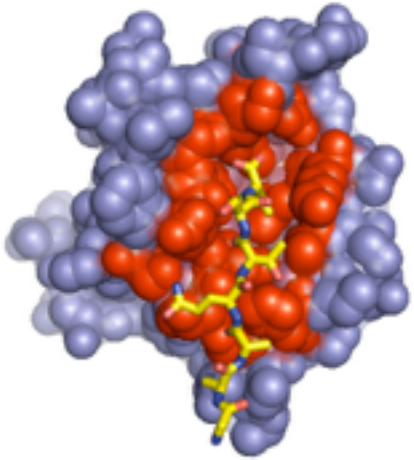
There are **some ideas**, of course….



Proteins as "3-D jigsaw puzzles"...
precise and locally exact

The principle of **spatial proximity**….

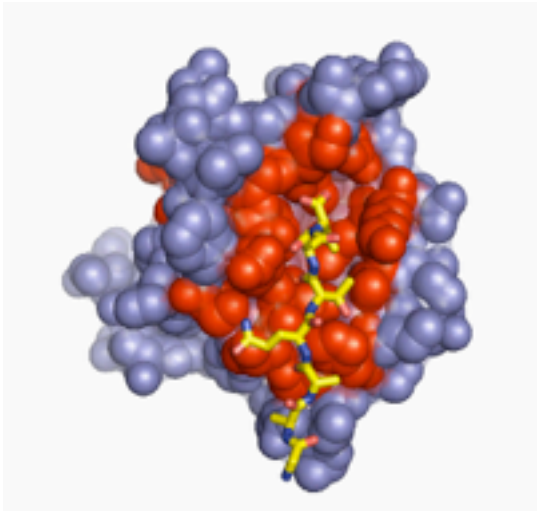
N — GEEDIPREPRRIVIHRGSTGLGFNIVGGEDGEGIFISFILAG-GPADLSGELRKGDQILSVNGVDLRNASHEQAAIALKNAGQTVTIIAQYKPEE — C



But….

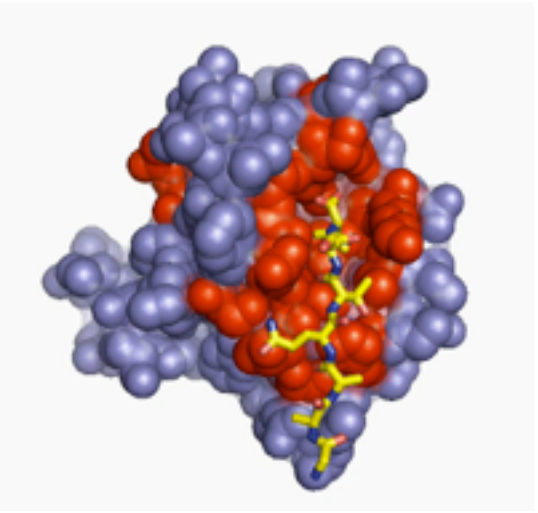
e.g. Skelton et al, JBC 278, p.7645 (2003)
Origins of PDZ domain ligand specificity. Structure determination and mutagenesis of the Erbin PDZ domain.

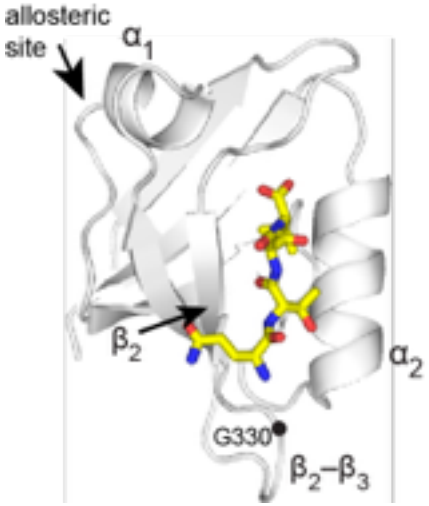
…proteins often encode the capacity for **long-range functional coupling**…



Garrard, Capaldo, Gao, Rosen, Macara, Tomchick, EMBO J. 22, 1125-33.
Peterson, Penkert, Volkmann, Prehoda, Mol. Cell 13, 665-76.

…proteins often encode the capacity for **long-range functional coupling**…

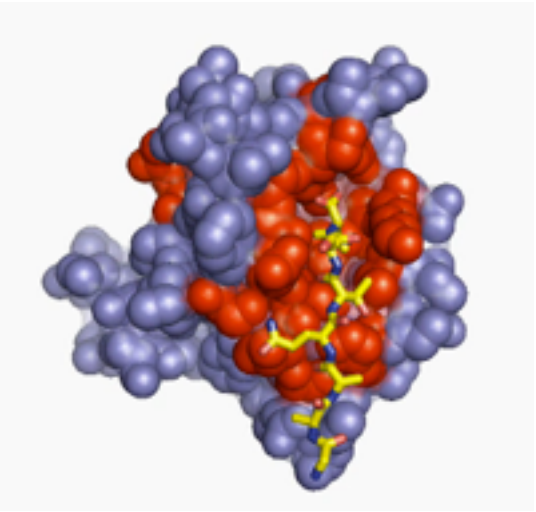


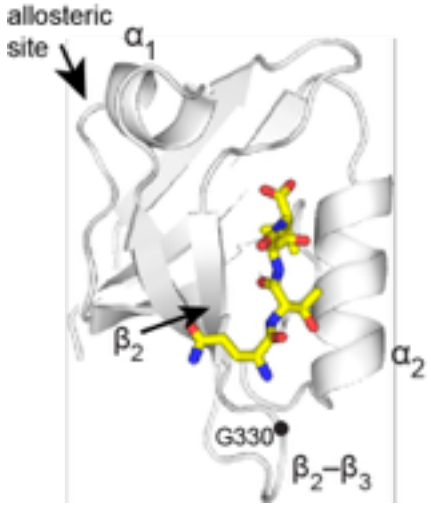


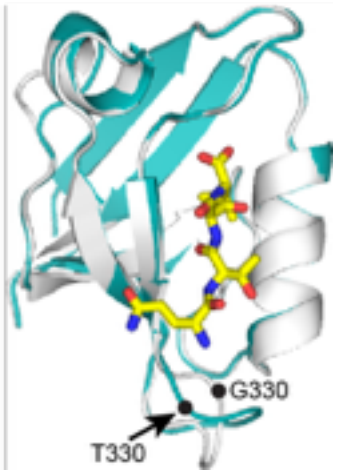
The PDZ domain….

A. Raman, K. I. White, manuscript in preparation

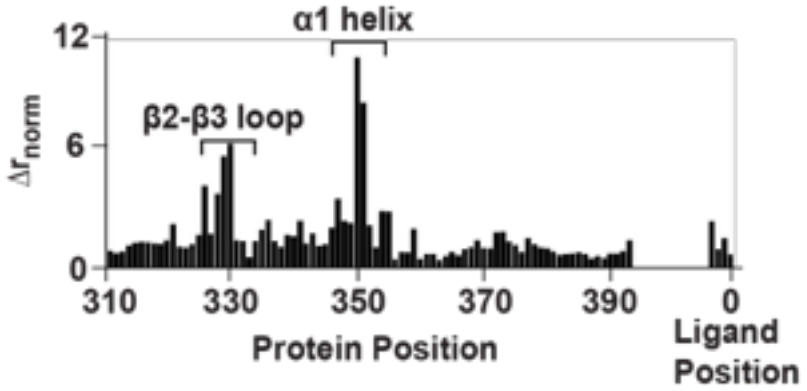
…proteins often encode the capacity for **long-range functional coupling**…





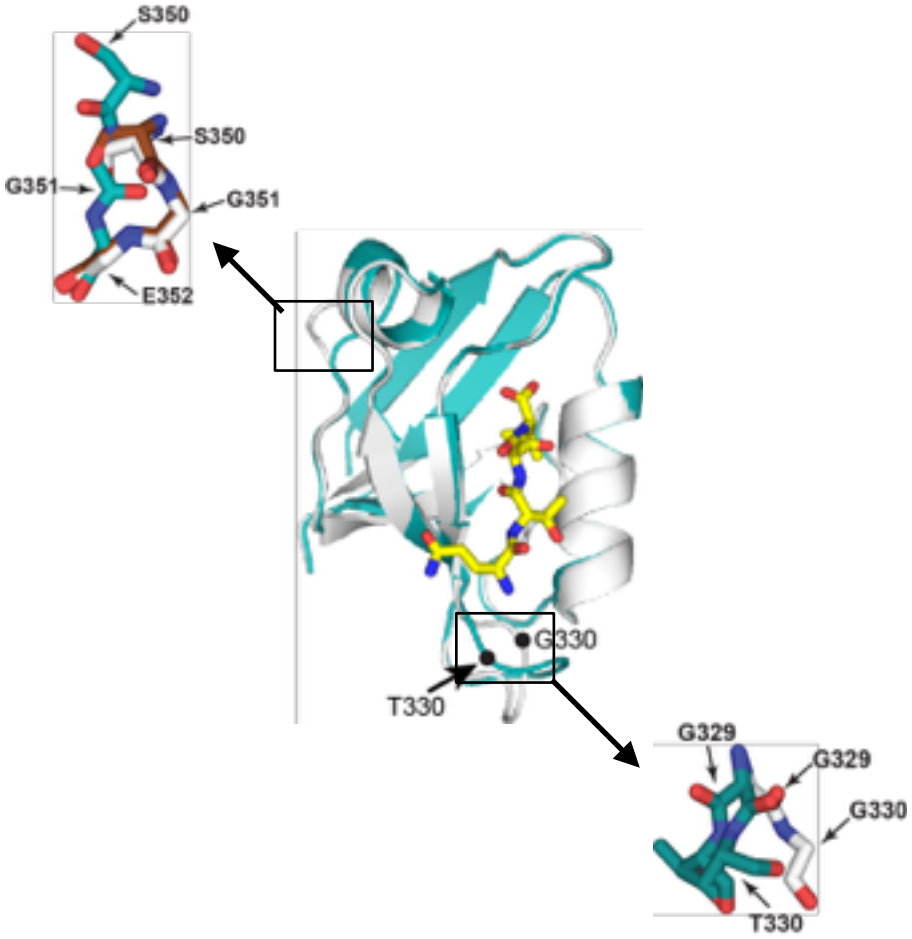


Point mutation in the beta2-beta3 loop…local and distant effects!



A. Raman, K. I. White, manuscript in preparation

…proteins often encode the capacity for **long-range functional coupling**…



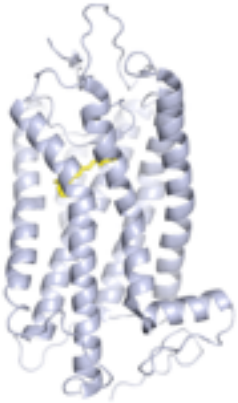
The transmission mechanism is not obvious….

