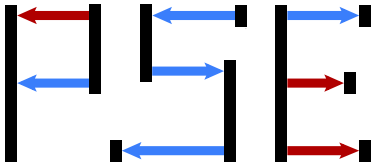


PROBABILISTIC STRUCTURES
IN EVOLUTION

DFG SPP 1590

COLLABORATIVE RESEARCH CENTER 1310

Predictability in Evolution

Evolutionary accessibility in random and structured fitness landscapes

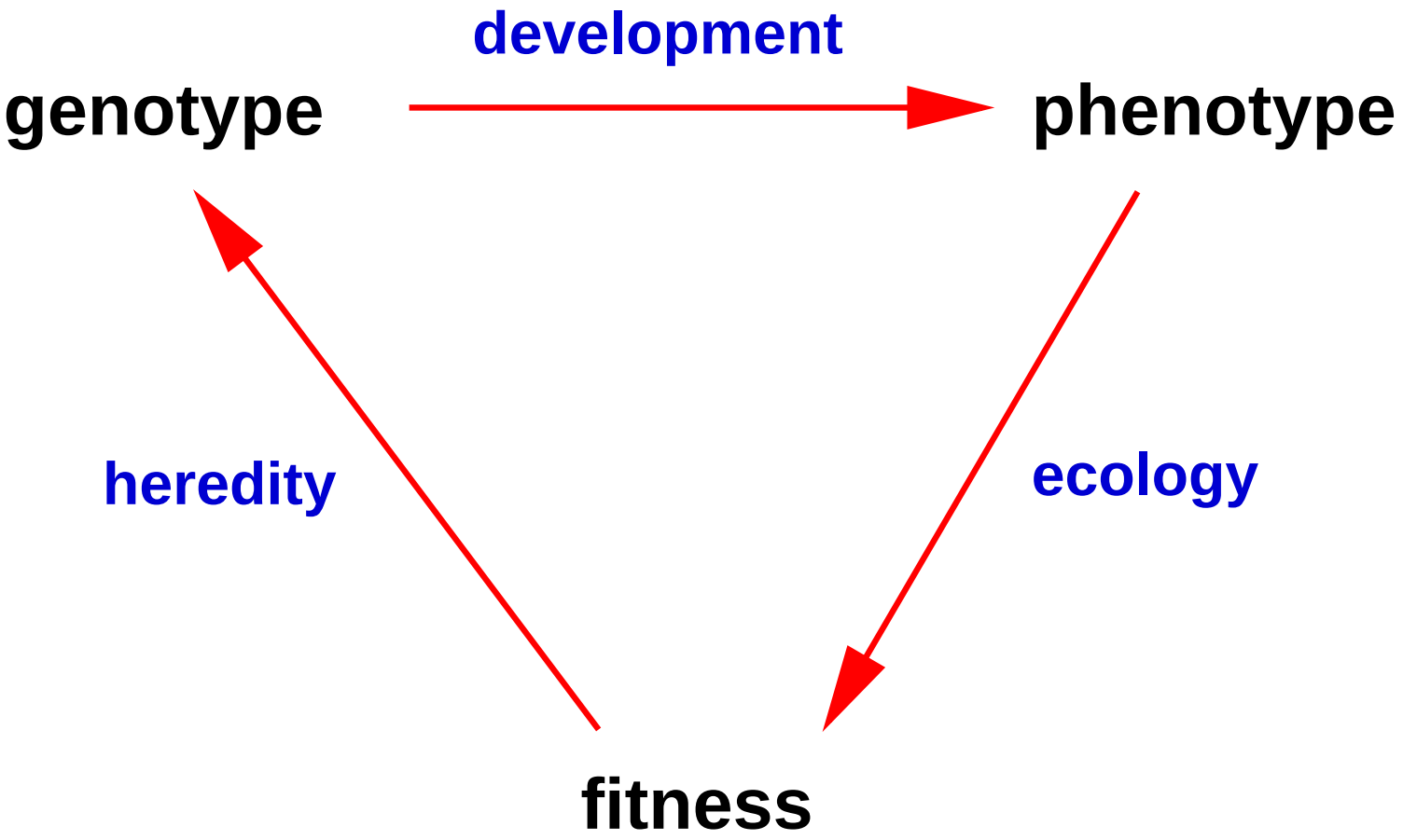
Joachim Krug
Institute for Biological Physics
University of Cologne

