Lecture 12: Epistasis and the General Problem

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

So…the full complexity of interactions within systems.

Sewall Wright
1889 - 1988

Ronald Fisher
1890 - 1962

Jeffries Wyman
1901 - 1995

Alan Fersht
1943 -

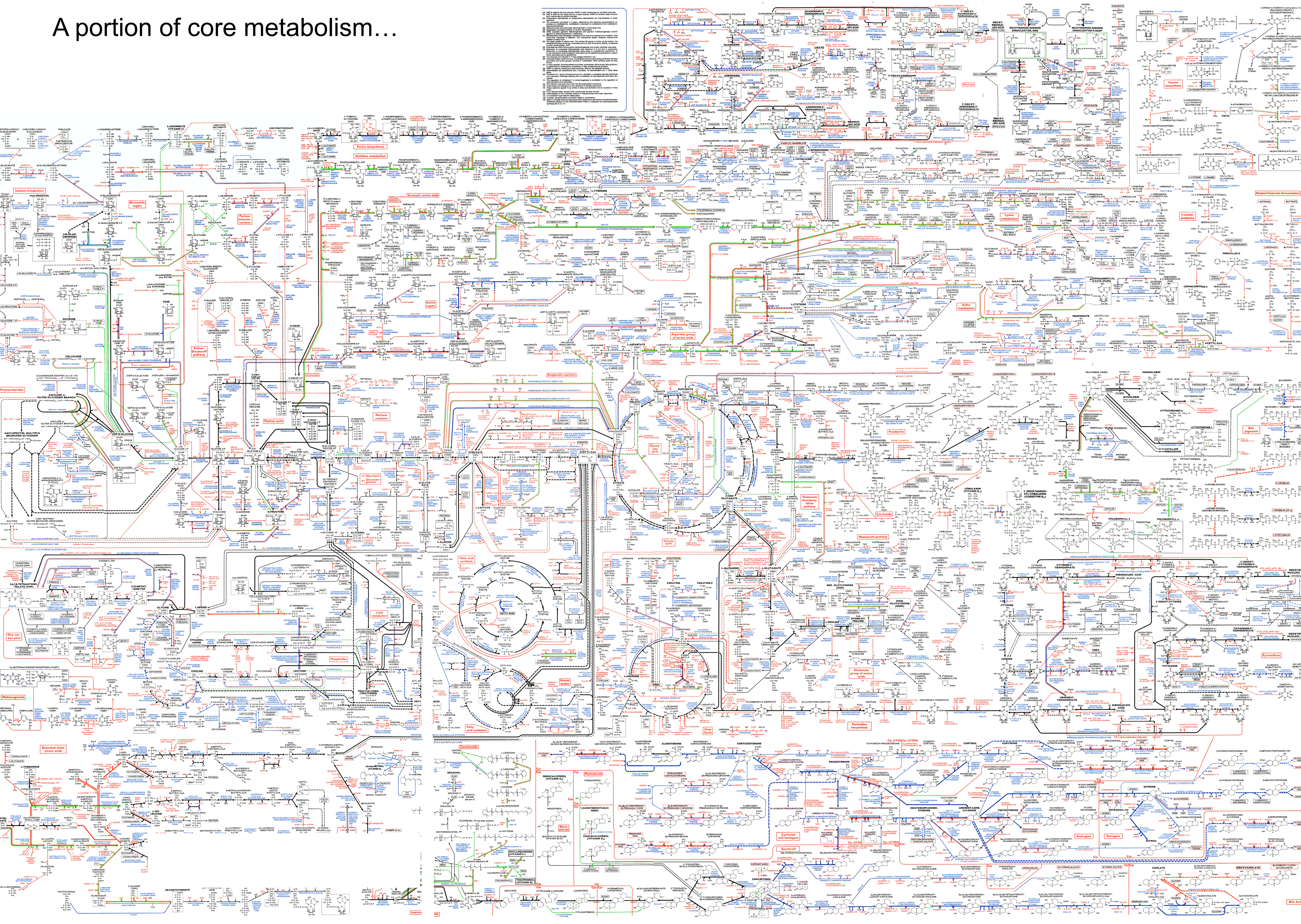
?

So, large-scale **non-linear dynamical systems**. First, we need a general theory of interactions…

	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	anharmonic 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

A portion of core metabolism…



the essential characteristics of **complex systems**…

<u>heterogeneity</u> and **<u>non-linearity</u>**…

the essential characteristics of **complex systems**…

<u>heterogeneity</u> and **<u>non-linearity</u>**…

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

complex systems show…

heterogeneity and **non-linearity**…

so, for biological systems, what are the **relevant parts**? And what is the **extent of non-linearity** in the interactions between parts?

complex systems show…

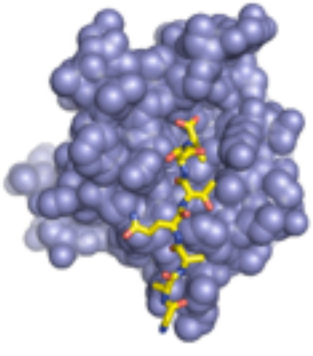
<u>heterogeneity</u> and **<u>non-linearity</u>**…

so, for biological systems, what are the **relevant parts**? And what is the
extent of non-linearity in the interactions between parts?

How can we design **a process** to systematically address these questions?
For next time…

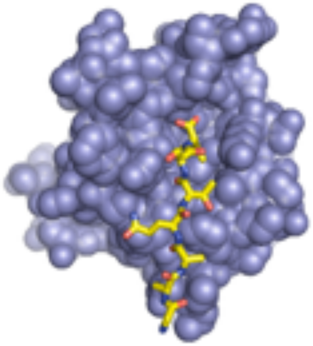
What is the **extent of non-linearity** in the interactions between parts? We need a general formalism for "interactions"…

A general theory of **epistasis**



…in the context of proteins for clarity, but the ideas are completely general.

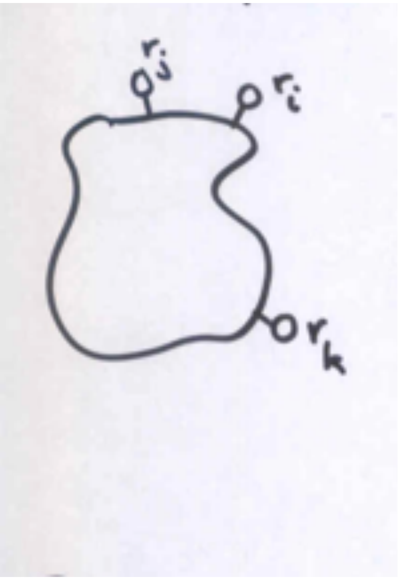
A general theory of **epistasis**



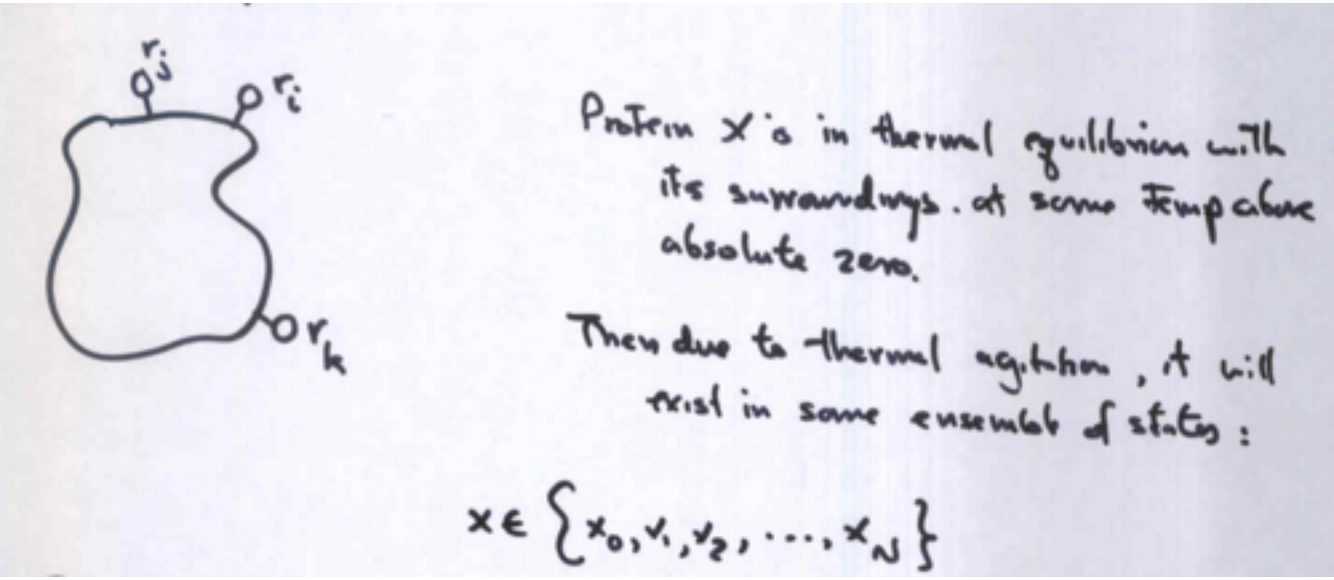
~100 amino acids

Epistasis is defined as the **non-independence** of the parts that make up a system.

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



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



Protein x is in thermal equilibrium with its surroundings. at some temp above absolute zero.

Then due to thermal agitation, it will exist in some ensemble of states:

$$x \in \{x_0, x_1, x_2, \ldots, x_N\}$$

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

Protein X is in thermal equilibrium with its surroundings. at some temp above absolute zero.

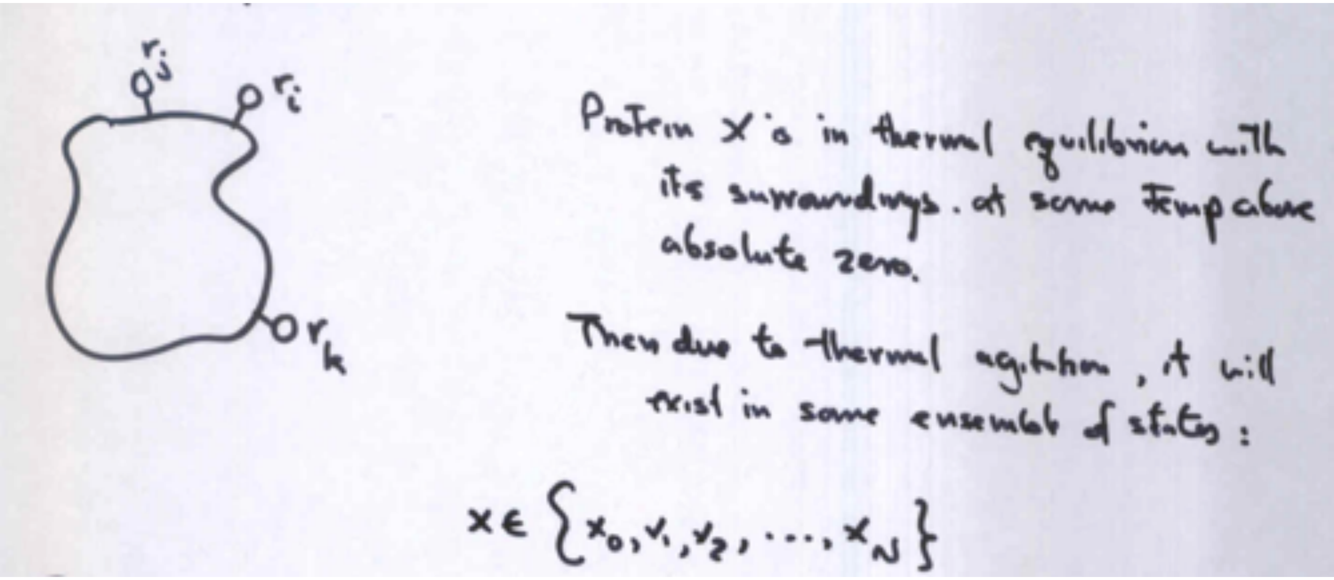
Then due to thermal agitation, it will exist in some ensemble of states:

$$x \in \{x_0, x_1, x_2, \dots, x_N\}$$

Take x_0 as the lowest energy state ... most populated at equilibrium. We know that:

$$\frac{P(x_n)}{P(x_0)} = e^{-\Delta G_n/kT}$$

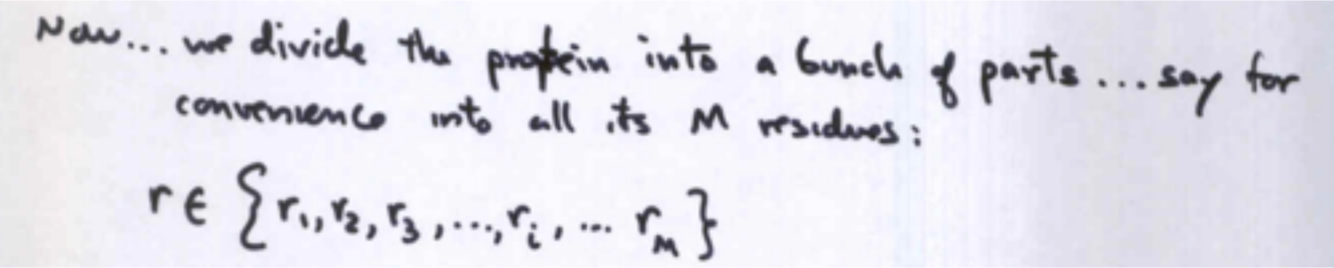
What does **non-independence** mean, exactly?



Protein X's in thermal equilibrium with its surroundings. at some temp above absolute zero.

Then due to thermal agitation, it will exist in some ensemble of states:

$$x \in \{x_0, x_1, x_2, \dots, x_N\}$$

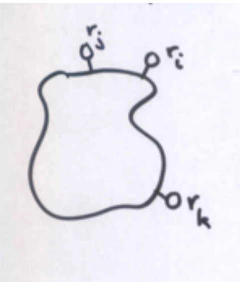


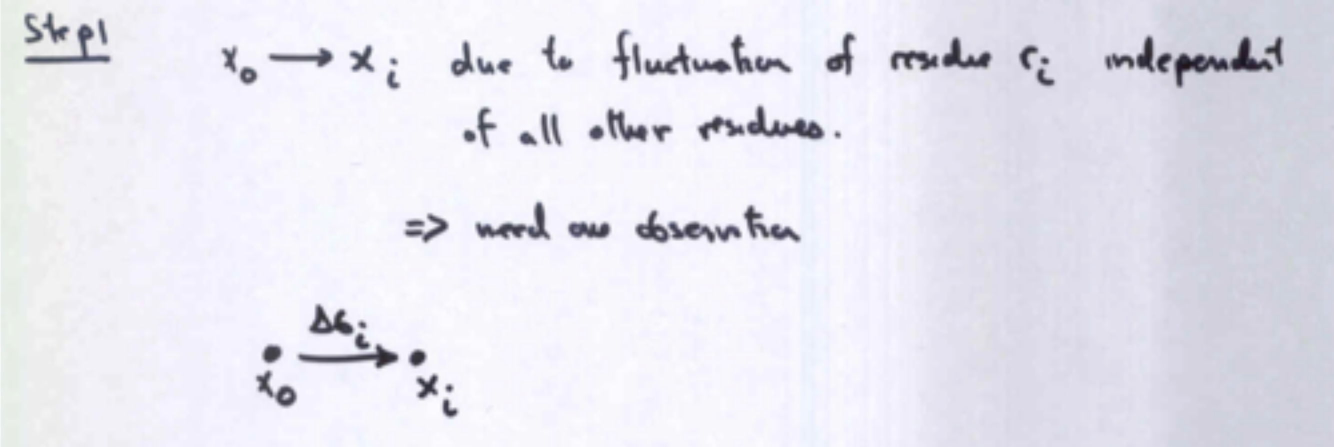
Now... we divide the protein into a bunch of parts... say for convenience into all its M residues:

$$r \in \{r_1, r_2, r_3, \dots, r_i, \dots r_M\}$$

How many observations of free energies of the **parts r** do we need to guarantee knowledge of **states x**?

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




Step 1

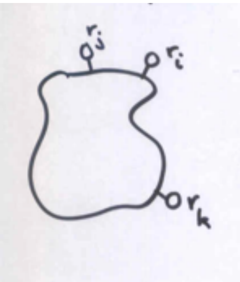
$x_0 \longrightarrow x_i$ due to fluctuation of residue r_i independent of all other residues.