A. Raman, K. I. White, manuscript in preparation

The biological role of **long-range coupling**…

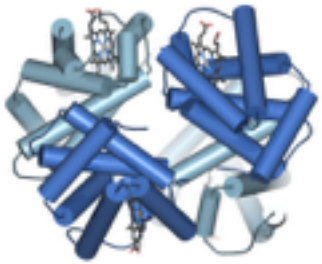
PDZ

GPCR



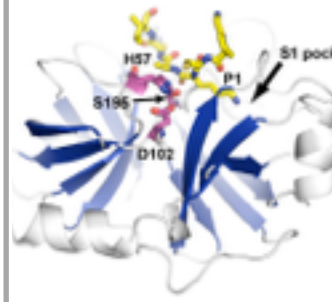
TonB-dependent transporters

hemoglobin

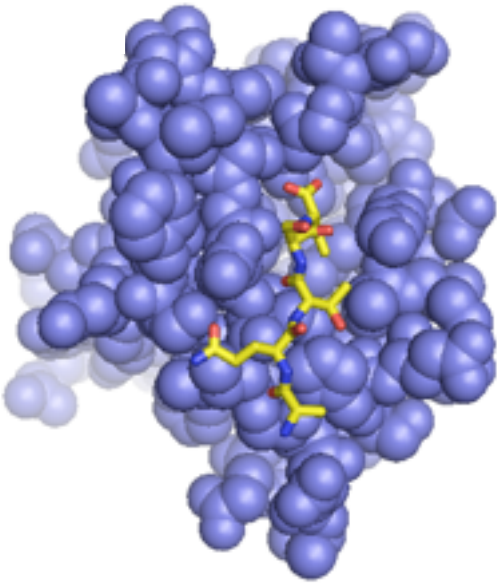


DHFR

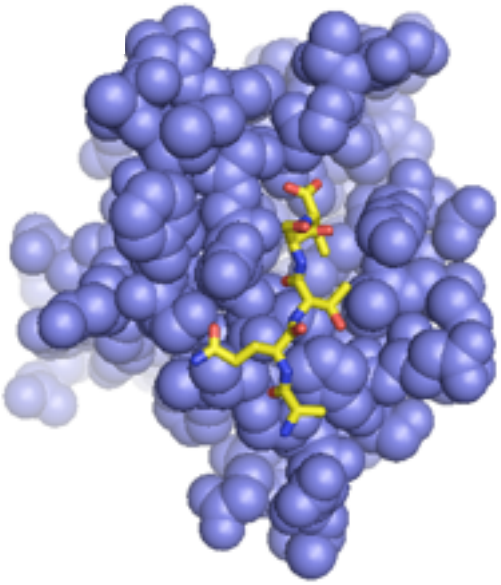
S1A protease



Long-range interactions also mediate **signal transmission**, **catalysis**, and **regulation**,….basic and defining features of protein families



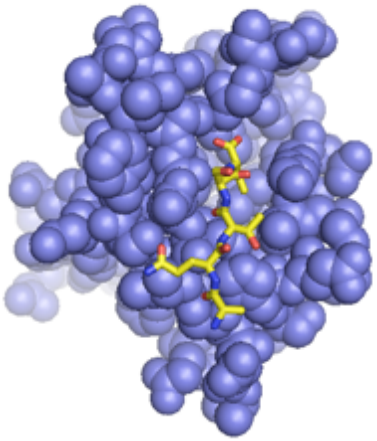
What is the **essence of the problem**?



What is the essence of the problem?

Due to marginal stability, the subtlety of the physical forces acting between atoms, and **the combinatorial complexity of non-linear interactions**, we don't have good models for the pattern of net forces between atoms.

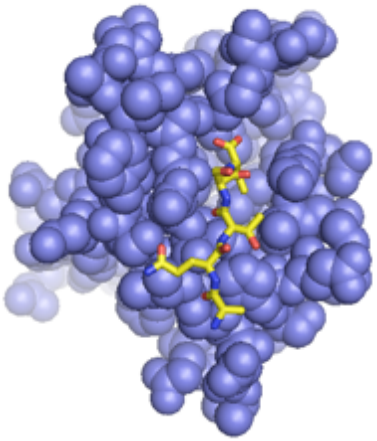
What does **non-independence** mean, exactly?



y_0 → the “wild-type” phenotype

here, for an equilibrium thermodynamic property of a protein,
but the framework is **general for any system** split up
into a bunch of operational parts…as long as….

What does **non-independence** mean, exactly?

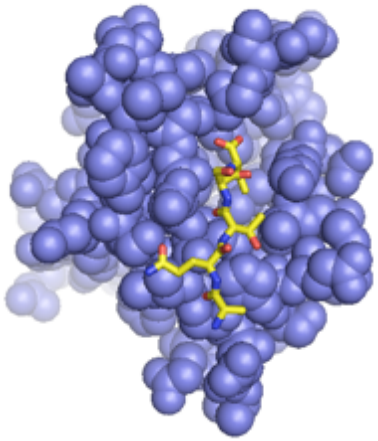


y_0 → the “wild-type” phenotype

↑

also, the zeroth order epistasis ϵ_0

What does **non-independence** mean, exactly?

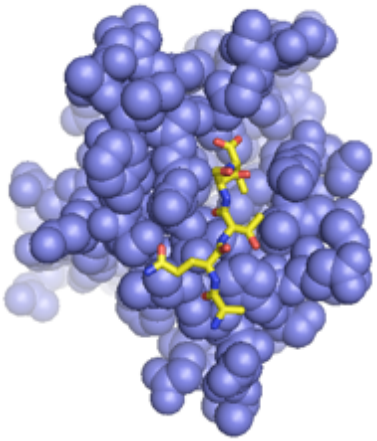


y_0 → the “wild-type” phenotype

in matrix form….

$$\epsilon_0 = [1]\, y_0$$

What does **non-independence** mean, exactly?

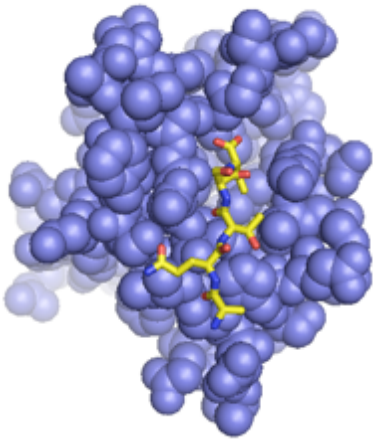


a single mutant phenotype

y_0 • ——→ • y_1

$\varepsilon_1 = y_1 - y_0$ a first-order epistasis

What does **non-independence** mean, exactly?

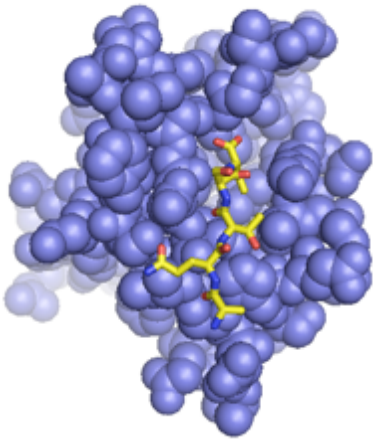


a single mutant phenotype

in matrix form….

$$\begin{bmatrix} \epsilon_0 \\ \epsilon_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \end{bmatrix}$$

What does **non-independence** mean, exactly?



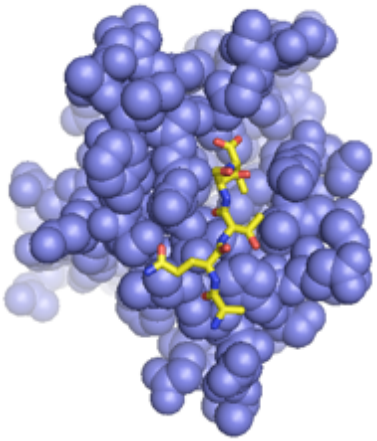
a single mutant phenotype

in matrix form….

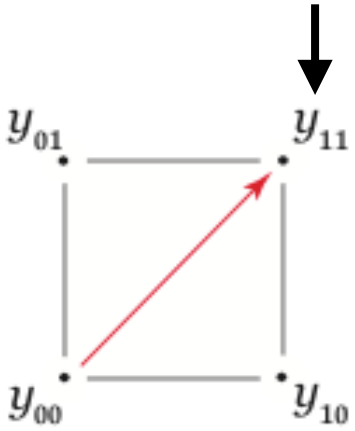
$$\begin{bmatrix} \epsilon_0 \\ \epsilon_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} y_0 \\ y_1 \end{bmatrix}$$

Do you see that we are re-parameterizing (or **transforming**) our representation of this system from a space of phenotypes to a space of epistases….

What does **non-independence** mean, exactly?



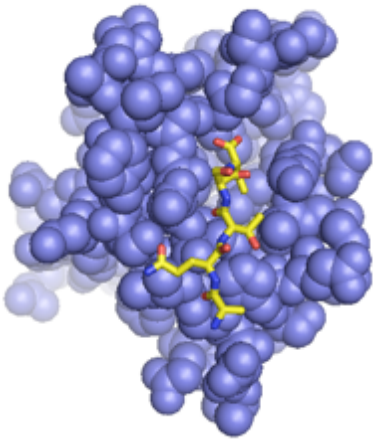
a double mutant phenotype



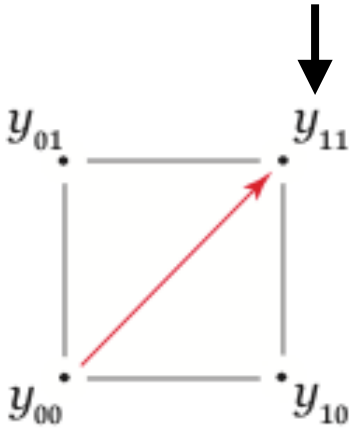
$$\varepsilon_{11} = (y_{11} - y_{01}) - (y_{10} - y_{00})$$

The **second order epistasis** of two mutations - the degree to which the effect of one mutation depends on the background of a second…

What does **non-independence** mean, exactly?



a double mutant phenotype

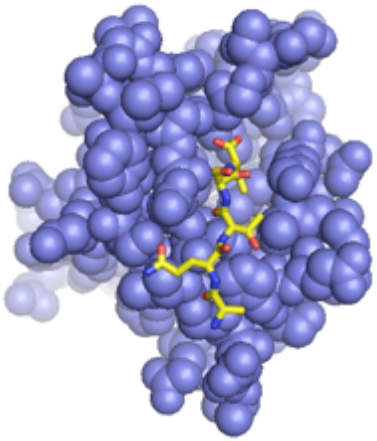


in matrix form….

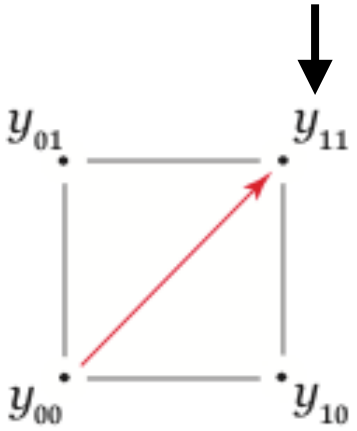
$$\begin{bmatrix} \epsilon_{00} \\ \epsilon_{01} \\ \epsilon_{10} \\ \epsilon_{11} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} y_{00} \\ y_{01} \\ y_{10} \\ y_{11} \end{bmatrix}$$

again, a **mapping** from the space of phenotypes to the space of epistases…

What does **non-independence** mean, exactly?



a double mutant phenotype



in matrix form….

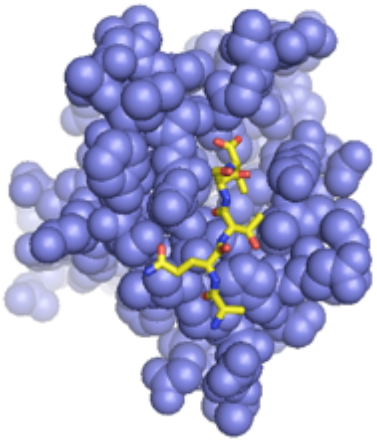
$$\begin{bmatrix} \epsilon_{00} \\ \epsilon_{01} \\ \epsilon_{10} \\ \epsilon_{11} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & -1 & -1 & 1 \end{bmatrix} \begin{bmatrix} y_{00} \\ y_{01} \\ y_{10} \\ y_{11} \end{bmatrix}$$

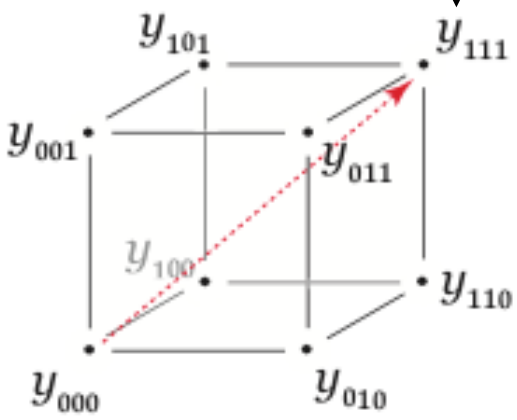
note the potential **value of the transform**….

(1) what if all mutants had no effect?
(2) what if the two mutations had an effect, but are independent?