based on arXiv:2311.174321

Frontiers in Statistical Physics, RRI Bengaluru, December 5, 2023

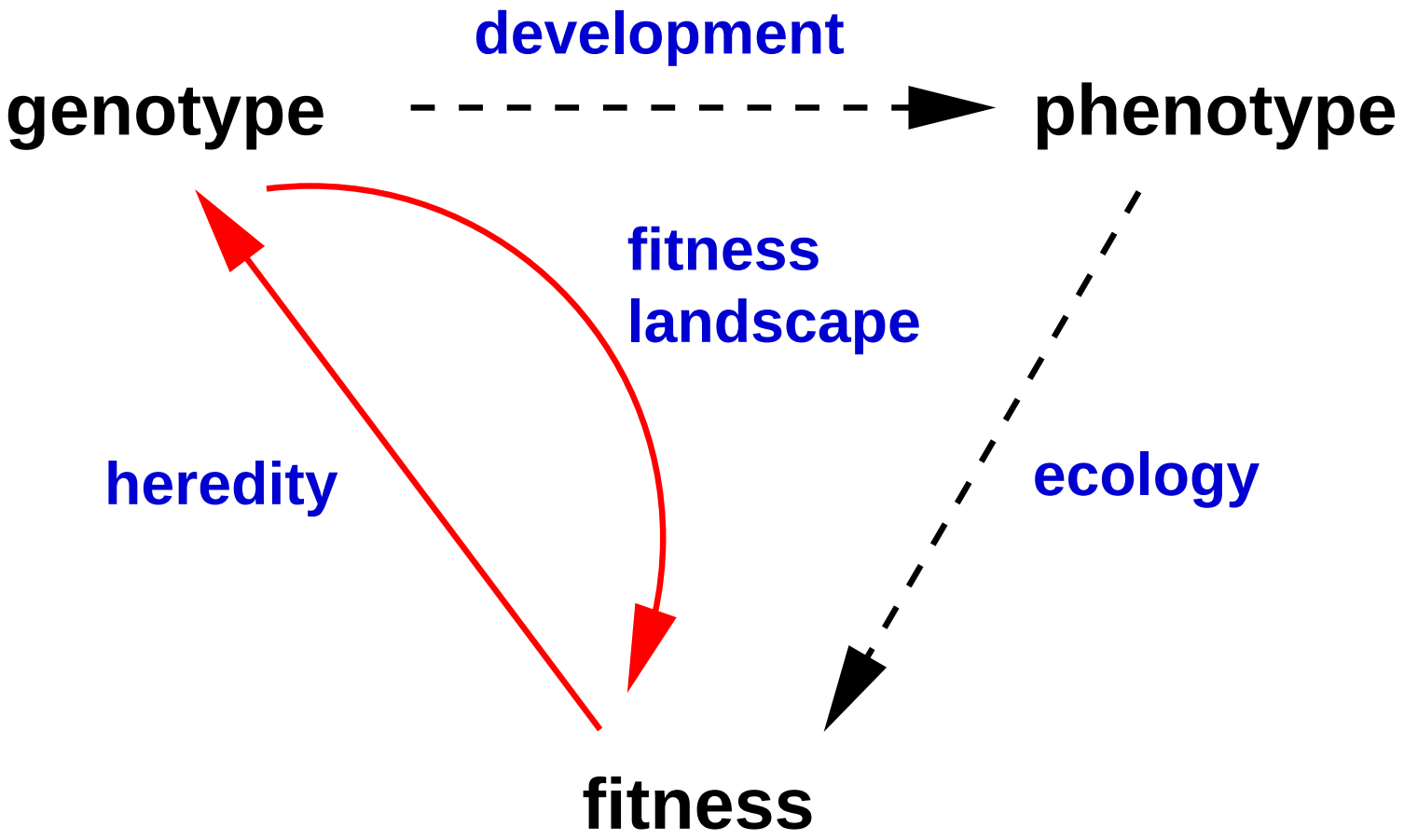
How evolution works

Courtesy Amitabh Joshi, JNCASR



How evolution works

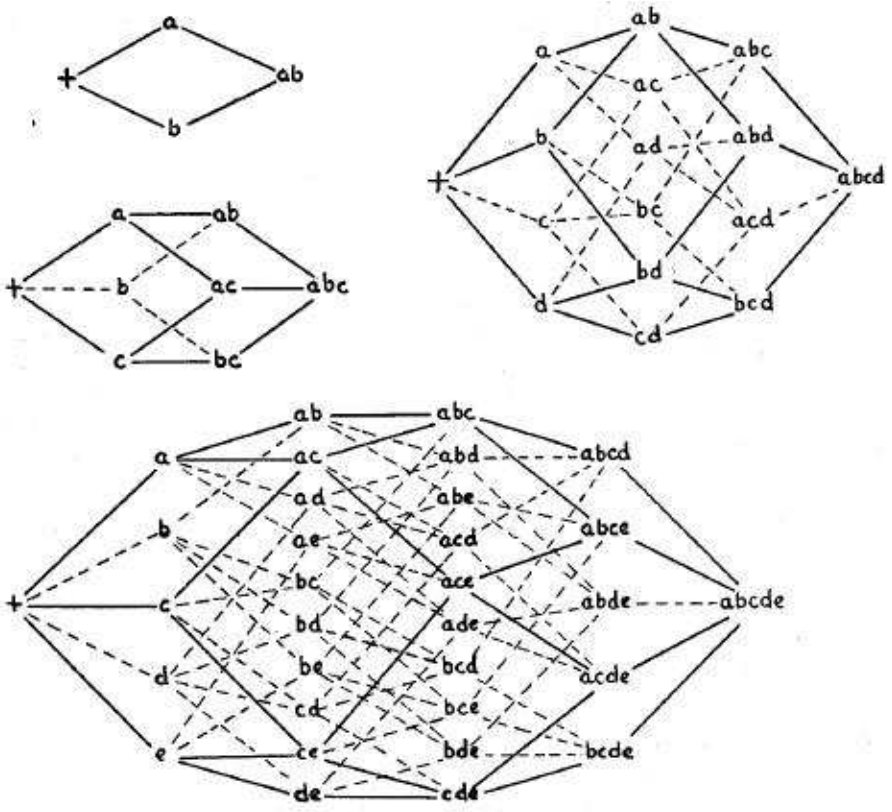
Courtesy Amitabh Joshi, JNCASR

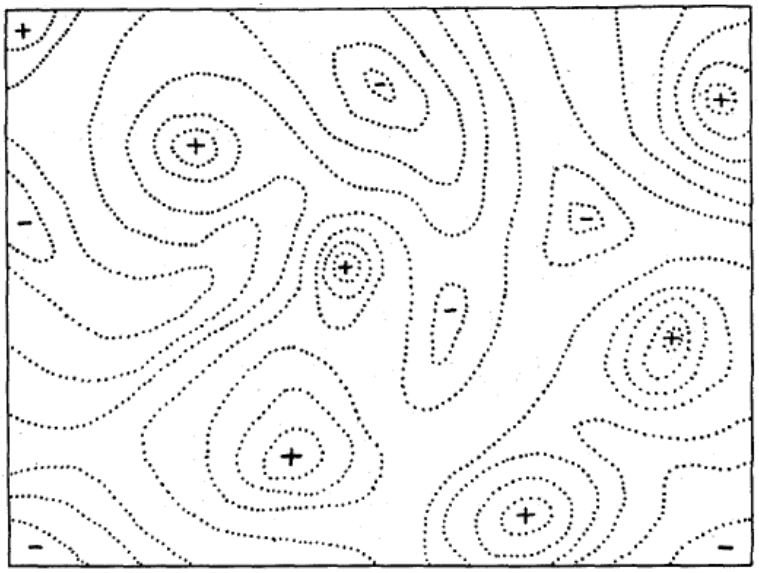


- Fitness landscape concept introduced by Sewall Wright (1932)

Fitness landscapes

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





- Sequence space
- Peaks and valleys

Mathematical setting

- Genotypes are sequences of length L

$$\sigma = (\sigma_1, \ldots, \sigma_L) \in \{0, \ldots, a-1\}^L, \quad a \geq 2 \quad \text{number of } \textit{alleles}$$

- The Hamming distance $d_H(\sigma, \tau)$ is the number of sites at which the two sequences differ

