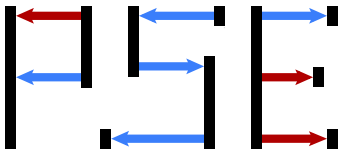


PROBABILISTIC STRUCTURES
IN EVOLUTION

DFG SPP 1590

COLLABORATIVE RESEARCH CENTER | SFB 680
Molecular Basis of
Evolutionary Innovations

Epistasis and genotypic complexity in Fisher's geometric model

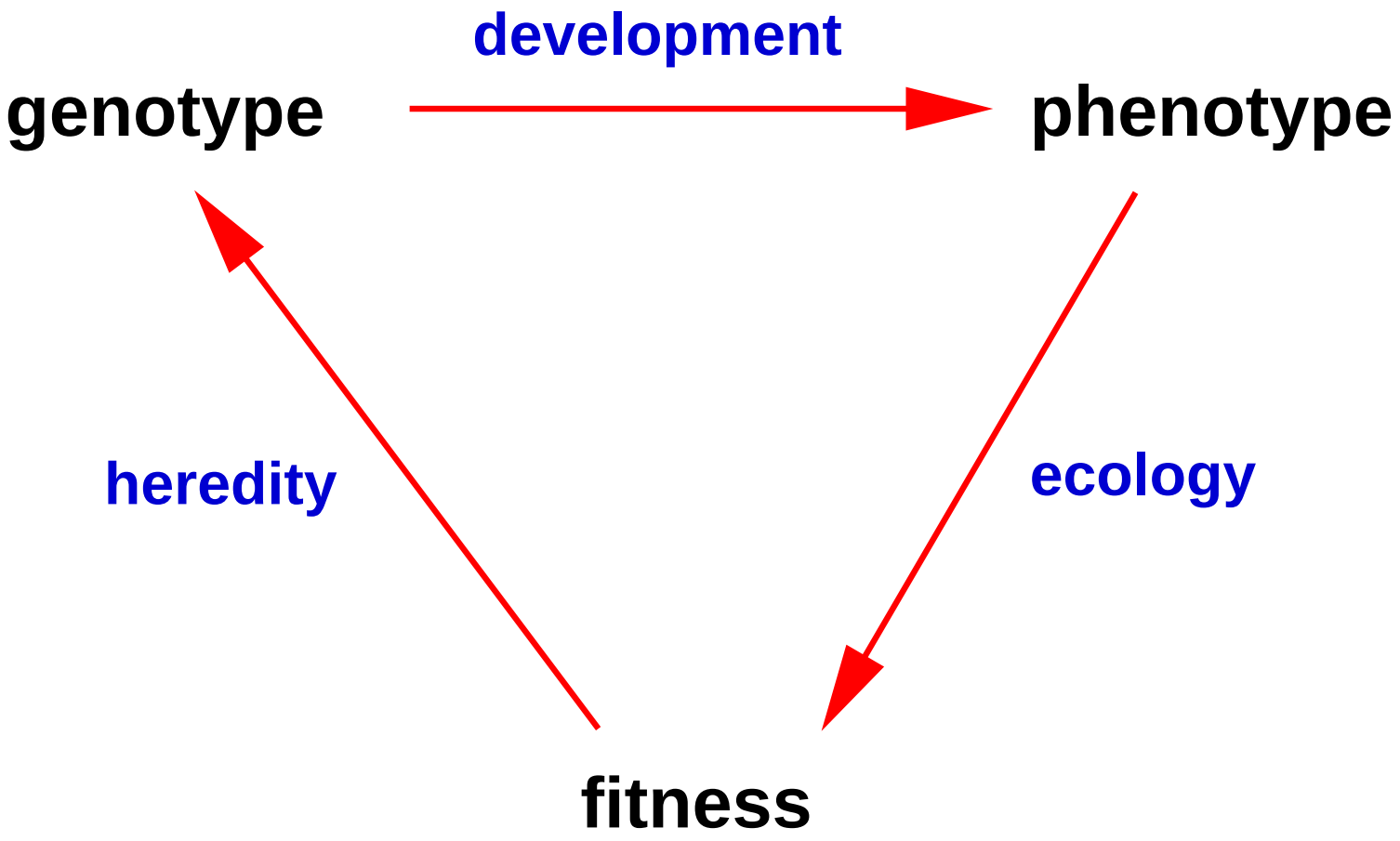
Joachim Krug
Institute for Theoretical Physics
University of Cologne

with Sungmin Hwang (Cologne), Su-Chan Park (Seoul),
Sijmen Schoustra and Arjan de Visser (Wageningen)

"Eco-Evolutionary Dynamics in Nature and the Lab", KITP, 29.8.2017

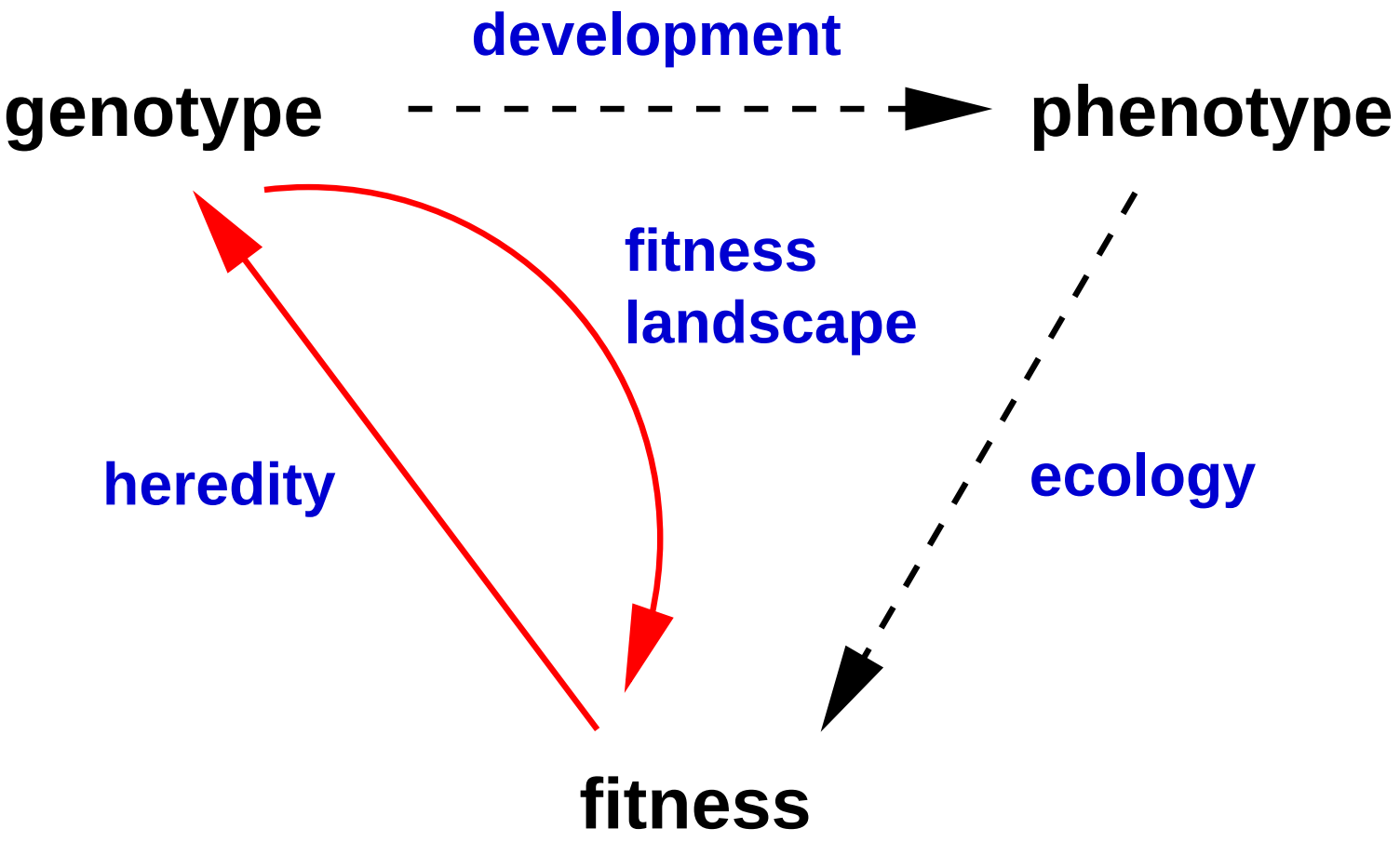
Evolution in a nutshell

Courtesy Amitabh Joshi



Evolution in a nutshell

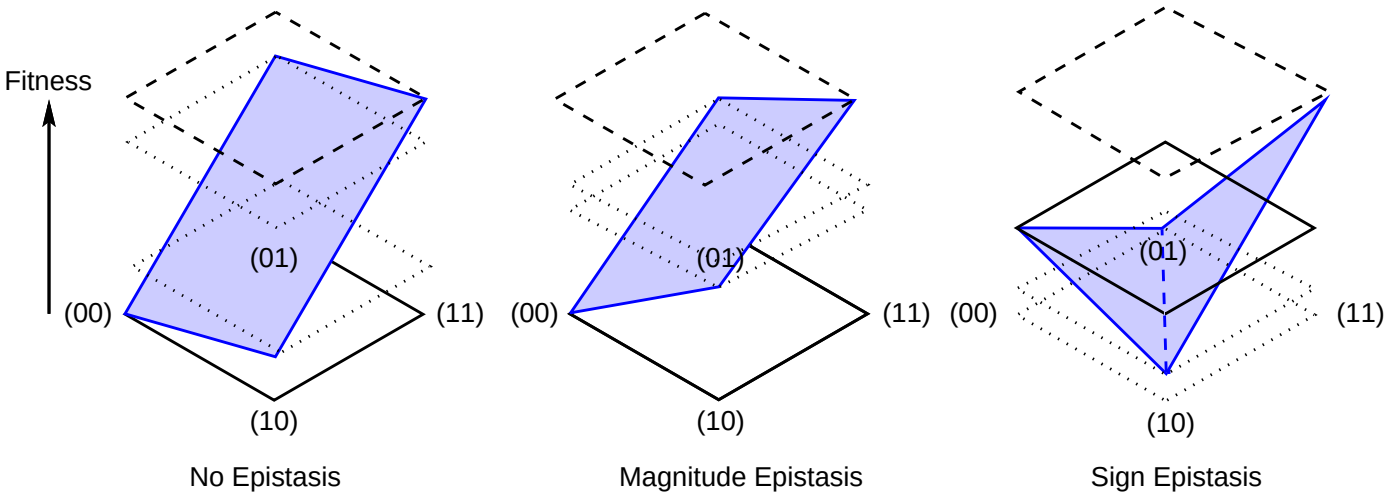
Courtesy Amitabh Joshi



Epistasis and sign epistasis

- General setting: L diallelic haploid loci τ_i at which a mutation can be present ($\tau_i = 1$) or absent ($\tau_i = 0$).
- A genotypic fitness landscape is a function on the set of 2^L genotypes
- Epistasis implies interactions between the effects of different mutations
- Sign epistasis: Mutation at a given locus is beneficial or deleterious depending on the state of other loci