What does **non-independence** mean, exactly?

a triple mutant phenotype



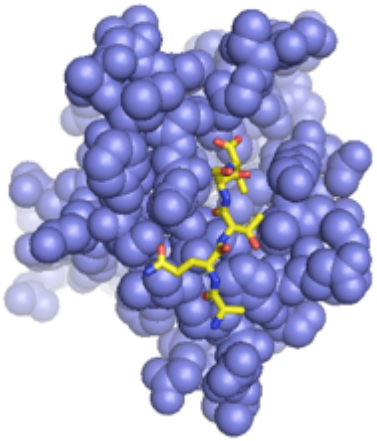


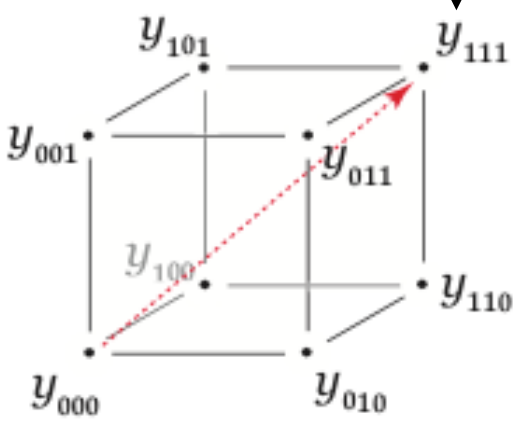
$$\varepsilon_{111} = (y_{111} - y_{101} - y_{110} + y_{100}) - (y_{011} - y_{001} - y_{010} + y_{000})$$

The **third order epistasis** - the degree to which a second order epistasis depends on a third mutation…

What does **non-independence** mean, exactly?

a triple mutant phenotype





in matrix form….

$$\begin{pmatrix} \varepsilon_{000} \\ \varepsilon_{001} \\ \varepsilon_{010} \\ \varepsilon_{011} \\ \varepsilon_{100} \\ \varepsilon_{101} \\ \varepsilon_{110} \\ \varepsilon_{111} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \\ -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \end{pmatrix} * \begin{pmatrix} y_{000} \\ y_{001} \\ y_{010} \\ y_{011} \\ y_{100} \\ y_{101} \\ y_{110} \\ y_{111} \end{pmatrix}$$

What does **non-independence** mean, exactly?

$$\begin{pmatrix} \varepsilon_{000} \\ \varepsilon_{001} \\ \varepsilon_{010} \\ \varepsilon_{011} \\ \varepsilon_{100} \\ \varepsilon_{101} \\ \varepsilon_{110} \\ \varepsilon_{111} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \\ -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \end{pmatrix} * \begin{pmatrix} y_{000} \\ y_{001} \\ y_{010} \\ y_{011} \\ y_{100} \\ y_{101} \\ y_{110} \\ y_{111} \end{pmatrix}$$

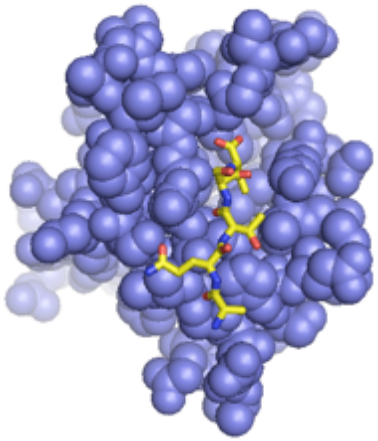
…again, note **the dramatic effect** of independence in simplifying the epistasis vector.

What does **non-independence** mean, exactly?

$$\begin{pmatrix} \varepsilon_{000} \\ \varepsilon_{001} \\ \varepsilon_{010} \\ \varepsilon_{011} \\ \varepsilon_{100} \\ \varepsilon_{101} \\ \varepsilon_{110} \\ \varepsilon_{111} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\ -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \\ -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \end{pmatrix} * \begin{pmatrix} y_{000} \\ y_{001} \\ y_{010} \\ y_{011} \\ y_{100} \\ y_{101} \\ y_{110} \\ y_{111} \end{pmatrix}$$

…again, note **the dramatic effect** of independence in simplifying the epistasis vector. In a sense, the fraction of non-zero terms in the epistasis vector is a measure of system complexity….the number of numbers I need to know to specify it.

What does **non-independence** mean, exactly?



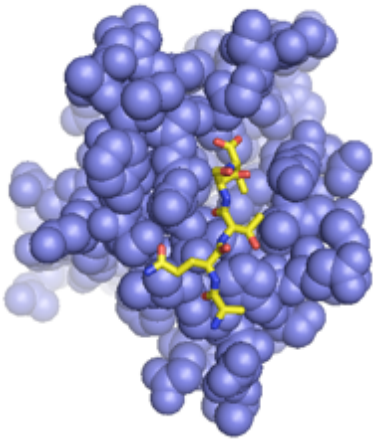
the phenotypes

$$\bar{\varepsilon} = \boldsymbol{G}\,\bar{y}$$

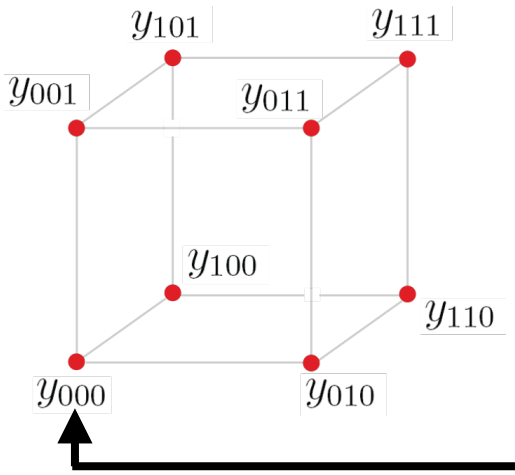
the epistases

A recursive **generative function** for n[th]-order epistasis….

$$\boldsymbol{G}_{n+1} = \begin{pmatrix} \boldsymbol{G}_n & 0 \\ -\boldsymbol{G}_n & \boldsymbol{G}_n \end{pmatrix} \quad \text{with} \quad \boldsymbol{G}_0 = 1$$

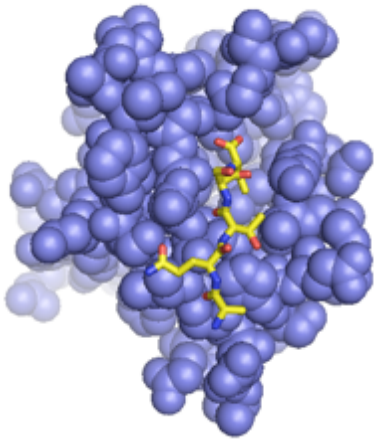


$$\bar{\varepsilon} = \boldsymbol{G}\,\bar{y}$$



In the **biochemical view of epistasis**…we take one particular genotype (the “wild-type”) as our reference…

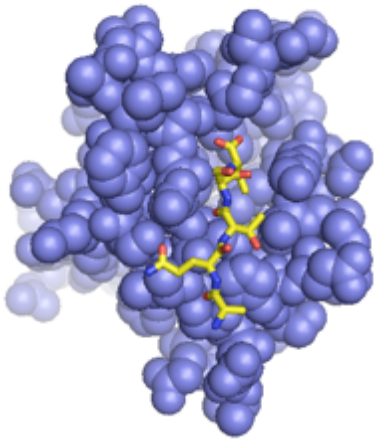
Background averaged epistasis…



$$\bar{\varepsilon} = \underline{\boldsymbol{V}\boldsymbol{H}}\,\bar{y}.$$

the background-averaged epistasis operator

Background averaged epistasis…



$$\bar{\varepsilon} = \boldsymbol{V}\boldsymbol{H}\bar{y}.$$

$$\begin{pmatrix} e_{***} \\ e_{**1} \\ e_{*1*} \\ e_{*11} \\ e_{1**} \\ e_{1*1} \\ e_{11*} \\ e_{111} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{pmatrix} * \begin{pmatrix} y_{000} \\ y_{001} \\ y_{010} \\ y_{011} \\ y_{100} \\ y_{101} \\ y_{110} \\ y_{111} \end{pmatrix}$$

so, for example…

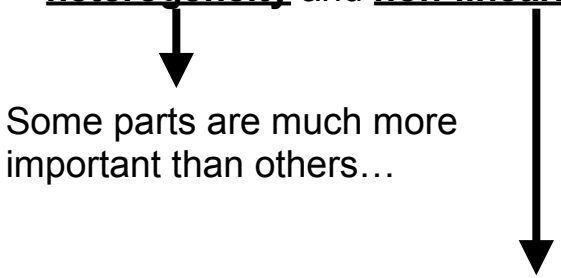
$$\varepsilon_{***} = \frac{(y_{111}+y_{101}+y_{110}+y_{100}+y_{011}+y_{001}+y_{010}+y_{000})}{8}$$

the essential characteristics of **complex systems**…

heterogeneity and **non-linearity**…

the essential characteristics of **complex systems**…

heterogeneity and **non-linearity**…

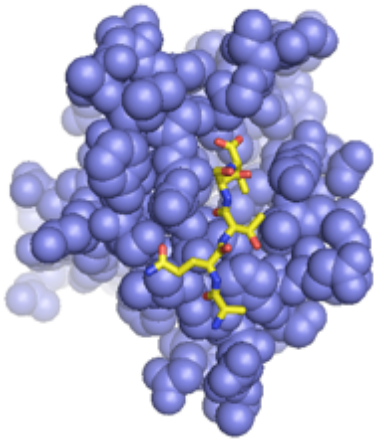


Some parts are much more important than others…

Parts don't act independently….the whole displays behaviors that are much more than the summed action of the parts

…and this information is held in **the epistasis vector**

So, three options….

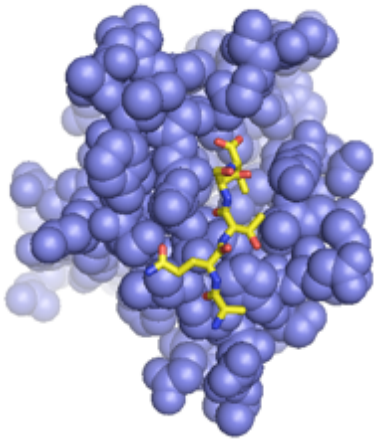


$$\bar{\varepsilon} = G\,\bar{y}$$ → single-reference epistasis

$$\bar{\varepsilon} = V\,H\,\bar{y}.$$ → background-averaged epistasis

$$\bar{\omega} = \Omega_{\text{epi}}\,\bar{y}$$

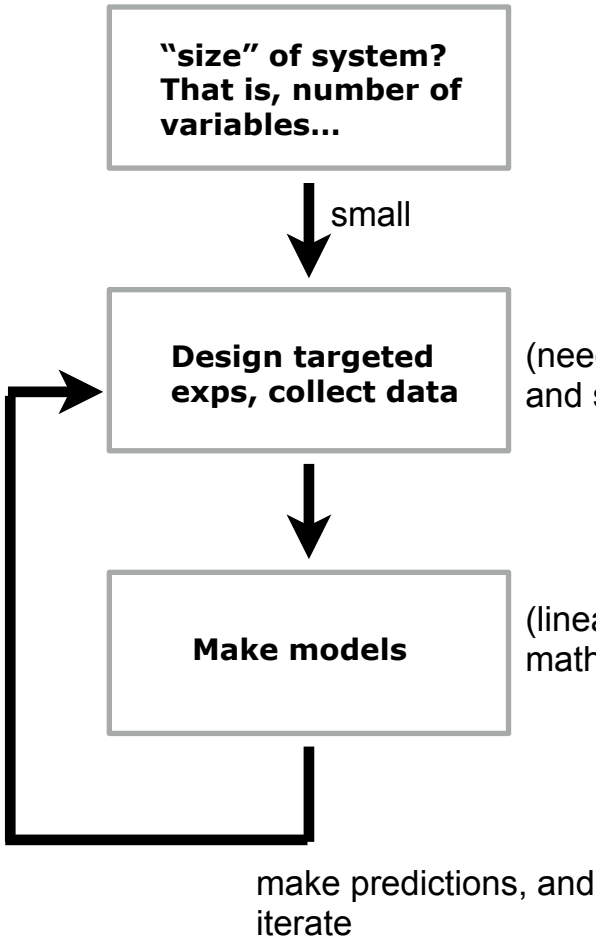
So, three options….



$$\bar{\omega} = \Omega_{\text{epi}}\, \bar{y}$$

(1) **Enumerate** the phenotype vector experimentally and then compute the importance and non-linearity of positions (in the epistasis vector).