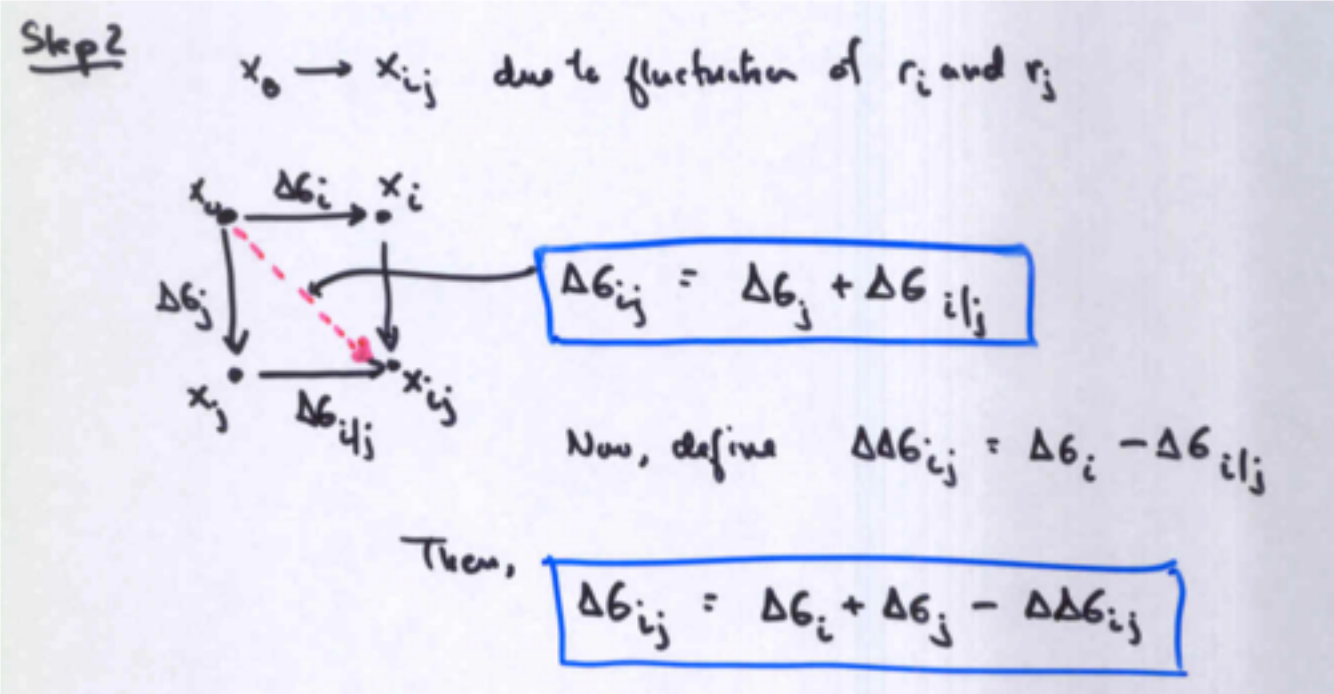
=> need one observation

$$x_0 \xrightarrow{\Delta G_i} x_i$$

….the effect of **a single part** taken independently

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





Step 2

$x_0 \rightarrow x_{ij}$ due to fluctuation of r_i and r_j

$$\begin{tikzcd}
x_0 \arrow[r, "\Delta G_i"] \arrow[d, "\Delta G_j"'] \arrow[dr, dashed] & x_i \arrow[d] \\
x_j \arrow[r, "\Delta G_{i|j}"'] & x_{ij}
\end{tikzcd}$$

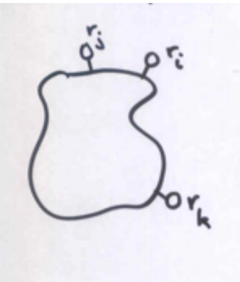
$$\boxed{\Delta G_{ij} = \Delta G_j + \Delta G_{i|j}}$$

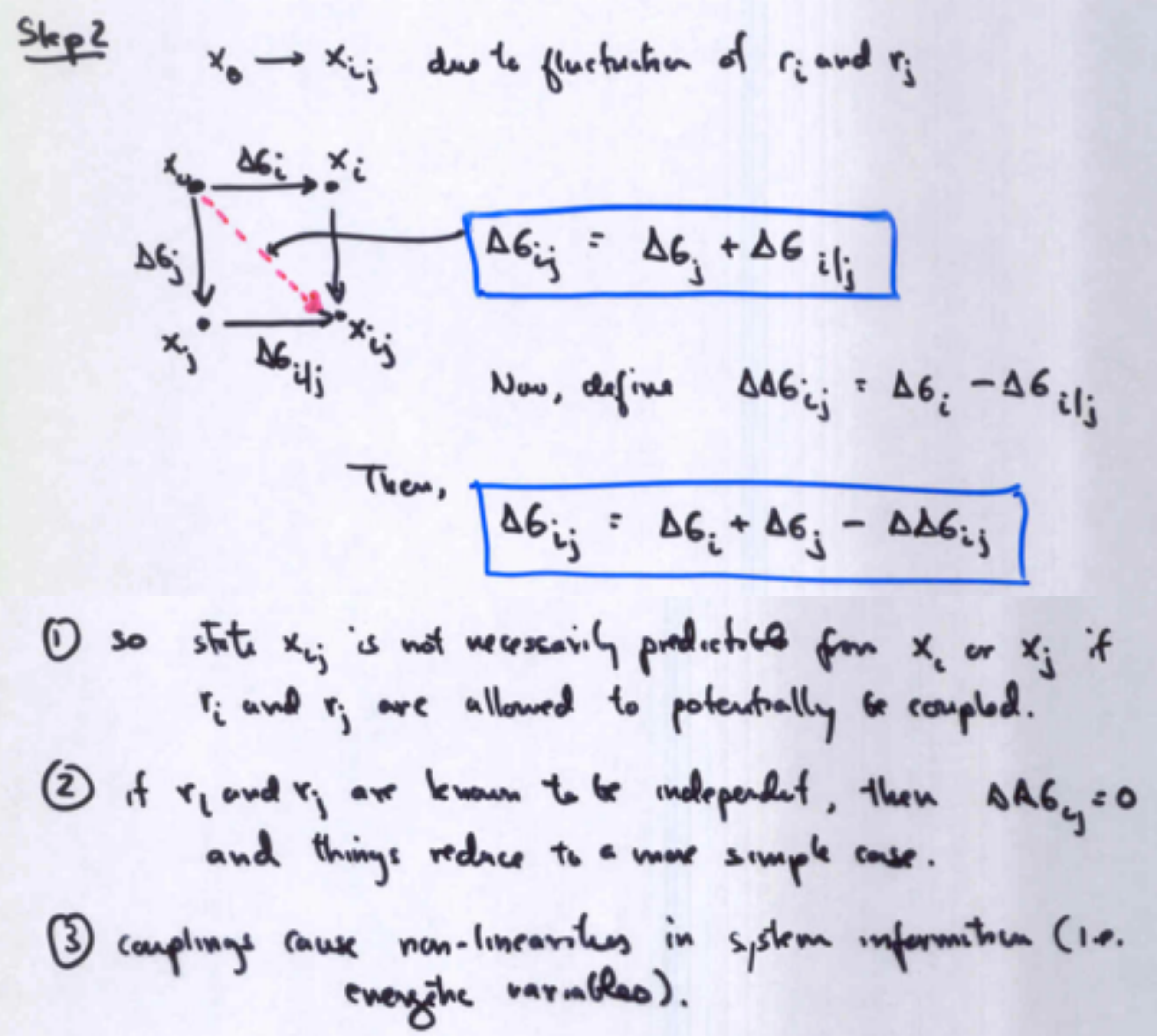
Now, define $\Delta\Delta G_{ij} = \Delta G_i - \Delta G_{i|j}$

Then, $$\boxed{\Delta G_{ij} = \Delta G_i + \Delta G_j - \Delta\Delta G_{ij}}$$

….the effect of **a two parts** acting together

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





Step 2

$x_0 \to x_{ij}$ due to fluctuation of r_i and r_j

$$\begin{tikzcd} x_0 \arrow[r, "\Delta G_i"] \arrow[d, "\Delta G_j"'] \arrow[dr, dashed] & x_i \arrow[d] \\ x_j \arrow[r, "\Delta G_{i|j}"'] & x_{ij} \end{tikzcd}$$

$$\boxed{\Delta G_{ij} = \Delta G_j + \Delta G_{i|j}}$$

Now, define $\Delta\Delta G_{ij} = \Delta G_i - \Delta G_{i|j}$

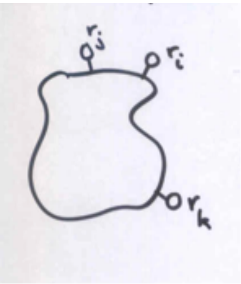
Then, $$\boxed{\Delta G_{ij} = \Delta G_i + \Delta G_j - \Delta\Delta G_{ij}}$$

① so state x_{ij} is not necessarily predictable from x_i or x_j if r_i and r_j are allowed to potentially be coupled.

② if r_i and r_j are known to be independent, then $\Delta\Delta G_{ij} = 0$ and things reduce to a more simple case.

③ couplings cause non-linearities in system information (i.e. energetic variables).

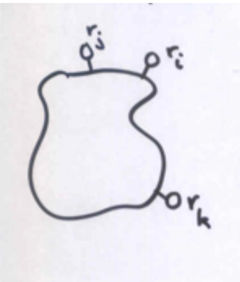
What does **non-independence** mean, exactly?

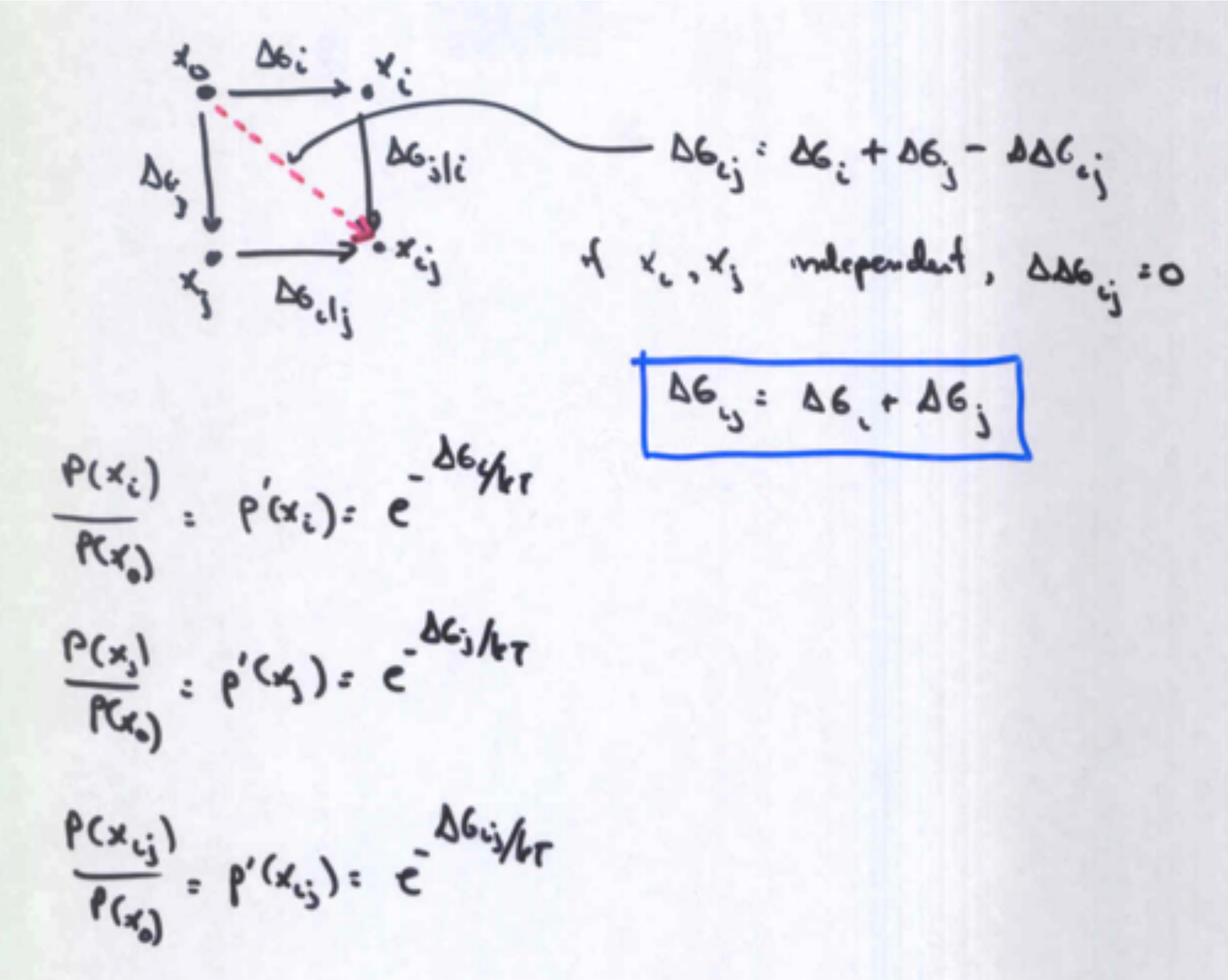


If we are measuring an **equilibrium state function**, then independence means additivity in that variable.

And, non-independence (or epistasis) means **non-additivity**

Thermodynamic additivity is **statistical independence**…





$$\begin{tikzcd}
x_0 \arrow[r, "\Delta G_i"] \arrow[d, "\Delta G_j"'] \arrow[rd, dashed] & x_i \arrow[d, "\Delta G_{j|i}"] \\
x_j \arrow[r, "\Delta G_{i|j}"'] & x_{ij}
\end{tikzcd}$$

$\Delta G_{ij} = \Delta G_i + \Delta G_j - \Delta\Delta G_{ij}$

if x_i, x_j independent, $\Delta\Delta G_{ij} = 0$

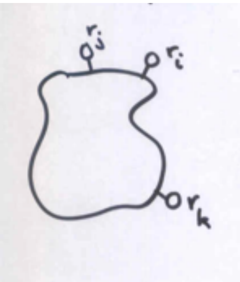
$$\boxed{\Delta G_{ij} = \Delta G_i + \Delta G_j}$$

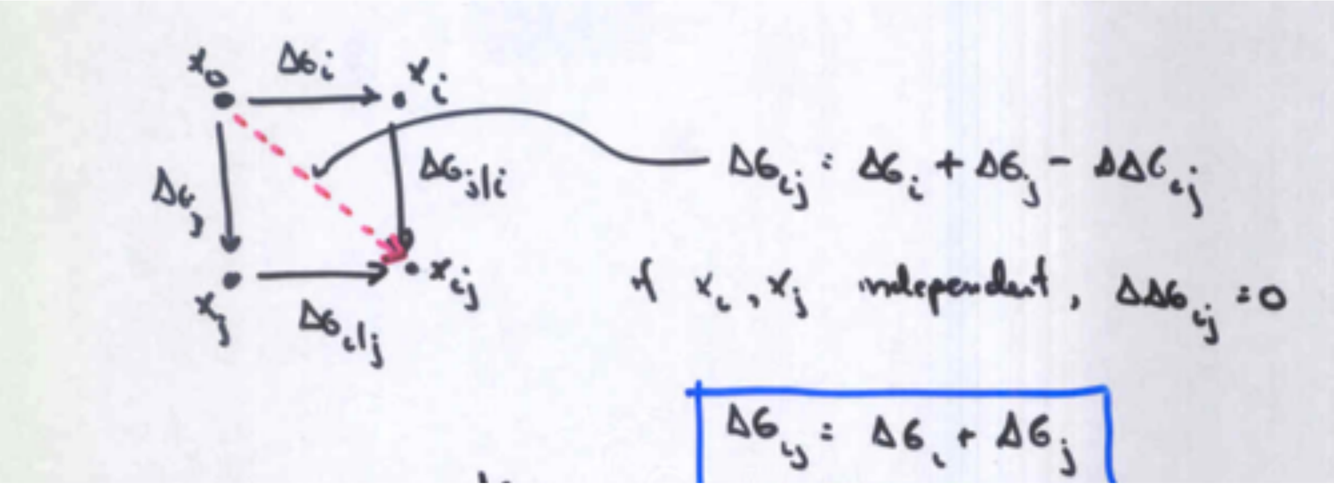
$$\frac{P(x_i)}{P(x_0)} = p'(x_i) = e^{-\Delta G_i/kT}$$

$$\frac{P(x_j)}{P(x_0)} = p'(x_j) = e^{-\Delta G_j/kT}$$

$$\frac{P(x_{ij})}{P(x_0)} = p'(x_{ij}) = e^{-\Delta G_{ij}/kT}$$

Thermodynamic additivity is **statistical independence**…



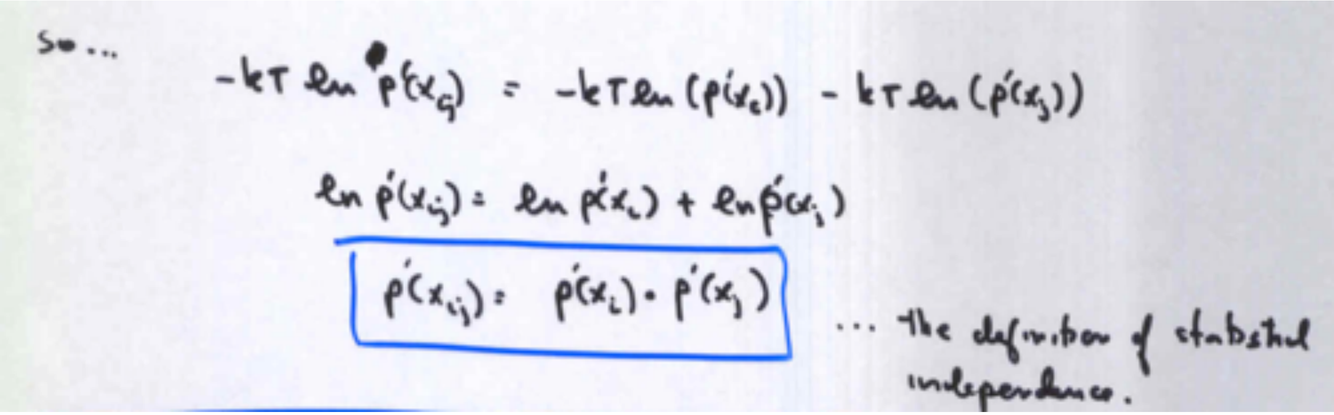


$$\begin{tikzcd}
x_0 \arrow[r, "\Delta G_i"] \arrow[d, "\Delta G_j"'] \arrow[dr, dashed] & x_i \arrow[d, "\Delta G_{j|i}"] \\
x_j \arrow[r, "\Delta G_{i|j}"'] & x_{ij}
\end{tikzcd}$$

$\Delta G_{ij} = \Delta G_i + \Delta G_j - \Delta\Delta G_{ij}$

if x_i, x_j independent, $\Delta\Delta G_{ij} = 0$

$$\boxed{\Delta G_{ij} = \Delta G_i + \Delta G_j}$$



So…

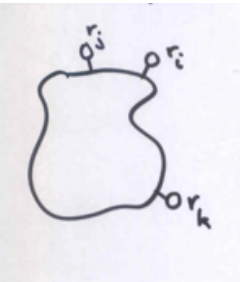
$$-kT \ln p'(x_{ij}) = -kT \ln(p'(x_i)) - kT \ln(p'(x_j))$$