Weinreich, Watson & Chao (2005)

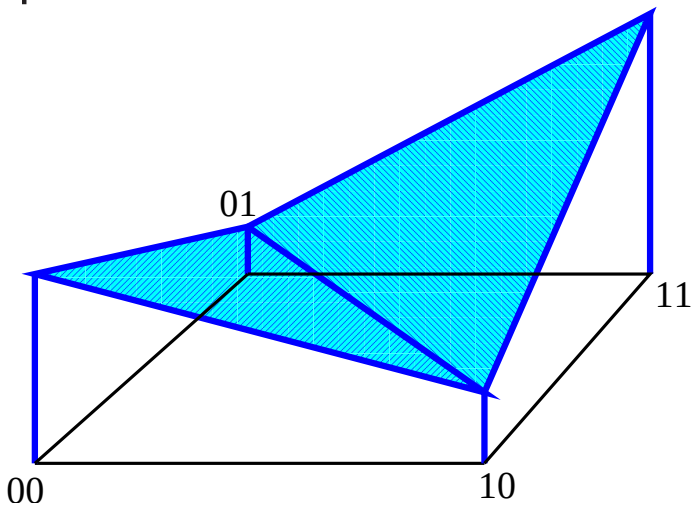


Genotypic complexity

- A genotypic fitness landscape is complex/rugged if it has multiple fitness maxima

- The existence of reciprocal sign epistasis is a necessary condition for the existence of multiple peaks

Poelwijk et al., JTB 2011



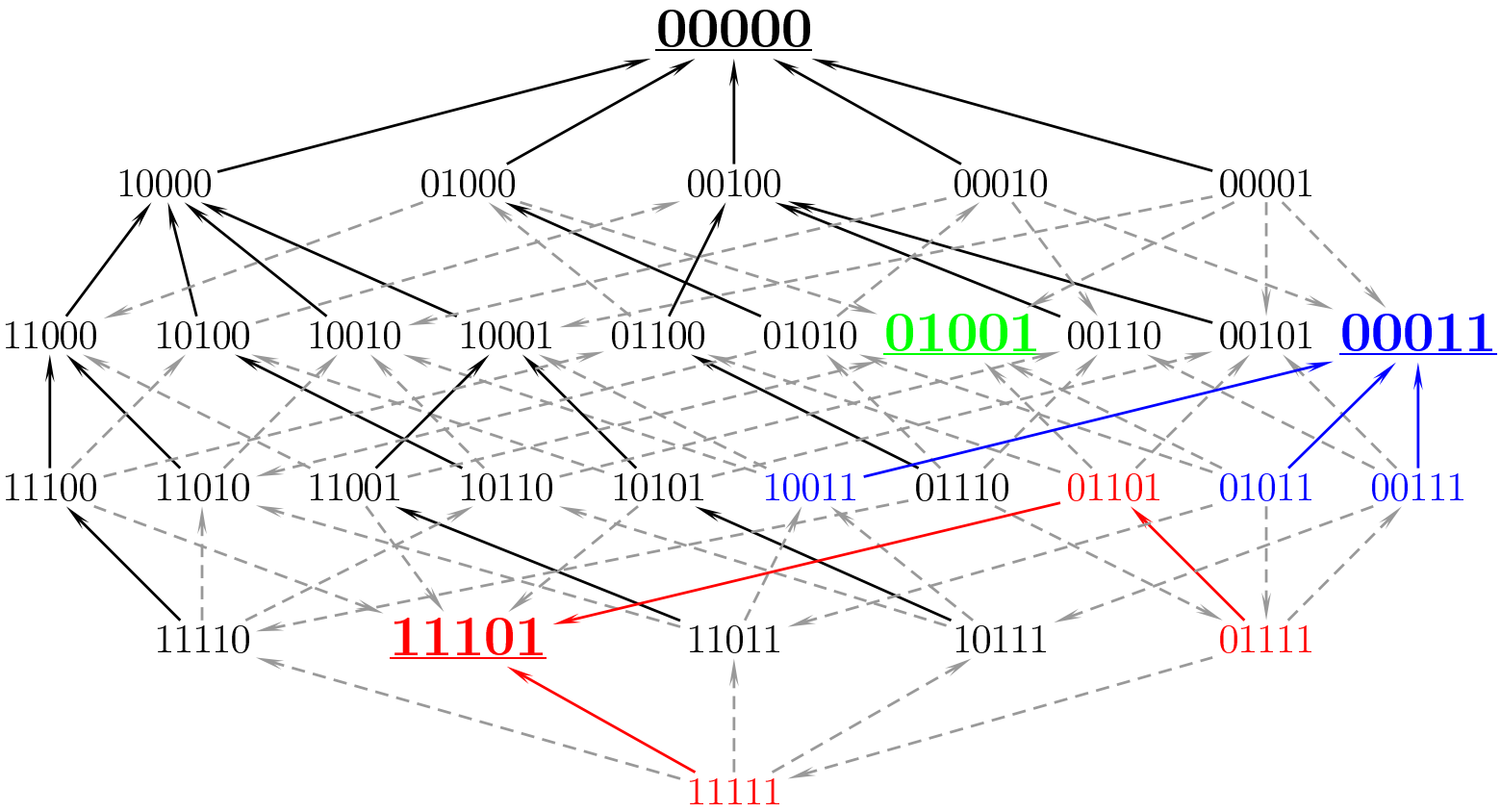
- Multi-peakedness is guaranteed if all instances of pairwise sign epistasis are reciprocal

Crona et al., JTB 2013

- Question for this talk: How does genotypic complexity arise from a nonlinear phenotype-fitness map?

Empirical example: *Aspergillus niger*

J.A.G.M. de Visser, S.C. Park, JK, American Naturalist **174**, S15 (2009)

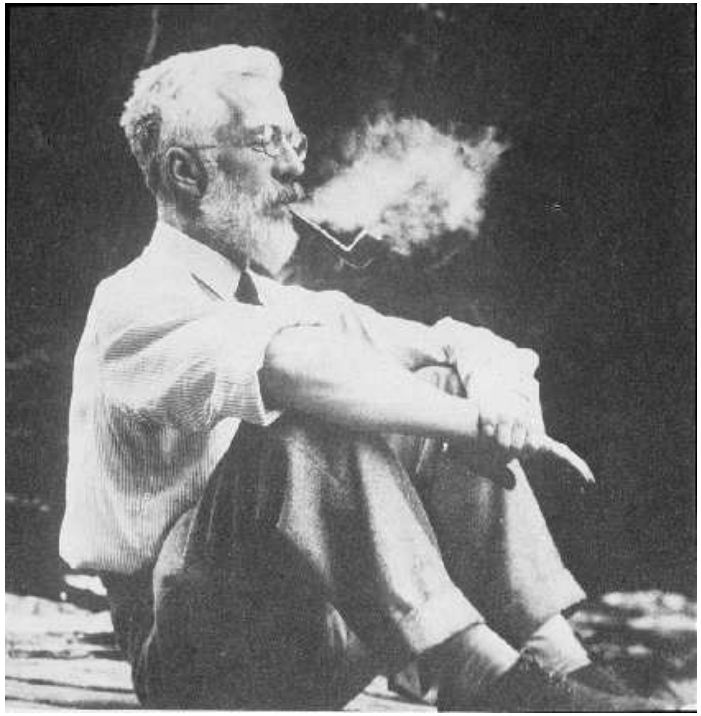


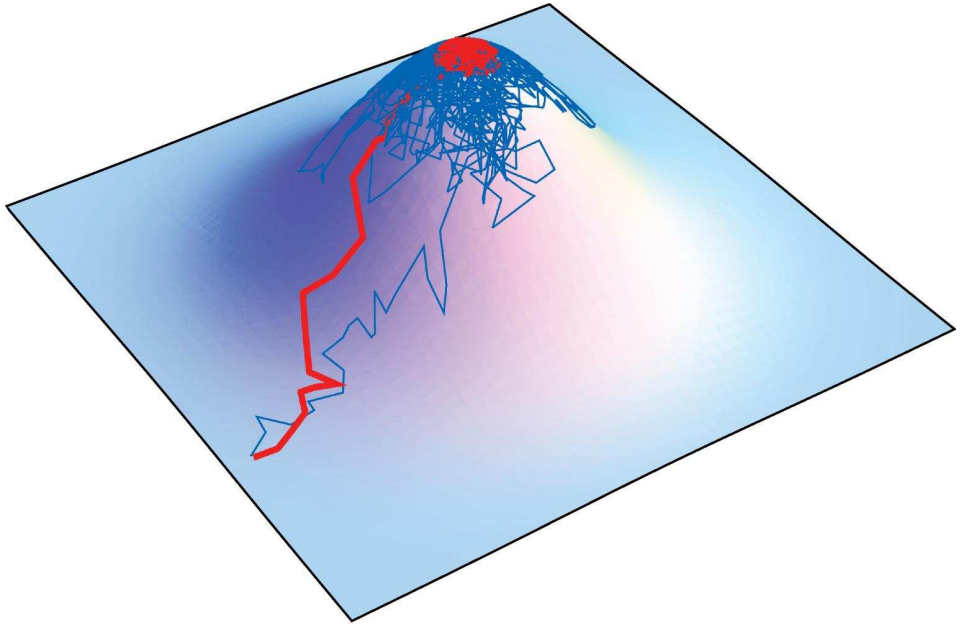
- Combinations of 8 individually deleterious marker mutations (one out of $\binom{8}{5} = 56$ five-dimensional subsets shown)
- Arrows point to increasing fitness, 3 local fitness maxima highlighted

Genotype-phenotype-fitness maps

“The statistical requirements of the situation, in which one thing is made to conform to another in a large number of different respects, may be illustrated geometrically...”

