SFB 680
Molecular Basis of
Evolutionary Innovations

Fitness landscapes and adaptive evolution

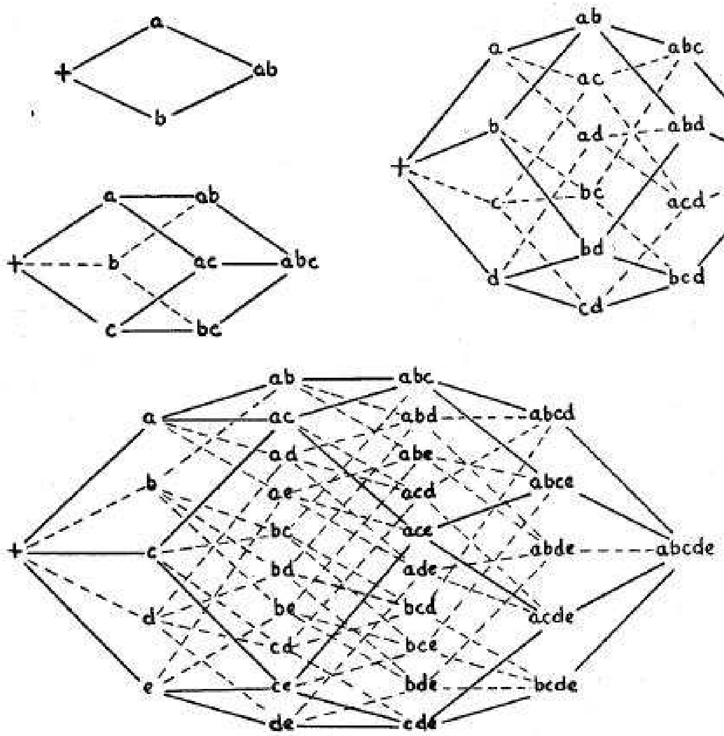
Joachim Krug
Institute for Theoretical Physics, University of Cologne, Germany

- Empirical fitness landscapes and measures of epistasis
- Accessible mutational pathways in random field models
- Adaptive walks

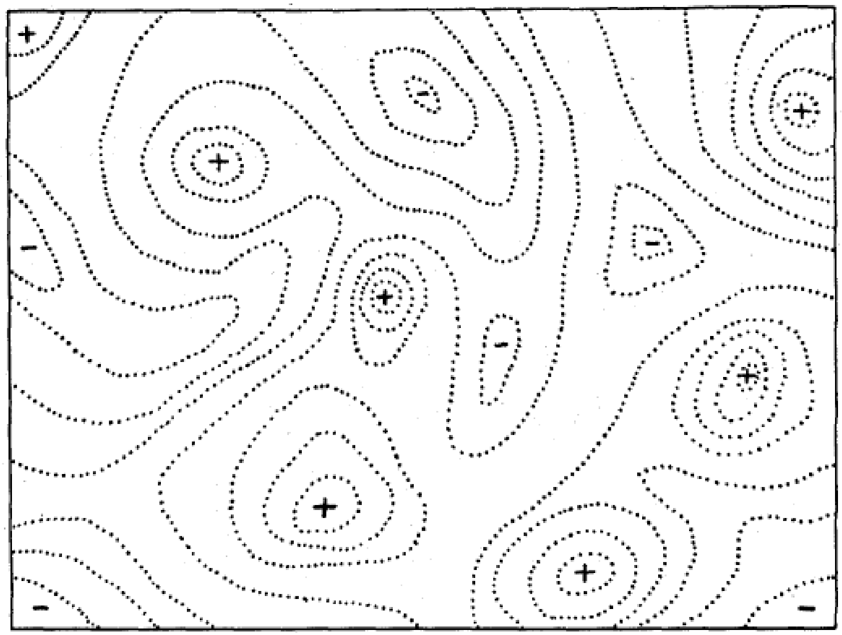
“New Directions in Probabilistic Models of Evolution“
Simons Institute for the Theory of Computing, Berkeley, May 2, 2014

Fitness landscapes

S. Wright, Proc. 6th Int. Congress of Genetics (1932)



genotype space



two-dimensional landscape cartoon

Mathematical setting

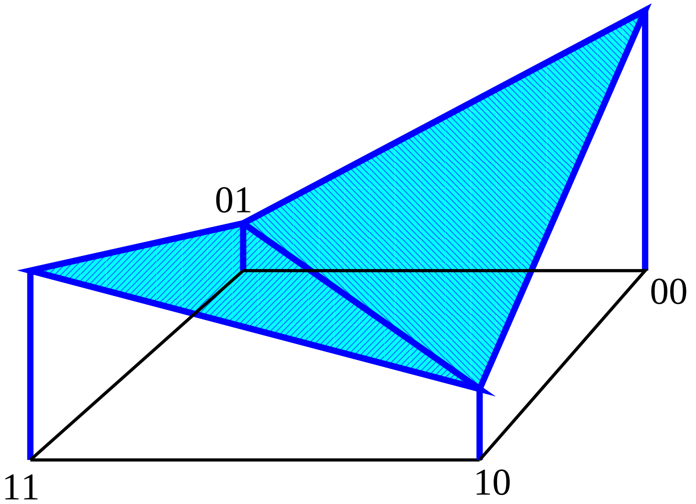
- Genotypes are binary sequences $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_L)$ with $\sigma_i \in \{0, 1\}$ or $\sigma_i \in \{-1, 1\}$ (presence/absence of mutation).

- A fitness landscape is a function $f(\sigma)$ on the space of 2^L genotypes

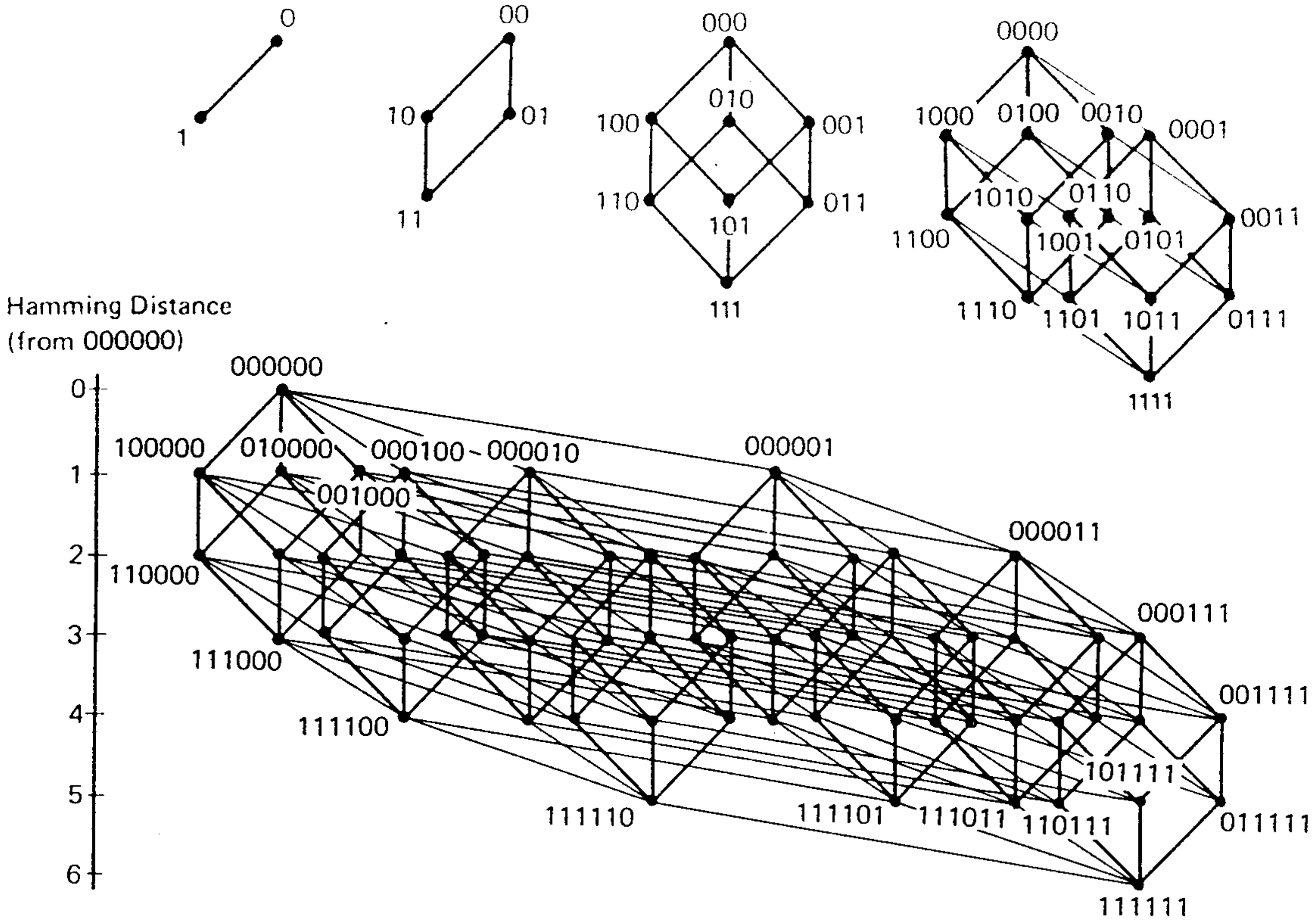
- Epistasis implies interactions between the effects of different mutations