For **small systems…**

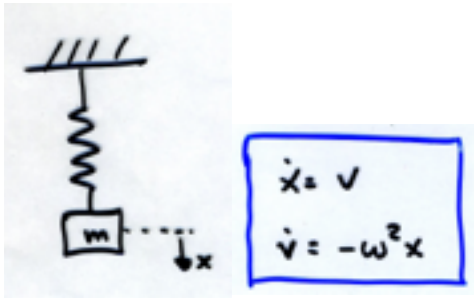


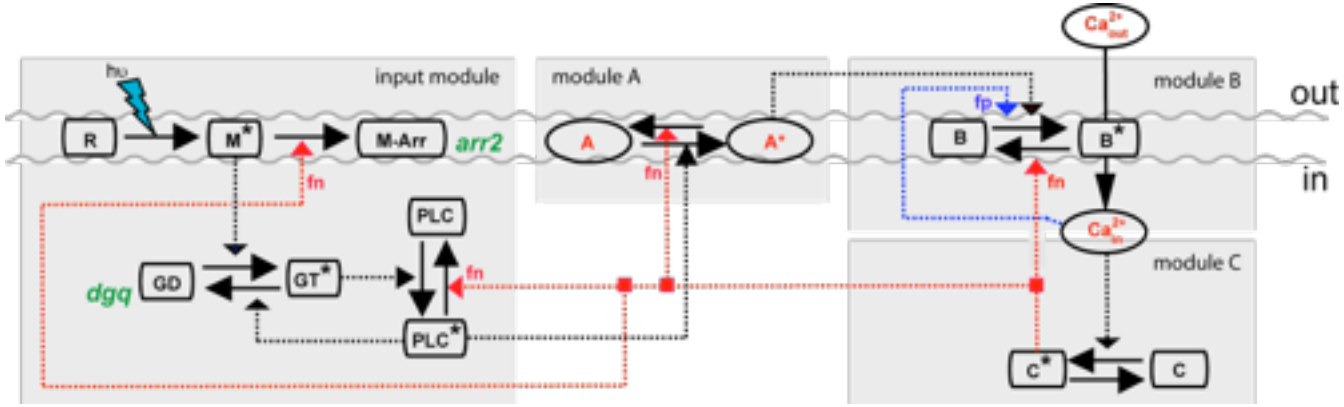
(need intuition, high-quality work, and some imagination)

(linear or non-linear, and use the tools of mathematical analysis to get insight)

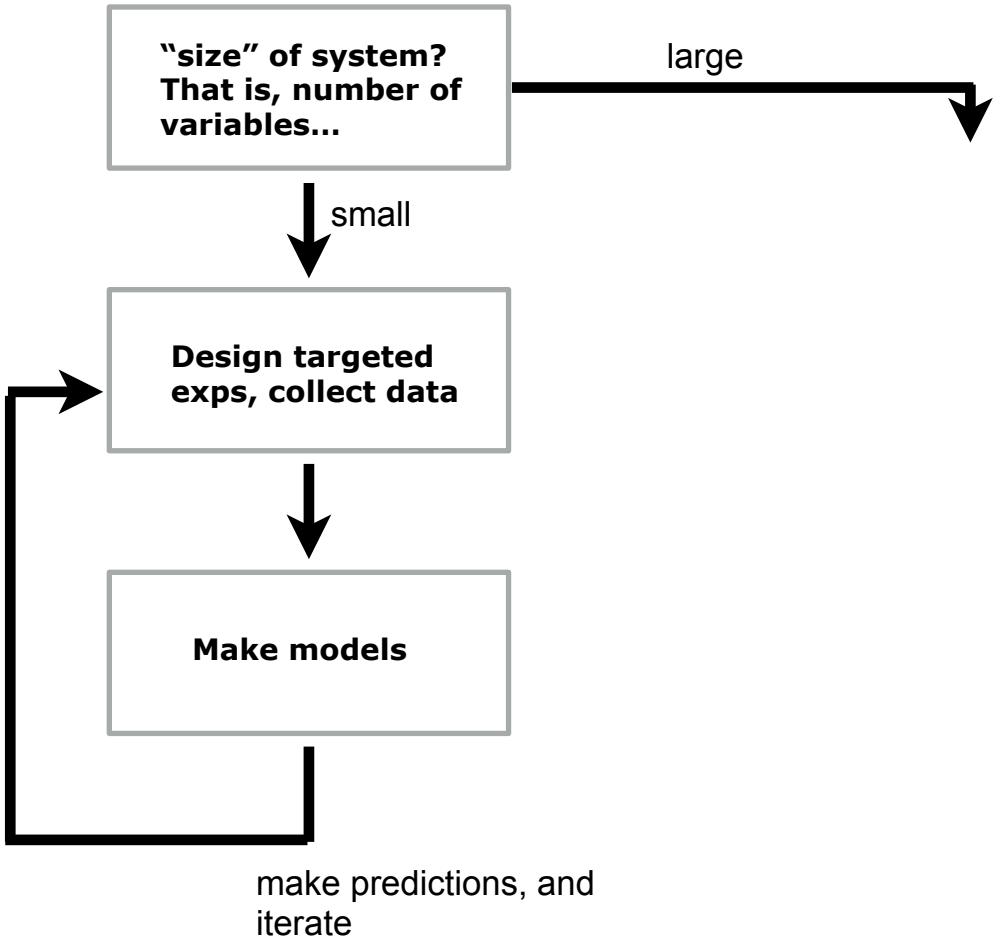
So, **for example**…

The linear harmonic oscillator, the MAPK bistable switch, and the relaxation oscillator…

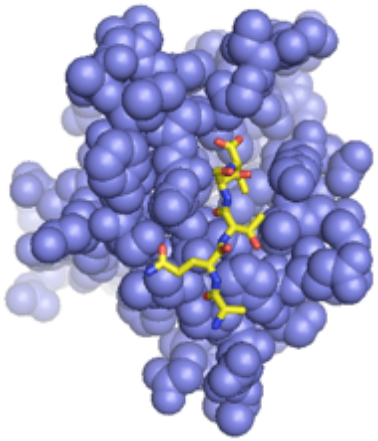




What about **large systems**?



So, three options….

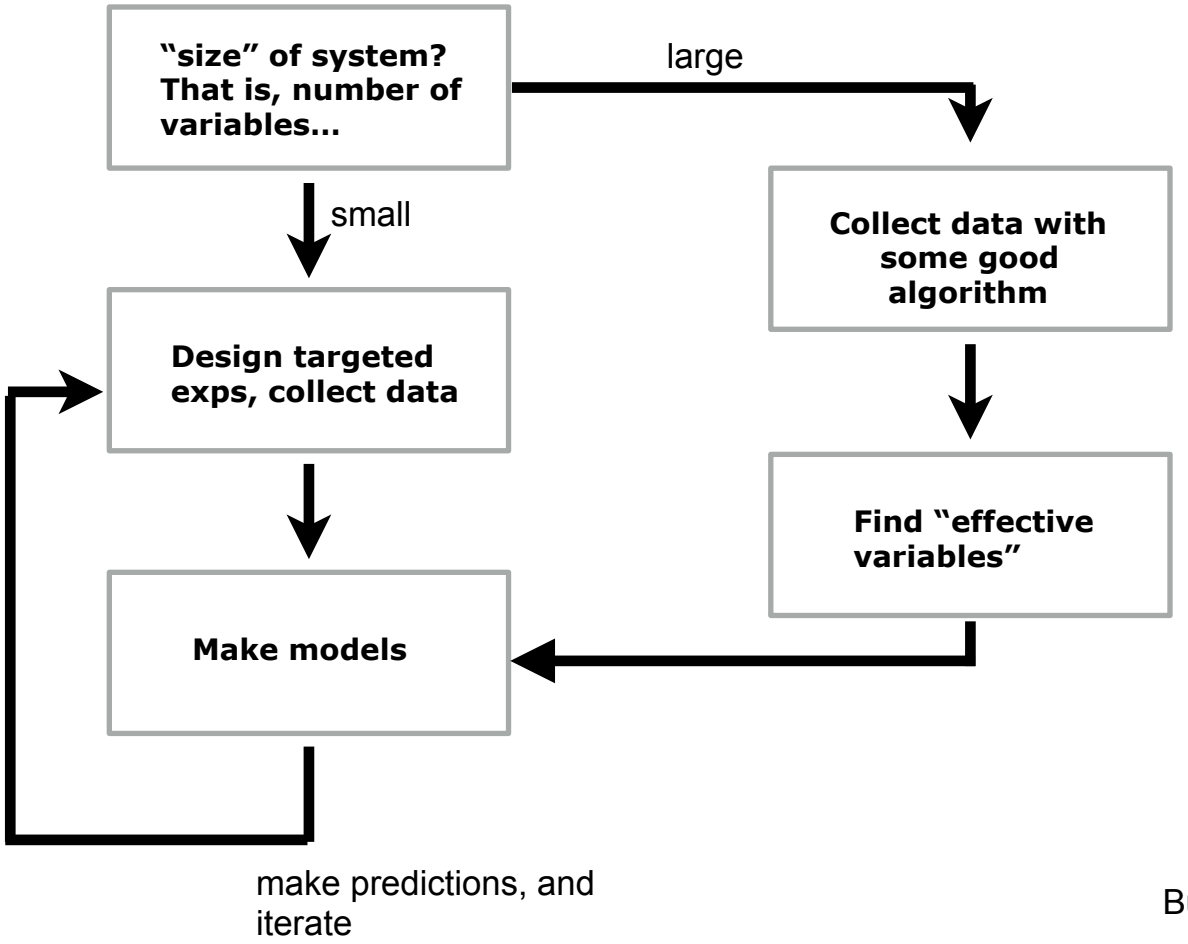


$$\bar{\omega} = \Omega_{\text{epi}}\, \bar{y}$$

(1) **Enumerate** the phenotype vector experimentally and then compute the importance and non-linearity of positions (in the epistasis vector).

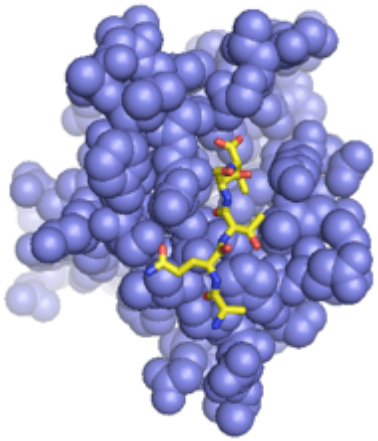
(2) Find the **right way to sub-sample** the phenotype vector so that it well-approximates the epistasis vector.

What about **large systems**?



But, **what algorithm?**

So, three options….



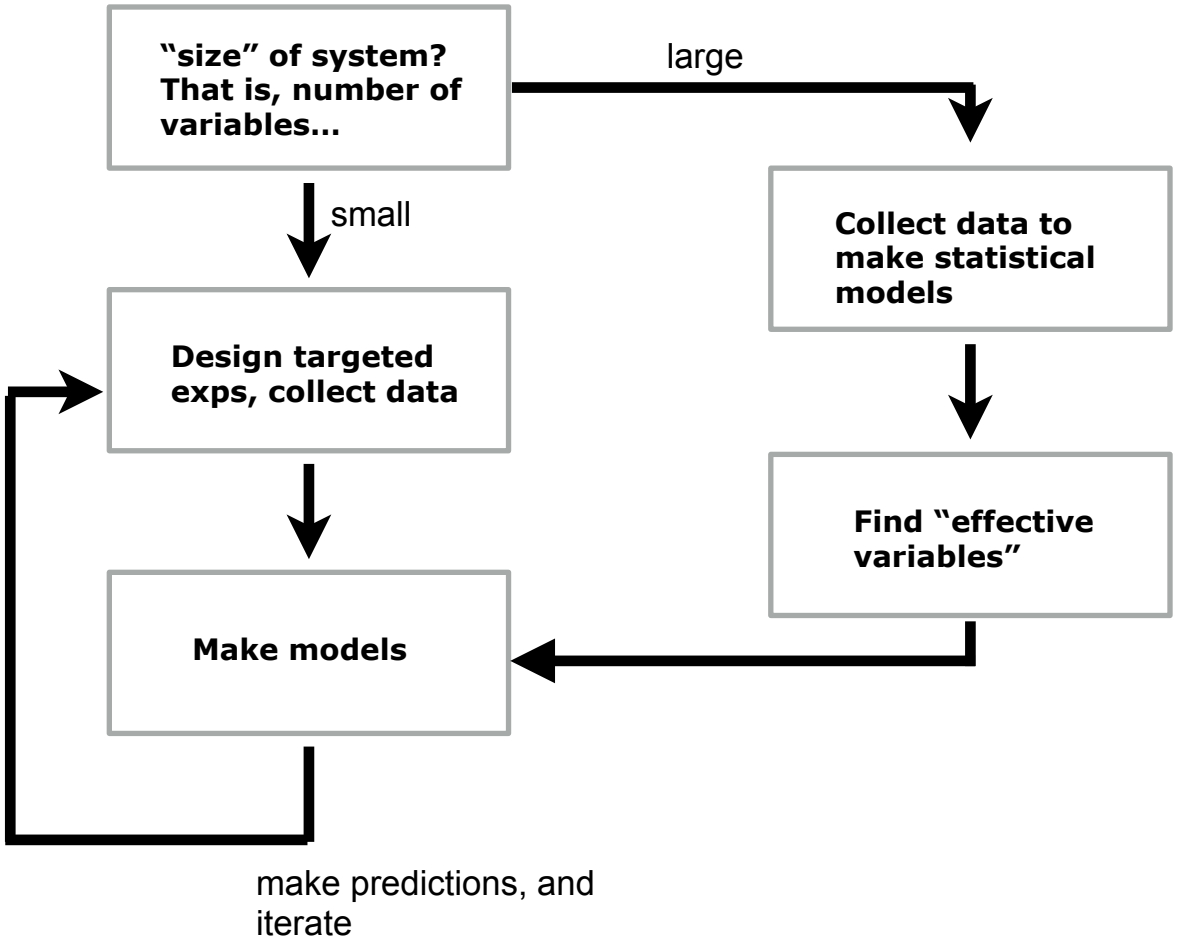
$$\bar{\omega} = \Omega_{\text{epi}}\, \bar{y}$$

(1) **Enumerate** the phenotype vector experimentally and then compute the importance and non-linearity of positions (in the epistasis vector).