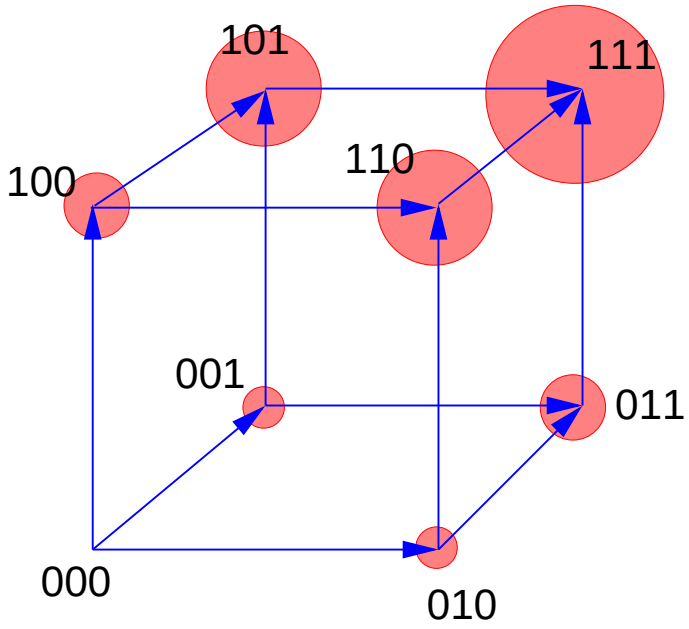
- A fitness landscape is a real-valued function

$$g : \{0, \ldots, a-1\}^L \to \mathbb{R}$$

- A path $\sigma^{(0)} \to \sigma^{(1)} \to \ldots \sigma^{(\ell)}$ with $d_H(\sigma^{(i+1)}, \sigma^{(i)}) = 1$ is called (evolutionarily) accessible if $g_{\sigma^{(i)}} > g_{\sigma^{(i-1)}} \quad \forall i$

- Binary alphabet ($a = 2$): $\sigma_i = 1$ ($\sigma_i = 0$) denotes the presence (absence) of a certain mutation at position i

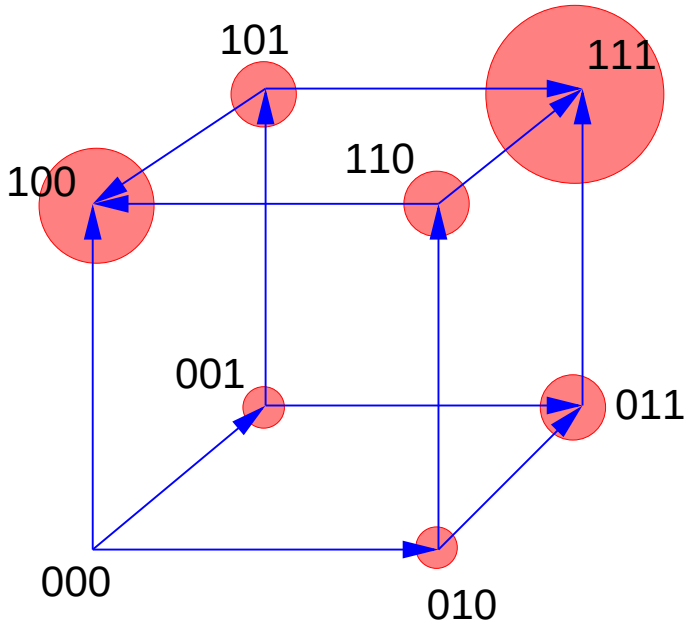
$L = 3$



- Fitness values represented by the size of the circles
- Fitness graph: Arrows point in the direction of increasing fitness

  De Visser et al. 2009, Crona et al. 2013
- Mutations $000 \rightarrow 111$ can occur in $3! = 6$ different orders corresponding to 6 possible direct pathways