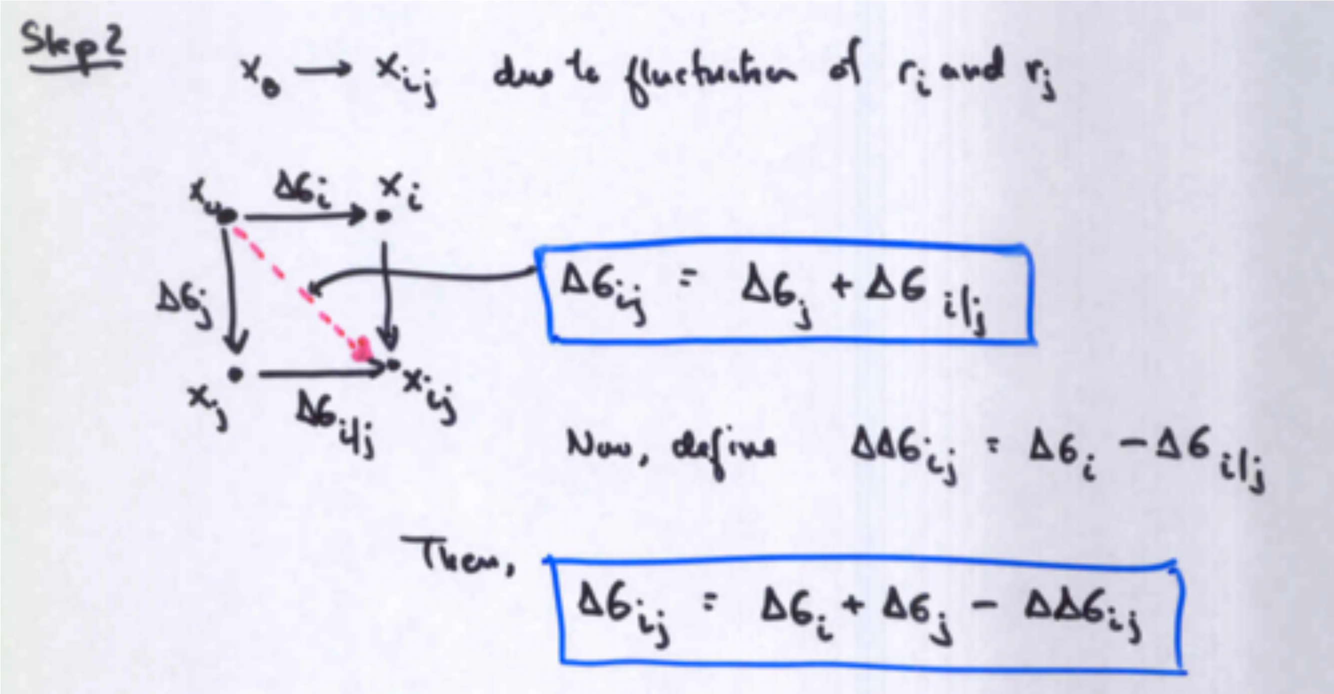
$$\ln p'(x_{ij}) = \ln p'(x_i) + \ln p'(x_j)$$

$$\boxed{p'(x_{ij}) = p'(x_i) \cdot p'(x_j)}$$

… the definition of statistical independence.

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





Step2

$x_0 \longrightarrow x_{ij}$ due to fluctuation of r_i and r_j

$$\begin{tikzcd}
x_0 \arrow[r, "\Delta G_i"] \arrow[d, "\Delta G_j"'] \arrow[dr, dashed] & x_i \arrow[d] \\
x_j \arrow[r, "\Delta G_{i|j}"'] & x_{ij}
\end{tikzcd}$$

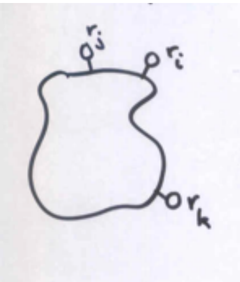
$$\boxed{\Delta G_{ij} = \Delta G_j + \Delta G_{i|j}}$$

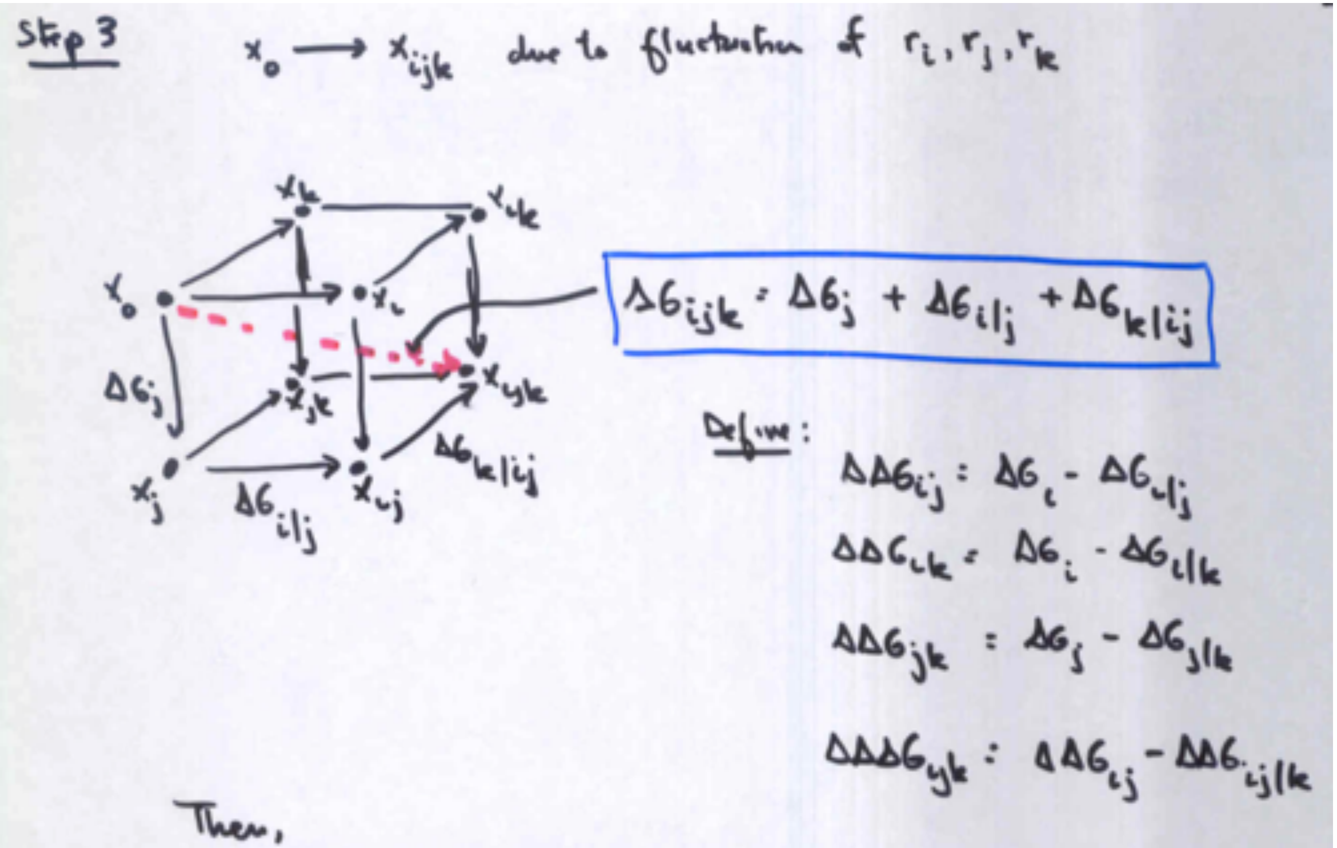
Now, define $\Delta\Delta G_{ij} = \Delta G_i - \Delta G_{i|j}$

Then, $$\boxed{\Delta G_{ij} = \Delta G_i + \Delta G_j - \Delta\Delta G_{ij}}$$

….the effect of **a two parts** acting together

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





Step 3

$x_0 \longrightarrow x_{ijk}$ due to fluctuation of r_i, r_j, r_k

$$\boxed{\Delta G_{ijk} = \Delta G_j + \Delta G_{i|j} + \Delta G_{k|ij}}$$

Define:

$$\Delta\Delta G_{ij} = \Delta G_i - \Delta G_{i|j}$$

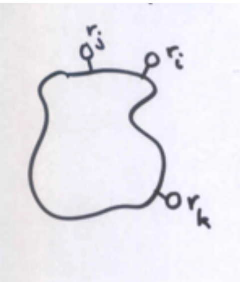
$$\Delta\Delta G_{ik} = \Delta G_i - \Delta G_{i|k}$$

$$\Delta\Delta G_{jk} = \Delta G_j - \Delta G_{j|k}$$

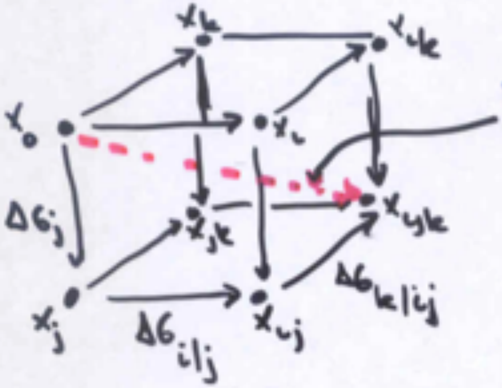
$$\Delta\Delta\Delta G_{ijk} = \Delta\Delta G_{ij} - \Delta\Delta G_{ij|k}$$

Then,

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



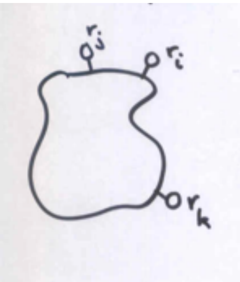
<u>Step 3</u> $x_0 \longrightarrow x_{ijk}$ due to fluctuation of r_i, r_j, r_k

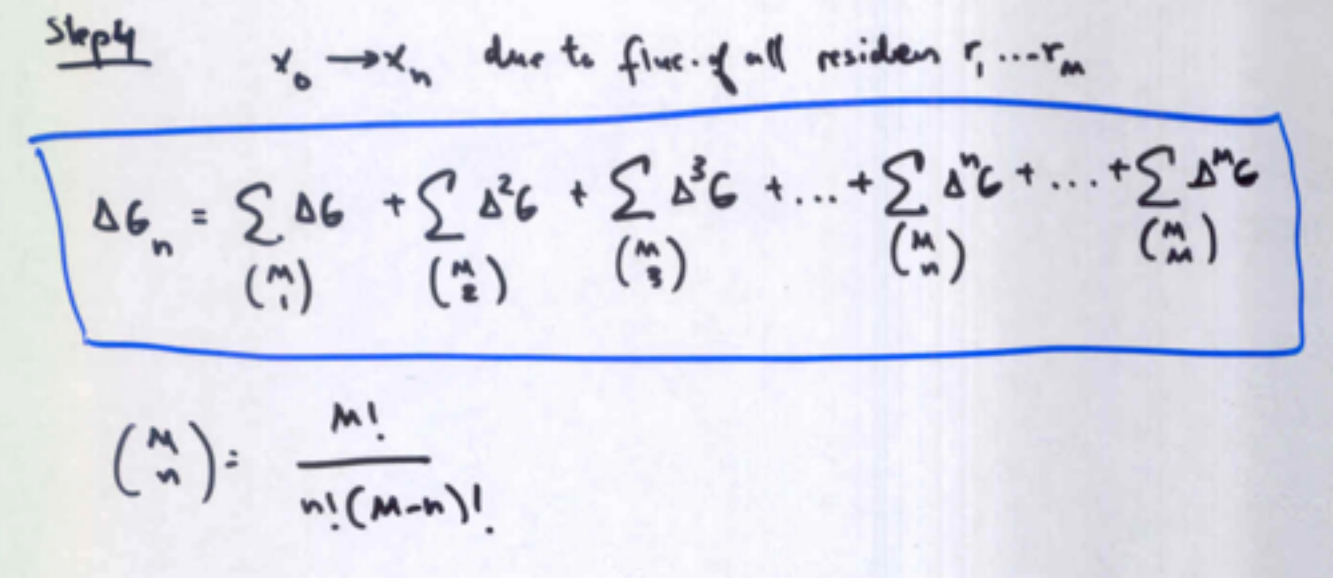


$$\boxed{\Delta G_{ijk} = \Delta G_j + \Delta G_{i|j} + \Delta G_{k|ij}}$$

$$\boxed{\Delta G_{ijk} = \underbrace{\Delta G_i + \Delta G_j + \Delta G_k}_{\text{Sum of fluc. of } r_i, r_j, r_k \text{ taken indep.}} \underbrace{- \Delta\Delta G_{ij} - \Delta\Delta G_{ik} - \Delta\Delta G_{jk}}_{\text{Sum of pairwise couplings of } r_i, r_j, r_k} \underbrace{+ \Delta\Delta\Delta G_{ijk}}_{\text{the one 3-way coupling of } r_i r_j r_k}}$$

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





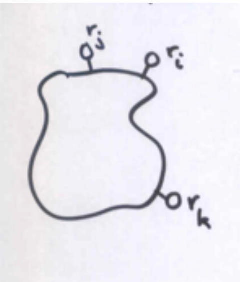
Step 4 $x_0 \rightarrow x_n$ due to fluc. of all residues $r_1 \ldots r_m$

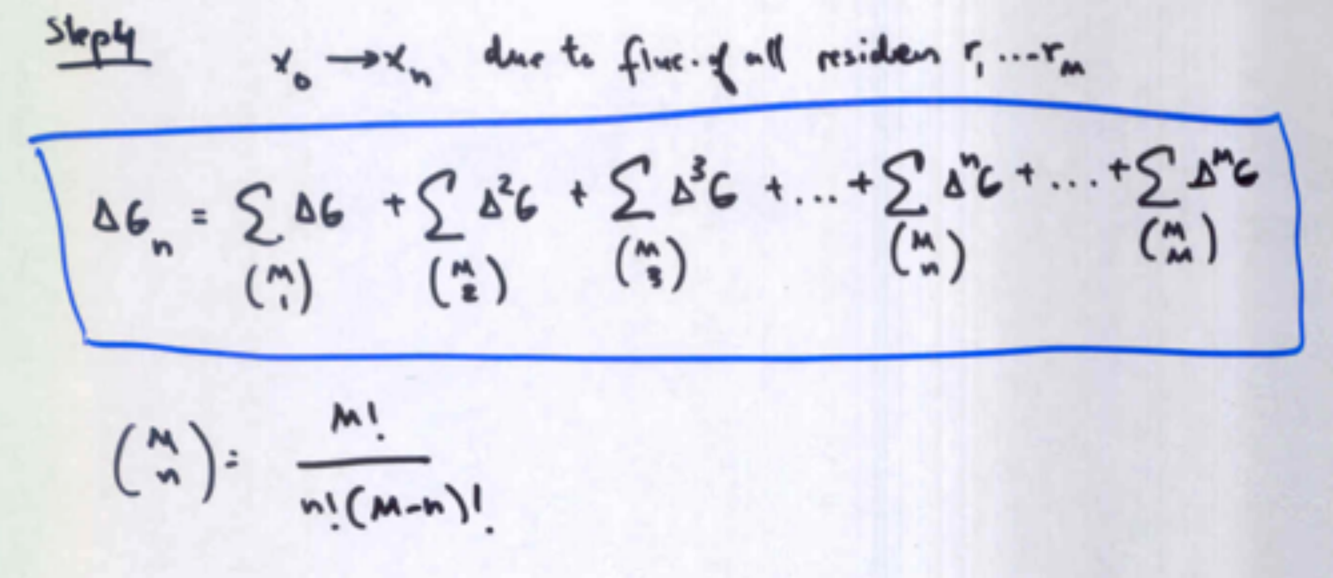
$$\boxed{\Delta G_n = \sum_{\binom{m}{1}} \Delta G + \sum_{\binom{m}{2}} \Delta^2 G + \sum_{\binom{m}{3}} \Delta^3 G + \ldots + \sum_{\binom{m}{n}} \Delta^n G + \ldots + \sum_{\binom{m}{m}} \Delta^m G}$$

$$\binom{m}{n} : \frac{m!}{n!(m-n)!}$$

….the case of **all parts** acting cooperatively

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

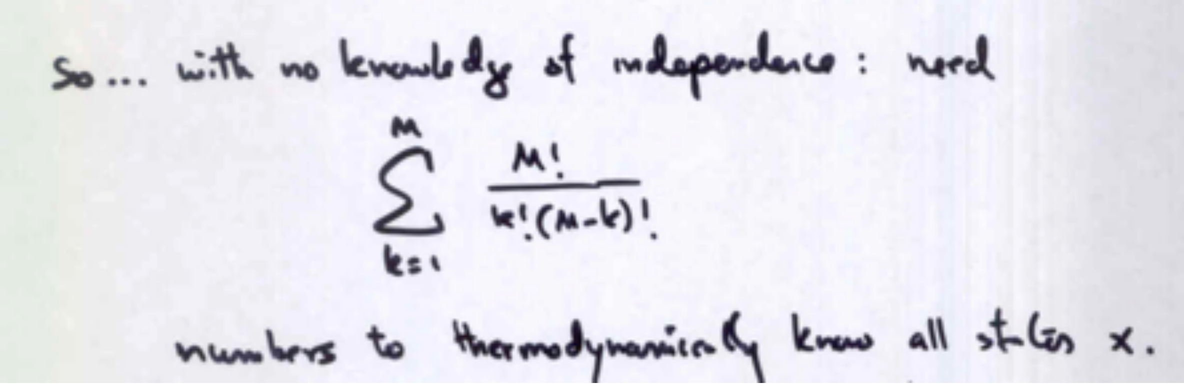




Step 4 $x_0 \to x_n$ due to fluc. of all residues $r_1 \ldots r_M$

$$\boxed{\Delta G_n = \sum_{\binom{M}{1}} \Delta G + \sum_{\binom{M}{2}} \Delta^2 G + \sum_{\binom{M}{3}} \Delta^3 G + \ldots + \sum_{\binom{M}{n}} \Delta^n G + \ldots + \sum_{\binom{M}{M}} \Delta^M G}$$

$$\binom{M}{n} = \frac{M!}{n!(M-n)!}$$



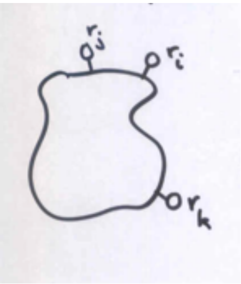
So… with no knowledge of independence: need

$$\sum_{k=1}^{M} \frac{M!}{k!(M-k)!}$$

numbers to thermodynamically know all states x.