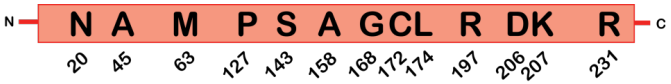
(2) Find the **right way to sub-sample** the phenotype vector so that it well-approximates the epistasis vector.

(3) Find a way to **directly estimate the epistasis vector** through some other kind of strategy…

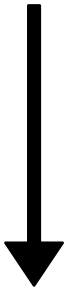
What about **large systems**?



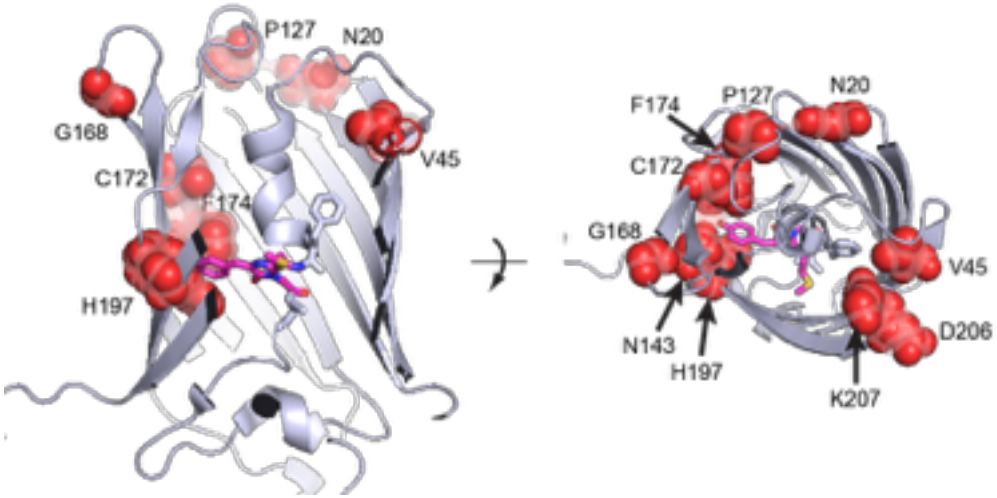
A "small-scale" experiment…to build some intuition



mKate2, ex/em 570nm/633nm



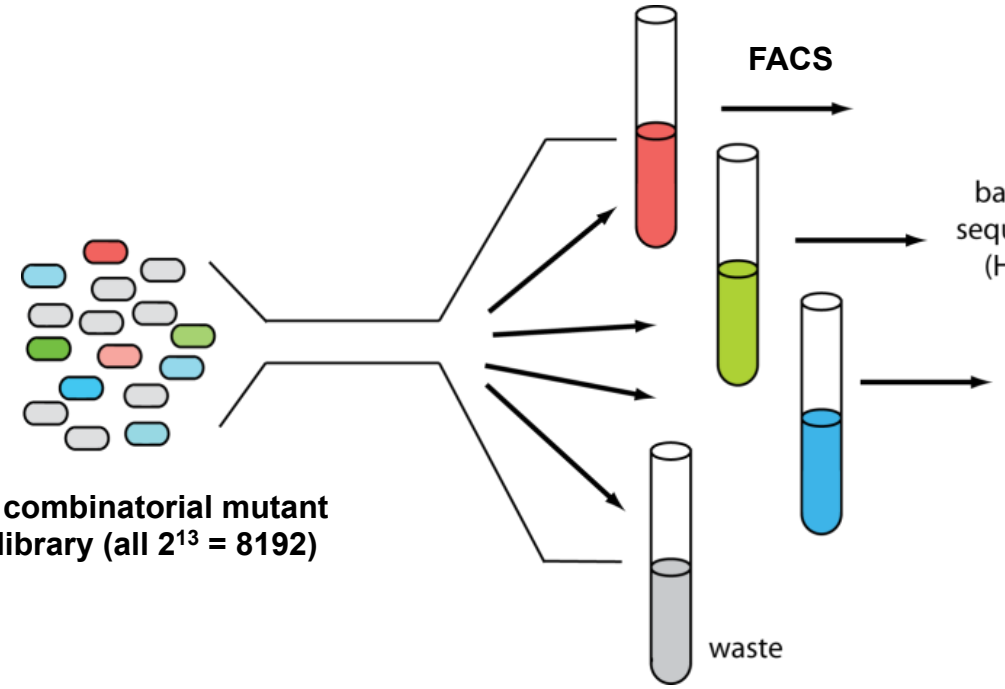
mTagBFP2, ex/em 405nm/456nm



2^{13} or 8,192 total genotypes…linking the red and blue proteins.

F. Poelwijk

Red FP to a Blue FP in 13 mutations....



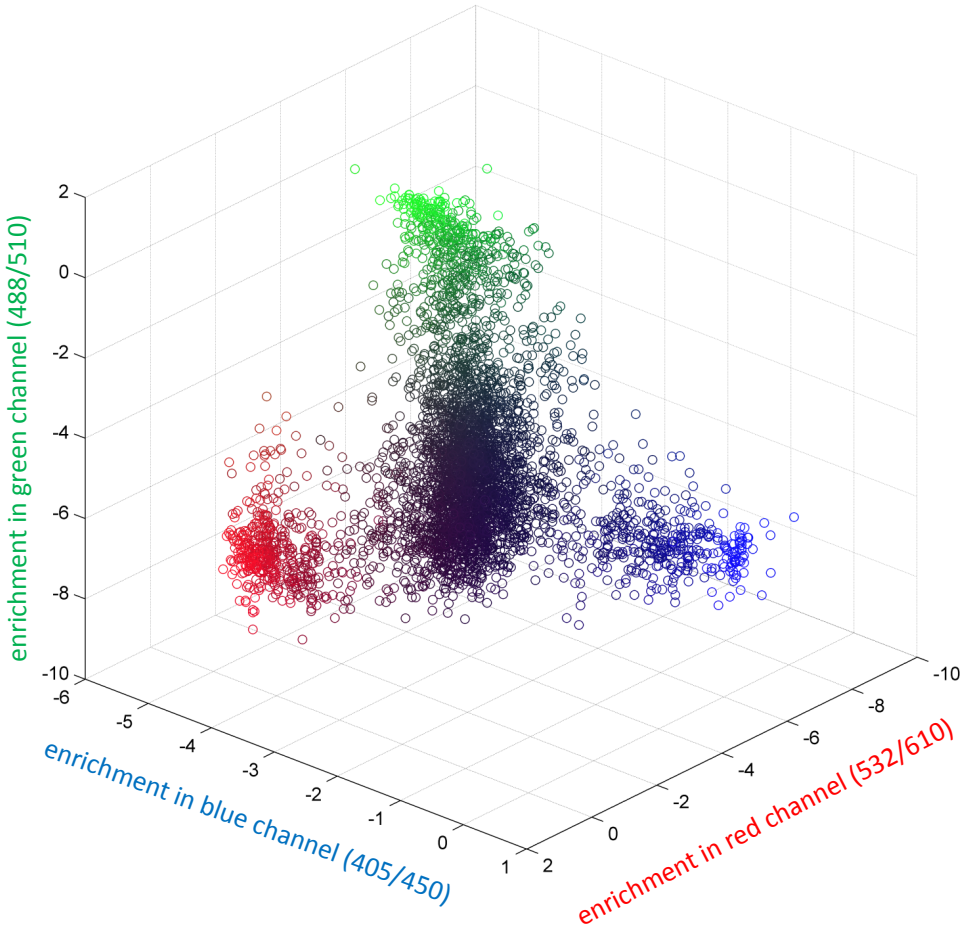
create combinatorial mutant library (all 2^{13} = 8192)

computational analysis (quality filter, error correction…)

$$E_i = \log \frac{N_i^{\text{sel}}}{N_i^{\text{unsel}}} \sim \text{brightness}$$

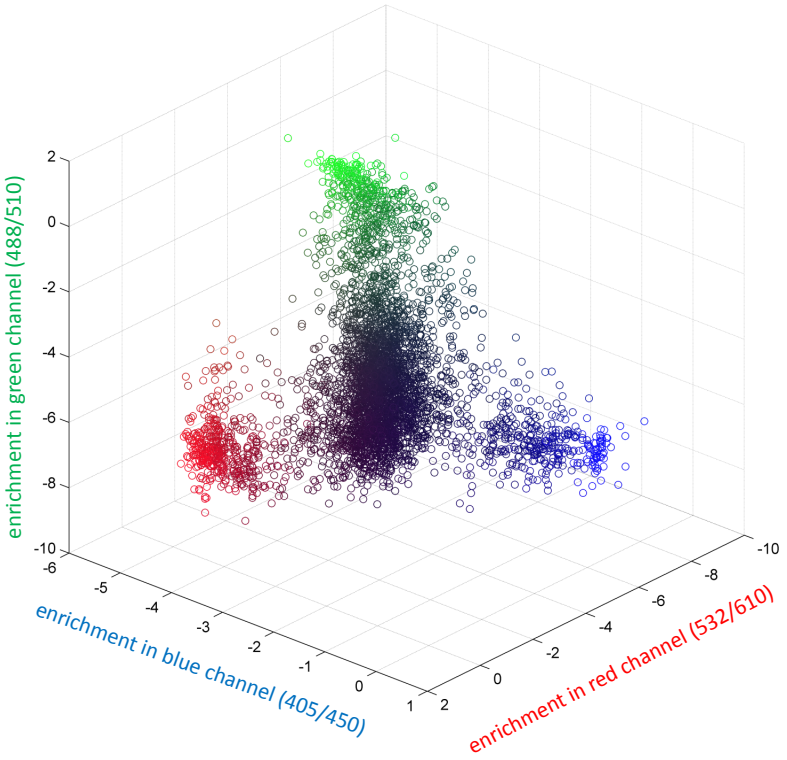
F. Poelwijk

Red FP to a Blue FP in 13 mutations....