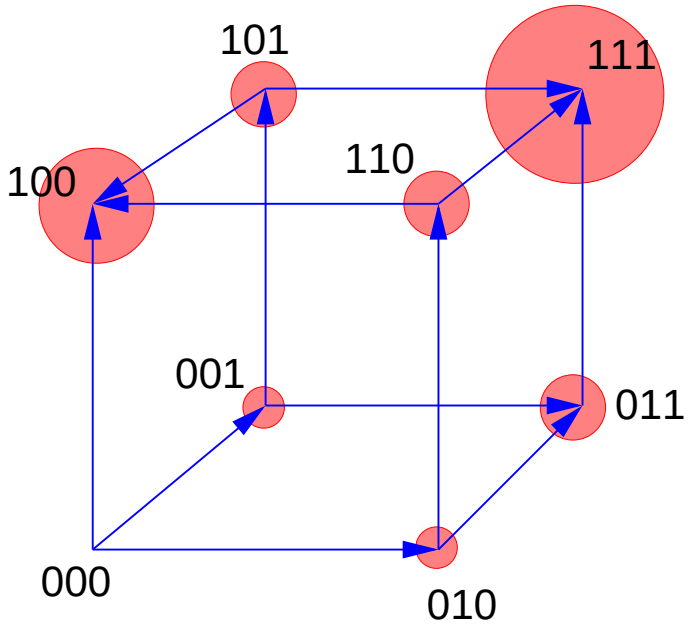
$L = 3$



- A local fitness peak at 100 has been added and 2 out of 6 direct paths to 111 become inaccessible

- In addition, there are one direct and two indirect paths $000 \rightarrow 100$

$L = 3$



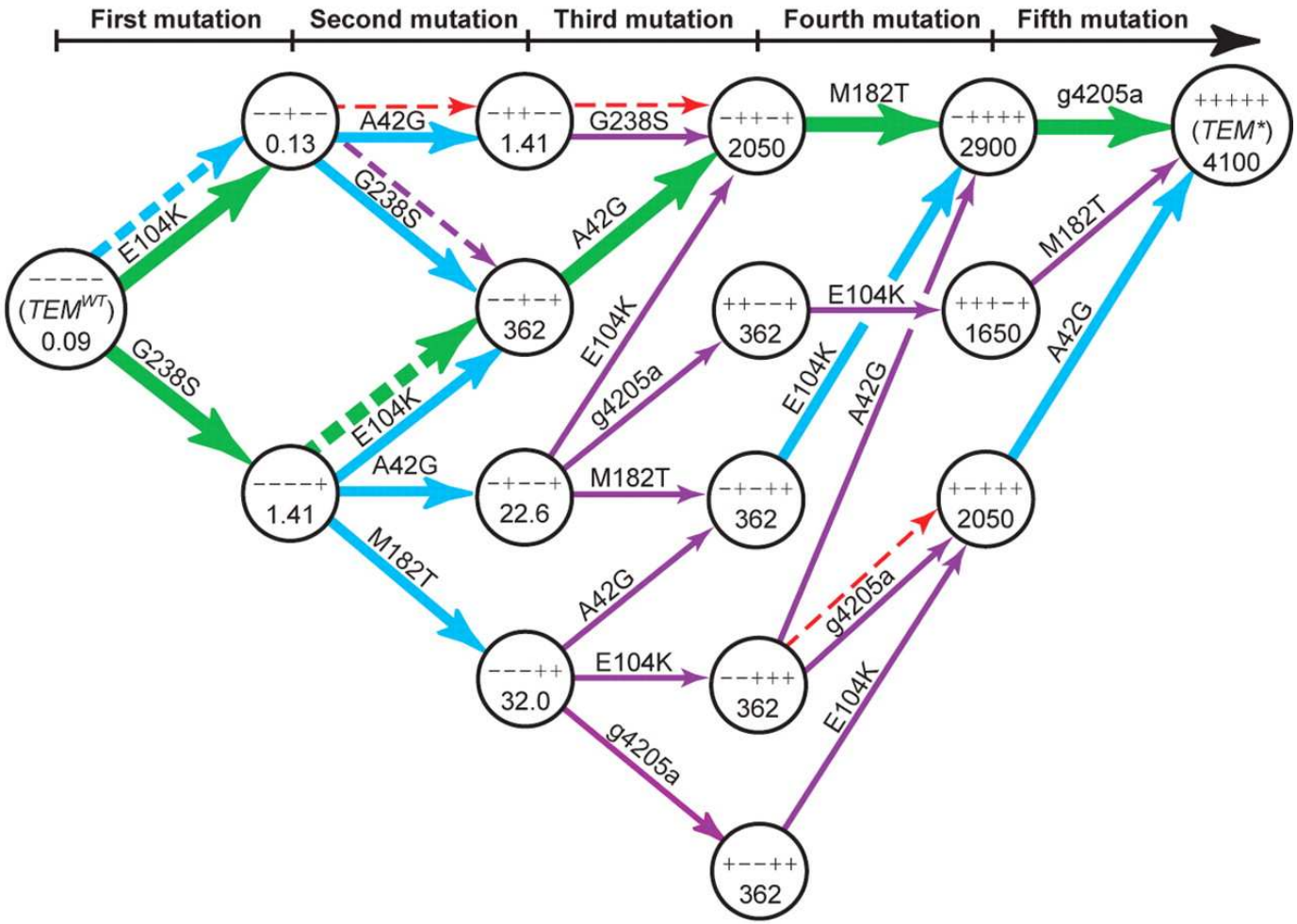
Questions for this talk

- How many accessible paths should we expect if the fitness values were random? J. Franke, A. Klözer, J.A.G.M. de Visser, JK, PLoS Comp. Biol. 2011

- How does accessibility depend on the landscape structure?