- Sign epistasis: Mutation at a given locus is beneficial or deleterious depending on the state of other loci Weinreich, Watson & Chao (2005)

- Reciprocal sign epistasis for $L = 2$:



Binary sequence spaces are hypercubes



Measures of epistasis

Local fitness optima

Haldane 1931, Wright 1932

- A genotype σ is a local optimum if $f(\sigma) > f(\sigma')$ for all one-mutant neighbors σ'
- In the absence of sign epistasis there is a single global optimum
- Reciprocal sign epistasis is a necessary but not sufficient condition for the existence of multiple fitness peaks Poelwijk et al. 2011, Crona et al. 2013

Selectively accessible paths

Weinreich et al. 2005

- A path of single mutations connecting two genotypes $\sigma \longrightarrow \sigma'$ with $f(\sigma) < f(\sigma')$ is selectively accessible if fitness increases monotonically along the path
- In the absence of sign epistasis all paths to the global optimum are accessible, and vice versa

Fourier/Walsh decomposition

Weinberger 1991, Stadler 1996

- Any fitness landscape can be decomposed into epistatic interactions of different orders

$$f(\sigma) = a^{(0)} + \sum_{j=1}^{L} a_j^{(1)} \sigma_j + \sum_{\substack{j,k=1 \\ j>k}}^{L} a_{jk}^{(2)} \sigma_j \sigma_k + \ldots + a^{(L)} \sigma_1 \sigma_2 \ldots \sigma_L$$

- For the symmetric alphabet $\sigma_i \in \{-1, 1\}$ this amounts to an expansion in eigenfunctions of the graph Laplacian on the L-dimensional hypercube

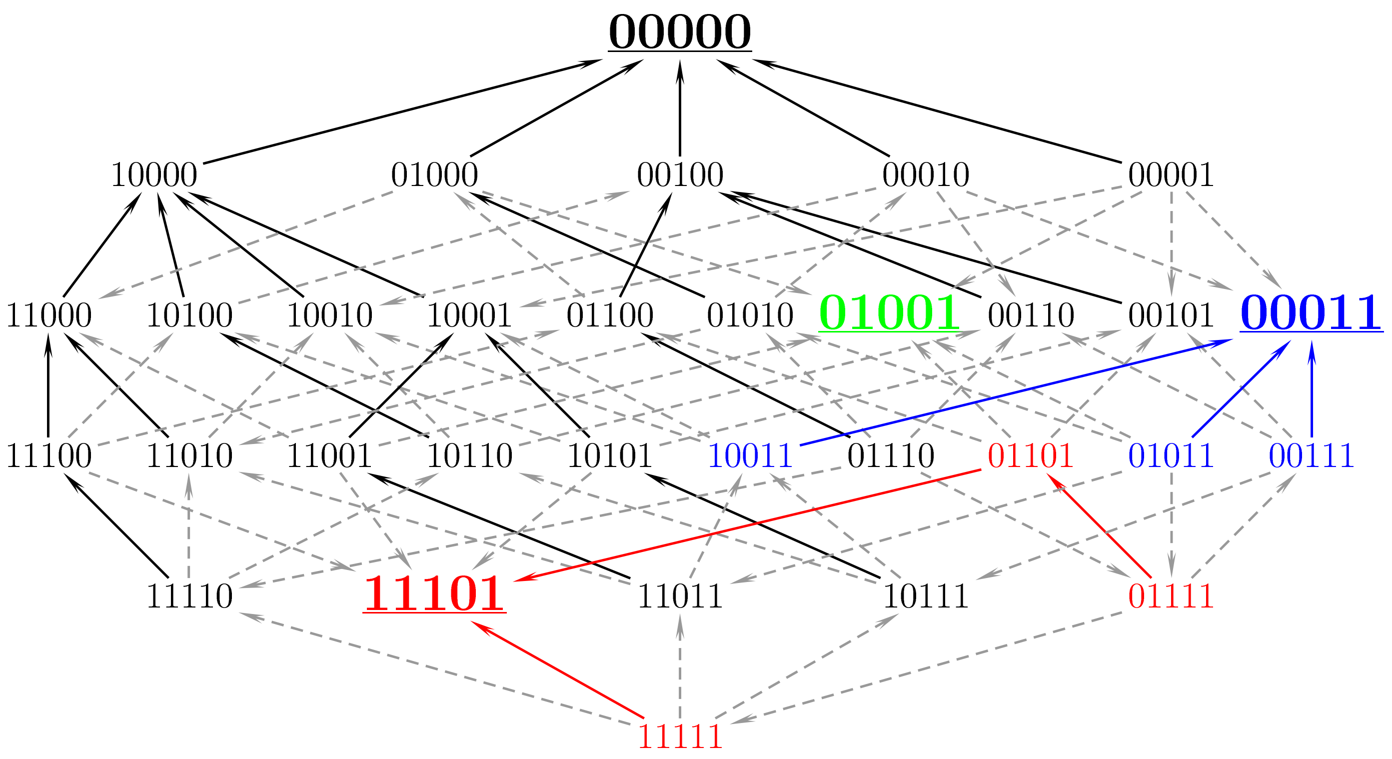
- Weight of epistatic interactions of order n is quantified by the "Fourier spectrum"

$$F_n = \frac{\beta_n}{\sum_{j=1}^{L} \beta_j} \quad \text{with} \quad \beta_n = \sum_{j=1}^{\binom{L}{n}} (a_j^{(n)})^2, \;\; n = 2, \ldots, L$$