F. Poelwijk

For Red FP to Blue FP:



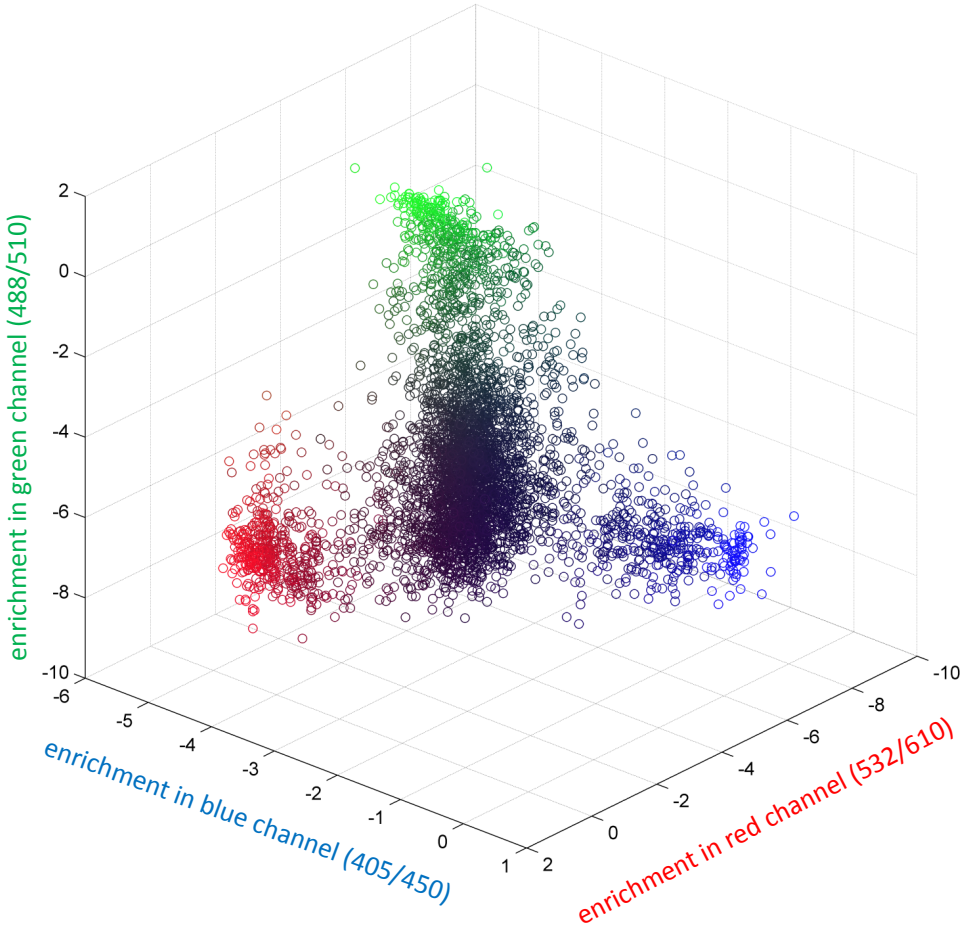
the phenotypes (2^{13})

$$\bar{\omega} = \boldsymbol{\Omega}_{\text{epi}}\, \bar{\boldsymbol{y}}$$

we can compute the epistases up to the 13th order!

F. Poelwijk

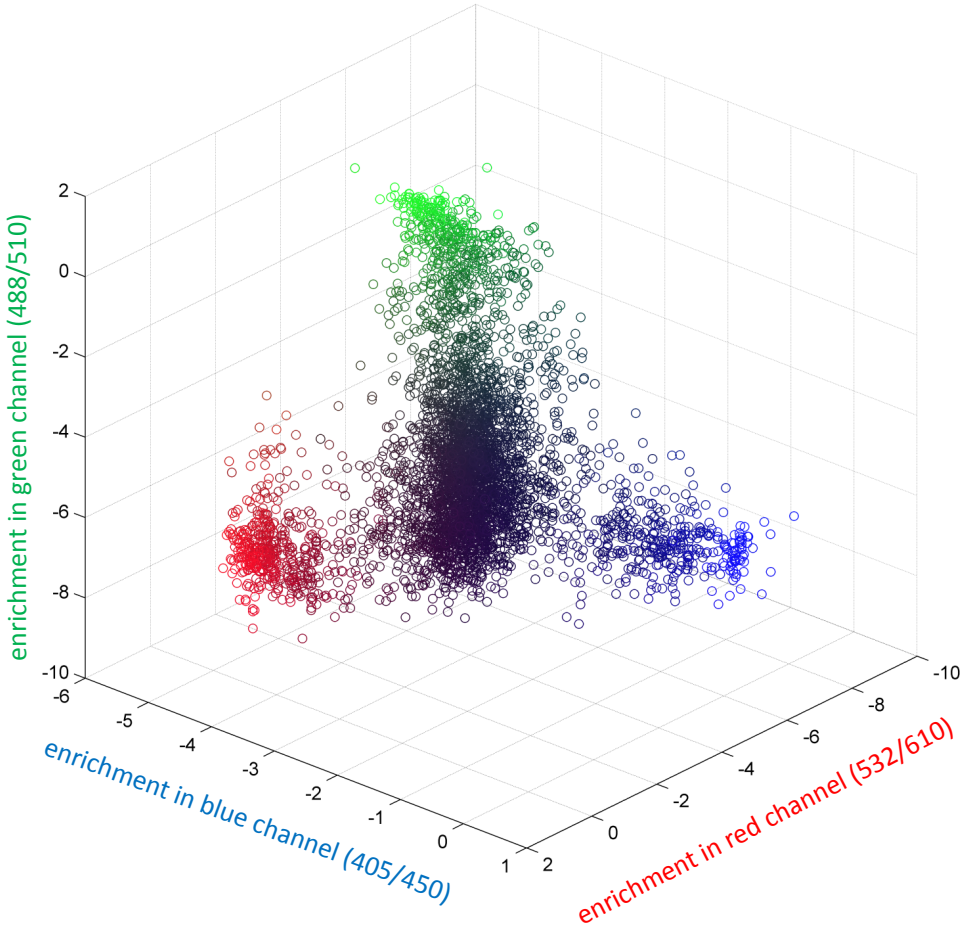
Red FP to a Blue FP in 13 mutations....



There are **three** main results:

F. Poelwijk

Red FP to a Blue FP in 13 mutations....

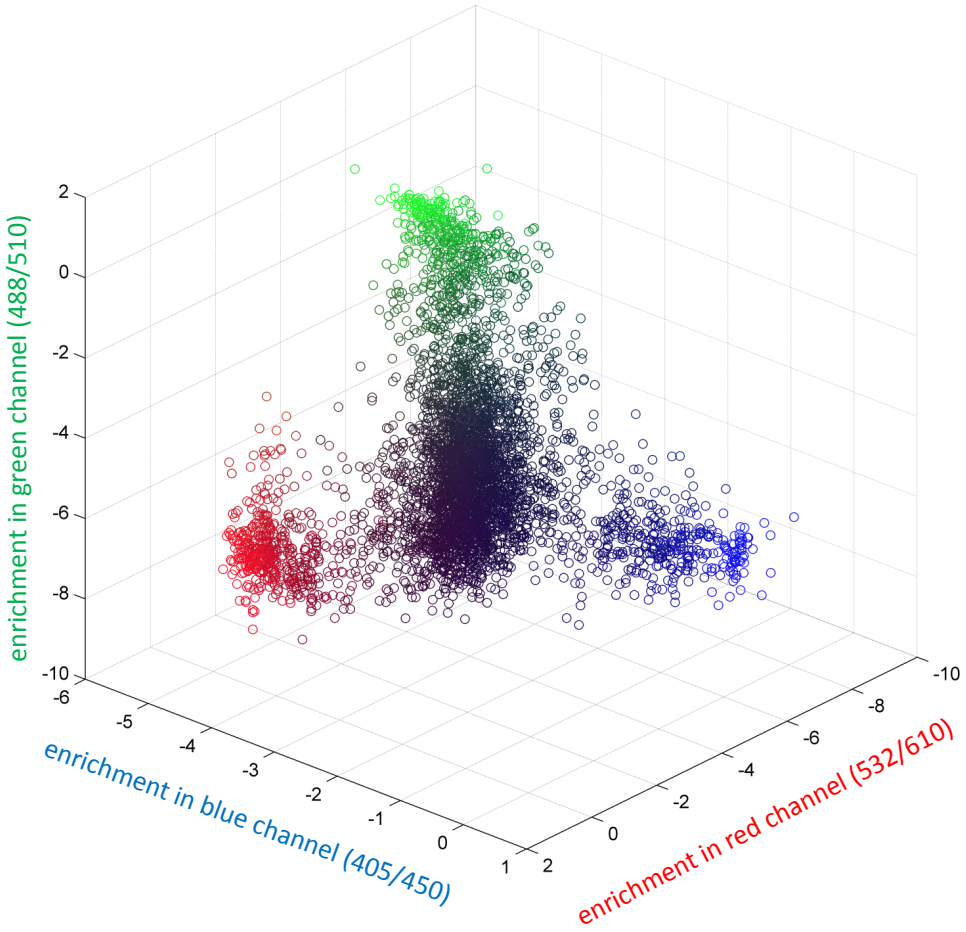


There are **three** main results:

(1) ~2,000 of 8,192 are discernably fluorescent (already tells you there is epistasis)

F. Poelwijk

Red FP to a Blue FP in 13 mutations....



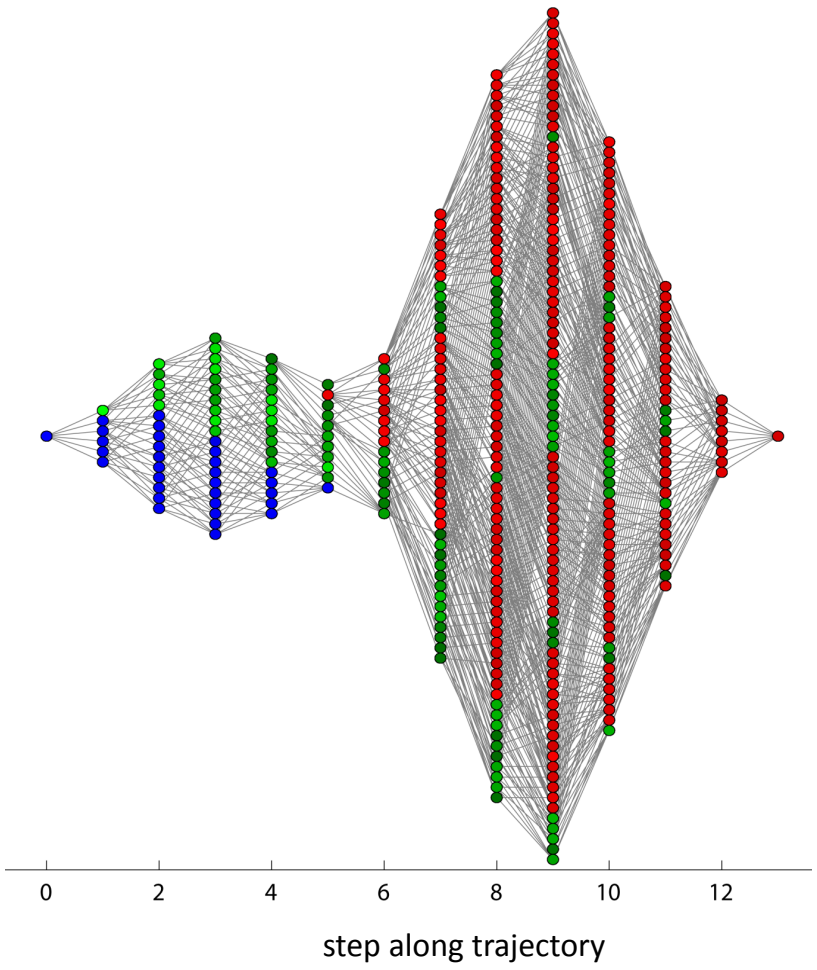
There are **three** main results:

(1) ~2,000 of 8,192 are discernably fluorescent (already tells you there is epistasis)

(2) There are non-trivial couplings up to the 6th-order between sequence positions!

F. Poelwijk

Red FP to a Blue FP in 13 mutations....



There are **three** main results:

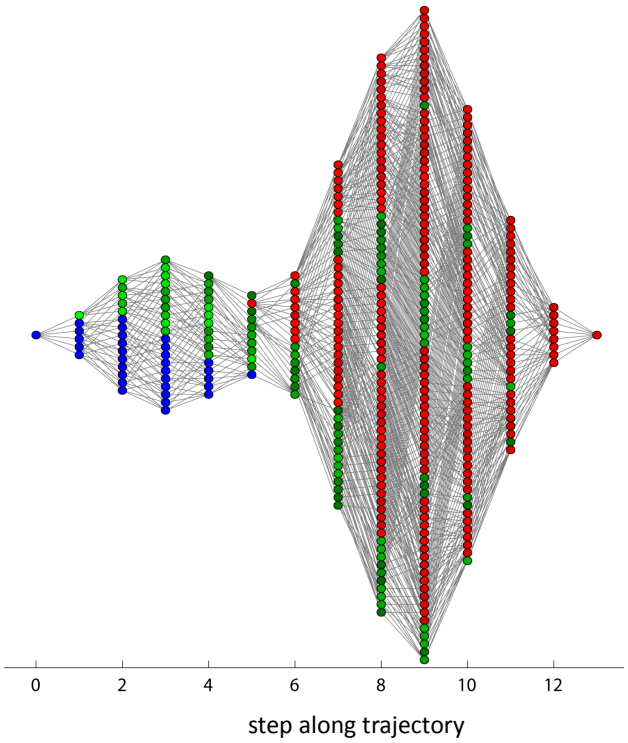
(1) ~2,000 of 8,192 are discernably fluorescent (already tells you there is epistasis)

(2) There are non-trivial couplings up to the 6th-order between sequence positions!