"Darwinian evolution can follow only very few mutational paths to fitter proteins"

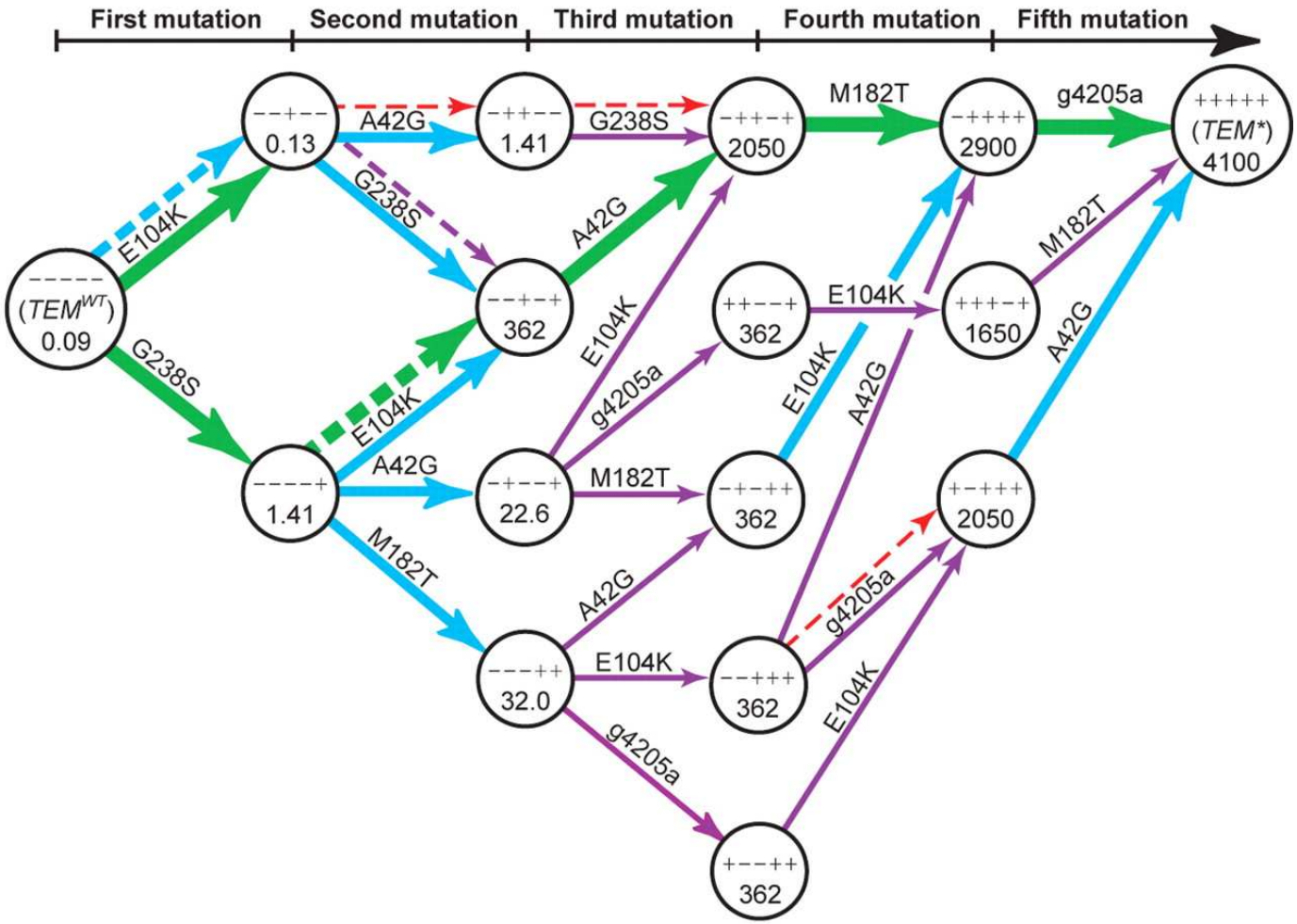
D.M. Weinreich et al., Science **312**, 111 (2006)



- 5 mutations in an enzyme increase antibiotic resistance by $\sim 4.5 \times 10^4$

"Darwinian evolution can follow only very few mutational paths to fitter proteins"

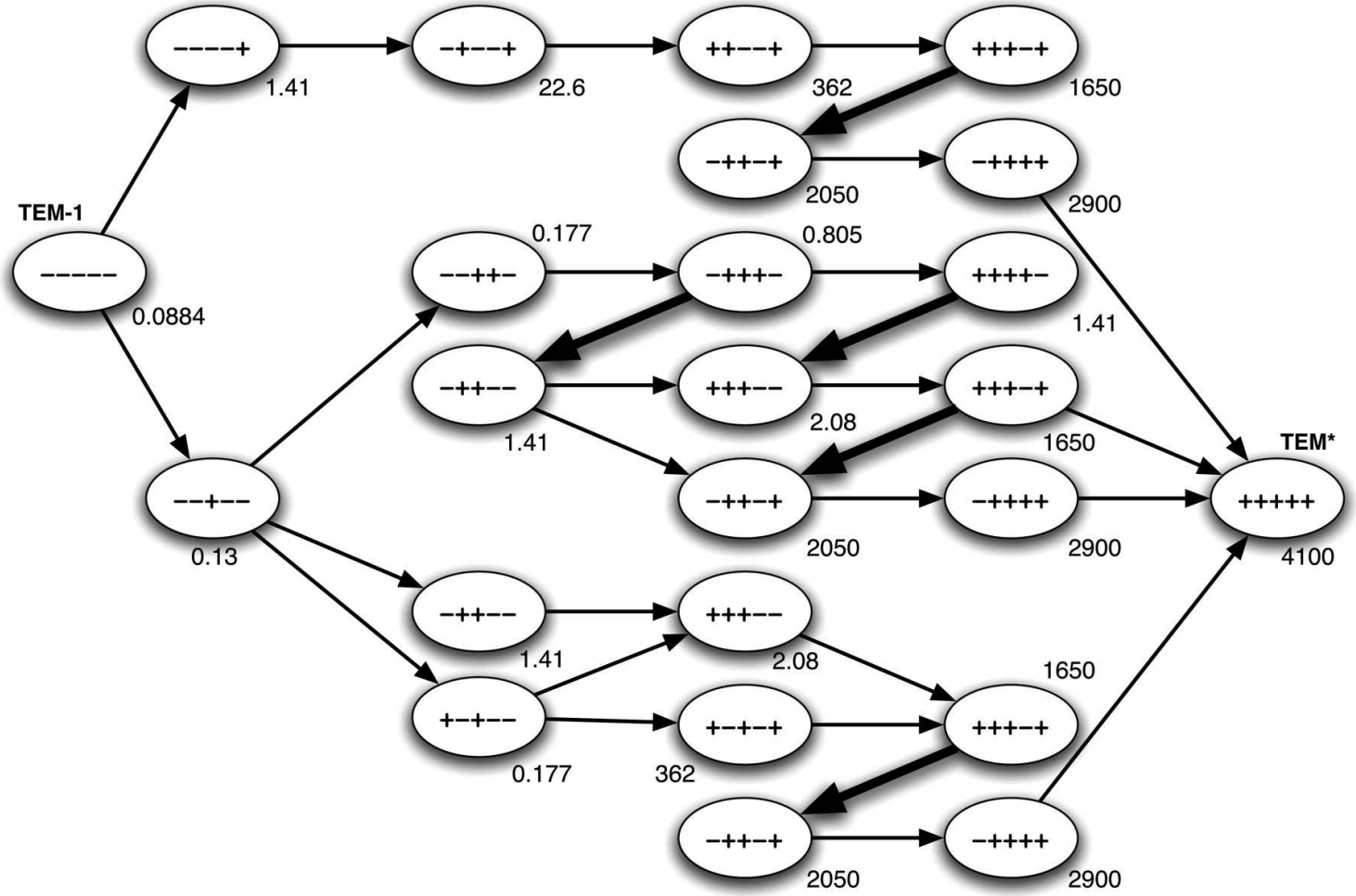
D.M. Weinreich et al., Science **312**, 111 (2006)