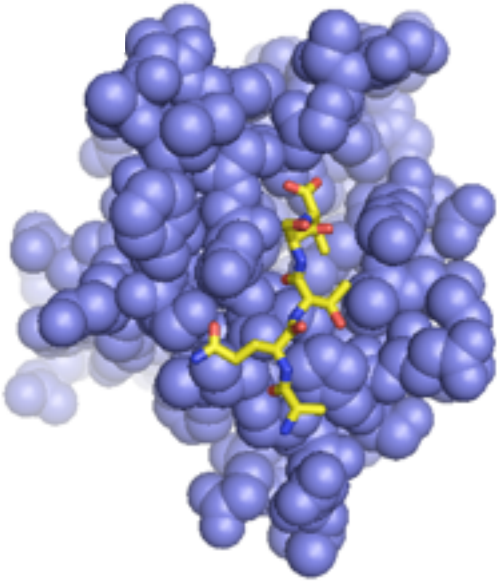
….the case of **all parts** acting cooperatively

A **warning**….!!

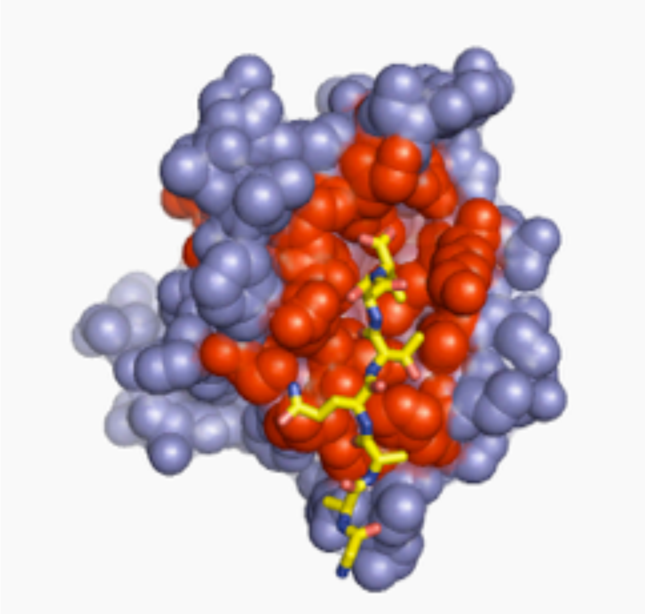


Note….epistasis is the non-independence of the parts that make up a system. But, it is sometimes **not obvious** what this means for parameters we might measure in general. Is independence necessarily additivity of a thing?

Needs to be validated…

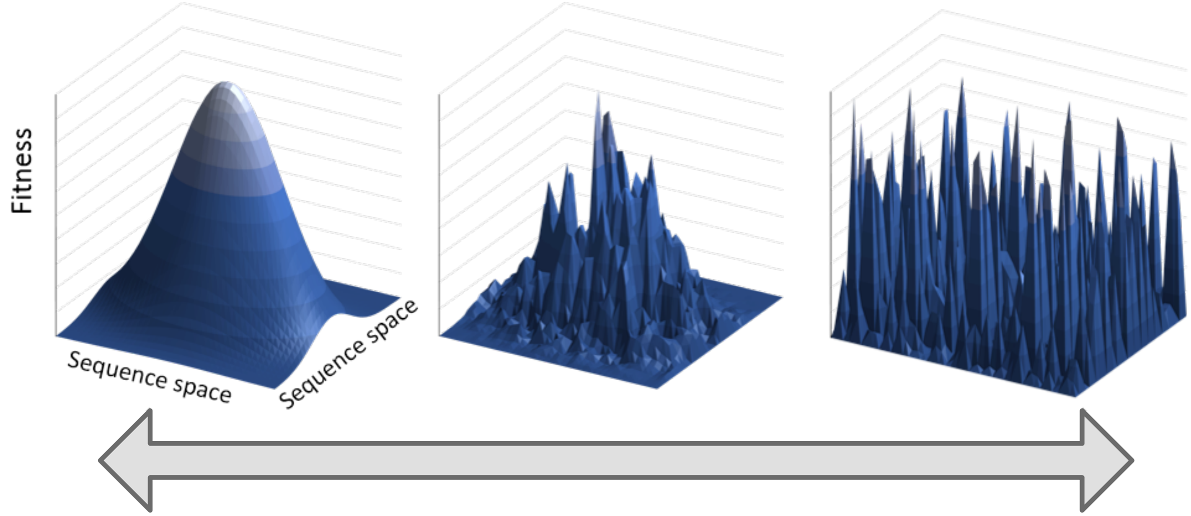


So…**a formal theory of epistasis** in systems. We will use point mutagenesis in proteins as a case study…



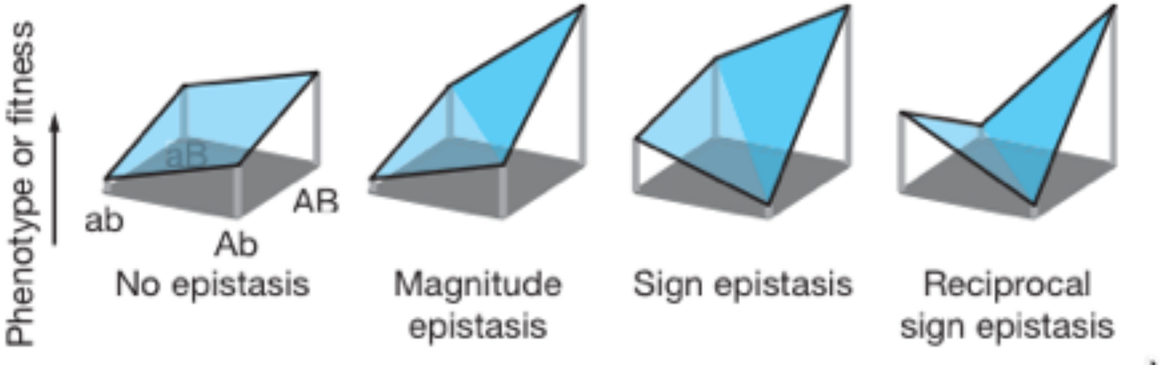
Epistasis is important for **function**….

…and for **evolution**



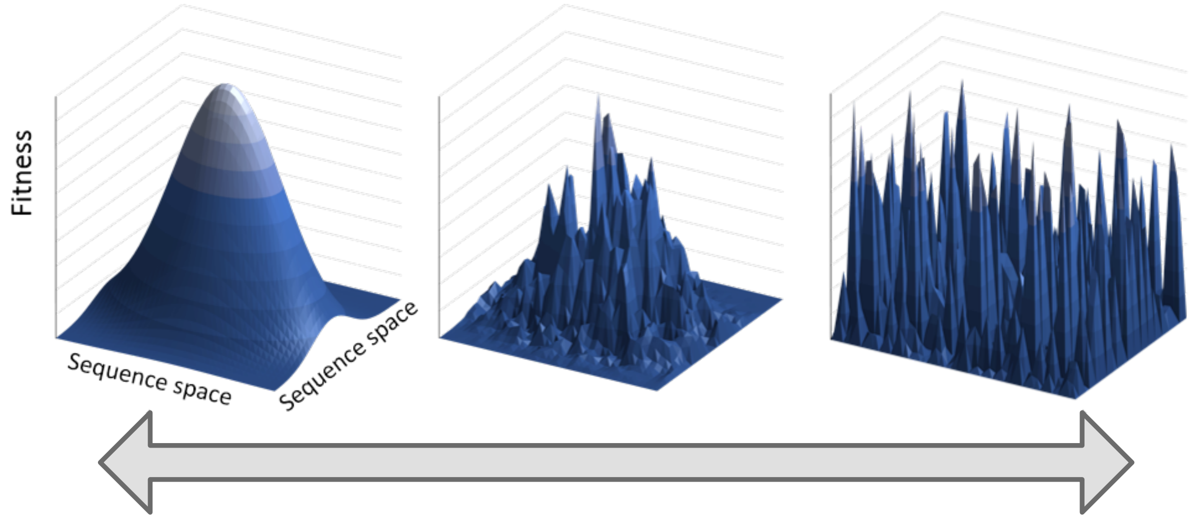
less epistatic

more epistatic



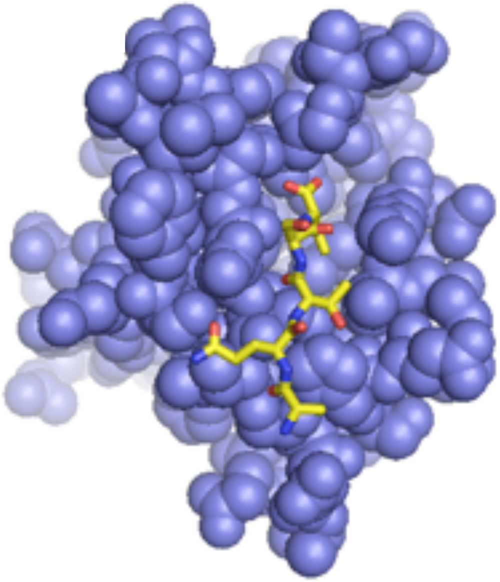
epistasis can effectively block off mutational paths
between functional genotypes…

…and for **evolution**



less epistatic

more epistatic



So…**a formal theory of epistasis** in systems. We will use point mutagenesis in proteins as a case study…

Orders of epistasis…

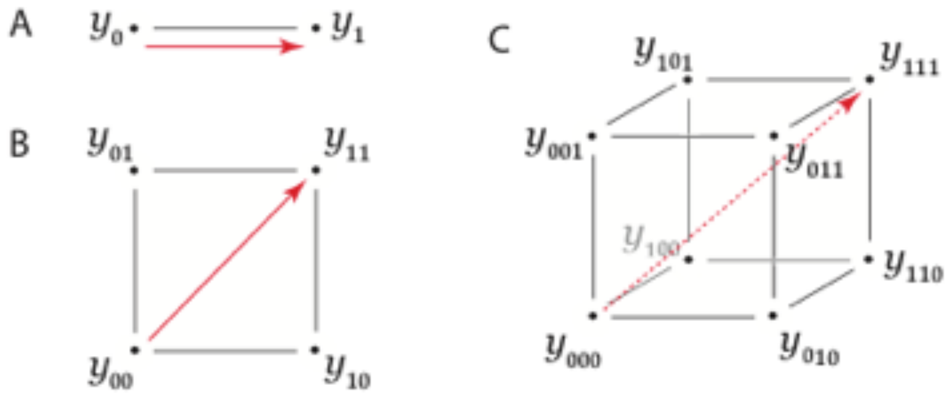
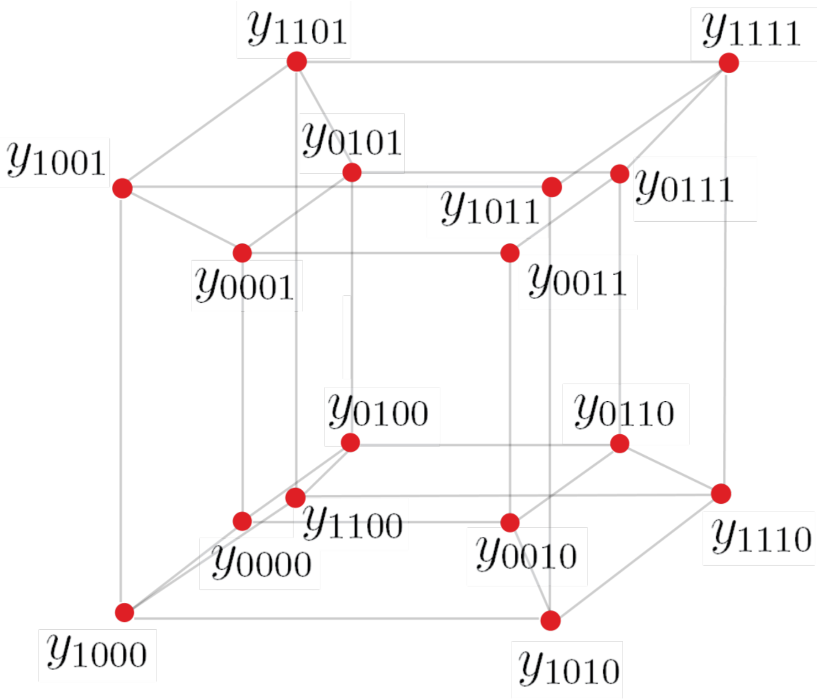


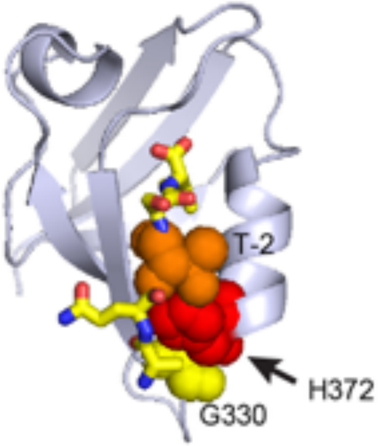
Fig. 1. Abstract representation of variants involved in (A) single mutant, (B) double mutant, and (C) triple mutant experiments. Genotypes are denoted as $g = \{g_N, ..., g_1\}$ with $g_i \in \{0, 1\}$, where '0' or '1' indicate the state of the mutable entity (e.g., the gene or the amino acid in the protein). Each genotype g has an associated phenotype y_g. The effect of a single, double, triple mutation is given by the red arrows. Two-way epistasis is defined as the differential effect of a mutation depending on the background in which it occurs, for example in (B) it is the phenotypic difference between mutating one position from '0' to '1' depending on the other position being either '0' or '1', i.e., $(y_{11} - y_{10}) - (y_{01} - y_{00})$.

Orders of epistasis…



The **fourth order epistasis** is the degree to which a third order epistasis depends on a fourth mutation…and so on. In general, *n*th order epistasis is the difference of two *n-1* dimensional epistases.

I. The biochemical view of epistasis…



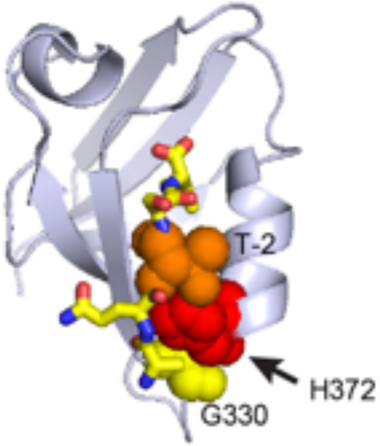
● wild-type binding free energy (in kcal/mol)

-8.17

y_0

In biochemistry the idea is to start with **one genotype** (the "wild-type") and use this as a reference for all effects of mutations….

The biochemical view of epistasis…



● wild-type binding free energy (in kcal/mol)

-8.17

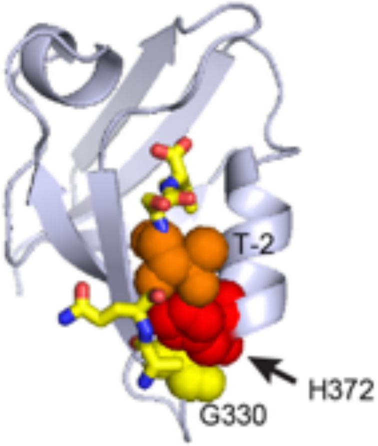
y_0

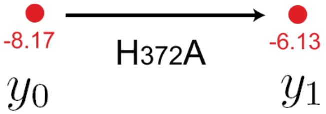
↑

the zeroth order term

In biochemistry the idea is to start with **one genotype** (the “wild-type”) and use this as a reference for all effects of mutations….

The biochemical view of epistasis…



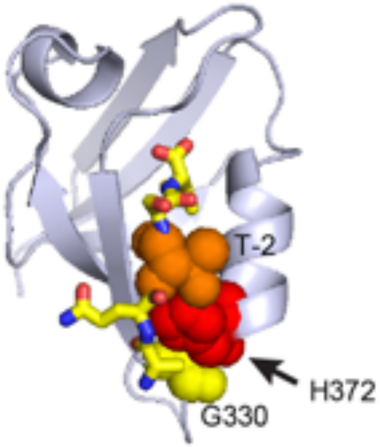


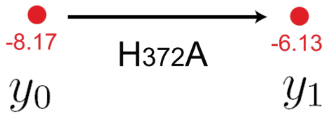
$$y_0 \xrightarrow{\text{H372A}} y_1$$

(in kcal/mol)

The single mutation experiment….