R.A. Fisher, The Genetical Theory of Natural Selection (1930)





O. Tenaillon, Annu. Rev. Ecol. Evol. Sys. (2014)

From simple phenotypes to complex genotypes

- Organism is characterized by n real-valued phenotypic traits x_i which form a vector $\vec{x} = (x_1, x_2, \ldots, x_n)$ in a n-dimensional Euclidean space

- Fitness is a (nonlinear) function $F(\vec{x})$ of the phenotype with a unique optimum at the origin $x_1 = x_2 = \ldots = x_n = 0$

- Universal pleiotropy: Mutations are isotropic random displacements in phenotypic space (univariate Gaussian)

- Additivity of phenotypes: Given two phenotypic mutations $\vec{m}_1$, $\vec{m}_2$, the phenotypic effect of the double mutant is $\vec{m}_{12} = \vec{m}_1 + \vec{m}_2$ Martin et al. 2007

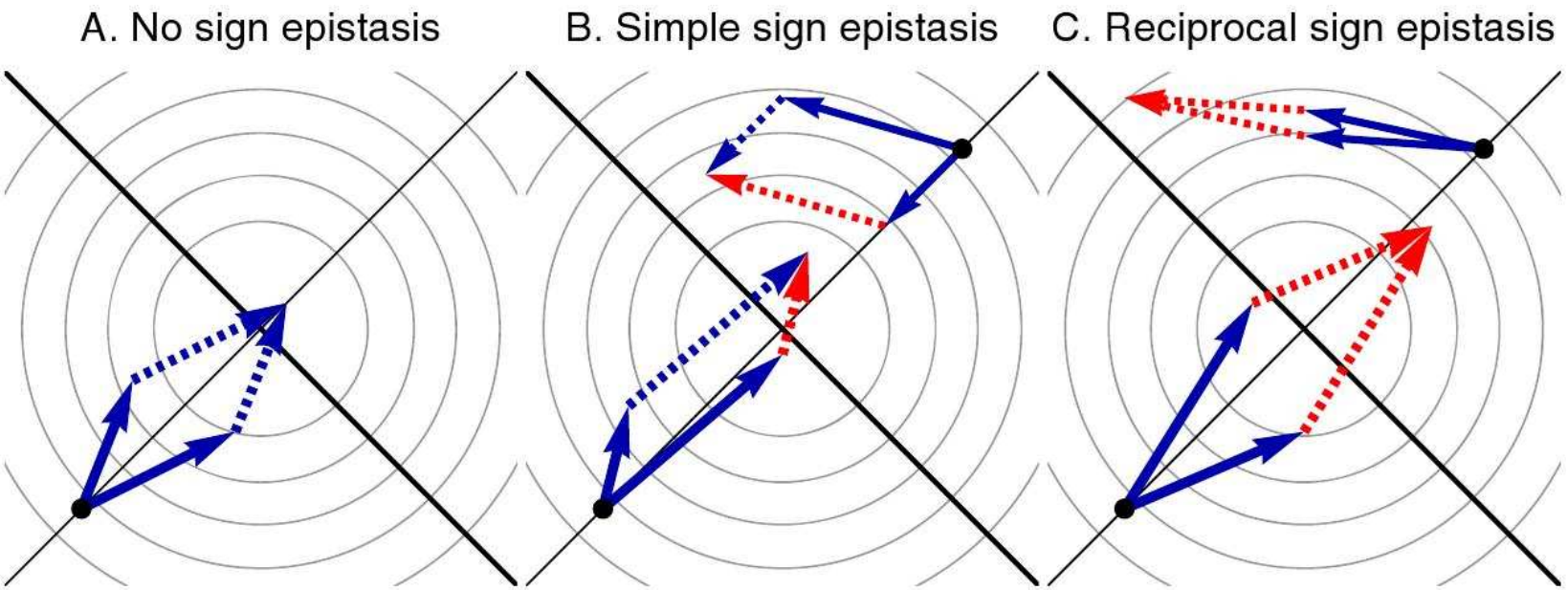
- Then the phenotypic landscape $F(\vec{x})$ induces a genotypic landscape

$$f(\tau_1, \ldots, \tau_L) = F\left(\vec{Q} + \sum_{i=1}^{L} \tau_i \vec{m}_i\right)$$

where $\vec{Q}$ represents the wildtype and the $\vec{m}_i$ are a fixed set of mutations

Sign epistasis in Fisher's geometric model

Blanquart et al., Evolution (2014)

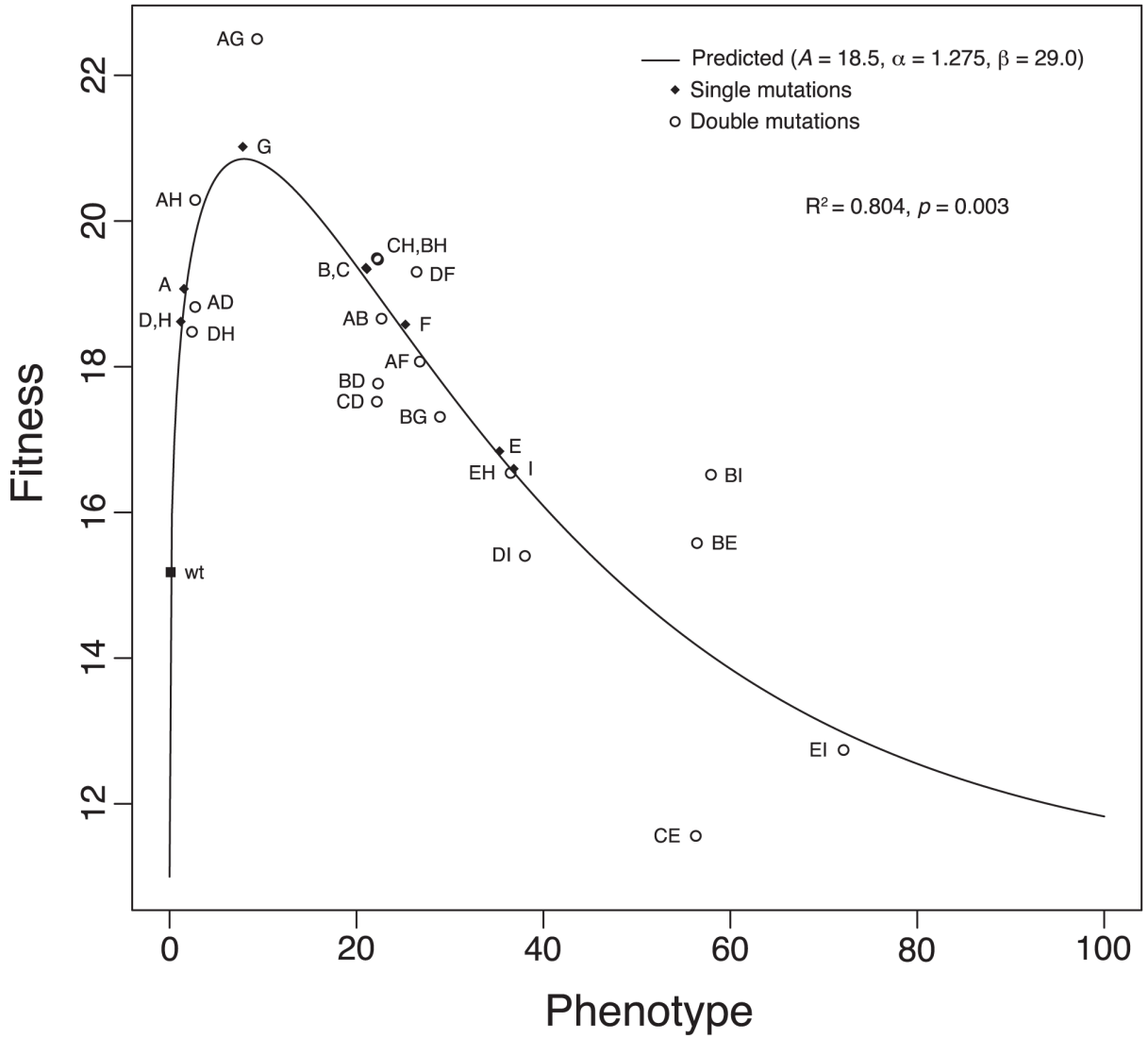


Two distinct mechanisms related to

- the overshooting of the phenotypic optimum or
- (for $n > 1$) the curvature of fitness isoclines (antagonistic pleiotropy)

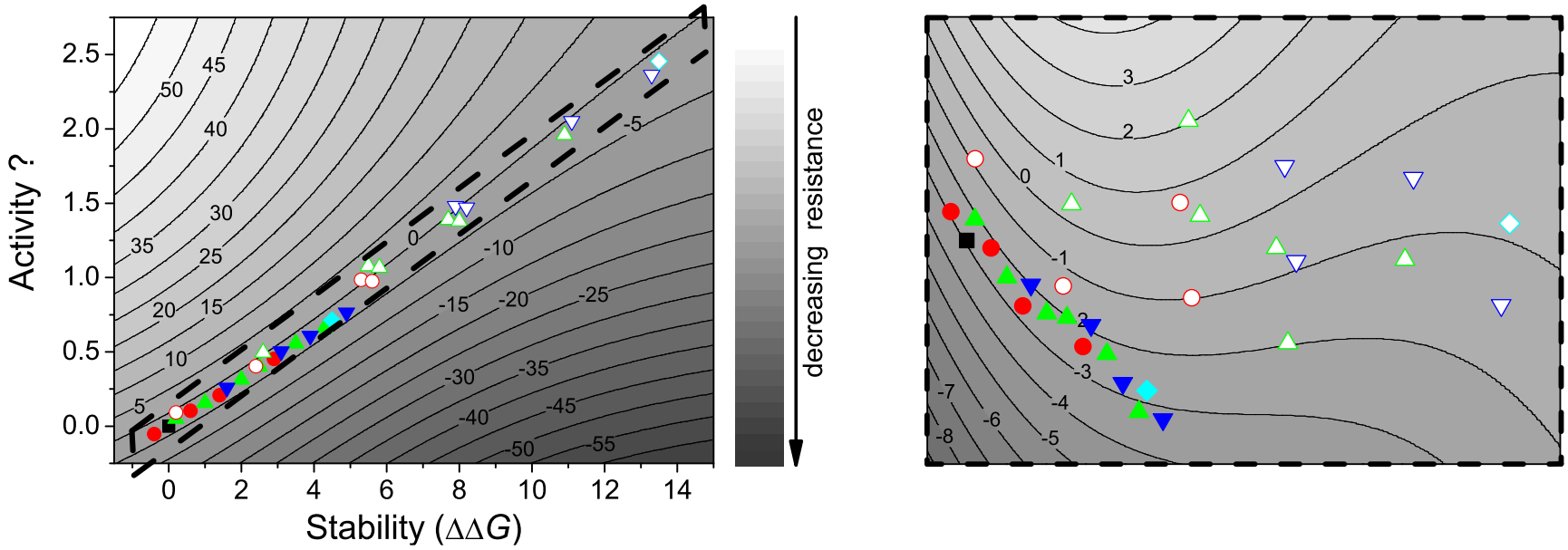
One-dimensional example: Bacteriophage ID 11

Rokyta et al., PLOS Genetics 2011



Two-dimensional example: TEM-1 β-lactamase

M.F. Schenk et al., Mol. Biol. Evol. (2013)



- Genotypic landscapes constructed from two sets of mutations increasing resistance against cefotaxime