- 18 out of $5! = 120$ direct mutational pathways are accessible...

Including backsteps

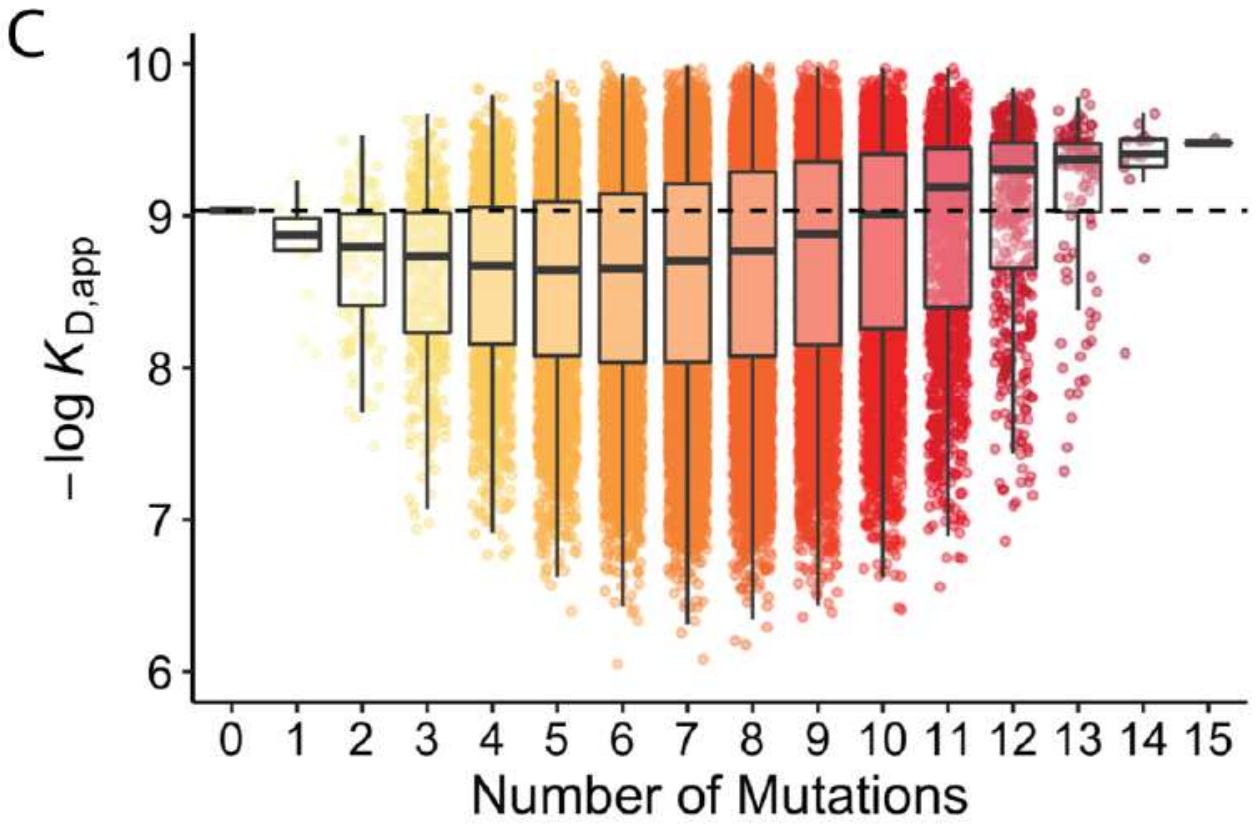
De Pristo et al. 2007



- ...and 27 out of 18651552840 indirect pathways

Affinity landscape of the SARS-CoV2 spike protein

Moulana et al., Nat. Comm. 2022



- All $2^{15} = 32768$ combinations of $L = 15$ mutations separating the ancestral Wuhan strain from Omicron BA.1
- None of the $15! \approx 1.3 \times 10^{12}$ direct paths is accessible