The biochemical view of epistasis…



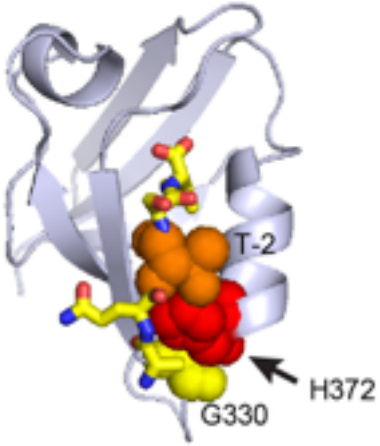


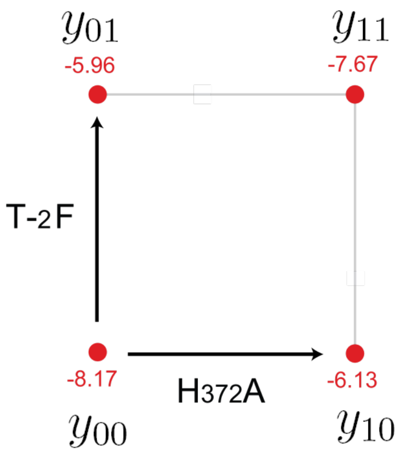
The effect of a single mutant….

$$\varepsilon_1 = y_1 - y_0$$

a first-order term

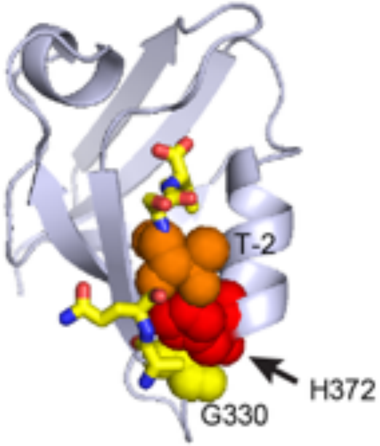
The biochemical view of epistasis…

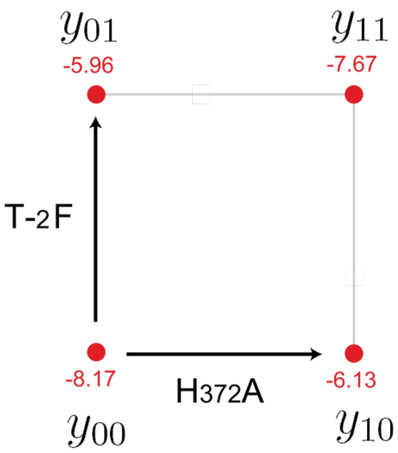




A double mutation experiment….

The biochemical view of epistasis…

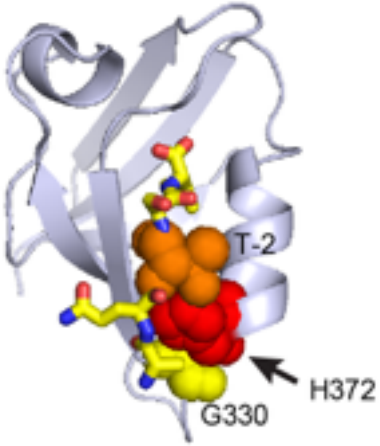


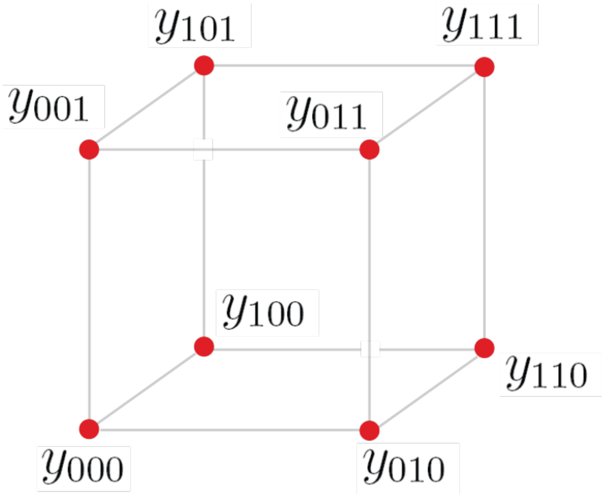


$$\varepsilon_{11} = (y_{11} - y_{01}) - (y_{10} - y_{00})$$

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

The biochemical view of epistasis…

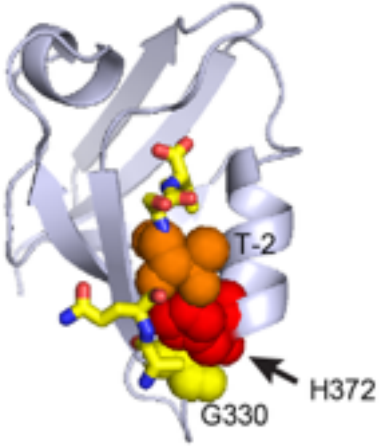


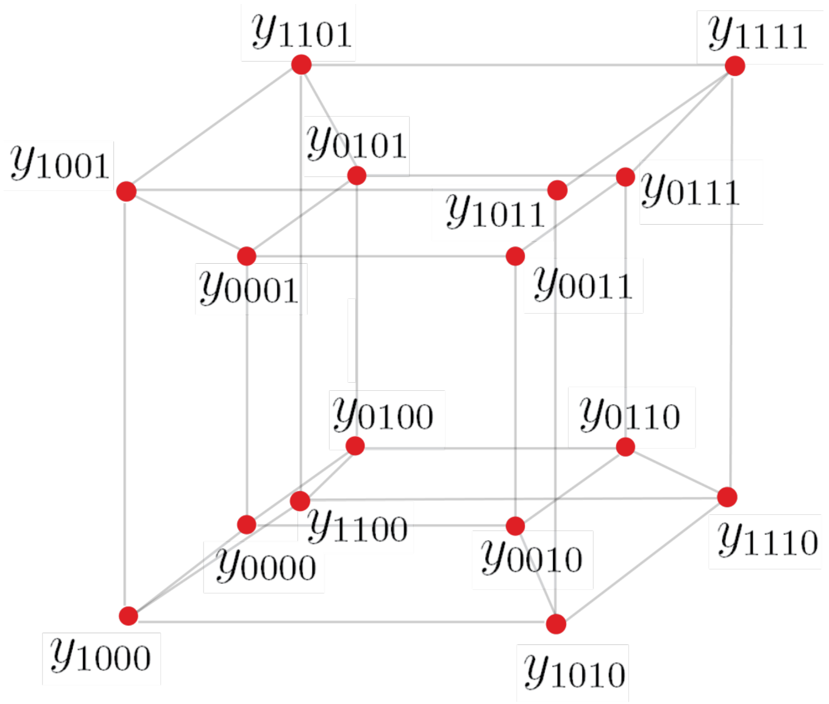


$$\varepsilon_{111} = (y_{111} - y_{101} - y_{110} + y_{100}) - (y_{011} - y_{001} - y_{010} + y_{000})$$

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

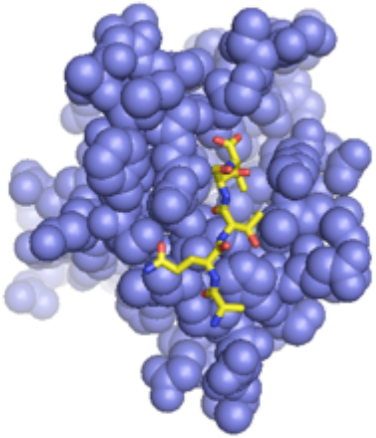
The biochemical view of epistasis…





The **fourth order epistasis** is the degree to which a third order epistasis depends on a fourth mutation…

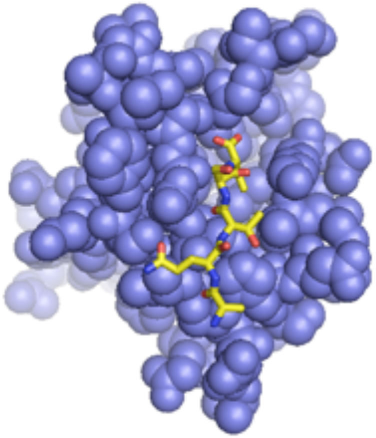
The biochemical view of epistasis…



$$\begin{aligned}\Delta^n G_{1,\dots,n} &= \Delta G^{\circ}_{1,\dots,n} + (-1)^1 \sum_{i_1<i_2<\dots<i_{n-1}}^{n} \Delta G^{\circ}_{i_1,i_2,\dots,i_{n-1}} \\ &\quad + (-1)^2 \sum_{i_1<i_2<\dots<i_{n-2}}^{n} \Delta G^{\circ}_{i_1,i_2,\dots,i_{n-2}} + \dots + (-1)^n \Delta G^{\circ}_0\end{aligned}$$

In **general**, there is a hierarchical expansion of terms….

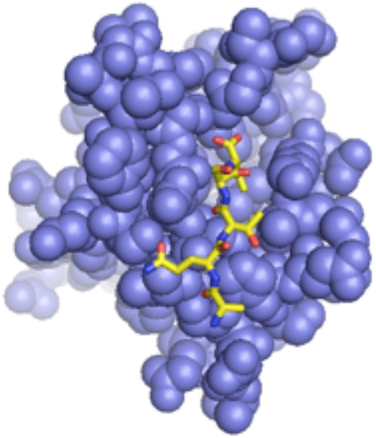
The biochemical view of epistasis…



In matrix form….

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

The biochemical view of epistasis…

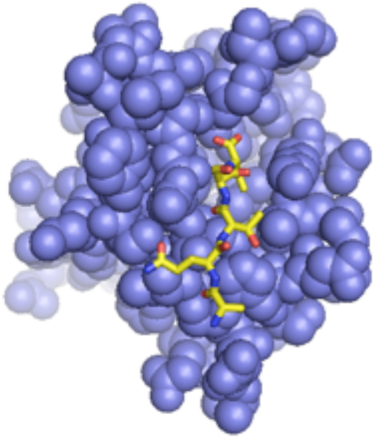


the phenotypes

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

the epistases

The biochemical view of epistasis…



the phenotypes

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

the epistases

…this is a **transform**, just like Fourier or Laplace. A one-to-one mapping between a space of phenotypes and epistases.

You've already been introduced to many transforms….

The concept of transforms

the phenotypes

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

the epistases

In general, a **transform** is a mapping of variables from one space to another. Why do this?

The concept of transforms

$$\underset{\substack{\uparrow \\ \text{the epistases}}}{\bar{\varepsilon}} = \boldsymbol{G}\, \overset{\substack{\text{the phenotypes} \\ \downarrow}}{\bar{\boldsymbol{y}}}$$

In general, a **transform** is a mapping of variables from one space to another. Why do this? Well…problems that are initially defined in one space - or one parametrization - might in fact be a lot simpler in another…

The concept of transforms

An example….

(LXXIV) times (XXVIII) → A multiplication problem in Roman numerals…

The concept of transforms

(LXXIV) times (XXVIII)

Step 1. Make a transformation

(74) times (28)

A multiplication problem in Hindu-Arabic numerals…

The concept of transforms

(LXXIV) times (XXVIII)

Step 1. Make a transformation ↓

(74) times (28)

Step 2. Solve it ↓

2072 → The answer

The concept of transforms

(LXXIV) times (XXVIII)

↓ **Step 1.** Make a transformation

(74) times (28)

↓ **Step 2.** Solve it

2072

↓ **Step 3.** Inverse transform

MMLXXII → The answer in Roman numerals…

The concept of transforms

$$\frac{dA}{dt} = -kA \quad , \quad A(0) = A_0$$

The concept of transforms

$$\frac{dA}{dt} = -kA \quad , \quad A(0) = A_0$$

Step 1. Make a transformation

$$\mathcal{L}\{f(t)\} = F(s) = \int_0^\infty f(t)\, e^{-st}\, dt$$

The LaPlace transform….

The concept of transforms

$$\frac{dA}{dt} = -kA \quad , \quad A(0) = A_0$$

Step 1. Make a transformation

$$\mathcal{L}\left\{\frac{dA}{dt}\right\} = \mathcal{L}\left\{-kA\right\}$$

Step 2. Solve it

$$\boxed{A(s) = \frac{A_0}{s+k}}$$

→ The LaPlace solution…

The concept of transforms

$$\frac{dA}{dt} = -kA \quad , \quad A(0) = A_0$$

Step 1. Make a transformation

$$\mathcal{L}\left\{\frac{dA}{dt}\right\} = \mathcal{L}\{-kA\}$$

Step 2. Solve it

$$\boxed{A(s) = \frac{A_0}{s+k}}$$

Step 3. Inverse transform

$$\mathcal{L}^{-1}(A(s)) = \mathcal{L}^{-1}\left[\frac{A_0}{s+k}\right]$$

So...

$f(t)$	$F(s)$
e^{-kt}	$\frac{1}{s+k}$

The concept of transforms

$$\frac{dA}{dt} = -kA \qquad , \quad A(0) = A_0$$

Step 1. Make a transformation

$$\mathcal{L}\left\{\frac{dA}{dt}\right\} = \mathcal{L}\left\{-kA\right\}$$

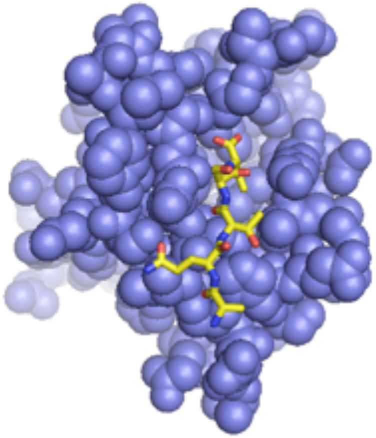
Step 2. Solve it

$$\boxed{A(s) = \frac{A_0}{s+k}}$$

Step 3. Inverse transform

$$A(t) = A_0 e^{-kt}$$ → The time-domain solution…

The biochemical view of epistasis…



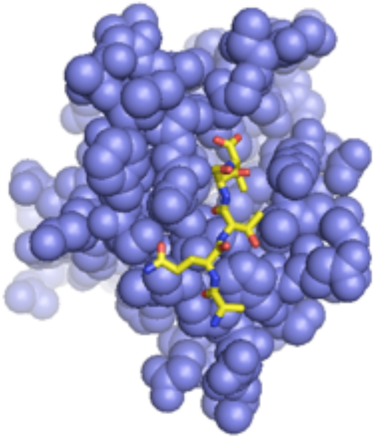
the phenotypes

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

the epistases

…this is a **transform**, just like Fourier or Laplace. A one-to-one mapping between a space of phenotypes and epistases. What kind of transform? wait…

The biochemical view of epistasis…

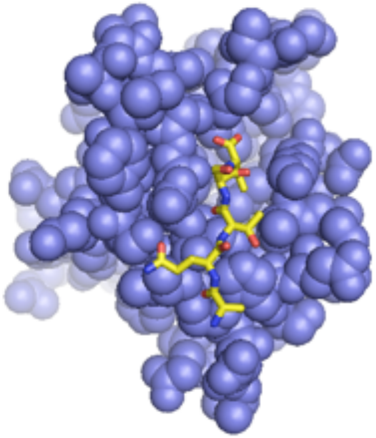


$$\bar{\varepsilon} = \boldsymbol{G}\,\bar{y}$$

↑

the epistasis operator

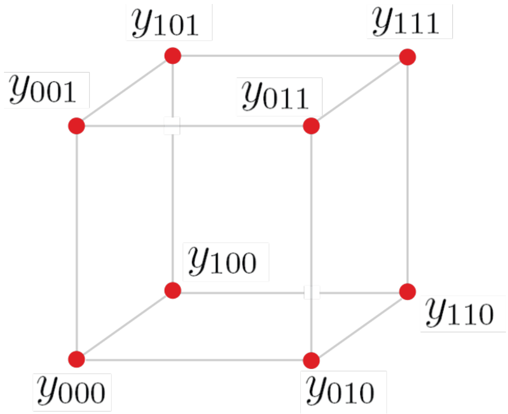
The biochemical view of epistasis…



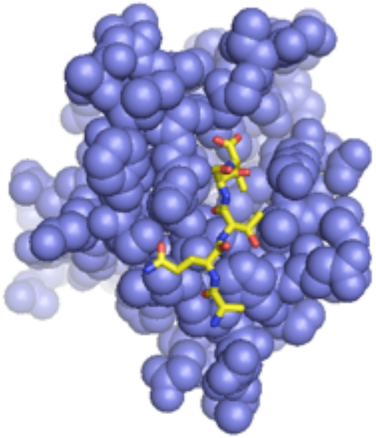
$$\bar{\boldsymbol{\varepsilon}} = \boldsymbol{G}\,\bar{\boldsymbol{y}}$$

for **3 mutations…**third order

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



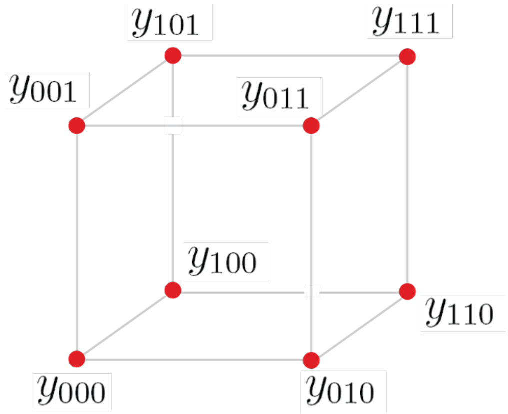
The biochemical view of epistasis…



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

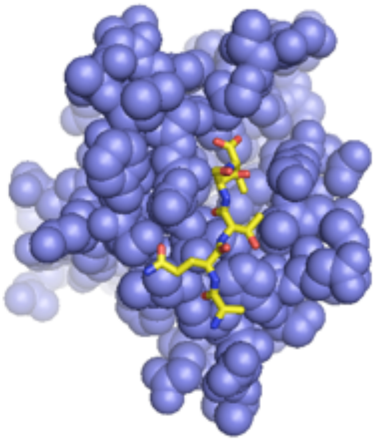
for **3 mutations…**third order

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



$$\varepsilon_{011} = (y_{011} - y_{001}) - (y_{010} - y_{000})$$

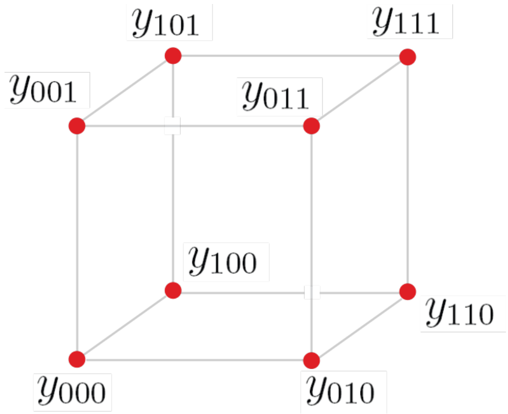
The biochemical view of epistasis…



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

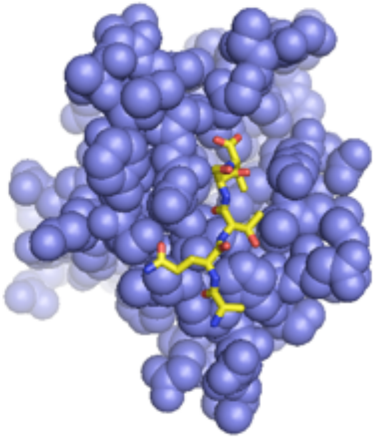
for **3 mutations…**third order

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



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

The biochemical view of epistasis…

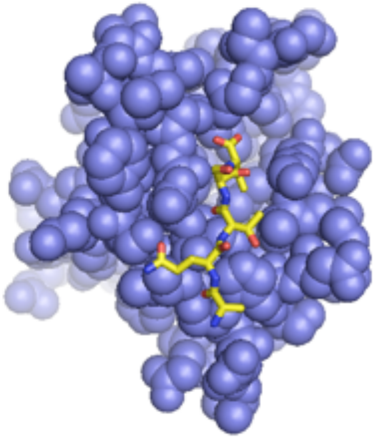


$$\bar{\boldsymbol{\varepsilon}} = \boldsymbol{G}\,\bar{\boldsymbol{y}}$$

A recursive **generative function** for nth-order epistasis….