- Data of large (open) and small (filled) effect landscapes are well described by a two-dimensional phenotype-fitness map without an optimal phenotype

Diminishing returns epistasis in Aspergillus nidulans

S. Schoustra, S. Hwang, JK and J.A.G.M. de Visser, Proc. Roy. Soc. B (2016)

Experimental system

- 244 beneficial mutants of *A. nidulans* collected from the boundary of growing colonies in complex (rich) or minimal (poor) medium
- Generated 55 pairwise combinations between mutations of similar effect using sexual crosses
- Goal: Quantify the dependence of pairwise epistatic interaction

$$\varepsilon_{ab} = \Delta f_{ab} - (\Delta f_a + \Delta f_b)$$

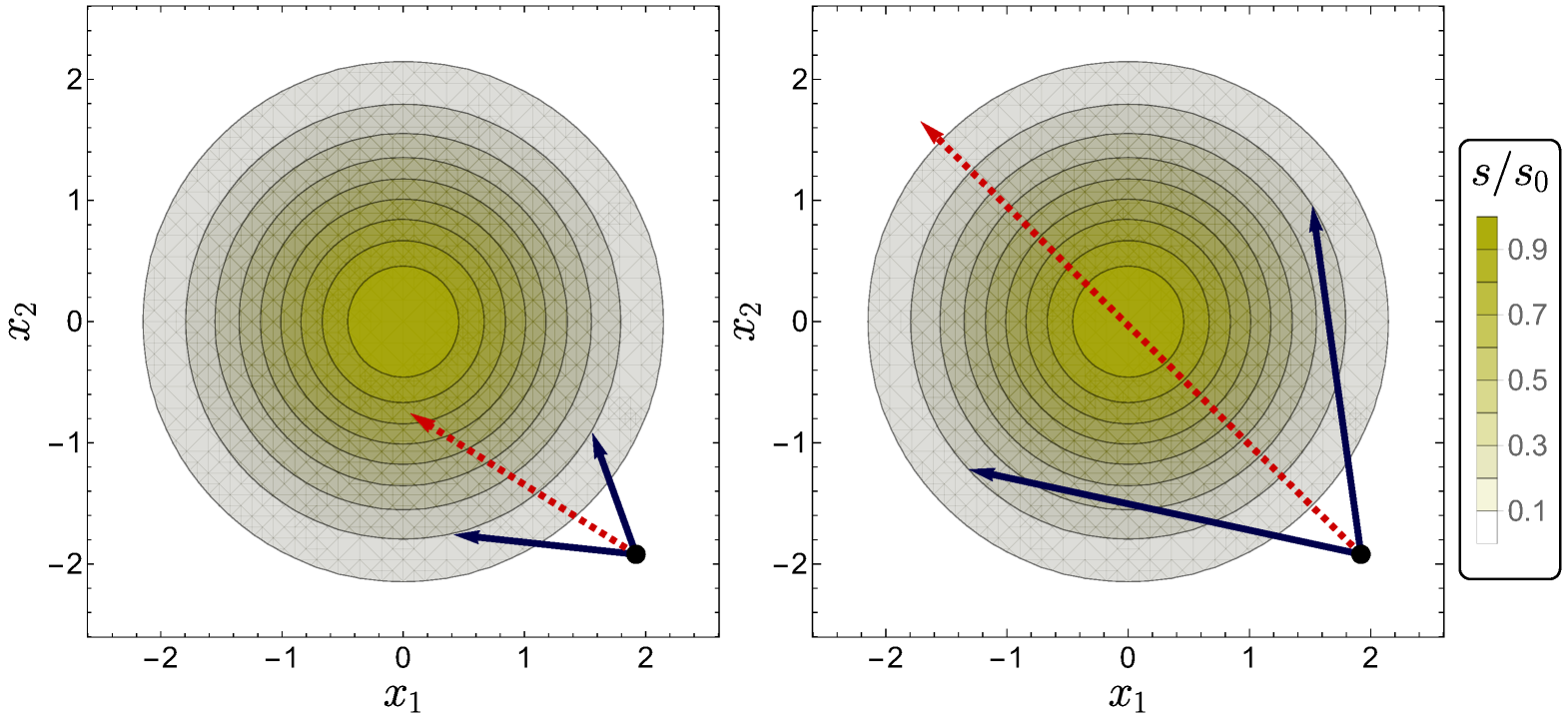
on the strength $s \approx \Delta f_a \approx \Delta f_b$ of single mutations

- Since double mutant fitness is determined by measuring the growth rate of colonies containing all four types, it can be detected only if

$$\Delta f_{ab} > \max\{\Delta f_a, \Delta f_b\} = s \text{ or } \varepsilon_{ab} > -s$$

- Data show that $\varepsilon_{ab} < 0$ and is negatively correlated with s

Diminishing returns epistasis from FGM

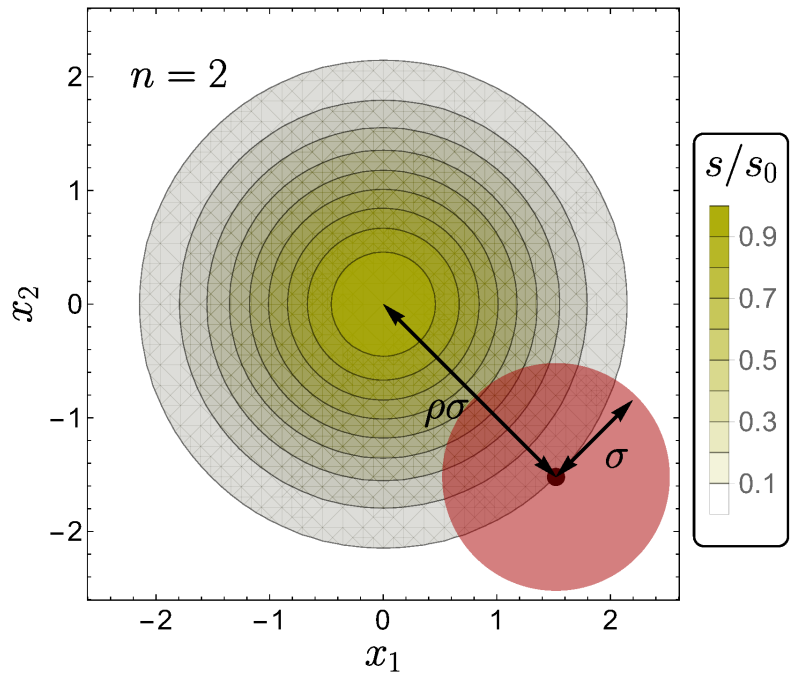


magnitude epistasis

sign epistasis

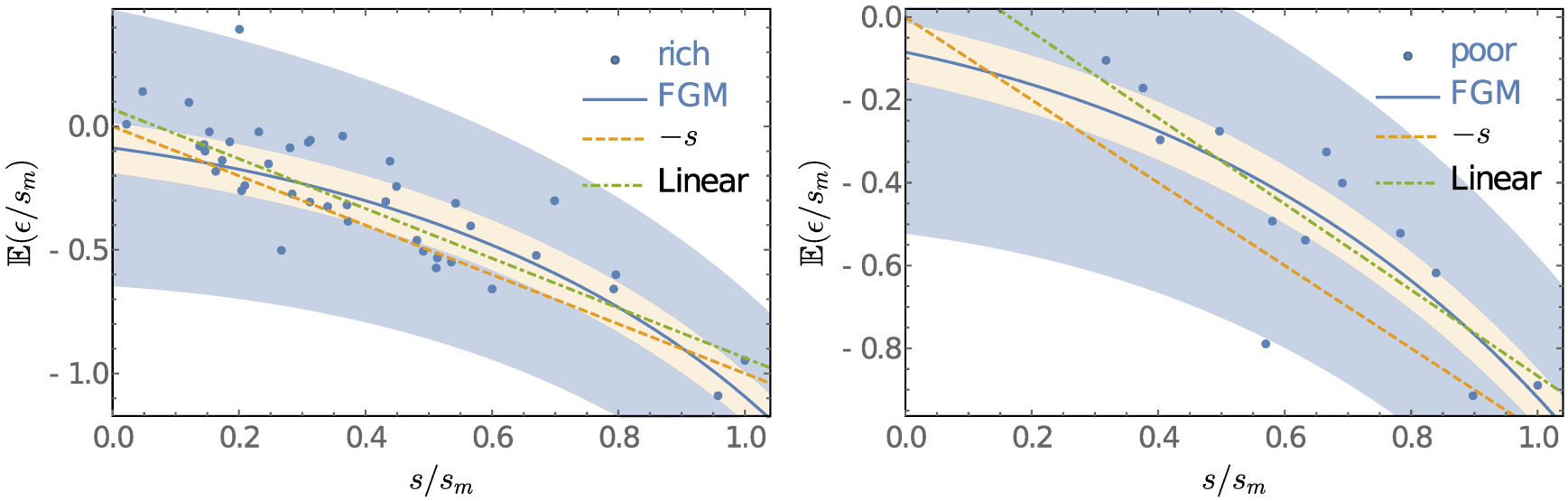
- Pairs of mutations with the same fitness effect differ widely in their epistatic interactions
- FGM contains a mechanism of intrinsic variability

FGM parameters and fitness function



- Mutations are drawn from an n-dimensional normal distribution with standard deviation σ
- Wild type resides at distance d from phenotypic optimum
  $\Rightarrow$ scaled distance d/σ
- Phenotypic fitness function $F(\vec{x}) = -s_0|\vec{x}/d|^2$

Fit of FGM to data



- Measurement error (inner pink region) is insufficient to explain variability
- Crowding of data points around the line $\varepsilon = -s$; outliers below this line originate from a tradeoff between germination and growth
- FGM parameters: $d/\sigma = 6.89,\ n = 19.3,\ s_0/s_m = 1.41$ (rich)
  $d/\sigma = 9.81,\ n = 34.8,\ s_0/s_m = 1.62$ (poor)
- How to interpret the differences in n?

Genotypic complexity of FGM

S. Hwang. S.-C. Park, JK, Genetics 206:1049-1079 (2017)

Phenotypic and genotypic complexity

- Fisher (1930) showed that the probability of a phenotypic mutation of size r to be beneficial is (for large n)

$$P_b = \frac{1}{\sqrt{2\pi}} \int_x^\infty dt\, e^{-t^2/2} = \frac{1}{2}\mathrm{erfc}(x/\sqrt{2})$$

  with $x = \sqrt{n}r/d = n\sigma/d$

- This implies that $P_b \to 0$ when $n \to \infty$ at fixed scaled distance to the phenotypic optimum.

- As a consequence, the probability that the phenotype is at a genotypic fitness maximum is

$$P_{\max} = (1 - P_b)^L \to 1$$

- Does this imply that the genotypic landscape becomes more rugged (complex) with increasing phenotypic dimension n?