Evolutionary accessibility

of random fitness landscapes

Accessibility percolation

review: arXiv:1903.11913

- Take fitness values to be i.i.d. $U[0,1]$ random variables

- A path of length ℓ between genotypes σ, τ with $g_\sigma - g_\tau = \beta \in [0,1]$ is accessible if all $\ell - 1$ intermediate fitness values are in (g_τ, g_σ) and increasingly ordered, which occurs with probability

$$P_{\beta,\ell} = \frac{\beta^{\ell-1}}{(\ell-1)!}$$

- The number of accessible paths is a non-negative integer-valued random variable $X_{\sigma,\tau}$

- Is there a sharp accessibility threshold β_c in $\mathbb{P}[X_{\sigma,\tau} \geq 1]$ when $L \to \infty$ and

$$\delta \equiv \lim_{L\to\infty} \frac{d_H(\sigma,\tau)}{L} > 0 \,?$$

Direct paths on the binary hypercube

P. Hegarty, A. Martinsson, Ann. Appl. Probab. 2014

- The total number of direct paths of length ℓ is $\ell!$, thus the expected number of accessible paths is

$$\mathbb{E}(X_{\sigma,\tau}) = \ell! P_{\beta,\ell} = \ell \beta^{\ell-1}$$

  which vanishes asymptotically for large ℓ when $\beta < 1$

- By Markov's inequality it then follows that $\lim_{\ell\to\infty} \mathbb{P}[X_{x,y} \geq 1] = 0$

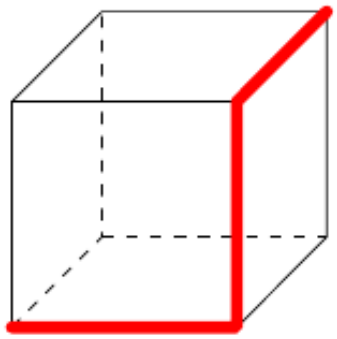
- Analysis of the second moment $\mathbb{E}(X^2_{\sigma,\tau})$ shows that, conversely, $\lim_{\ell\to\infty} \mathbb{P}[X_{\sigma,\tau} \geq 1] = 1$ for $\beta = \beta_\ell$ with $1 - \beta_\ell < \frac{\ln \ell}{\ell}$

- The directed hypercube is "marginally accessible" in the sense that percolation occurs at $\beta_c = 1^-$

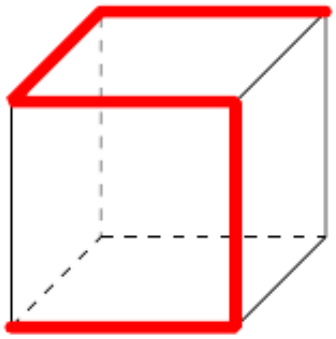
Indirect paths on the binary hypercube

Berestycki et al. 2014; Martinsson 2015; Li 2018

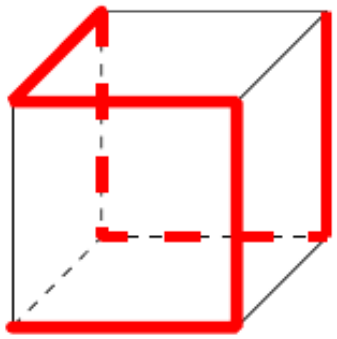
- Paths on the 3-cube with p backsteps and length $\ell = 3 + 2p$



$p = 0, \ell = 3$



$p = 2, \ell = 5$



$p = 4, \ell = 7$

- The accessibility threshold $\beta_c(\delta) < 1$ is the solution of

$$\lim_{L\to\infty} [\mathbb{E}(X_{\sigma,\tau})]^{1/L} = \sinh(\beta)^{\delta} \cosh(\beta)^{1-\delta} = 1$$

- The expectation $\mathbb{E}(X_{\sigma,\tau})$ "tells the truth"