and overall strength of epistasis is $F_{\text{sum}} = \sum_{n \geq 2} F_n$

Empirical example: The *Aspergillus niger* fitness landscape

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

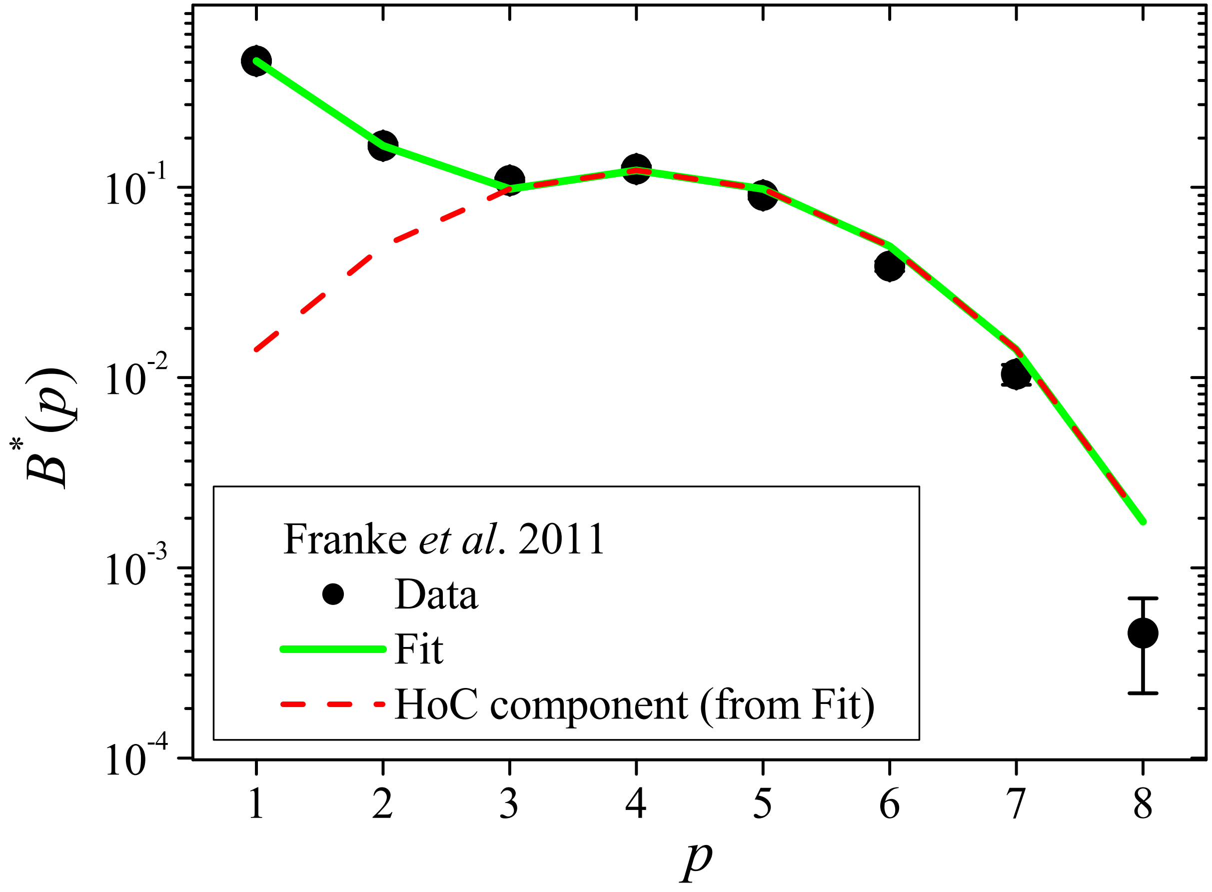


- Combinations of 8 individually deleterious marker mutations (one out of $\binom{8}{5} = 56$ five-dimensional subsets shown)

- 3 local fitness optima, 25 out of 120 paths are accessible

Fourier spectrum of the *A. niger* landscape

J. Neidhart, I.G. Szendro, JK, JTB **332**, 218 (2013)



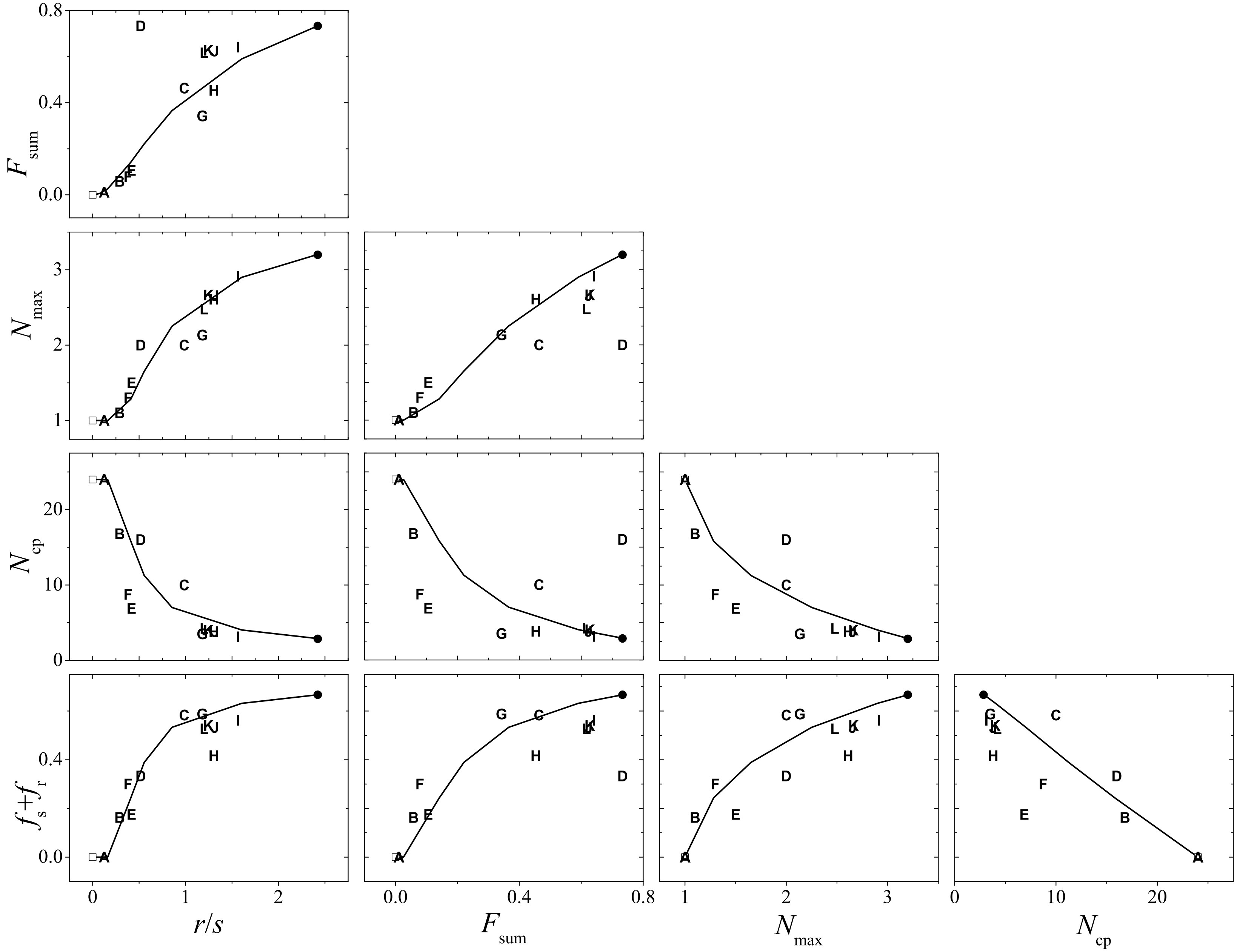
- Pairwise interactions ($p = 2$) and a random (HoC) component

A metaanalysis of empirical data sets

I.G. Szendro et al., JSTAT P01005 (2013)

ID	System (*organism*/gene)	L	Available combinations	Fitness (proxy)	Direction of mutations	Known effects
A	*Methylobacterium extorquens*	4	16/16	Growth rate	Beneficial	Combined
B	*Escherichia coli*	5	32/32	Fitness	Beneficial	Combined
C-D	Dihydrofolate reductase	4	16/16	Resistance/ Growth rate	Beneficial	Individual/ Combined
E	β-lactamase	5	32/32	Resistance	Beneficial	Combined
F	β-lactamase	5	32/32	Resistance	Beneficial	Combined
G	*Saccharomyces cerevisiae*	6	64/64	Growth rate	Deleterious	Individual
H	*Aspergillus niger*	8	186/256	Growth rate	Deleterious	Individual
I-J	Terpene synthase	9	418/512	Enzymatic specificity	–	–

Comparison of epistasis measures



Random field models of fitness landscapes

Null model: House-of-cards

- In the house-of-cards model fitness is assigned randomly to genotypes

  Kingman 1978, Kauffman & Levin 1987

- What is the expected number of shortest, selectively accessible paths n_{acc} from an arbitrary genotype at distance d to the global optimum?

- The total number of paths is $d!$, and a given path consists of d independent, identically distributed fitness values $f_0, \ldots., f_{d-1}$.