Fraction of sign epistasis

- For phenotypic mutation vectors $\vec{m}_1$ and $\vec{m}_2$ define the quantities

$$R_{1,2} = \frac{1}{n}\left(|\vec{Q}+\vec{m}_{1,2}|^2 - |\vec{Q}|^2\right), \quad R = \frac{1}{n}\left(|\vec{Q}+\vec{m}_1+\vec{m}_2|^2 - |\vec{Q}|^2\right)$$

  which determine whether the respective mutations are deleterious (>0) or beneficial (<0)

- For large n their joint distribution under FGM is

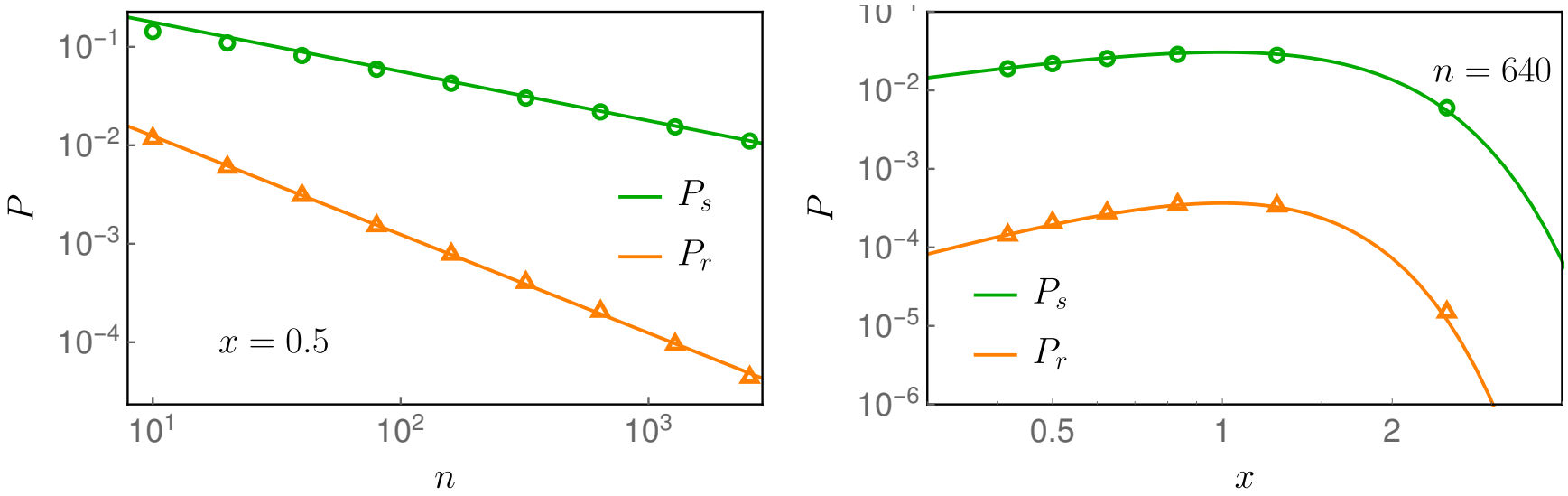
$$\mathscr{P}(R_1,R_2,R) \sim \exp\left[-\frac{1}{8}n(R_1+R_2-R)^2 - \frac{1}{2}x^2[(R_1-1)^2+(R_2-1)^2]\right]$$

  where mutation vectors have scale $\sigma = 1$ for simplicity

- Large n enforces $R = R_1 + R_2$ and hence suppresses epistasis

- Fraction of simple P_s and reciprocal P_r sign epistasis can be computed by integrating $\mathscr{P}$ over suitable domains

Fraction of sign epistasis: Asymptotic results

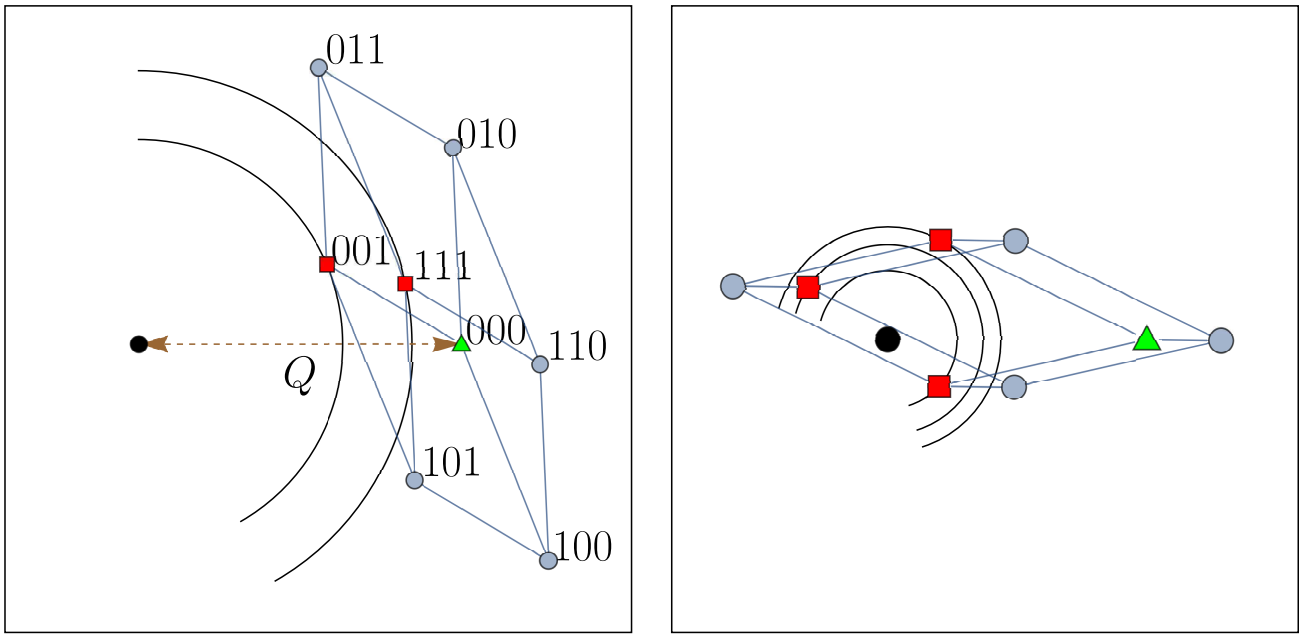


- For $n \to \infty$ at fixed x the leading behavior is

$$P_s \approx \frac{4xe^{-x^2/2}}{\pi\sqrt{n}}, \qquad P_r \approx \frac{2x^2e^{-x^2}}{\pi n}$$

- Sign epistasis becomes rare for large n and generally $P_r \ll P_s$
- Similar behavior obtains for beneficial single mutations

Genotype-phenotype-fitness map for multiple loci



- The mapping

$$\tau \to \vec{z}(\tau) = \vec{Q} + \sum_{i=1}^{L} \tau_i \vec{m}_i$$

projects L-dimensional hypercube onto n-dimensional phenotype space

- Figure shows the wild type phenotype (green triangle) and genotypic fitness maxima (red squares) for $L = 3, n = 2$

Number of genotypic maxima

- A common global quantifier of genotypic complexity is the expected number of genotypic fitness maxima $\langle \mathcal{N} \rangle$

- Experience with random field models shows that in many cases

$$\langle \mathcal{N} \rangle \sim \exp[\Sigma^* L] \quad \text{for} \quad L \to \infty$$

  which defines the genotypic complexity $\Sigma^* \geq 0$

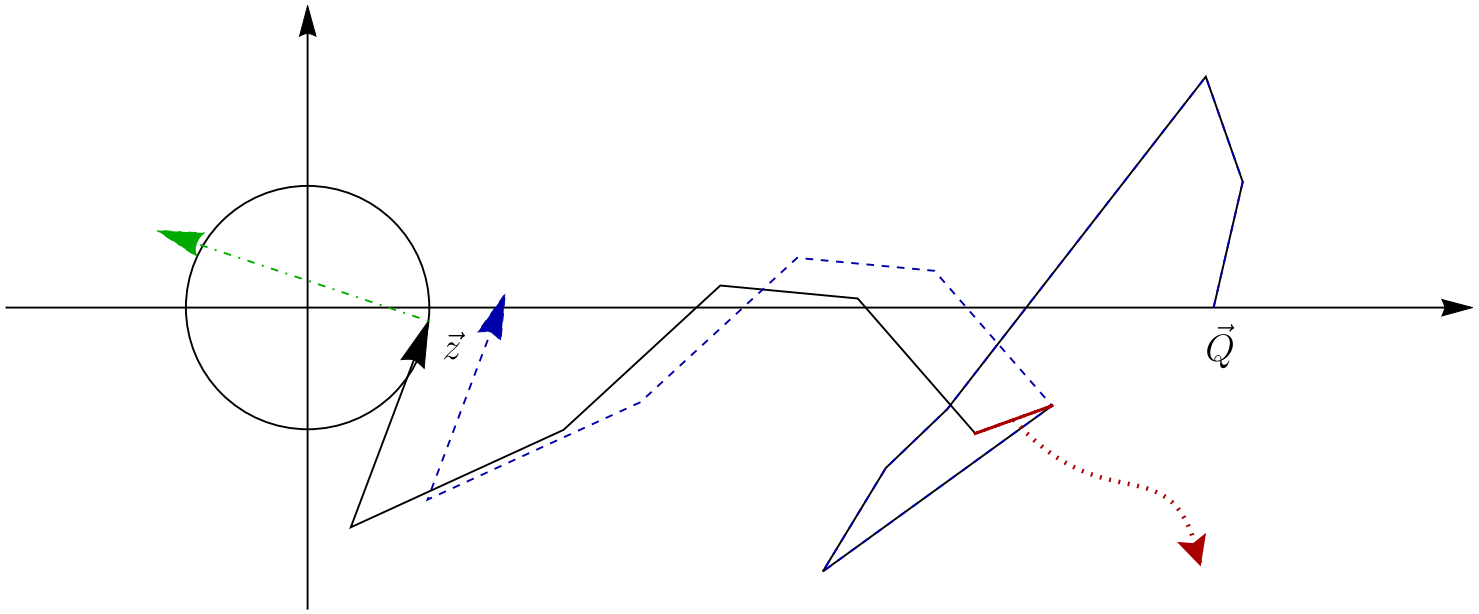
- Within FGM, a genotype $\tau = (\tau_1, \tau_2, ..., \tau_L)$ with phenotype

$$\vec{z} = \vec{Q} + \sum_{i=1}^{L} \tau_i \vec{m}_i$$

  is a fitness maximum iff $|\vec{z}| < |\vec{z} + (1 - 2\tau_j)\vec{m}_j|$ for all $j = 1, ..., L$