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

The biochemical view of epistasis…

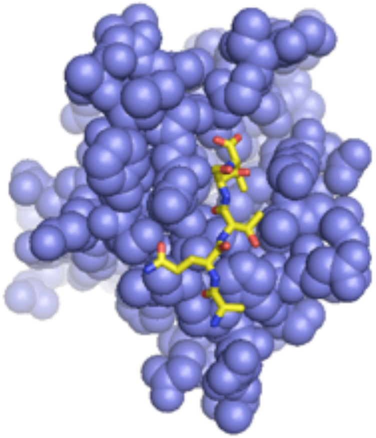


$$\boldsymbol{G}_{n+1} = \begin{pmatrix} \boldsymbol{G}_n & 0 \\ -\boldsymbol{G}_n & \boldsymbol{G}_n \end{pmatrix} \quad \text{with} \quad \boldsymbol{G}_0 = 1$$

for **n = 0…**zeroth order (wild-type)

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

The biochemical view of epistasis…

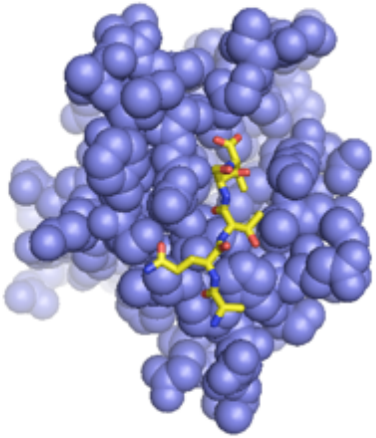


$$\boldsymbol{G}_{n+1} = \begin{pmatrix} \boldsymbol{G}_n & 0 \\ -\boldsymbol{G}_n & \boldsymbol{G}_n \end{pmatrix} \quad \text{with} \quad \boldsymbol{G}_0 = 1$$

for **n = 1…**first order (a single mutation experiment)

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

The biochemical view of epistasis…

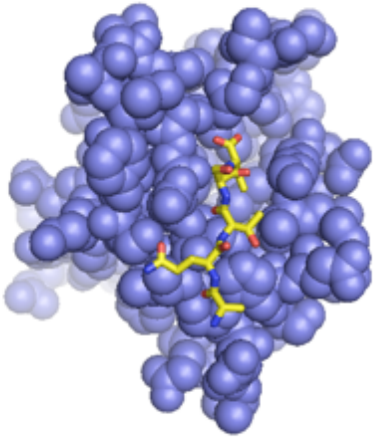


$$\boldsymbol{G}_{n+1} = \begin{pmatrix} \boldsymbol{G}_n & 0 \\ -\boldsymbol{G}_n & \boldsymbol{G}_n \end{pmatrix} \quad \text{with} \quad \boldsymbol{G}_0 = 1$$

for **n = 2…**second order (a double mutation experiment)

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

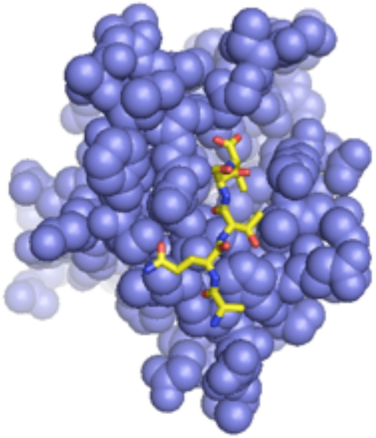
The biochemical view of epistasis…



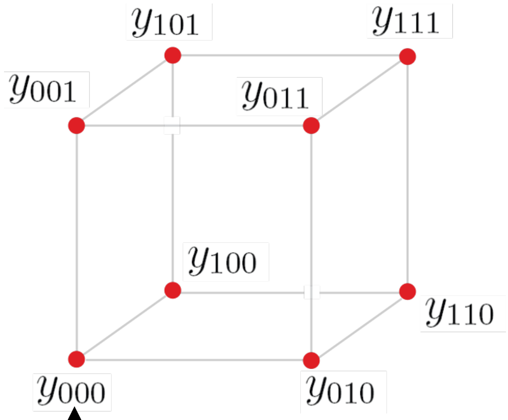
$$\boldsymbol{G}_{n+1} = \begin{pmatrix} \boldsymbol{G}_n & 0 \\ -\boldsymbol{G}_n & \boldsymbol{G}_n \end{pmatrix} \quad \text{with} \quad \boldsymbol{G}_0 = 1$$

for **n = 3…**third order (a triple mutation experiment)

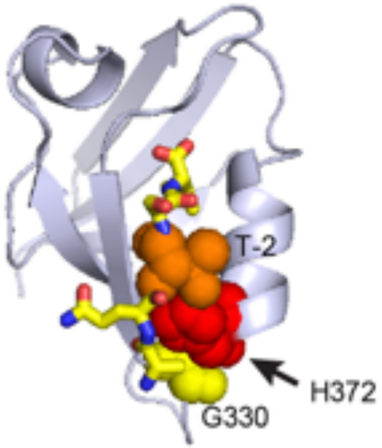
$$\begin{pmatrix} \varepsilon_{000} \\ \varepsilon_{001} \\ \varepsilon_{010} \\ \varepsilon_{011} \\ \varepsilon_{100} \\ \varepsilon_{101} \\ \varepsilon_{110} \\ \varepsilon_{111} \end{pmatrix} = \left(\begin{array}{|cccc|cccc|} \hline 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 1 & 0 & 0 & 0 & 0 \\ \cline{1-4} -1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & -1 & 1 & 0 & 0 \\ 1 & 0 & -1 & 0 & -1 & 0 & 1 & 0 \\ -1 & 1 & 1 & -1 & 1 & -1 & -1 & 1 \\ \hline \end{array}\right) * \begin{pmatrix} y_{000} \\ y_{001} \\ y_{010} \\ y_{011} \\ y_{100} \\ y_{101} \\ y_{110} \\ y_{111} \end{pmatrix}$$

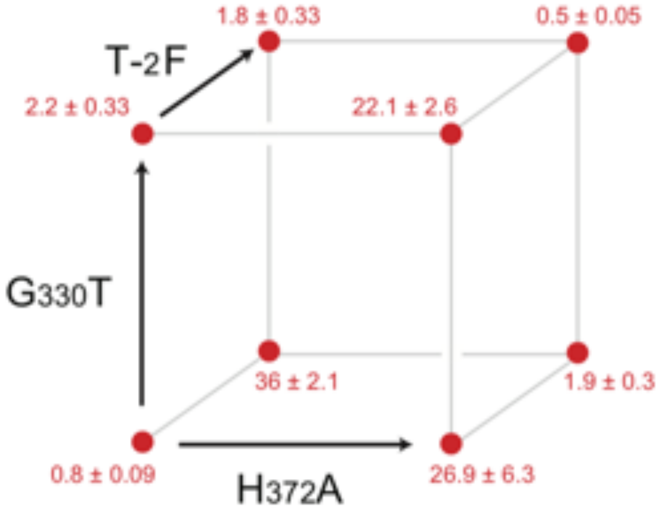


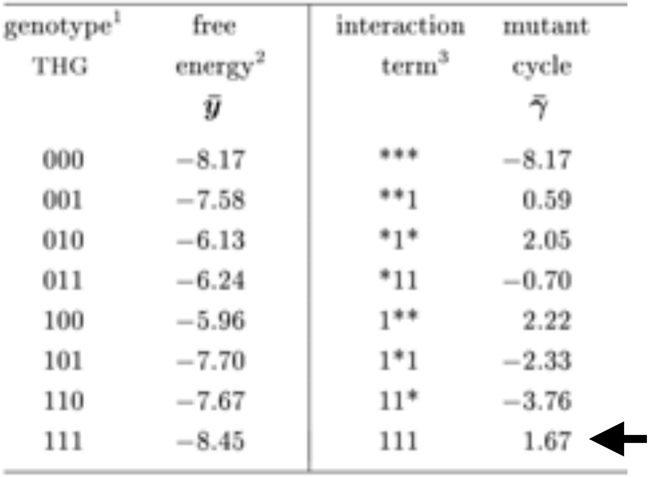
$$\bar{\boldsymbol{\varepsilon}} = \boldsymbol{G}\,\bar{\boldsymbol{y}}$$



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



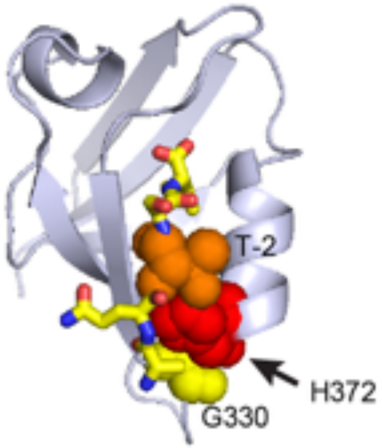


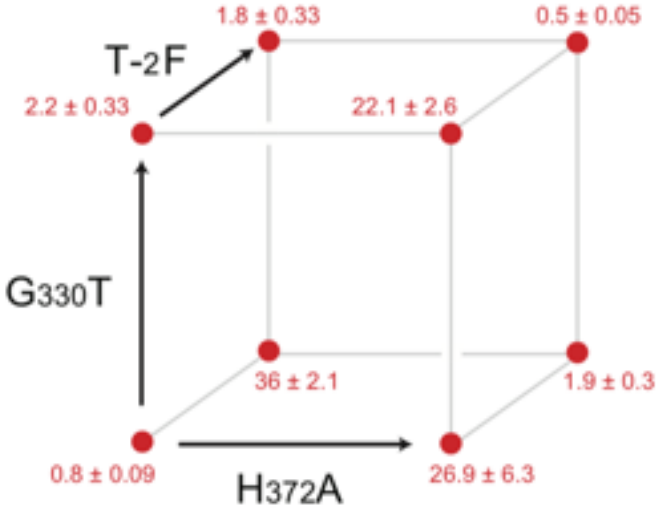


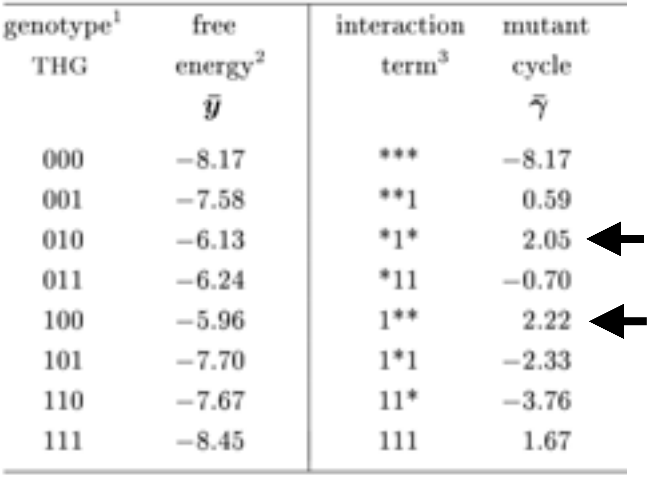
genotype[1] THG	free energy[2] $\bar{y}$	interaction term[3]	mutant cycle $\bar{\gamma}$
000	−8.17	***	−8.17
001	−7.58	**1	0.59
010	−6.13	*1*	2.05
011	−6.24	*11	−0.70
100	−5.96	1**	2.22
101	−7.70	1*1	−2.33
110	−7.67	11*	−3.76
111	−8.45	111	1.67

In this view…

(1) Large third order effect….all three mutations represent a collectively cooperative group.





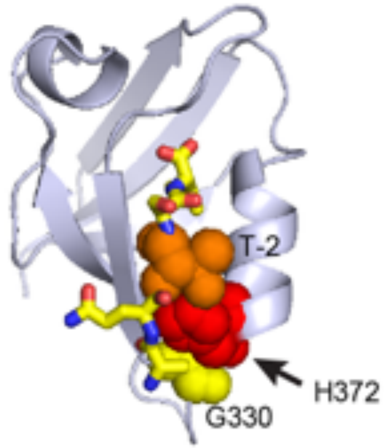


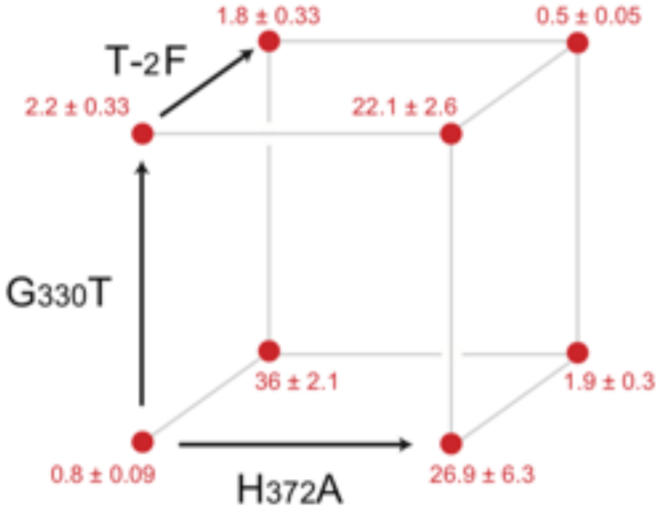
genotype[1] THG	free energy[2] $\bar{y}$	interaction term[3]	mutant cycle $\bar{\gamma}$
000	−8.17	***	−8.17
001	−7.58	**1	0.59
010	−6.13	*1*	2.05
011	−6.24	*11	−0.70
100	−5.96	1**	2.22
101	−7.70	1*1	−2.33
110	−7.67	11*	−3.76
111	−8.45	111	1.67

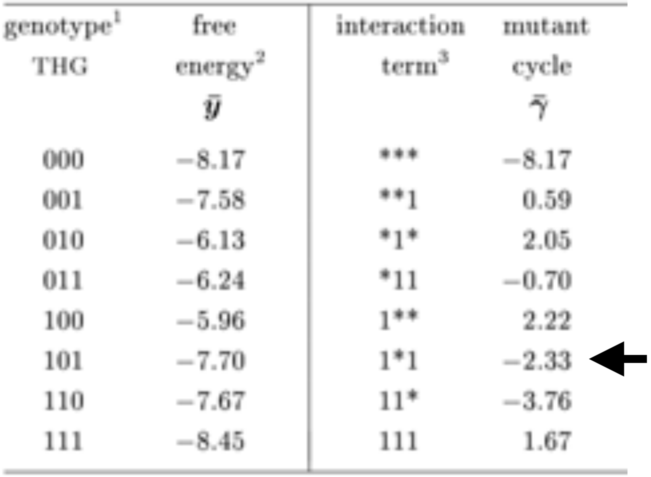
In this view…

(1) Large third order effect….all three mutations represent a collectively cooperative group.

(2) Large first order effects of the H372A and T-2F single mutations







genotype^1 THG	free energy^2 $\bar{y}$	interaction term^3	mutant cycle $\bar{\gamma}$
000	−8.17	***	−8.17
001	−7.58	**1	0.59
010	−6.13	*1*	2.05
011	−6.24	*11	−0.70
100	−5.96	1**	2.22
101	−7.70	1*1	−2.33
110	−7.67	11*	−3.76
111	−8.45	111	1.67

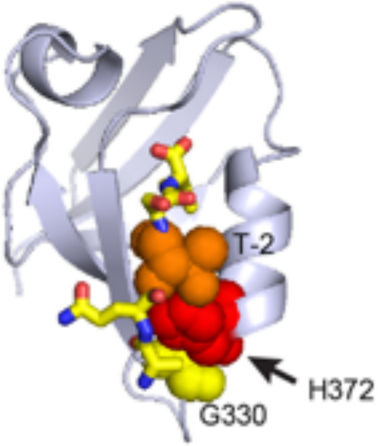
In this view…

(1) Large third order effect….all three mutations represent a collectively cooperative group.