- A path is accessible iff $f_0 < f_1 \ldots. < f_{d-1}$

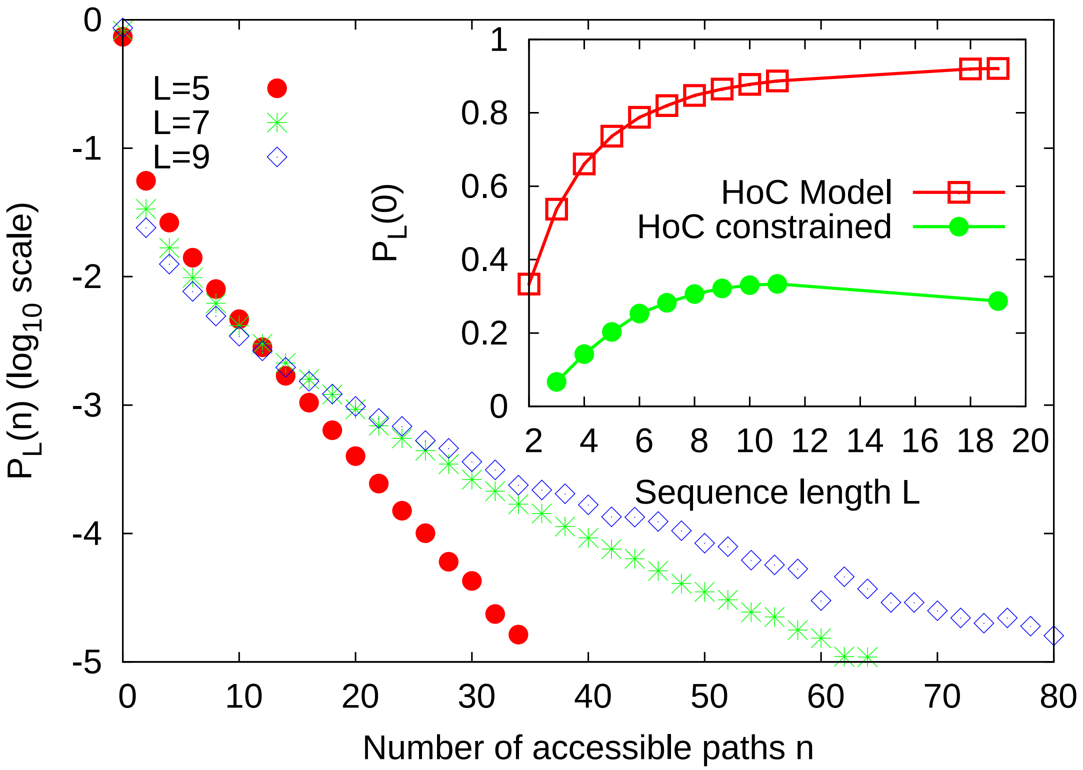
- Since all $d!$ permutations of the d random variables are equally likely, the probability for this event is $1/d!$

$$\Rightarrow \mathbb{E}(n_{\mathrm{acc}}) = \frac{1}{d!} \times d! = 1$$

- This holds in particular for the $L!$ paths from the reversal genotype of the global optimum.

Distribution of number of accessible paths from reversal genotype

J. Franke et al., PLoS Comp. Biol. **7** (2011) e1002134



- "Condensation of probability" at $n_{\mathrm{acc}} = 0$
- Characterize the distribution $P_L(n)$ by $\mathbb{E}(n_{\mathrm{acc}})$ and the probability $P_L(0)$ that no path is accessible $\Rightarrow$ define accessibility as $\overline{P}_L \equiv 1 - P_L(0)$

“Accessibility percolation” as a function of initial fitness

- When fitnesses are drawn from the uniform distribution and the fitness of the initial genotype is f_0, then
  Hegarty & Martinsson, arXiv:1210.4798

$$\lim_{L \to \infty} \overline{P}_L = \begin{cases} 0 \quad \text{for} \quad f_0 > \dfrac{\ln L}{L} \\ 1 \quad \text{for} \quad f_0 < \dfrac{\ln L}{L}, \end{cases}$$

- This implies in particular that $\lim_{L \to \infty} \overline{P}_L = 0$ for the HoC model with unconstrained initial fitness

- If arbitrary paths with backsteps are allowed, the accessibility threshold becomes independent of L and is conjectured to be $1 - \frac{1}{2}\sinh^{-1}(2) \approx 0.27818\ldots$
  Berestycki, Brunet, Shi, arXiv:1401.6894

- On a regular tree of height h and branching number b the accessibility threshold for $h, b \to \infty$ occurs at $h/b = e$
  Nowak & Krug, EPL 2013; Roberts & Zhao, ECP 2013

Landscapes with tunable ruggedness

Kauffman's NK-model

Kauffman & Weinberger 1989

- Each locus interacts randomly with $K \leq L-1$ other loci:

$$f(\sigma) = \sum_{i=1}^{L} f_i(\sigma_i | \sigma_{i_1}, \ldots, \sigma_{i_K})$$

f_i: Uncorrelated RV's assigned to each of the 2^{K+1} possible arguments

- $K = 0$: Non-epistatic $K = L-1$: House-of-cards

Rough Mt. Fuji model

Aita et al. 2000; Neidhart et al., arXiv:1402.3065

- Non-epistatic ("Mt. Fuji") landscape perturbed by a random component:

$$f(\sigma) = -\theta d(\sigma, \sigma^{(0)}) + \eta(\sigma)$$

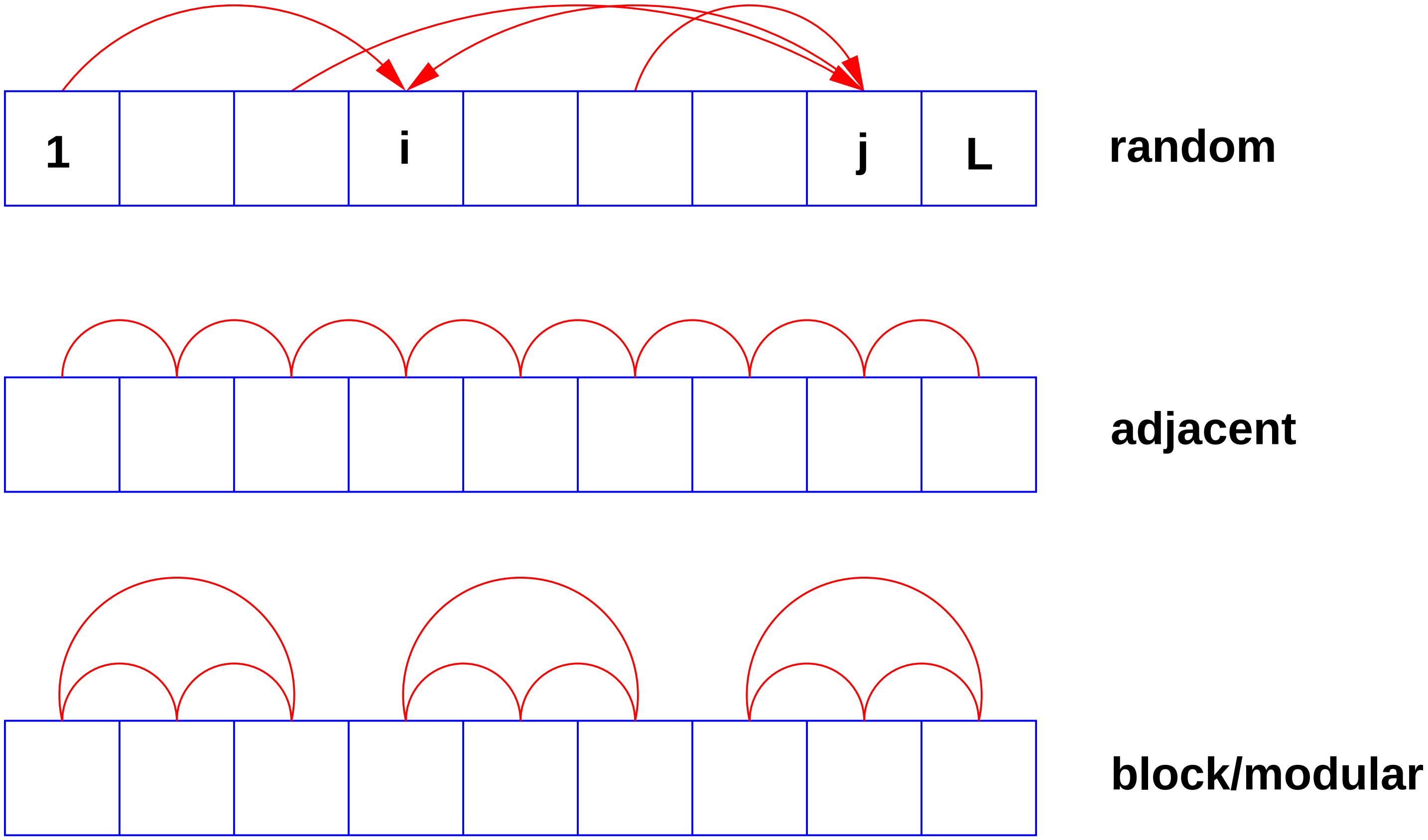
η: (Gaussian) RV's with unit variance $d(\sigma, \sigma')$: Hamming distance

- $\lim_{L\to\infty} \overline{P}_L = 1$ for any $\theta > 0$

Hegarty & Martinsson 2012

“Genetic architecture” in Kauffman’s NK-model

- Different schemes for choosing the interaction partners:



- Which properties of the fitness landscape are sensitive to this choice?