- This is true with unit probability if the corresponding phenotype is optimal, i.e. if $\vec{z} = 0 \quad \Rightarrow \quad$ genotypic maxima arise from near-optimal phenotypes

Number of genotypic maxima: Geometry



- Composition of mutation vectors defines a random walk in phenotype space with endpoint $\vec{z}$
- Removal of a leg of the walk (dashed) or addition of further mutation vectors (dash-dotted) should not move the endpoint into the circle of radius $|\vec{z}|$
- To generate genotypic maxima, the walk needs to be “stretched” towards the origin

Number of genotypic maxima: Asymptotics

- Expected number of maxima for large L is given by $\langle \mathcal{N} \rangle \sim L^{-(1+n/2)} \exp[\Sigma^* L]$ where Σ^* is the solution of the variational problem

$$\Sigma^* = \max_{\phi \in [0,1]} \left\{ -\phi \log \phi - (1-\phi) \log(1-\phi) - \frac{q^2}{2\phi} \right\}$$

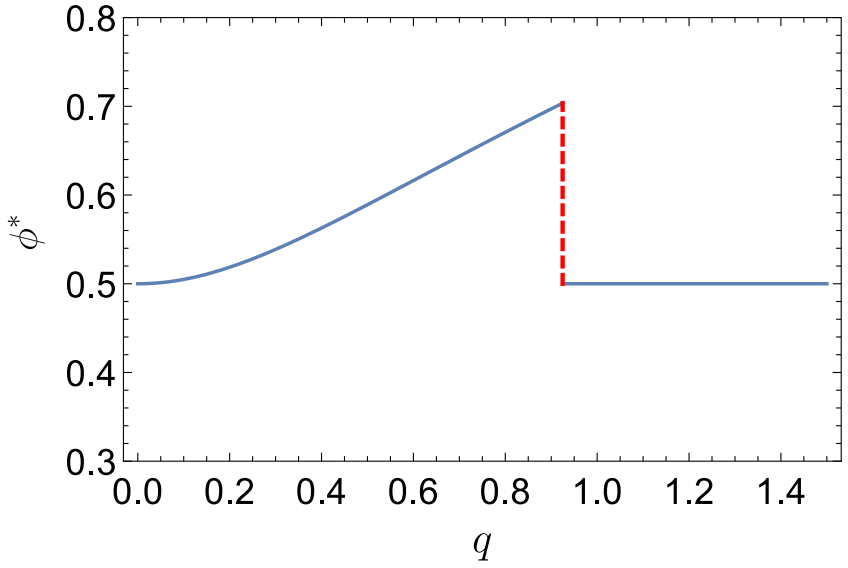
  with

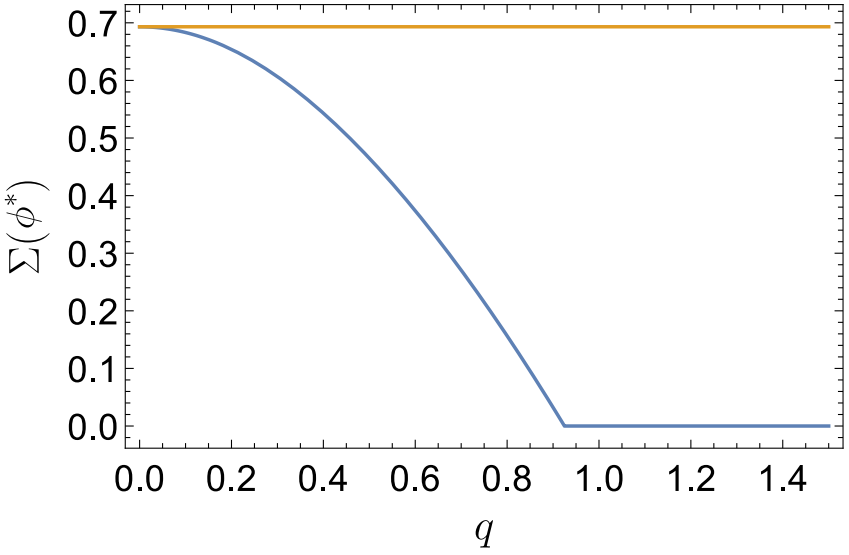
  - ϕ: fraction of mutations that are present (= have $\tau_i = 1$)
  - $q = |\vec{Q}|/L$: scaled distance of the wild type phenotype to the optimum

- Variational problem encodes a tradeoff between the abundance of genotypes ("entropy") and their likelihood to reach the phenotypic optimum ("energy")

- The number of maxima decreases with increasing phenotypic dimension, but to leading (exponential) order it is independent of n

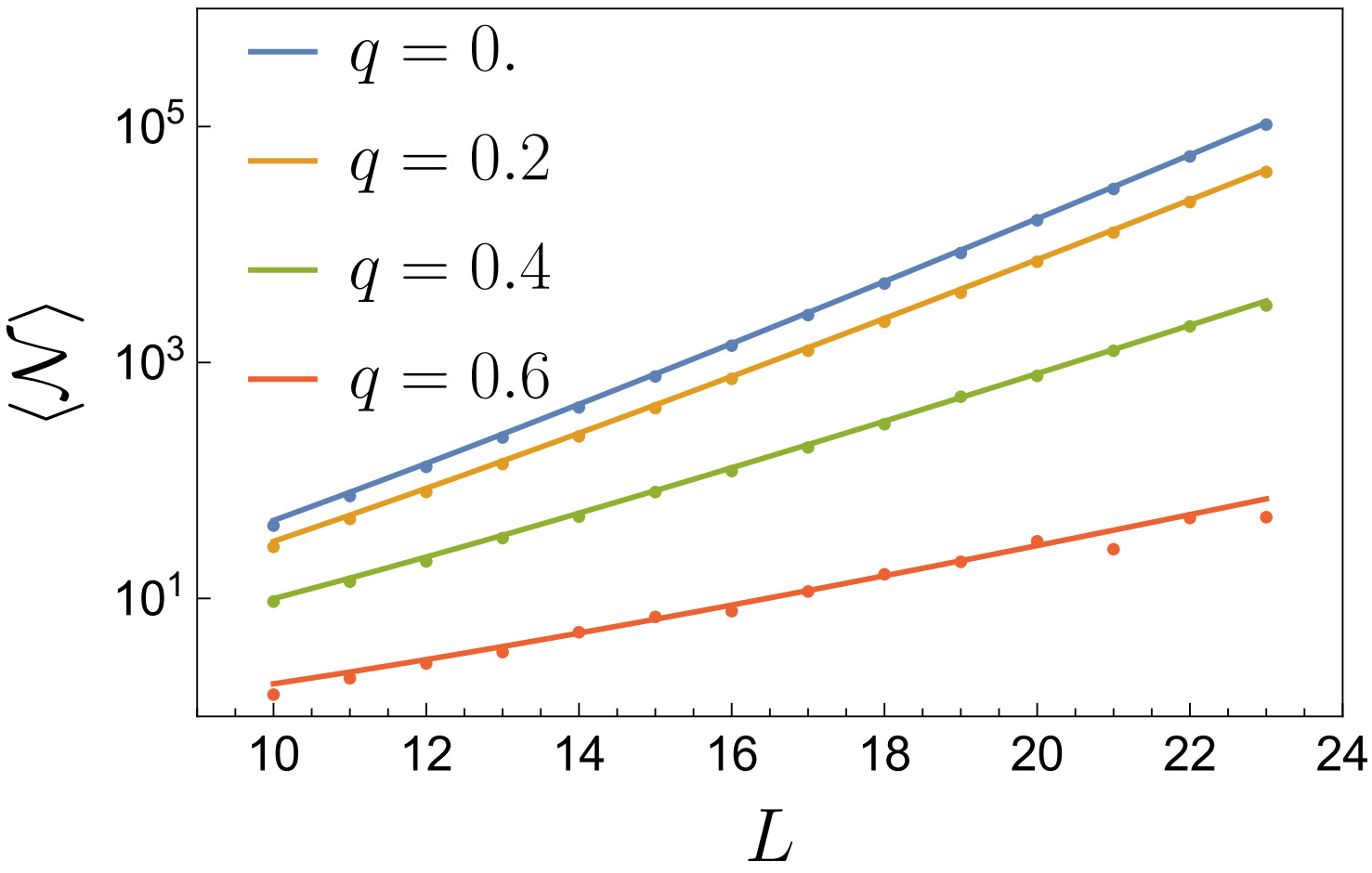
Number of genotypic maxima: Phase transition





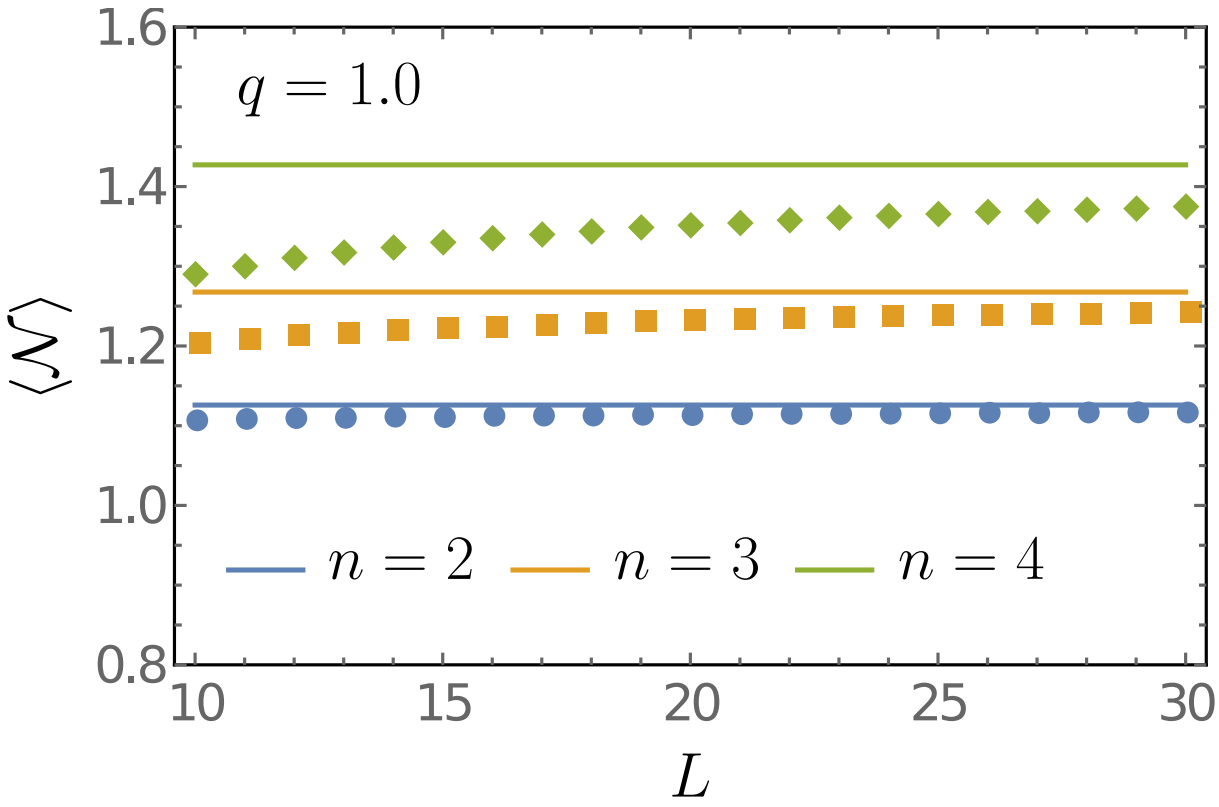
- $\Sigma^*(q=0) = \ln 2 \Rightarrow \langle \mathcal{N} \rangle \sim \frac{2^L}{L^{1+n/2}}$, to be compared to an uncorrelated random fitness landscape ("house-of-cards model") with $\langle \mathcal{N} \rangle \sim \frac{2^L}{L}$
- Σ^* vanishes at a first order phase transition at $q = q_c \approx 0.924809$
- For $q > q_c$ the number of maxima reaches a finite limit for $L \to \infty$ which however grows exponentially with n