(2) Large first order effects of the H372A and T-2F single mutations

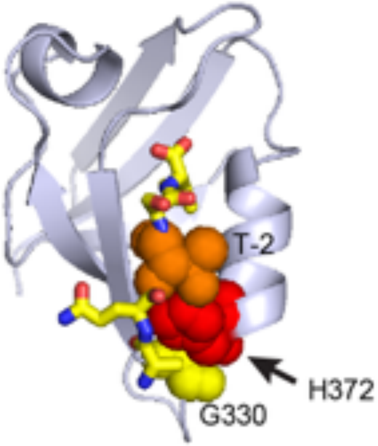
(3) Large second order epistatic coupling between G330T and T-2F

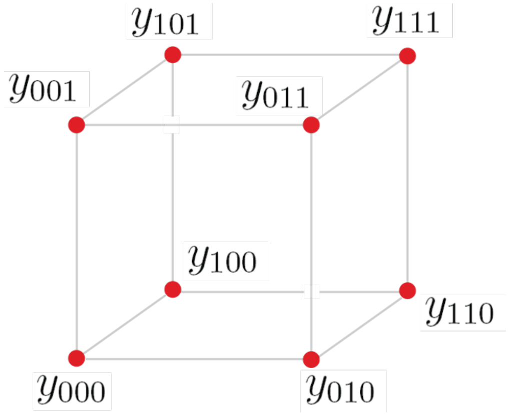
Another view of epistasis…background averaging!



In this view….there is **no reference sequence**. Values are averaged over all genotypes. For example…

Another view of epistasis…background averaging!

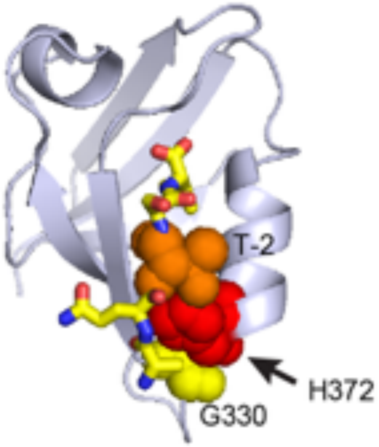


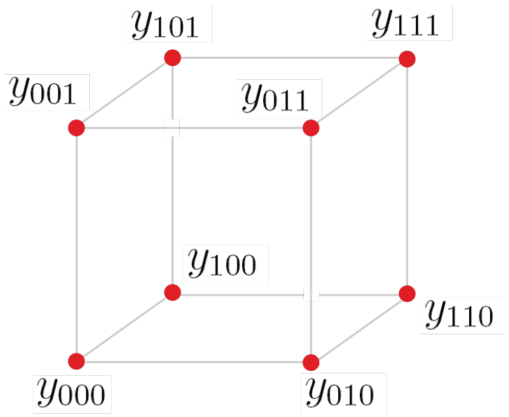


In biochemistry, the second order epistasis would be:

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

Another view of epistasis…background averaging!





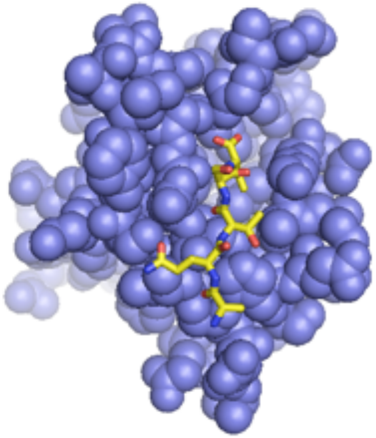
In biochemistry, the second order epistasis would be:

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

In this view, it is an average in the background of two genotypes… y_{0**} and y_{1**}

$$\varepsilon_{*11} = \frac{[(y_{111} - y_{101}) - (y_{110} - y_{100})] + [(y_{011} - y_{001}) - (y_{010} - y_{000})]}{2}$$

Background averaged epistasis…



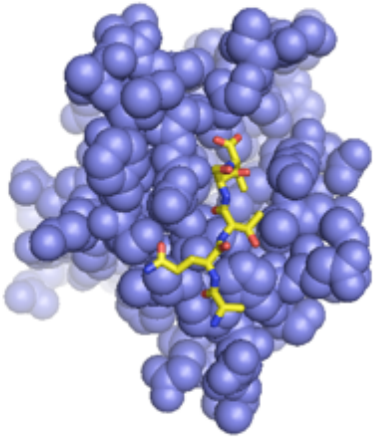
Has a matrix form too….

$$\bar{e} = \boldsymbol{H}\bar{y}$$

the background-averaged epistasis operator

…also a **transform**!

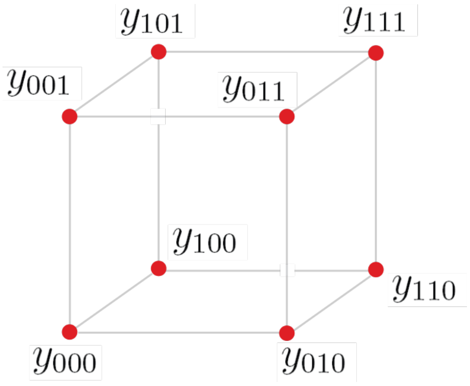
Background averaged epistasis…



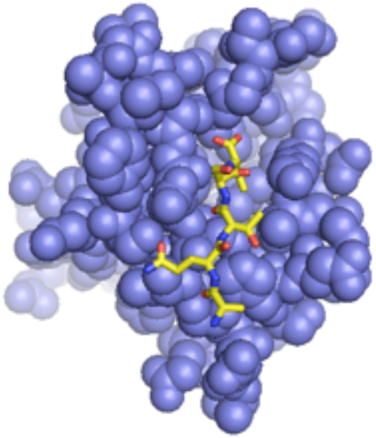
Has a matrix form too….

$$\bar{\boldsymbol{e}} = \boldsymbol{H}\bar{\boldsymbol{y}}$$

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



Background averaged epistasis…

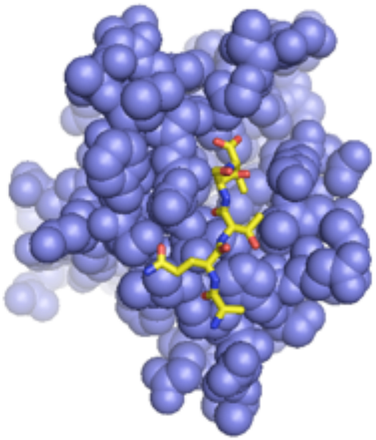


$$\begin{pmatrix} e_{***} \\ e_{**1} \\ e_{*1*} \\ \boxed{e_{*11}} \\ e_{1**} \\ e_{1*1} \\ e_{11*} \\ e_{111} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1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

$$\varepsilon_{*11} = \frac{[(y_{111} - y_{101}) - (y_{110} - y_{100})] + [(y_{011} - y_{001}) - (y_{010} - y_{000})]}{2}$$

Makes sense?

Background averaged epistasis…



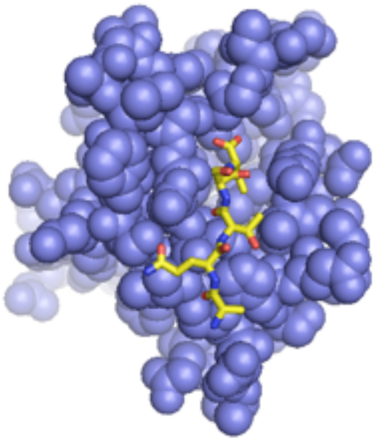
$$\begin{pmatrix} e_{***} \\ e_{**1} \\ e_{*1*} \\ \boxed{e_{*11}} \\ e_{1**} \\ e_{1*1} \\ e_{11*} \\ e_{111} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1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

$$\varepsilon_{*11} = \frac{[(y_{111}-y_{101})-(y_{110}-y_{100})]+[(y_{011}-y_{001})-(y_{010}-y_{000})]}{\boxed{2}}$$

$\uparrow$

There is a weighting for number of terms
over which the average is taken…

Background averaged epistasis…

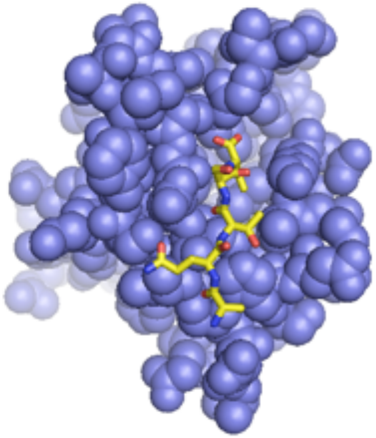


$$\begin{pmatrix} \boxed{e_{***}} \\ e_{**1} \\ e_{*1*} \\ e_{*11} \\ e_{1**} \\ e_{1*1} \\ e_{11*} \\ e_{111} \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1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

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

There is a weighting for number of terms over which the average is taken…

Background averaged epistasis…



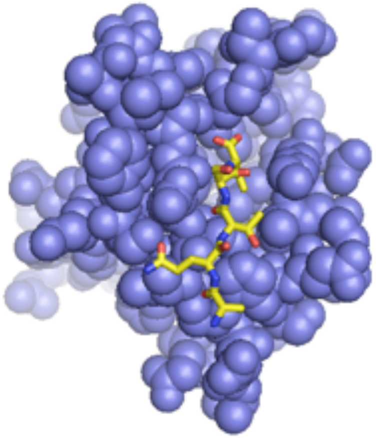
$$\begin{pmatrix} \boxed{e_{***}} \\ e_{**1} \\ e_{*1*} \\ e_{*11} \\ e_{1**} \\ e_{1*1} \\ e_{11*} \\ e_{111} \end{pmatrix} = \begin{pmatrix} \begin{array}{|cccccccc|} \hline 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline \end{array} \\ \begin{array}{cccccccc} 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \end{array} \end{pmatrix} * \begin{pmatrix} \boxed{\begin{matrix} y_{000} \\ y_{001} \\ y_{010} \\ y_{011} \\ y_{100} \\ y_{101} \\ y_{110} \\ y_{111} \end{matrix}} \end{pmatrix}$$

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

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

…the background averaged epistasis

Background averaged epistasis…

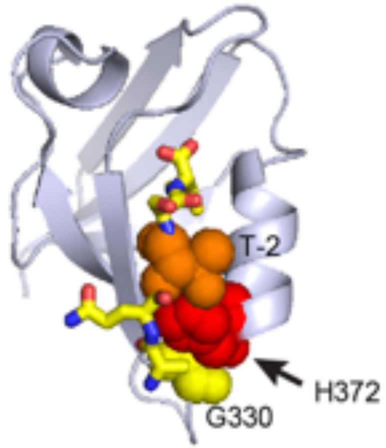


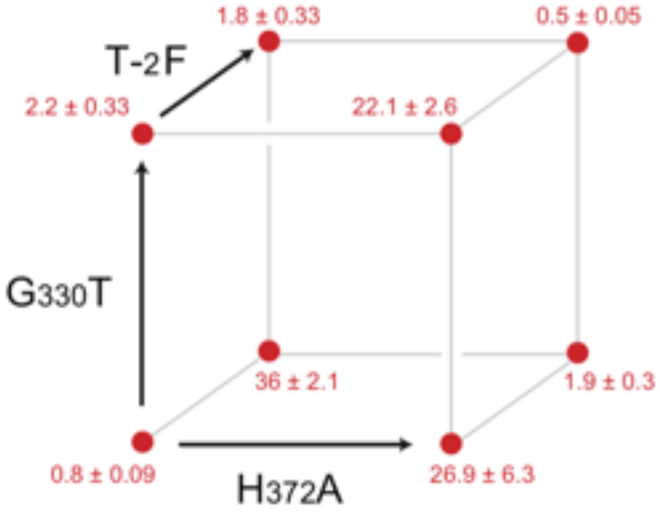
$$\bar{\varepsilon} = \boldsymbol{V}\boldsymbol{H}\bar{y}.$$

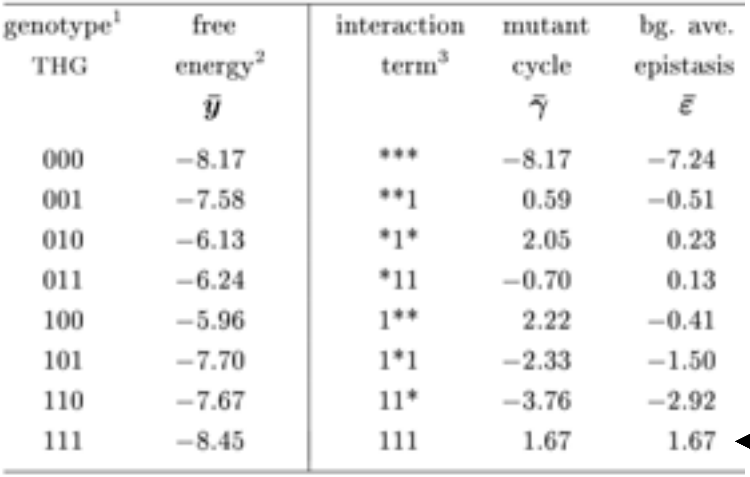
the recursive **generative function** for n^{th}-order background averaged epistasis…

$$\boldsymbol{V}_{n+1} = \begin{pmatrix} \frac{1}{2}\boldsymbol{V}_n & 0 \\ 0 & -\boldsymbol{V}_n \end{pmatrix} \quad \text{with} \quad \boldsymbol{V}_0 = 1$$

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



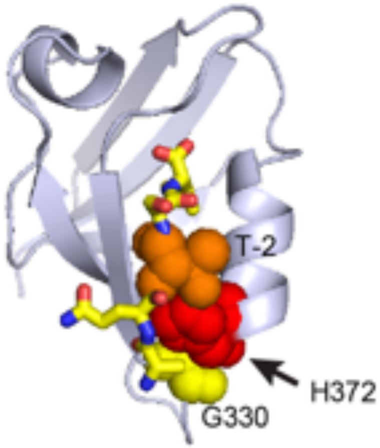


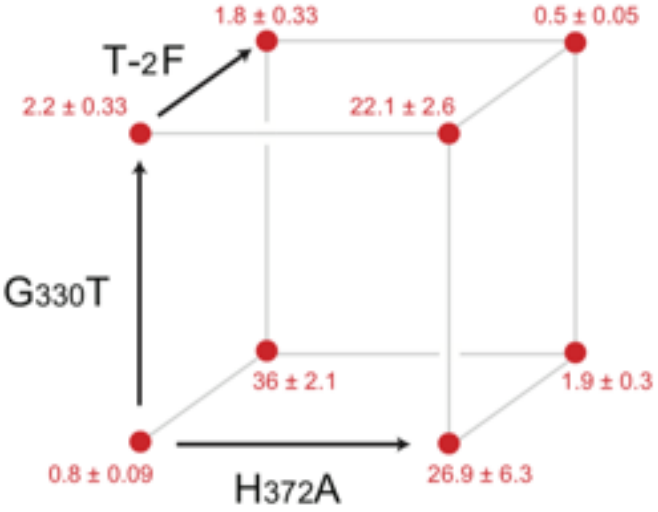


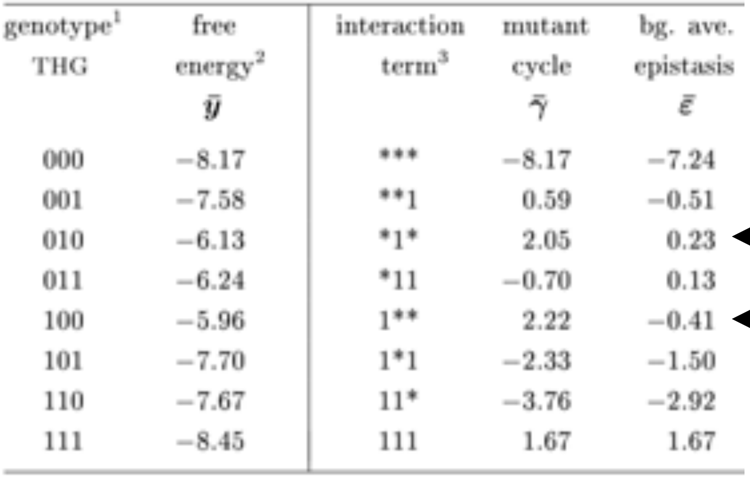
genotype[1] THG	free energy[2] $\bar{y}$	interaction term[3]	mutant cycle $\bar{\gamma}$	bg. ave. epistasis $\bar{\varepsilon}$
000	−8.17	***	−8.17	−7.24
001	−7.58	**1	0.59	−0.51
010	−6.13	*1*	2.05	0.23
011	−6.24	*11	−0.70	0.13
100	−5.96	1**	2.22	−0.41
101	−7.70	1*1	−2.33	−1.50
110	−7.67	11*	−3.76	−2.92
111	−8.45	111	1.67	1.67

So…

(1) Large third order effect….all three mutations represent a collectively cooperative group.







genotype[1] THG	free energy[2] $\bar{y}$	interaction term[3]	mutant cycle $\bar{\gamma}$	bg. ave. epistasis $\bar{\varepsilon}$
000	−8.17	***	−8.17	−7.24
001	−7.58	**1	0.59	−0.51
010	−6.13	*1*	2.05	0.23
011	−6.24	*11	−0.70	0.13
100	−5.96	1**	2.22	−0.41
101	−7.70	1*1	−2.33	−1.50
110	−7.67	11*	−3.76	−2.92
111	−8.45	111	1.67	1.67

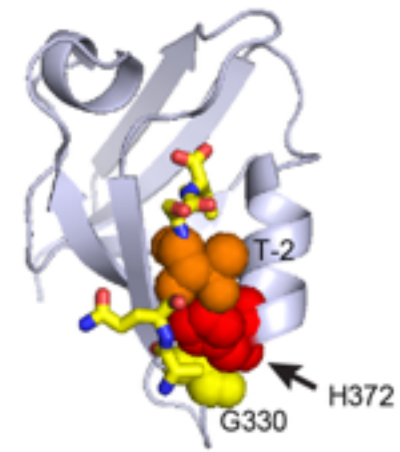
So…

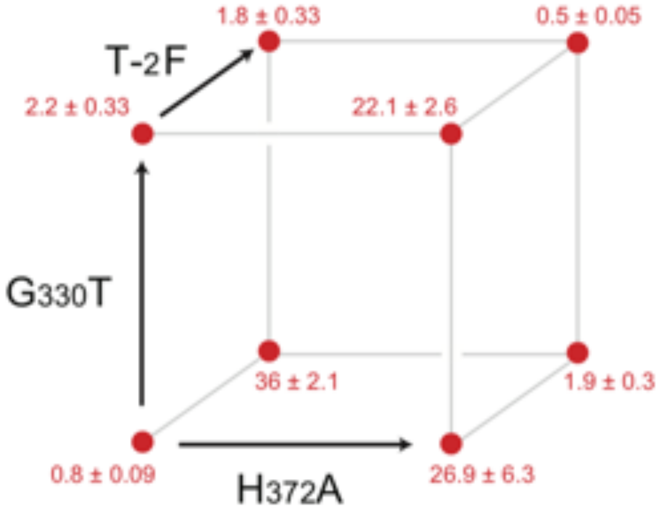
(1) Large third order effect….all three mutations represent a collectively cooperative group.