“Genetic architecture” in Kauffman’s NK-model

- Fitness correlation function is manifestly independent of the neighborhood scheme P.R.A. Campos, C. Adami, C.O. Wilke (2002)
- This implies independence also for the Fourier spectrum of the landscape, which can be computed exactly

  J. Neidhart, I.G. Szendro, JK, JTB 2013
- In the block model, the mean number of local maxima is given exactly by

  $$\mathbb{E}(n_{\max}^{\text{block}}) = \frac{2^L}{(K+2)^{L/(K+1)}}$$

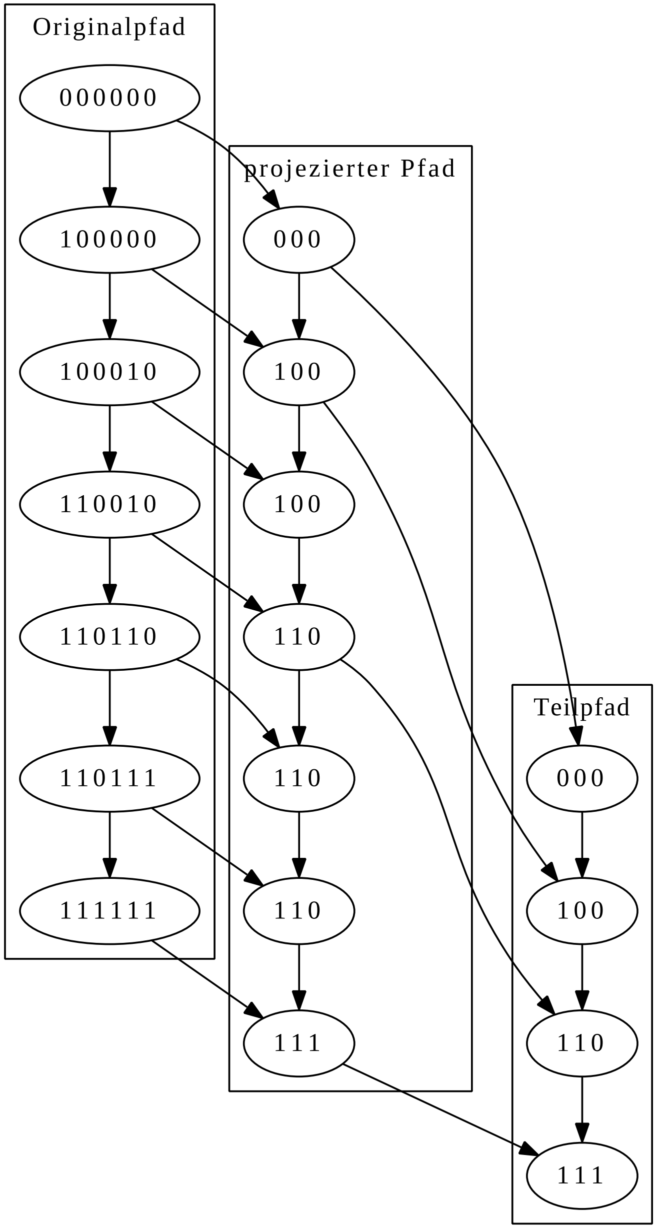
  A.S. Perelson, C.A. Macken (1995)

  which is very close (but not identical) to rigorous results for the adjacent model Durrett & Limic (2003), Limic & Pemantle (2004)
- Mean number of accessible paths in the block model:

  $$\mathbb{E}(n_{\text{acc}}^{\text{block}}) = \frac{L!}{[(K+1)!]^{L/(K+1)}}$$

  B. Schmiegelt, JK 2013

Path decomposition for the block model



Evolutionary accessibility in the block model

B. Schmiegelt, JK, J. Stat. Phys. **154**, 334 (2014)

- A given pathway spanning the whole landscape is accessible iff all subpaths within the $B = L/(K+1)$ blocks are accessible

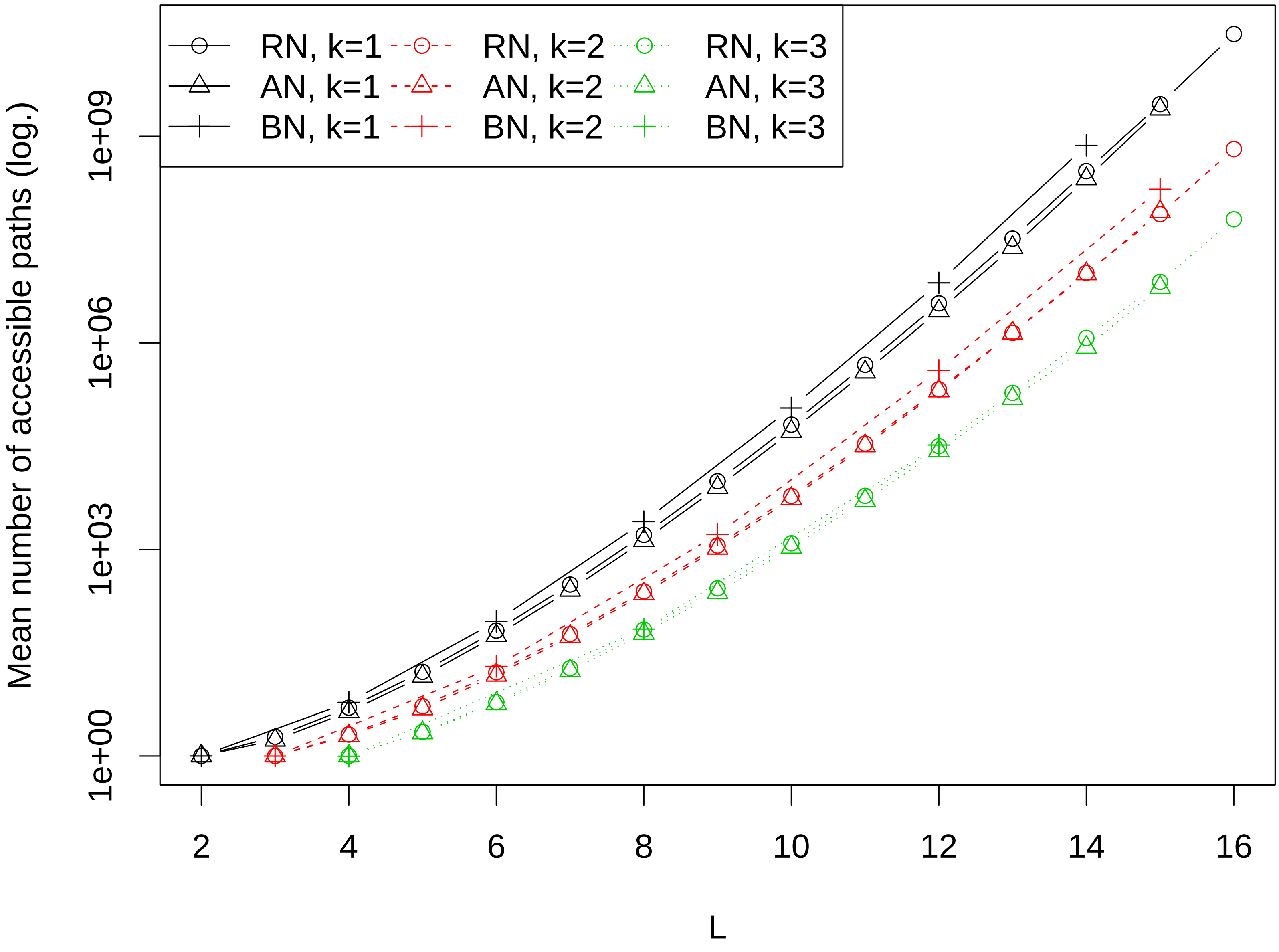
- Each combination of accessible subpaths can be combined into $\frac{L!}{[(K+1)!]^B}$ global paths

$$\Rightarrow \quad n_{\mathrm{acc}}^{\mathrm{block}} = \frac{L!}{[(K+1)!]^B} \prod_{i=1}^{B} n_{\mathrm{acc}}^{(i)}$$