Number of genotypic maxima: Regime I ($q < q_c$)



- Comparison to simulations for $n = 1$

Number of genotypic maxima: Regime II ($q > q_c$)



- Horizontal lines show the analytic expression

$$\langle \mathcal{N} \rangle \approx \mathcal{N}_> \equiv \left(\frac{q - q_0}{q} \exp\left[\frac{1}{q/q_0 - 1} \right] \right)^{n-1} \quad \text{with} \quad q_0 = \frac{1}{\sqrt{2\pi}} \approx 0.399$$

Digression: FGM as a spin glass model

- For a parabolic phenotypic fitness function $F(\vec{x}) = -|\vec{x}|^2$ the genotypic fitness landscape becomes

$$f(\tau) = -|\vec{Q}|^2 - 2\sum_{i=1}^{L}(\vec{Q}\cdot\vec{m}_i)\tau_i - \sum_{i,j=1}^{L}(\vec{m}_i\cdot\vec{m}_j)\tau_i\tau_j$$

  which corresponds to an antiferromagnetic Hopfield model with n continuous patterns and random fields of strength $\sim |\vec{Q}|$

- The linear part dominates for large $|\vec{Q}| \Rightarrow$ fitness landscape is less rugged when wildtype phenotype is far from the origin

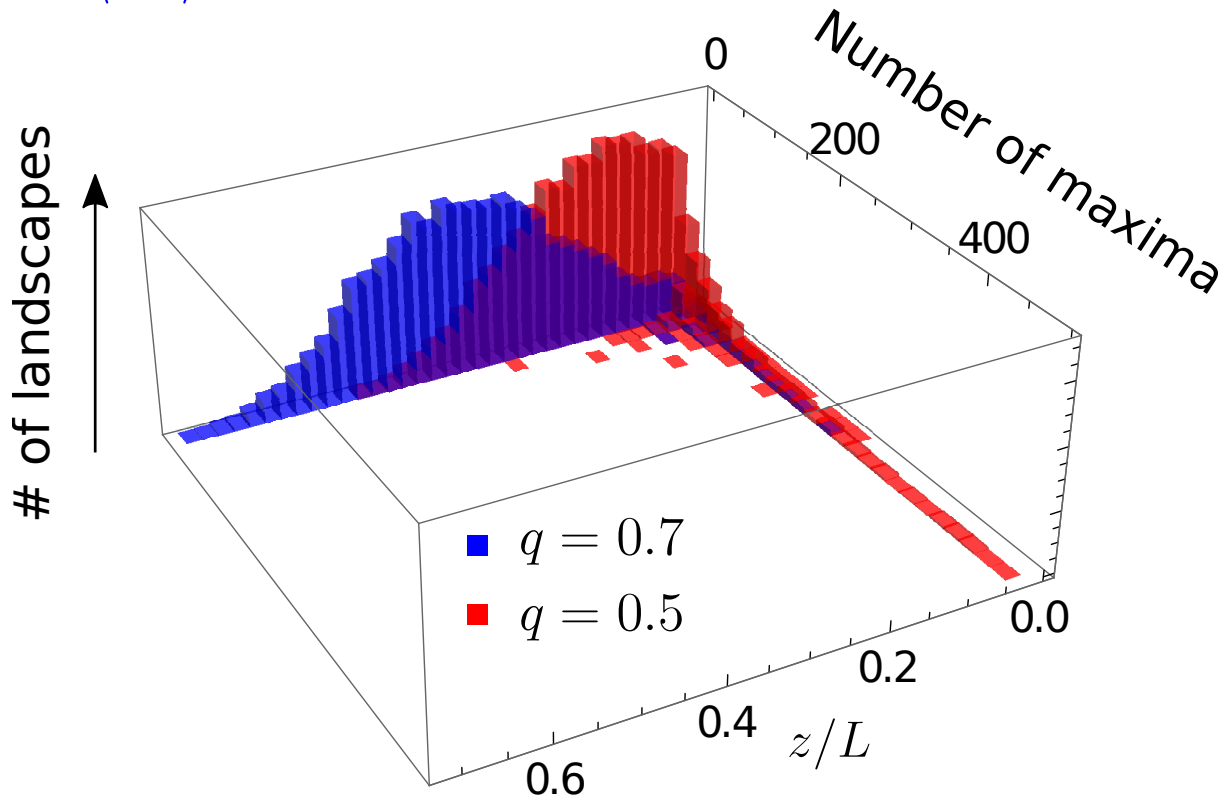
- The model displays a zero temperature phase transition at $q = q_0 < q_c$ where the extensive part of the ground state entropy S_0 vanishes

  S. Hwang, D. Dean, JK (unpublished)

- Since $S_0 \approx \langle \ln \mathcal{N} \rangle$, for $q_0 < q < q_c$ mean and typical values of $\mathcal{N}$ differ

Coexistence and rare events

- In the coexistence region $q_0 < q < q_c$, $\langle \mathcal{N} \rangle$ is dominated by rare realizations with exponentially many maxima, whereas for typical realizations $\langle \mathcal{N} \rangle \approx \mathcal{N}_>$

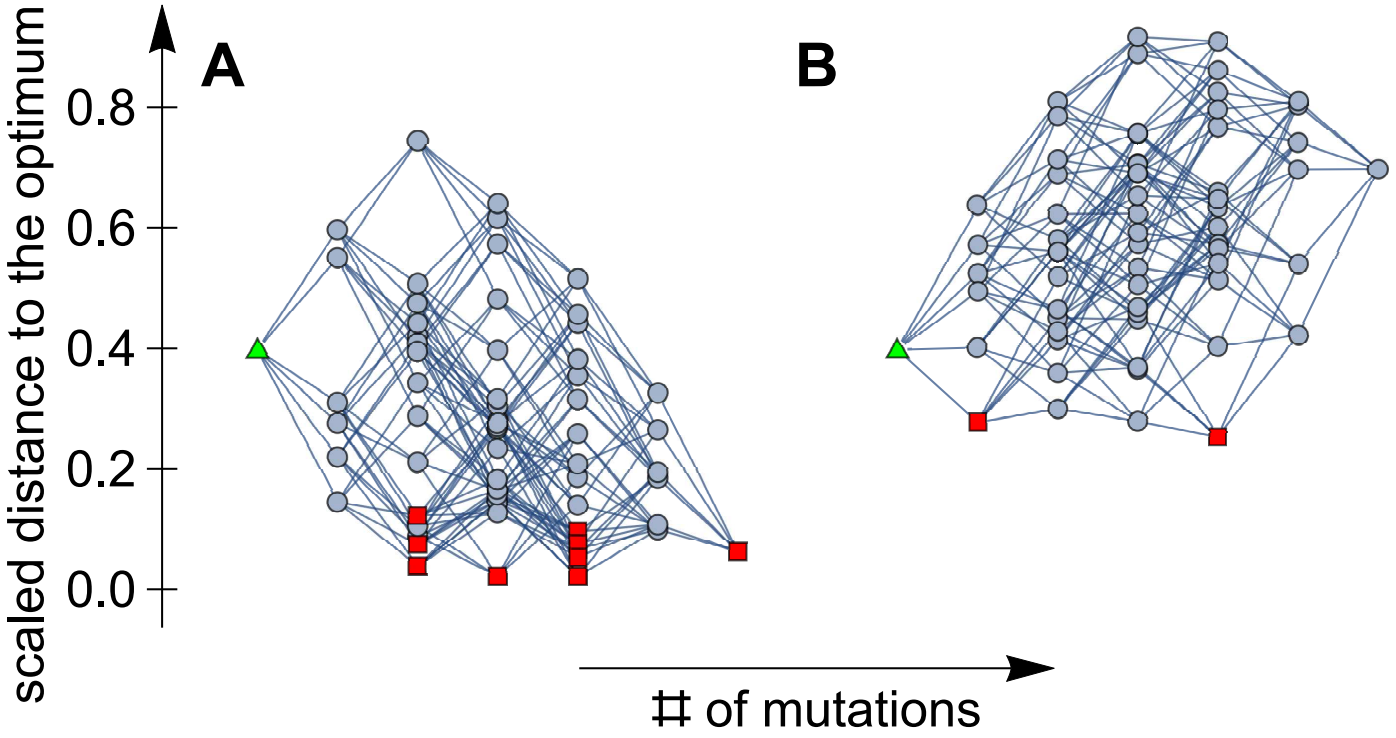


- These rare realizations are those for which the phenotypic displacements approach close to the optimum $z = 0$

Coexistence and rare events

- Coexistence leads to a large heterogeneity of landscape structures that has been observed in previous work

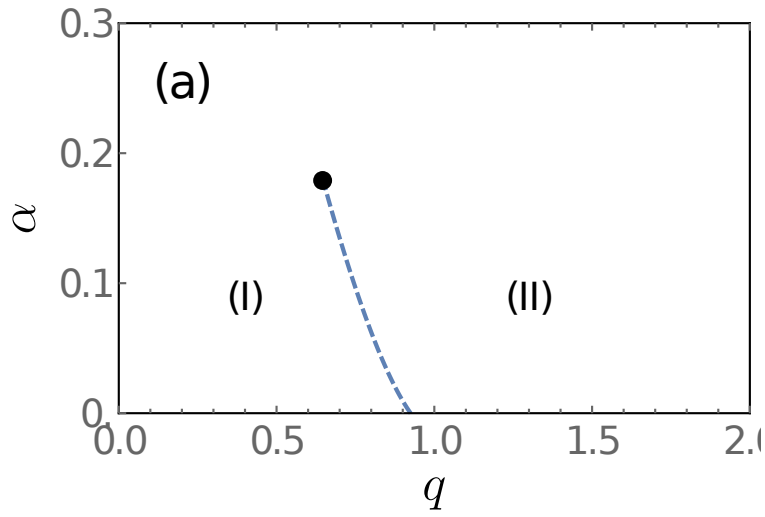
Blanquart et al. 2014; Blanquart and Bataillon 2016