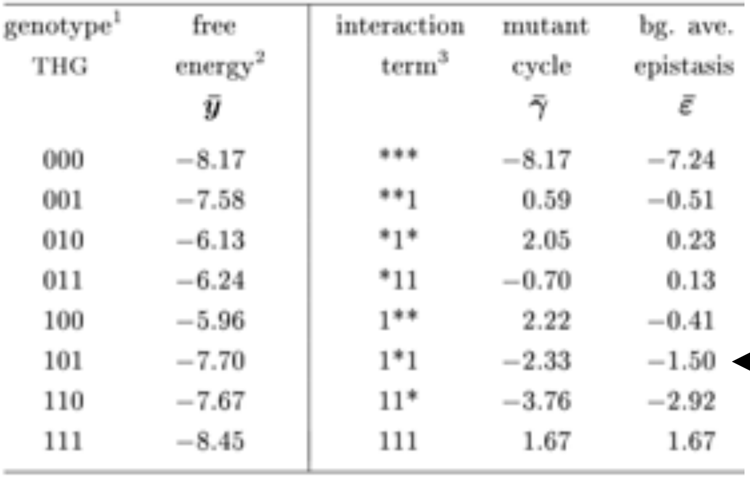
(2) First order effects of all mutations essentially negligible.

this is because of higher order epistasis…H372A has a 2.05 kcal/mol effect in the wild-type background and a -1.71 kcal/mol effect in the T-2F background!

It's not a general determinant of affinity!







genotype[1] THG	free energy[2] $\bar{y}$	interaction term[3]	mutant cycle $\bar{\gamma}$	bg. ave. epistasis $\bar{\varepsilon}$
000	−8.17	***	−8.17	−7.24
001	−7.58	**1	0.59	−0.51
010	−6.13	*1*	2.05	0.23
011	−6.24	*11	−0.70	0.13
100	−5.96	1**	2.22	−0.41
101	−7.70	1*1	−2.33	−1.50
110	−7.67	11*	−3.76	−2.92
111	−8.45	111	1.67	1.67

So…

(1) Large third order effect….all three mutations represent a collectively cooperative group.

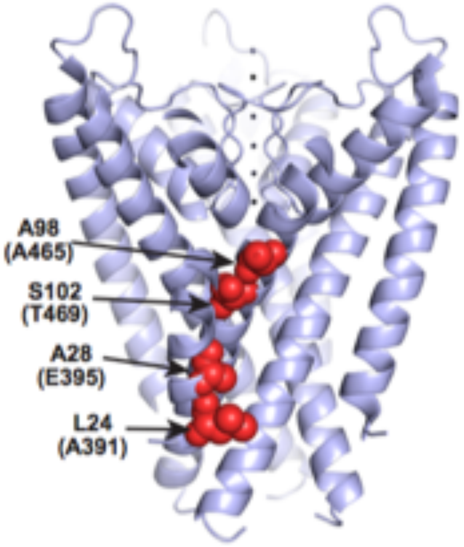
(2) First order effects of all mutations essentially negligible.

(3) Coupling of G330T and T-2F is lower

this is because of higher order epistasis…H372A has a 2.05 kcal/mol effect in the wild-type background and a -1.71 kcal/mol effect in the T-2F background!

It's not a general determinant of affinity!

Another case….



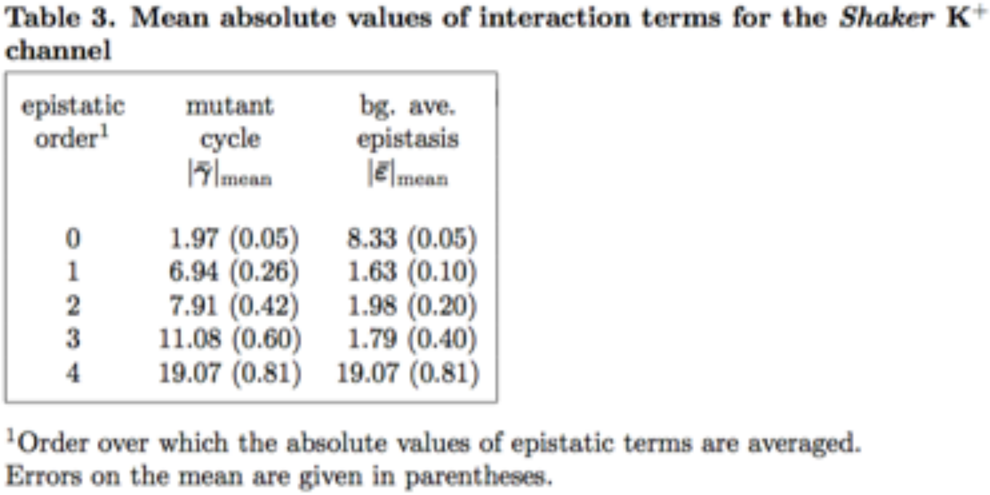


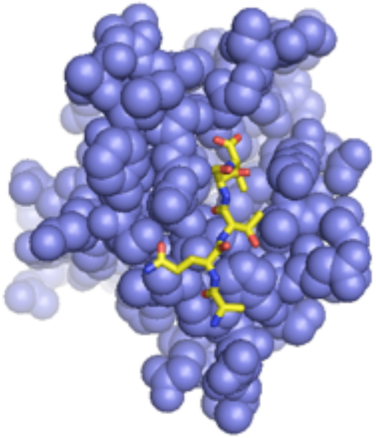
Table 3. Mean absolute values of interaction terms for the *Shaker* K^+ channel

epistatic order[1]	mutant cycle $\lvert\bar{\gamma}\rvert_{\text{mean}}$	bg. ave. epistasis $\lvert\bar{\varepsilon}\rvert_{\text{mean}}$
0	1.97 (0.05)	8.33 (0.05)
1	6.94 (0.26)	1.63 (0.10)
2	7.91 (0.42)	1.98 (0.20)
3	11.08 (0.60)	1.79 (0.40)
4	19.07 (0.81)	19.07 (0.81)

[1] Order over which the absolute values of epistatic terms are averaged. Errors on the mean are given in parentheses.

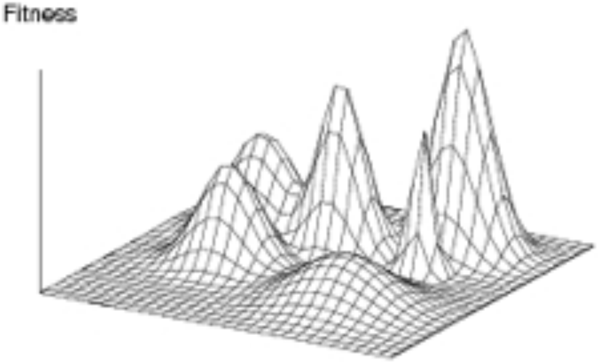
A high-order cooperativity in gating of the **K+ channel pore**….

The bottom line…

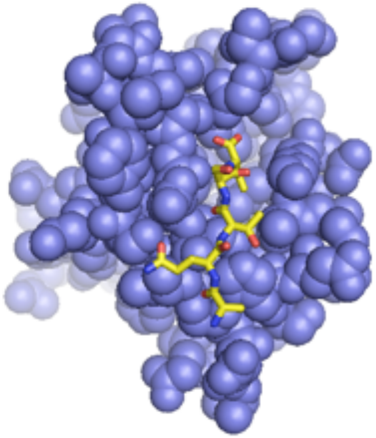


$$\bar{\varepsilon} = \boldsymbol{G}\,\bar{y}$$

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

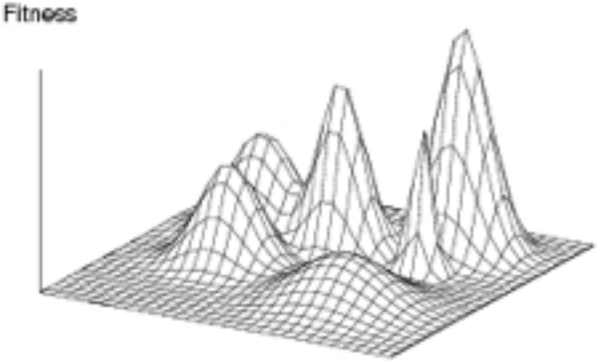


The bottom line…



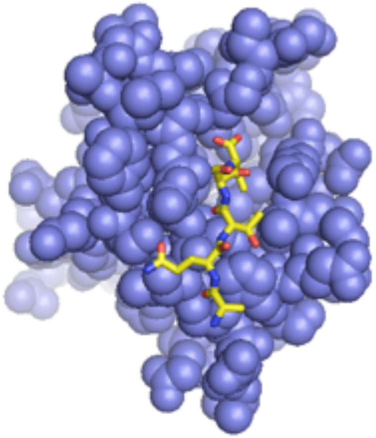
$$\bar{\varepsilon} = G\,\bar{y}$$

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



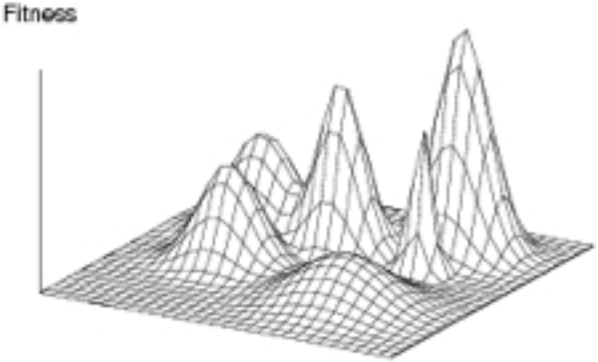
…its like doing a **Taylor's** (local) expansion of the fitness landscape around a particular point (the reference genotype). A detailed analysis of a particular solution.

The bottom line…

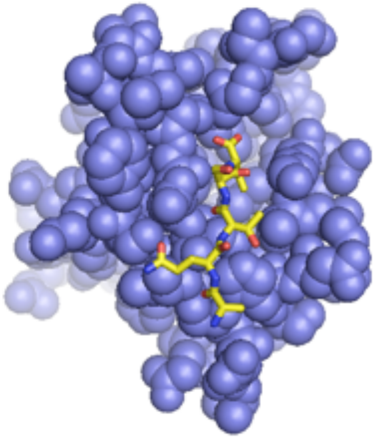


$$\bar{\varepsilon} = \boldsymbol{V}\boldsymbol{H}\bar{y}.$$

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

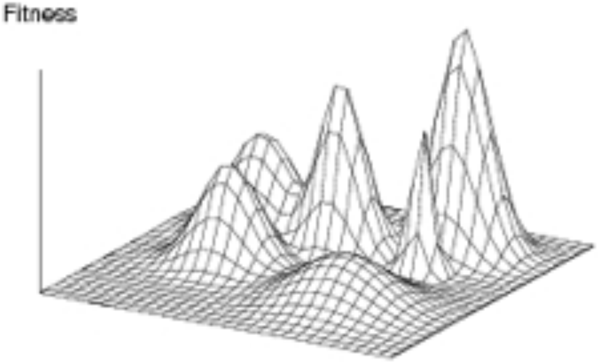


The bottom line…

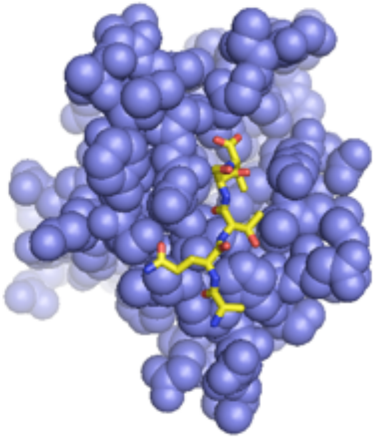


$$\bar{\varepsilon} = \boldsymbol{V}\boldsymbol{H}\bar{y}.$$

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



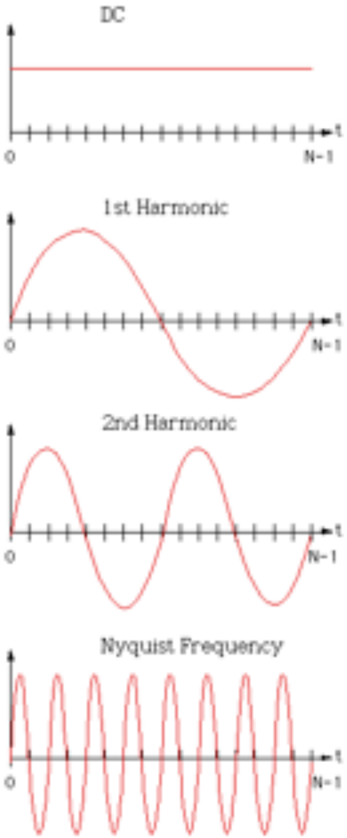
…its like doing a generalized **Fourier** expansion of the fitness landscape. A global analysis of all possible solutions given the design process.



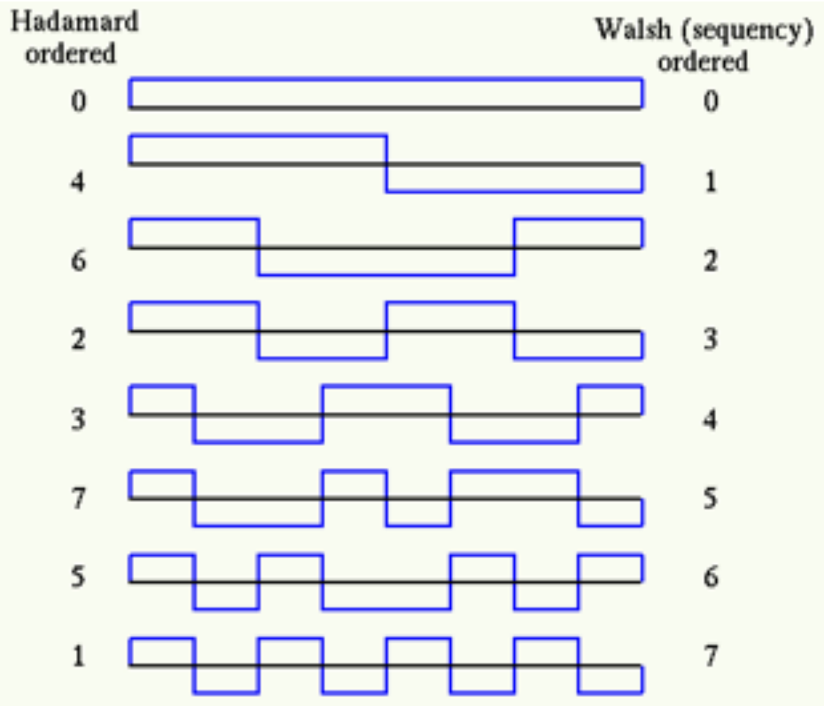
$$\bar{\varepsilon} = \boldsymbol{V}\boldsymbol{H}\bar{y}.$$

in fact, this mapping has a name….it is a generalized form of a Fourier transform called the **Hadamard transform**.

Fourier series…

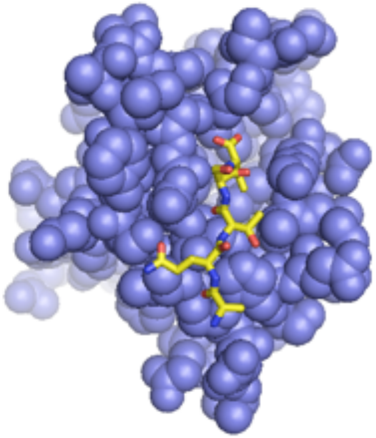


Hadamard series…



a representation based on just **square** functions…
with components just +1 and -1.

The bottom line…



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

…in general, **background averaging** seems like the right thing to do if we want to learn the general rules for specifying systems.

Ok…so **background averaged** epistasis is what we want?

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

But…background averaging is a ridiculous concept in general. In examples, we analyzed single mutations at like 4 positions (2^4). For 20 possible amino acids in a protein with 100 positions?

background averaged epistasis…

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

So…the bottom line is that we need **a general strategy** to learn the epistasis vector from a practical set of experiment. Can we do this?

heterogeneity and **non-linearity…**

For most systems, we don't know either the relevant parts, or the non-linear interactions between them. Why?

heterogeneity and **non-linearity…**

For most systems, we don't know either the relevant parts, or the non-linear interactions between them. Why?

(1) We do experiments in **specific model systems** under specific, carefully controlled experimental conditions in the laboratory….what does this do?

(2) We do experiments typically one the scale of **one part at a time**…single mutations, single gene knockouts, single drug applications, single cell recordings…what does this do?

….what is **an optimal strategy** to address this problem?

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

So, large-scale **non-linear dynamical systems**. Next time….the problem of proteins

	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	anharmonic 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