(3) But, the space of fluorescent proteins is fully connected by a path of single mutations....

F. Poelwijk

Red FP to a Blue FP in 13 mutations....



$$\bar{\omega} = \Omega_{\text{epi}} \bar{y}$$

background averaged epistasis

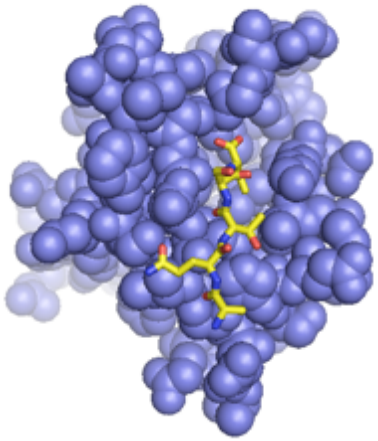
$$\Omega_{\text{epi}} = \boldsymbol{V}\boldsymbol{H}$$

$$\Omega_{\text{epi}} = \boldsymbol{V}\boldsymbol{X}^{T}\boldsymbol{H}$$

single reference epistasis

F. Poelwijk

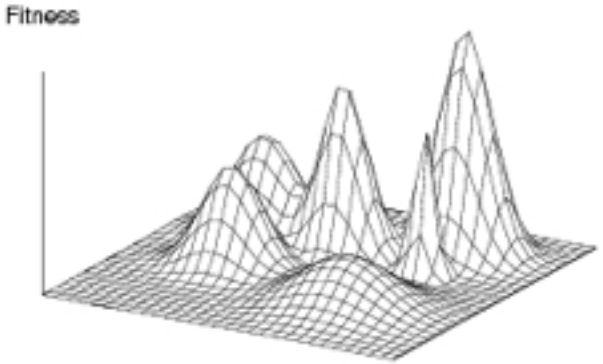
The bottom line…



$$\bar{\omega} = \Omega_{\text{epi}}\,\bar{y}$$

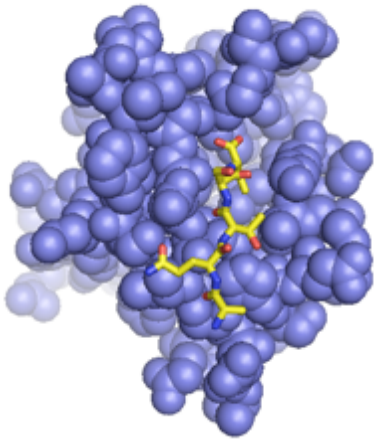
$$\Omega_{\text{epi}} = \boldsymbol{V}\boldsymbol{X}^{T}\boldsymbol{H}$$

this view of epistasis is a "**local**" one…taking a single genotype as an arbitrary reference.



…its like doing a **Taylor's** (local) expansion of the fitness landscape around a particular point (the reference genotype). A detailed analysis of a particular solution.

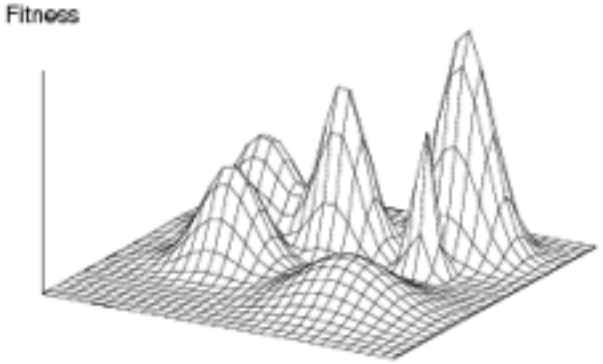
The bottom line…



$$\bar{\omega} = \Omega_{\text{epi}}\,\bar{y},$$

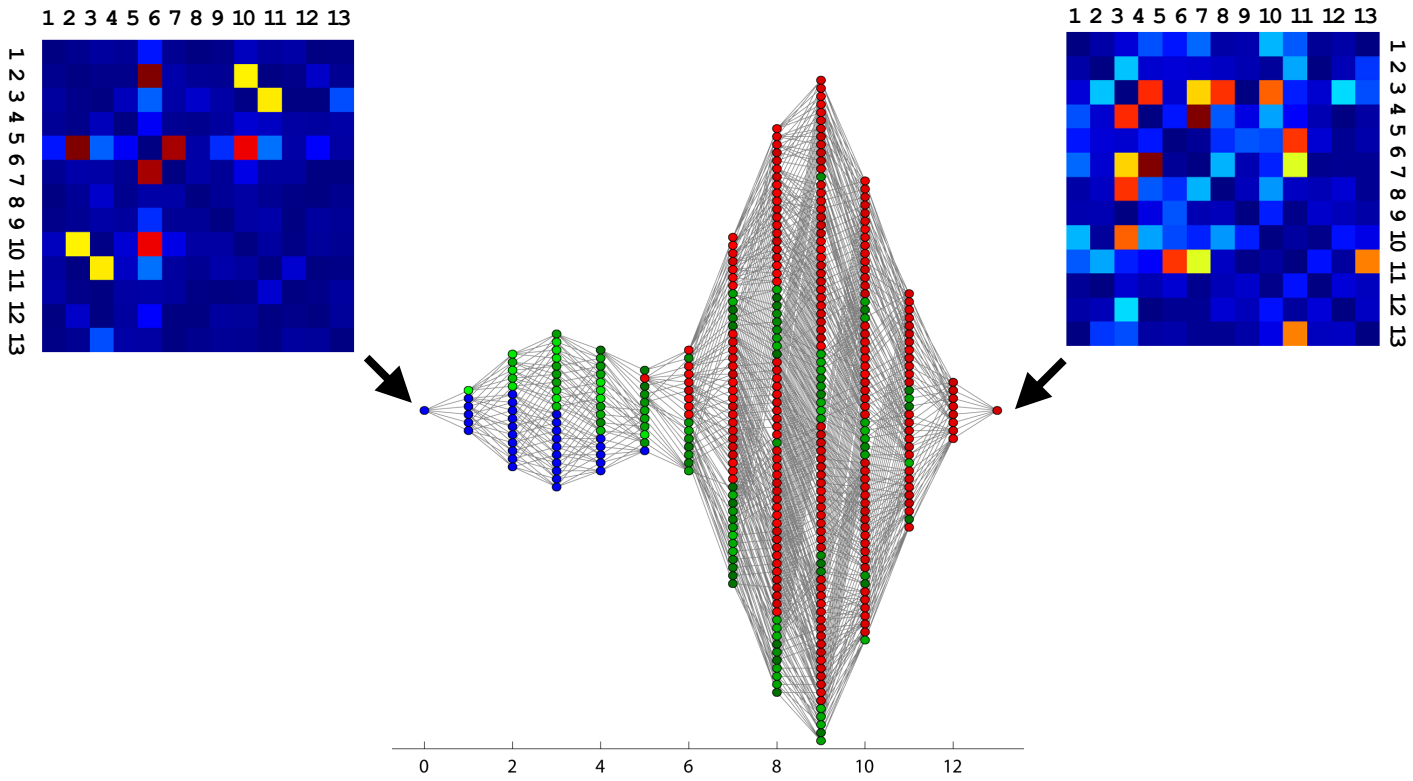
$$\Omega_{\text{epi}} = \boldsymbol{V}\boldsymbol{H}$$

the background averaged view of epistasis is a "**global**" one…taking averages over all possible genotypes.



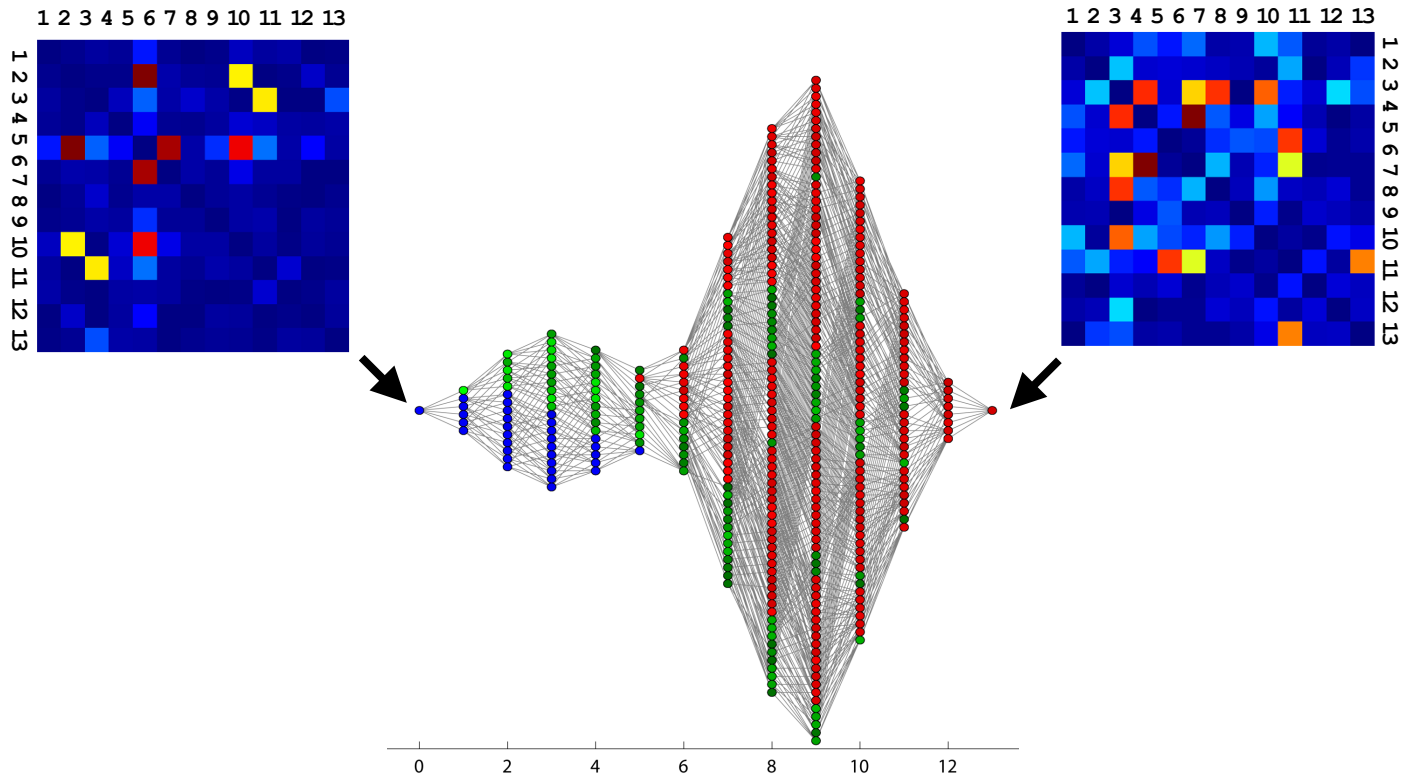
…its like doing a generalized **Fourier** expansion of the fitness landscape. A global analysis of all possible solutions given the design process.