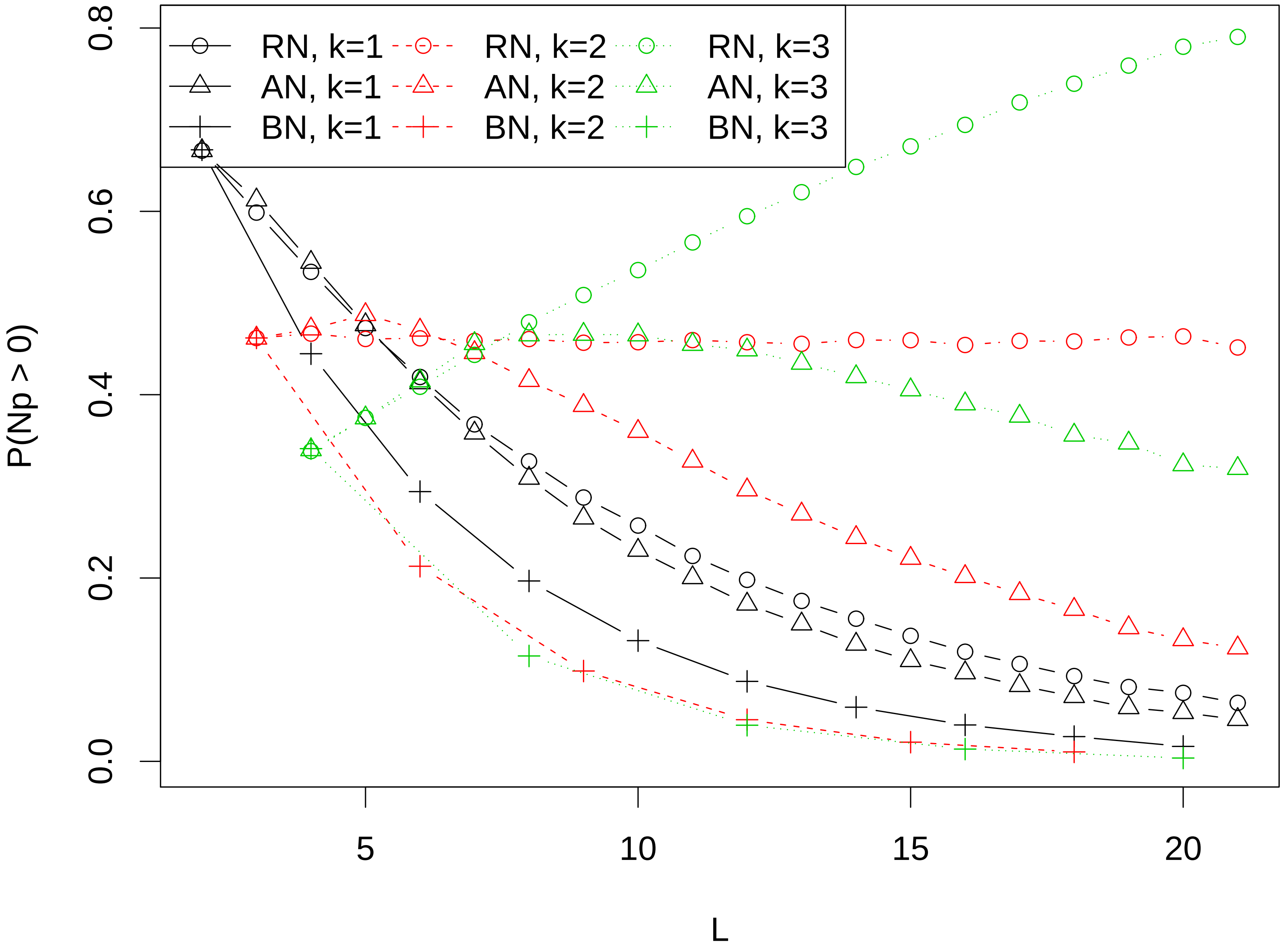
- Since the blocks are HoC-landscapes of size $K+1$, the expected number of accessible paths is $\mathbb{E}(n_{\mathrm{acc}}^{\mathrm{block}}) = \frac{L!}{[(K+1)!]^B}$ and the accessibility is $\overline{P}_L^{\mathrm{block}} = [\overline{P}_{K+1}^{\mathrm{HoC}}]^{\frac{L}{K+1}}$ which approaches zero exponentially fast in L for any K

- Full distribution of $n_{\mathrm{acc}}^{\mathrm{block}}$ can be computed in terms of the HoC distributions, explicit results for $K=1$ and $K=2$.

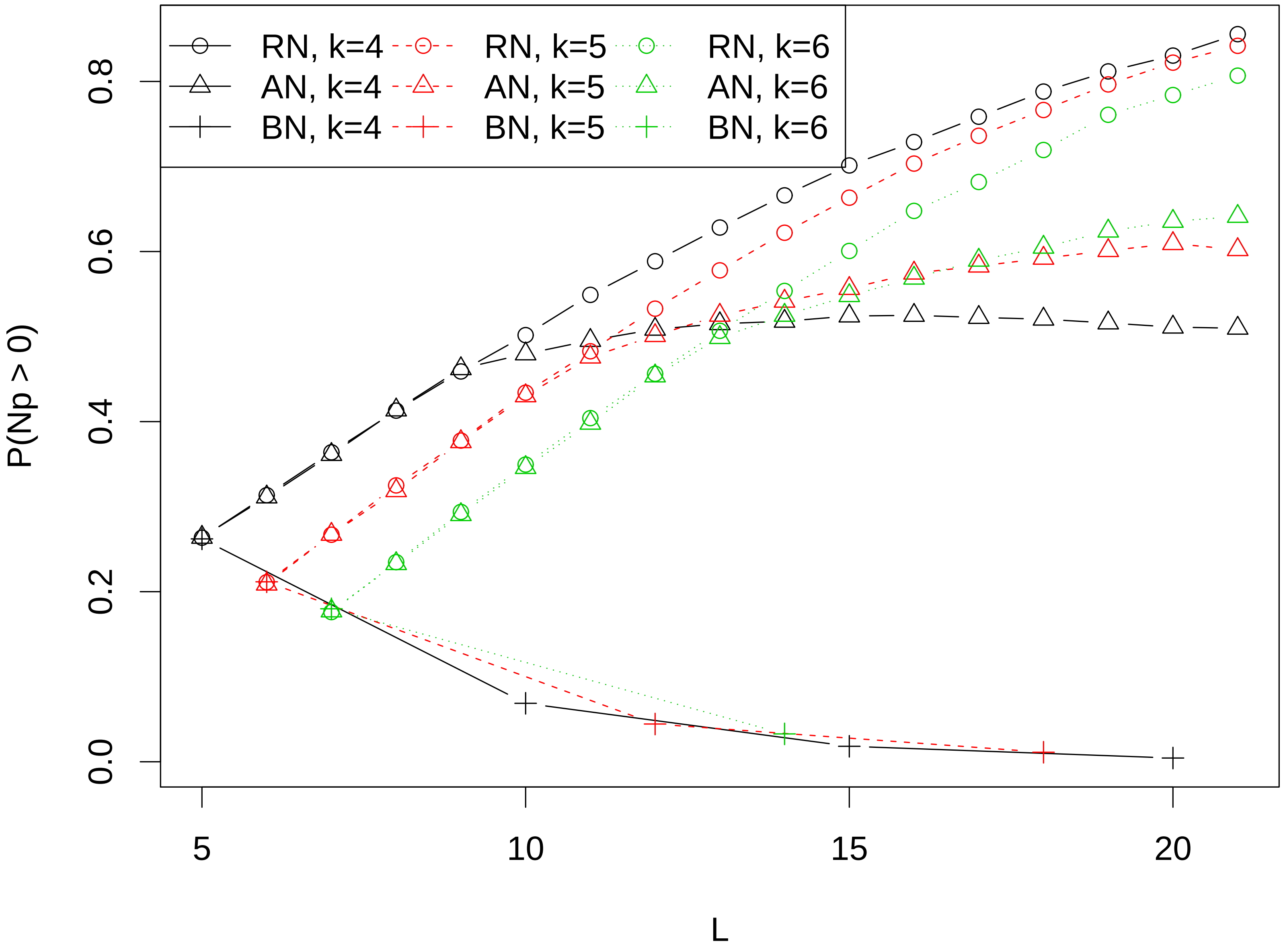
Mean number of paths is insensitive to genetic architecture