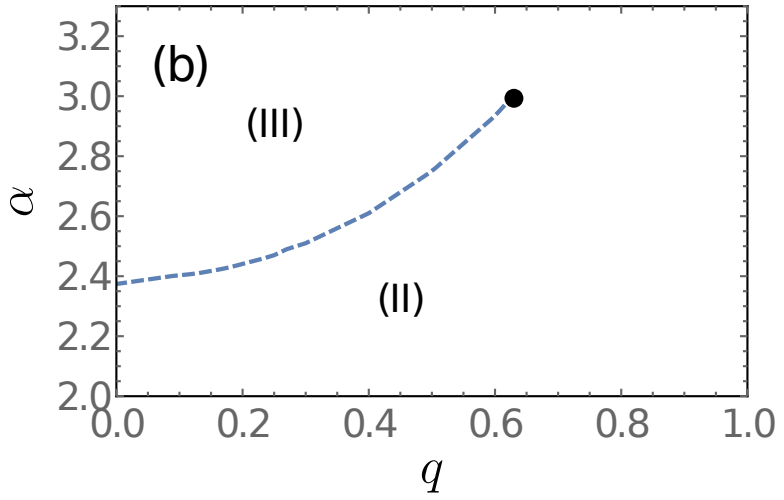


Two landscape realizations for $q = 0.5, L = 6, n = 2$

Joint limit of phenotypic and genotypic dimensions

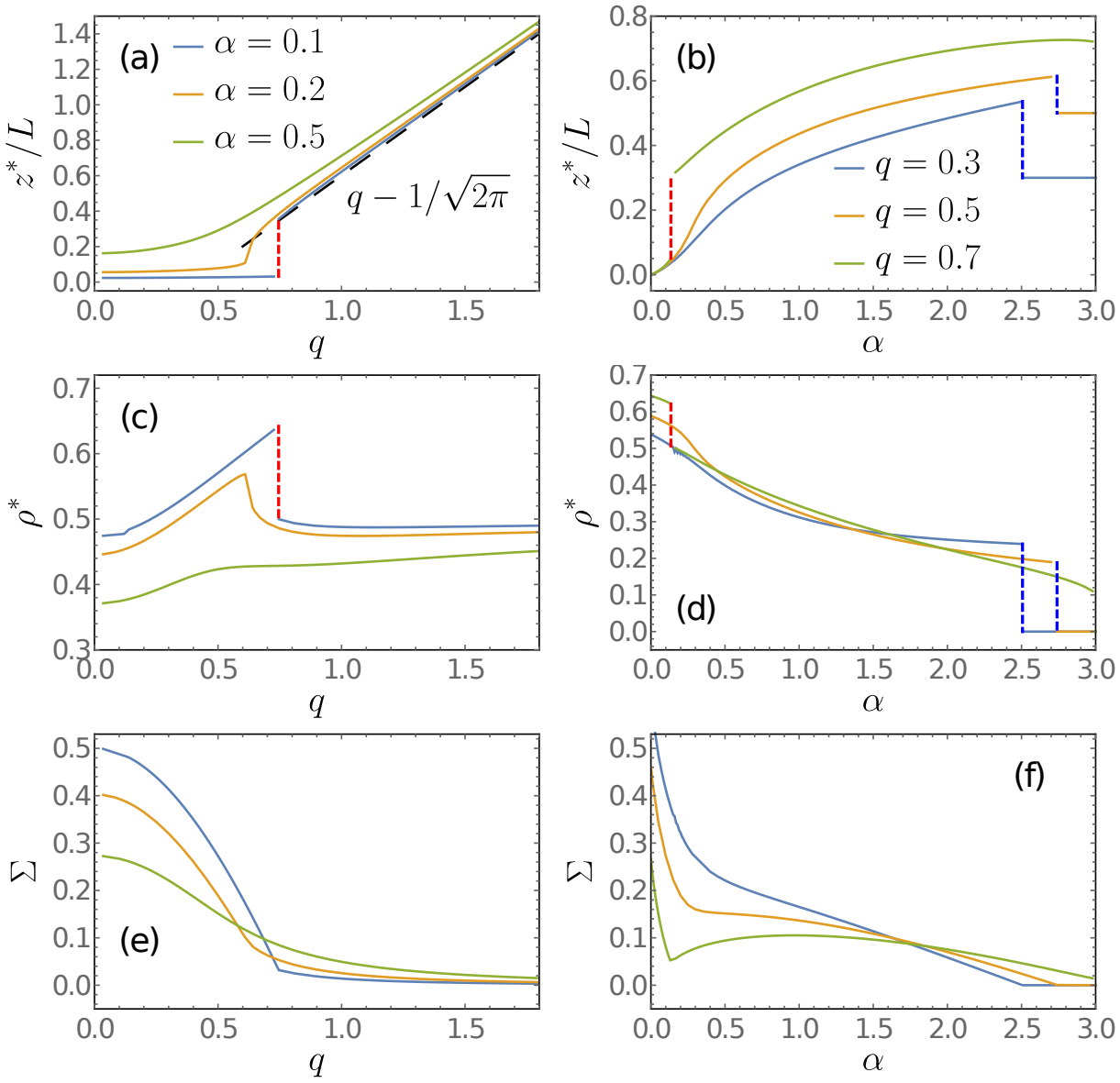
- Phenotypic dimension affects the number of genotypic fitness maxima to leading (exponential) order when $n \sim L$
- The joint limit $n, L \to \infty$ at fixed $\alpha = n/L > 0$ leads to a three-dimensional variational problem that can only be treated numerically
- Phase diagram in the (q, α)-plane comprises two lines of first-order transitions with critical endpoints



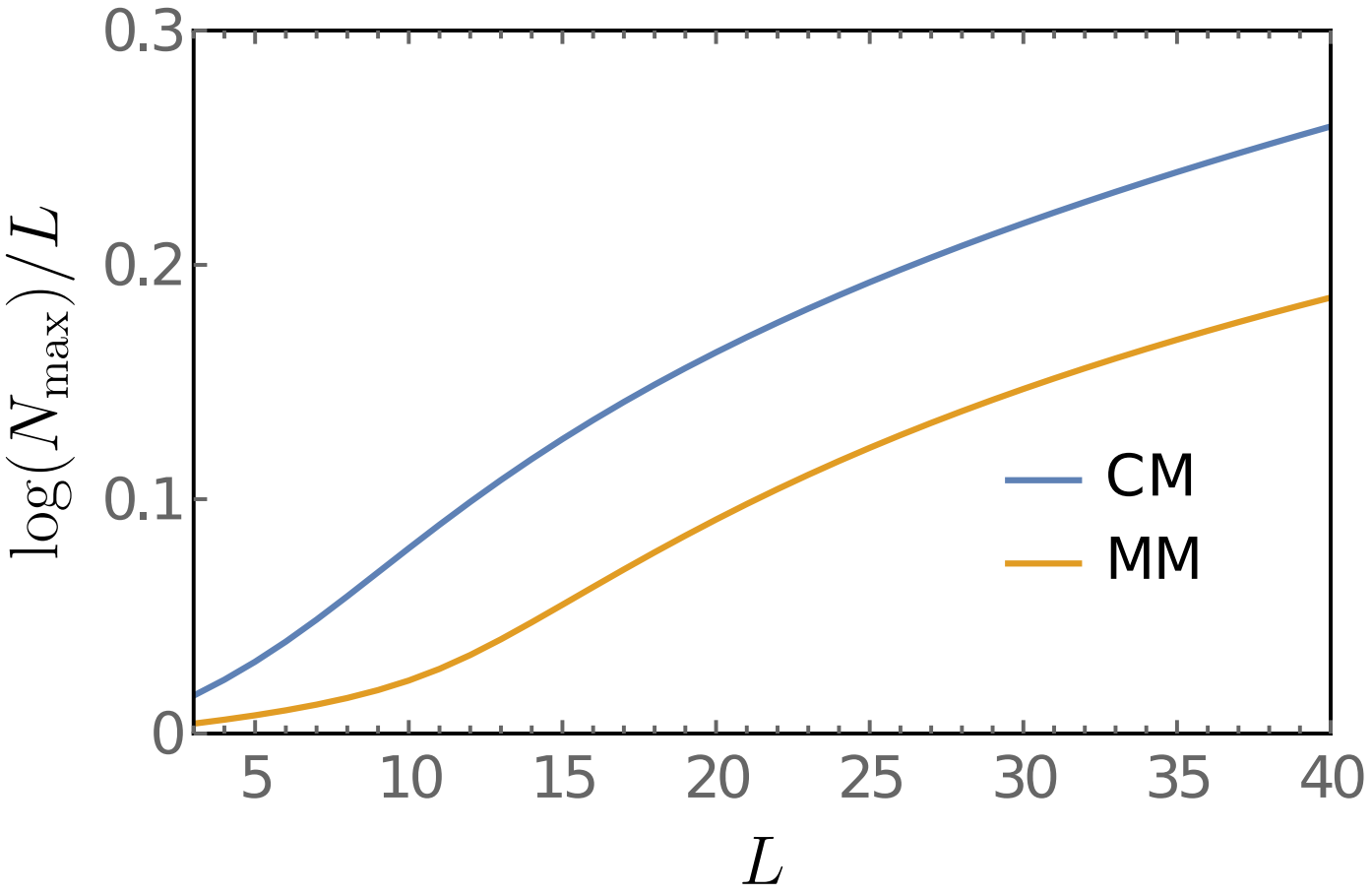


- For $n \gg L$ the landscape becomes essentially additive (regime III)

Joint limit of phenotypic and genotypic dimensions



Genotypic complexity of the A. nidulans landscapes



- Rich medium landscape (CM) is more rugged, despite having lower phenotypic dimension

Conclusions

- Fisher's geometric model is a good example of a "proof-of-concept" model

  Servedio et al., PLOS Biol. 2014

- It demonstrates how genotypic complexity can be explained in terms of additive phenotypes combined with a simple nonlinear phenotype-fitness map

- The relationship between complexity of the genotypic fitness landscape and the phenotypic dimension is complicated and often non-monotonic

- The model also provides a framework for condensing experimental data into a few phenomenological parameters, but their interpretation generally is not straightforward

- The role of rare events and sample-to-sample fluctuations remains to be better understood

Thank you !