So…**single reference (biochemical) epistasis**



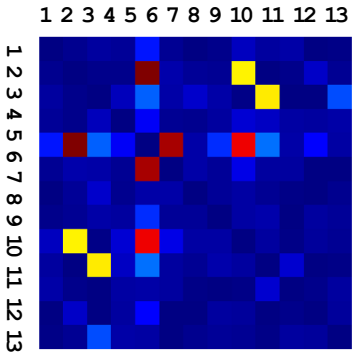
…just the **2nd order (pairwise)** terms in epistasis

So…**single reference (biochemical) epistasis**

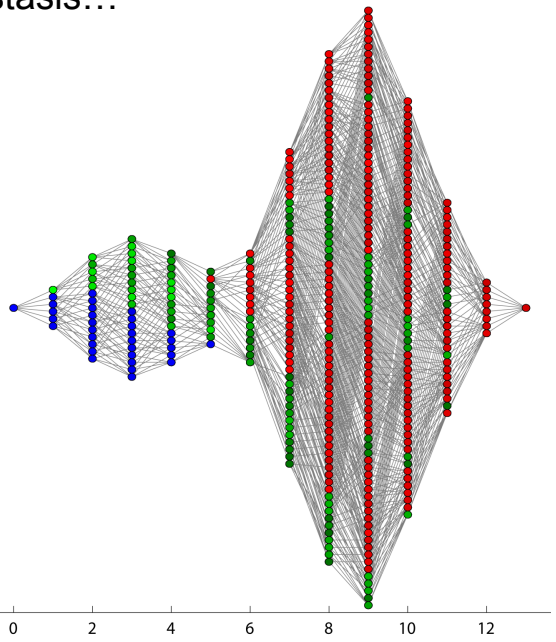


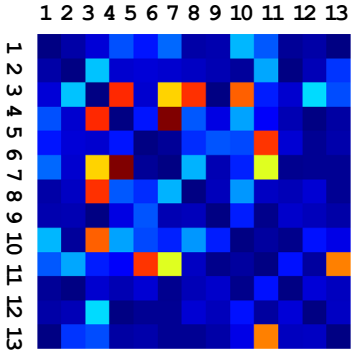
…depends strongly on reference genotype. This reveals the local epistatic structure around these genotypes

And **background averaged** epistasis…

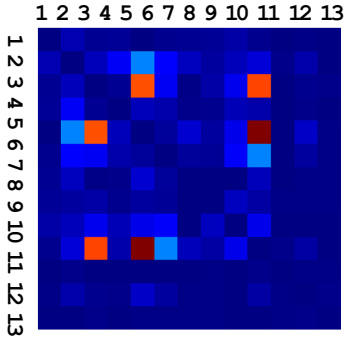


Blue reference



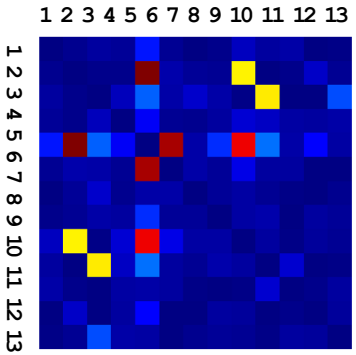


Red reference

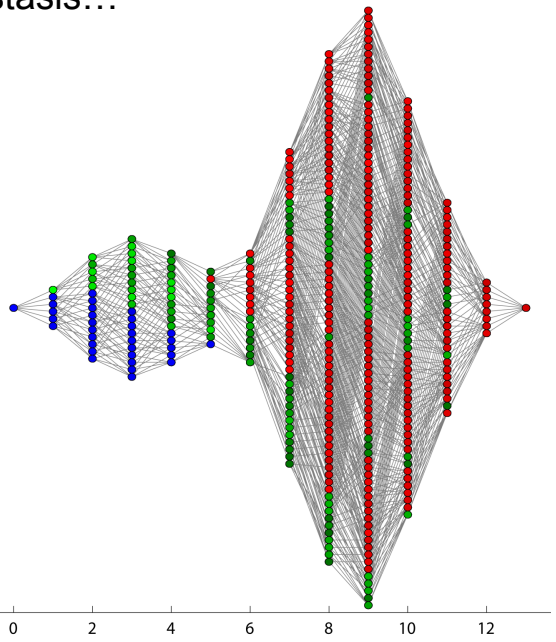


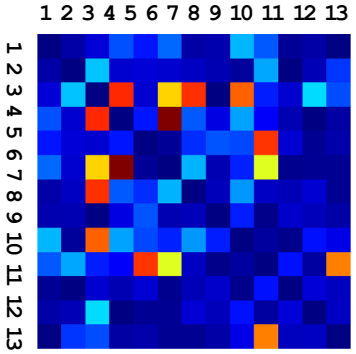
No reference

And **background averaged** epistasis…

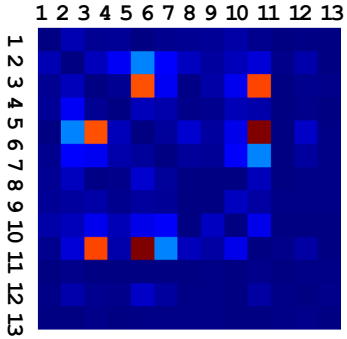


Blue reference





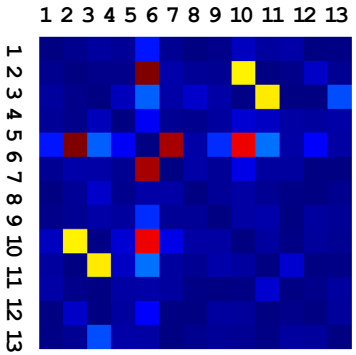
Red reference



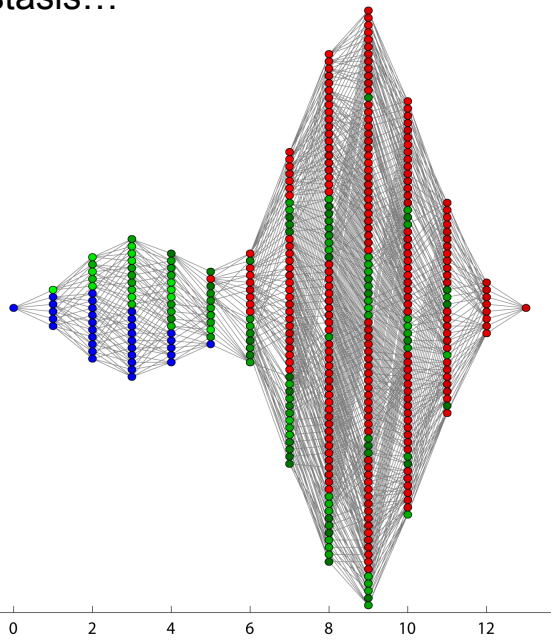
No reference

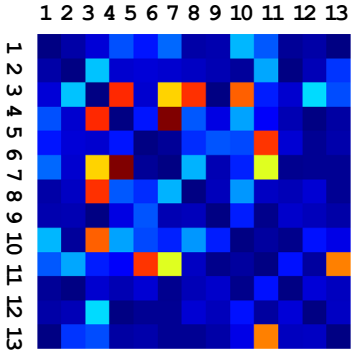
This is the **2nd order epistasis** averaged over the space of functional sequences….

And **background averaged** epistasis…

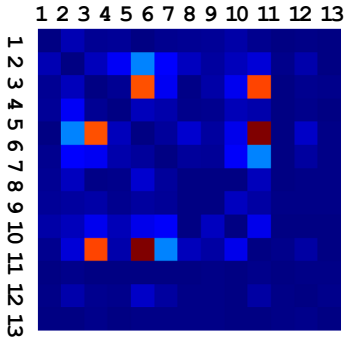


Blue reference



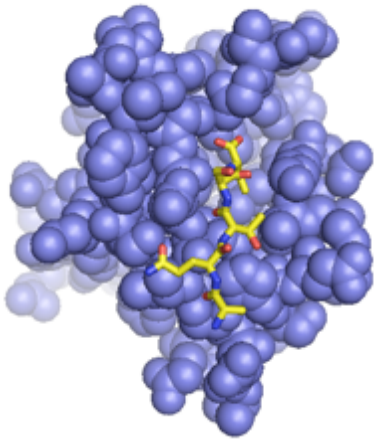


Red reference



No reference

Note the **sparsity of epistasis** and the connectivity of solutions!

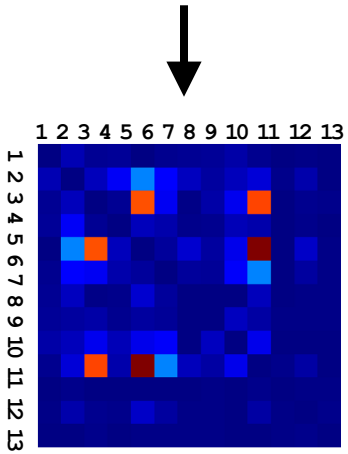


(1) how can we get this background averaged epistasis for the **general case**?

(2) What do we learn from **just the second order terms**? Remember that there are higher-order terms there…that's why the local pairwise terms are different!

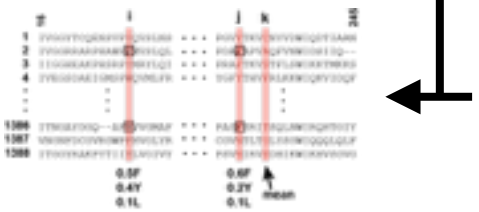
How can we get **background averaged** epistasis?

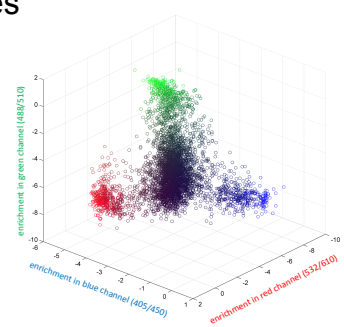
Average of epistasis from fluorescence data over all 8192 genotypes



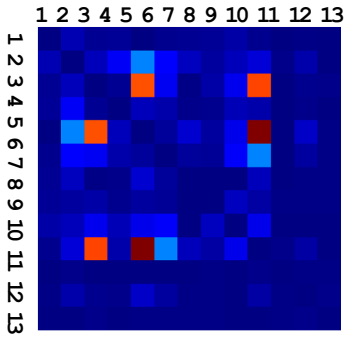
How can we get **background averaged** epistasis?

Alignment of 2000 functional genotypes

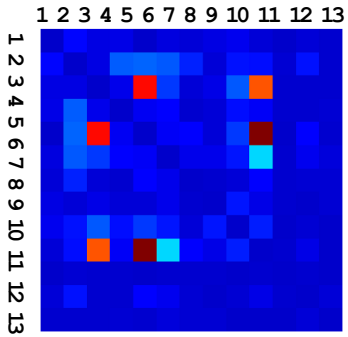




Average of epistasis from fluorescence data over all 8192 genotypes

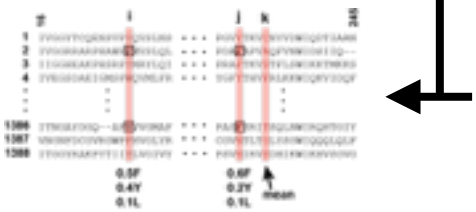


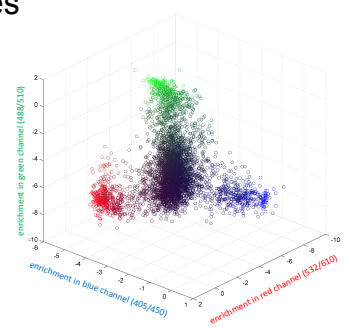
$$C_{i,j} = f_{i,j} - f_i f_j$$



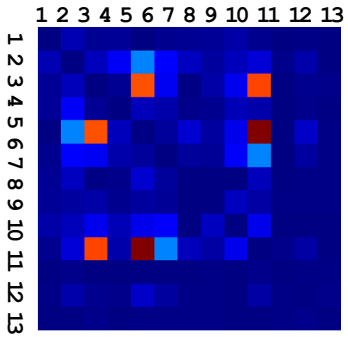
How can we get **background averaged** epistasis?

Alignment of 2000 functional genotypes

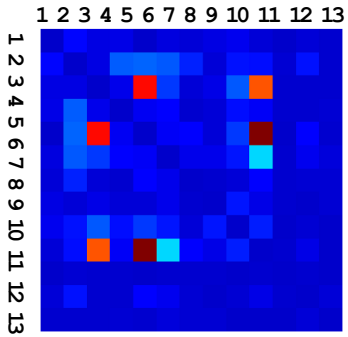




Average of epistasis from fluorescence data over all 8192 genotypes



$$C_{i,j} = f_{i,j} - f_i f_j$$



Experimental average over genotypes that satisfy some functional selection…

Statistical average over genotypes that satisfy some functional selection…

So, next time, the **statistical approach to epistasis**…

	n = 1	n = 2 or 3	n >> 1	continuum
Linear	exponential growth and decay single step conformational change fluorescence emission pseudo first order kinetics	second order reaction kinetics linear harmonic oscillators simple feedback control sequences of conformational change	electrical circuits molecular dynamics systems of coupled harmonic oscillators equilibrium thermodynamics diffraction, Fourier transforms	Diffusion Wave propagation quantum mechanics viscoelastic systems
Nonlinear	fixed points bifurcations, multi stability irreversible hysteresis overdamped oscillators	anharmomic oscillators relaxation oscillations predator-prey models van der Pol systems Chaotic systems	systems of non-linear oscillators non-equilibrium thermodynamics protein structure/ function neural networks the cell ecosystems	Nonlinear wave propagation Reaction-diffusion in dissipative systems Turbulent/chaotic flows

adapted from S. Strogatz