...but accessibility is very sensitive....



...at least for system sizes that can be simulated



Adaptive walks

Adaptive walks

- An adaptive walk is a Markov chain on sequence space that is constrained to move to genotypes of larger fitness and terminates at local fitness maxima

- Three flavors of adaptive walks differing in their transition probabilities:

  Random Adaptive Walk (RAW) Macken & Perelson 1989
  All fitter genotypes are chosen with equal probability

  Greedy Adaptive Walks (GAW) Orr 2003
  The most fit genotype is chosen deterministically

  True Adaptive Walk (TAW)
  Transition rate is proportional to the fitness difference between the resident and mutant genotype Gillespie 1983, Orr 2002

- Quantities of interest: Average length ℓ and achieved fitness (height) f^*

Walk length in the HoC landscape

- RAW's and GAW's are fully determined by the rank ordering of the fitness landscape. Their properties are independent of the fitness distribution and only depend on the number of uphill directions L in the initial state.

- RAW: $\ell \approx \ln(L) + 1.1$ for large L Flyvbjerg & Lautrup 1992

- GAW: $\ell \to e - 1 \approx 1.71828\ldots$ Orr 2003

- TAW length asymptotics depends on the extreme value index κ of the fitness distribution according to Neidhart & Krug 2011, Jain 2011

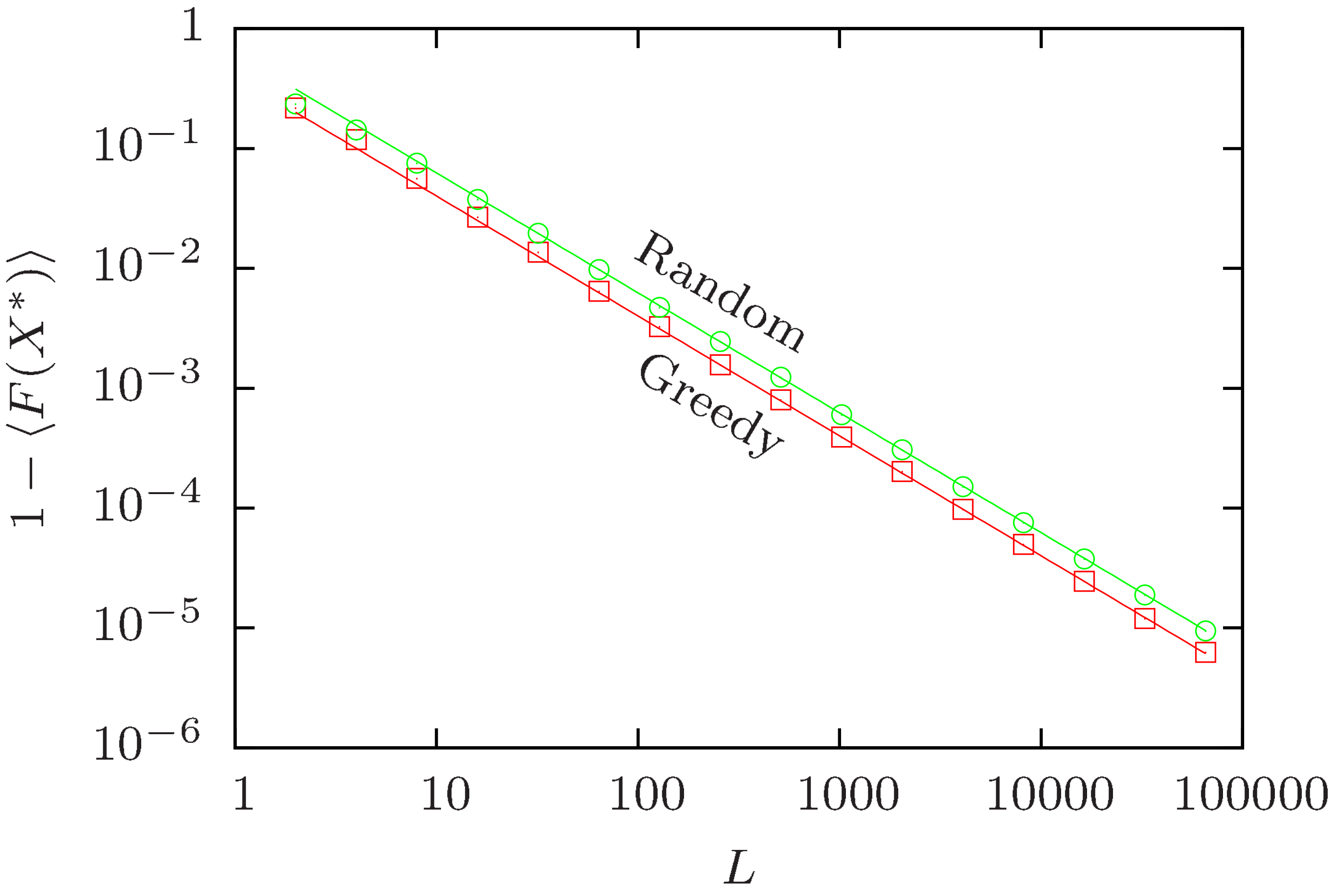
$$\ell \approx \frac{1-\kappa}{2-\kappa} \ln(L) + c_\kappa \quad \text{for} \quad \kappa < 1$$

  where $\kappa > 0$, $\kappa = 0$ and $\kappa < 0$ correspond to the Fréchet, Gumbel and Weibull classes, respectively.

- The TAW becomes effectively random (greedy) for $\kappa \to -\infty$ ($\kappa \to 1$)

Walk height in the HoC landscape

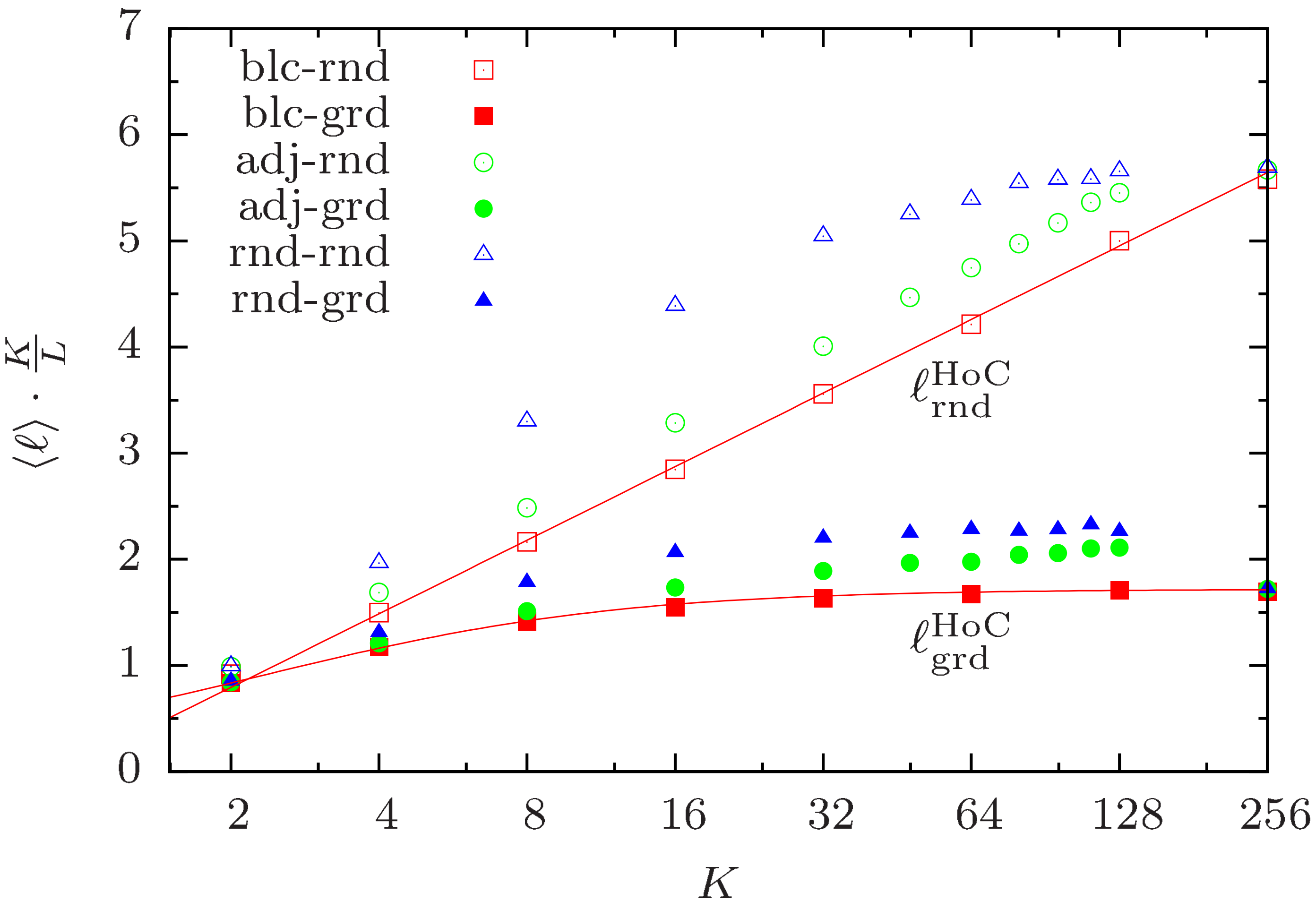
S. Nowak (unpublished)



- For uniform fitness distribution the expected final fitness is of the form $1-\mathbb{E}(f^*)\approx\frac{\beta}{L}$ with $\beta_{\mathrm{RAW}}\approx 0.6243..$ and $\beta_{\mathrm{GAW}}\approx 0.4003...$

Walk length in NK landscapes ($L = 256$)

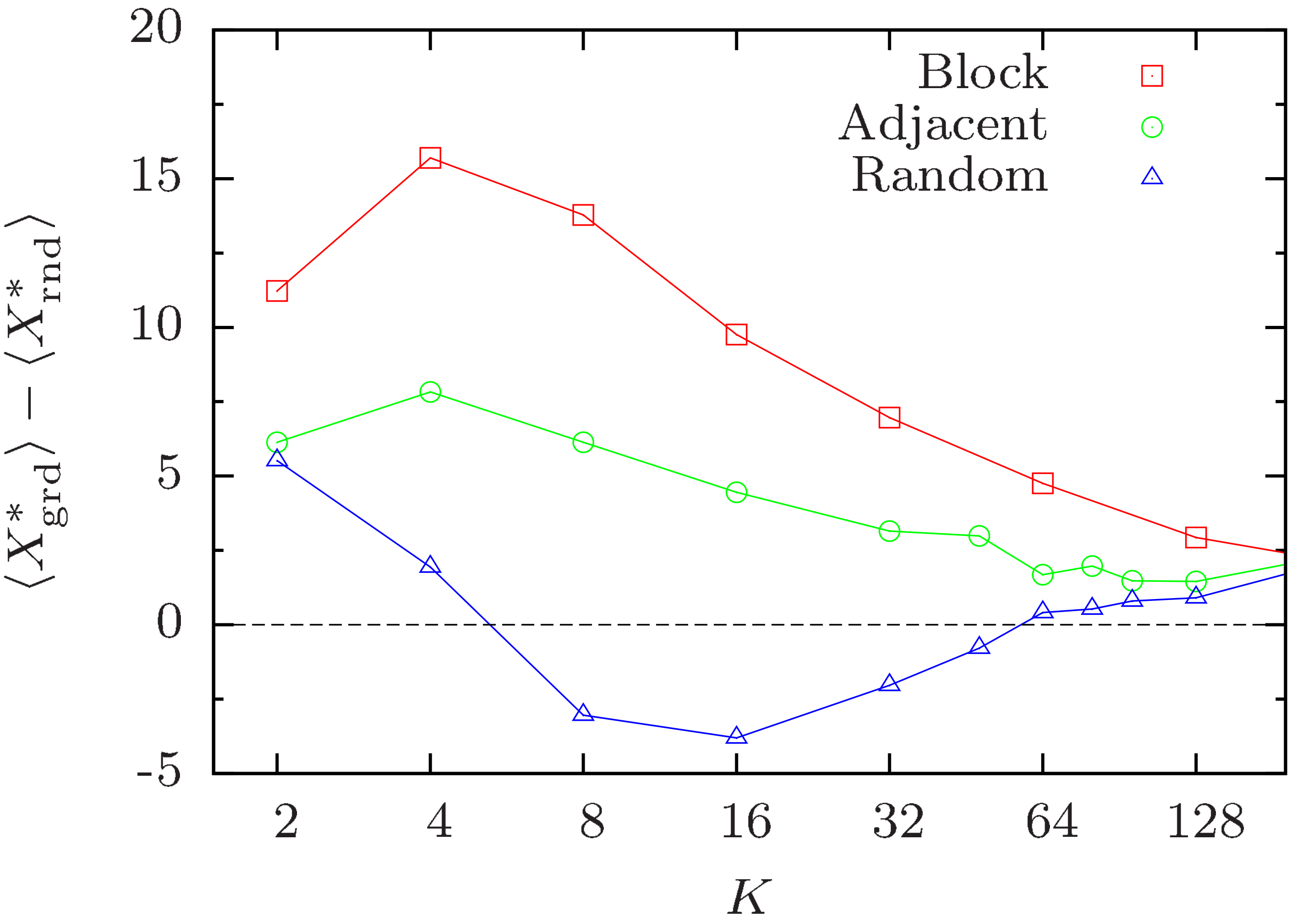
S. Nowak (unpublished)



- Walk length in block model is additive over blocks: $\ell = \frac{L}{K+1} \ell_{\mathrm{HoC}}(K+1)$

Walk height in NK landscapes ($L = 256$)

S. Nowak (unpublished)



- Fitness difference between GAW and RAW for normal fitness distribution

Summary

- Increasing number of empirical fitness landscapes provide insights into patterns of epistasis

- Random landscape models are useful to explore the effect of genotypic dimensionality, but conclusions are not clear-cut so far:
  - number of accessible pathways generally increases combinatorially, but
  - probability for existence of pathways may vanish for large L

- Static view focused on landscape structure is complemented by dynamic view of accessibility in term of adaptive walks and more complex evolutionary dynamics

Summary

- Increasing number of empirical fitness landscapes provide insights into patterns of epistasis

- Random landscape models are useful to explore the effect of genotypic dimensionality, but conclusions are not clear-cut so far:
  - number of accessible pathways generally increases combinatorially, but
  - probability for existence of pathways may vanish for large L

- Static view focused on landscape structure is complemented by dynamic view of accessibility in term of adaptive walks and more complex evolutionary dynamics

Thanks to:
Jasper Franke, Johannes Neidhart, Stefan Nowak, Benjamin Schmiegelt,
Ivan Szendro (Cologne)
Arjan de Visser (Wageningen)