Multiallelic fitness landscapes

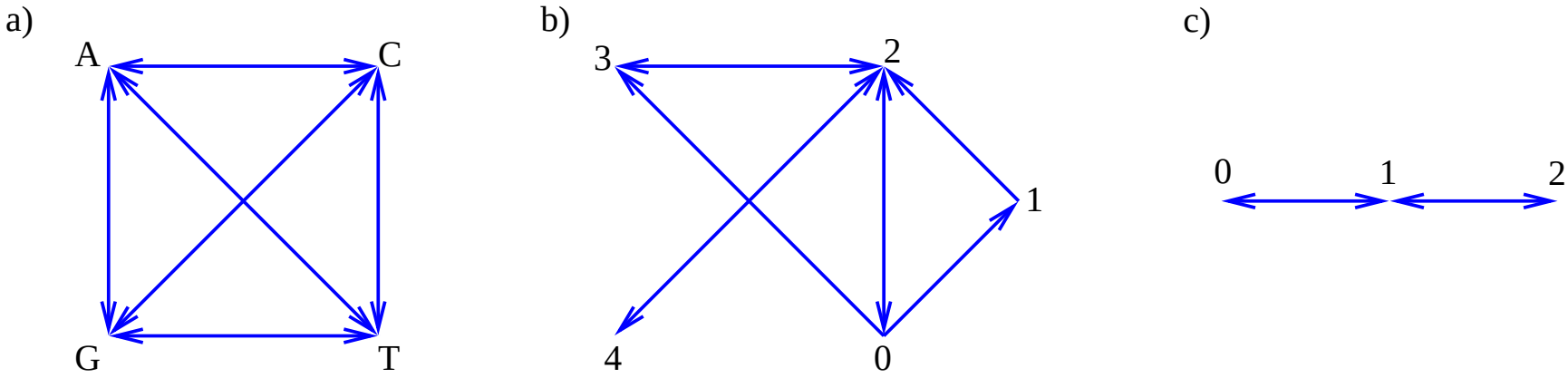
B. Schmiegelt, JK, J. Math. Biol. 2023

- Generalize the binary hypercube $\{0,1\}^L$ to Hamming graphs $\{0,\ldots,a-1\}^L$ with $a > 2$ alleles
- Biologically relevant cases are $a = 4$ (DNA, RNA) and $a = 20$ (proteins)
- Allowed mutational transitions between alleles are encoded by the $a \times a$ adjacency matrix $\mathbf{A}$ of the mutation graph
- Consider a sequence of initial and endpoints $\boldsymbol{\sigma}^{(L)}, \boldsymbol{\tau}^{(L)}$ such that the fraction of sites at which $\sigma_i^{(L)} = k$ and $\tau_i^{(L)} = l$ is given by p_{kl} for $L \to \infty$
- Theorem: The accessibility threshold β_c is given by the solution β^* of

$$\lim_{L\to\infty} [\mathbb{E}(X_{\boldsymbol{\sigma},\boldsymbol{\tau}})]^{1/L} = \prod_{k,l=0}^{a-1} [(e^{\beta \mathbf{A}})_{kl}]^{p_{kl}} = 1$$

for most (but not all) mutation graphs. In general, β^* is a lower bound on β_c, and there are no accessible paths if $\beta^* > 1$

Examples of mutation graphs



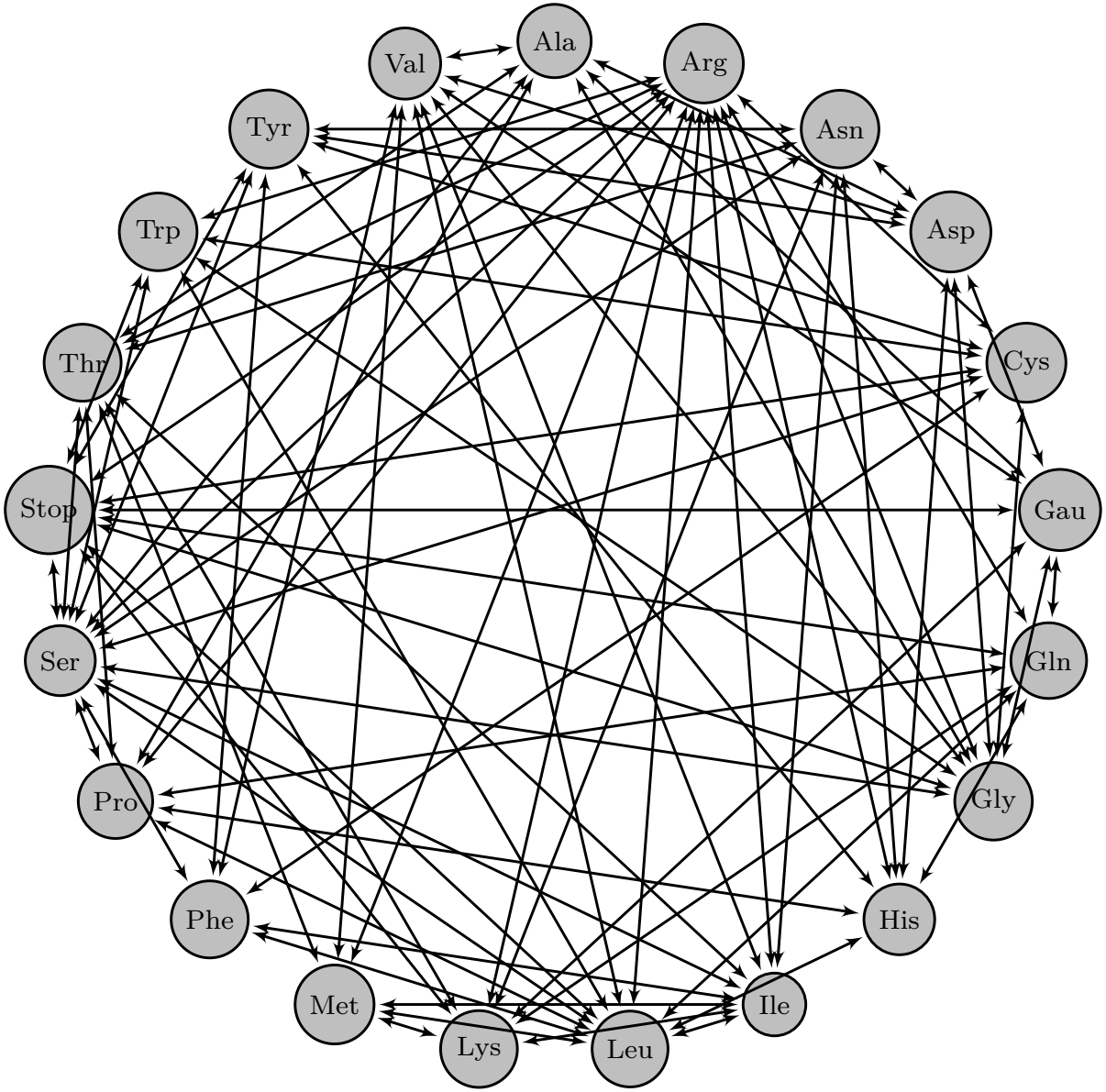
a) Nucleotide mutation graph $(a = 4)$:

$$\beta_c(\delta = 1) = \ln\left(\frac{1}{\sqrt{2}} + \sqrt{\sqrt{2} - \frac{1}{2}}\right) \approx 0.509$$

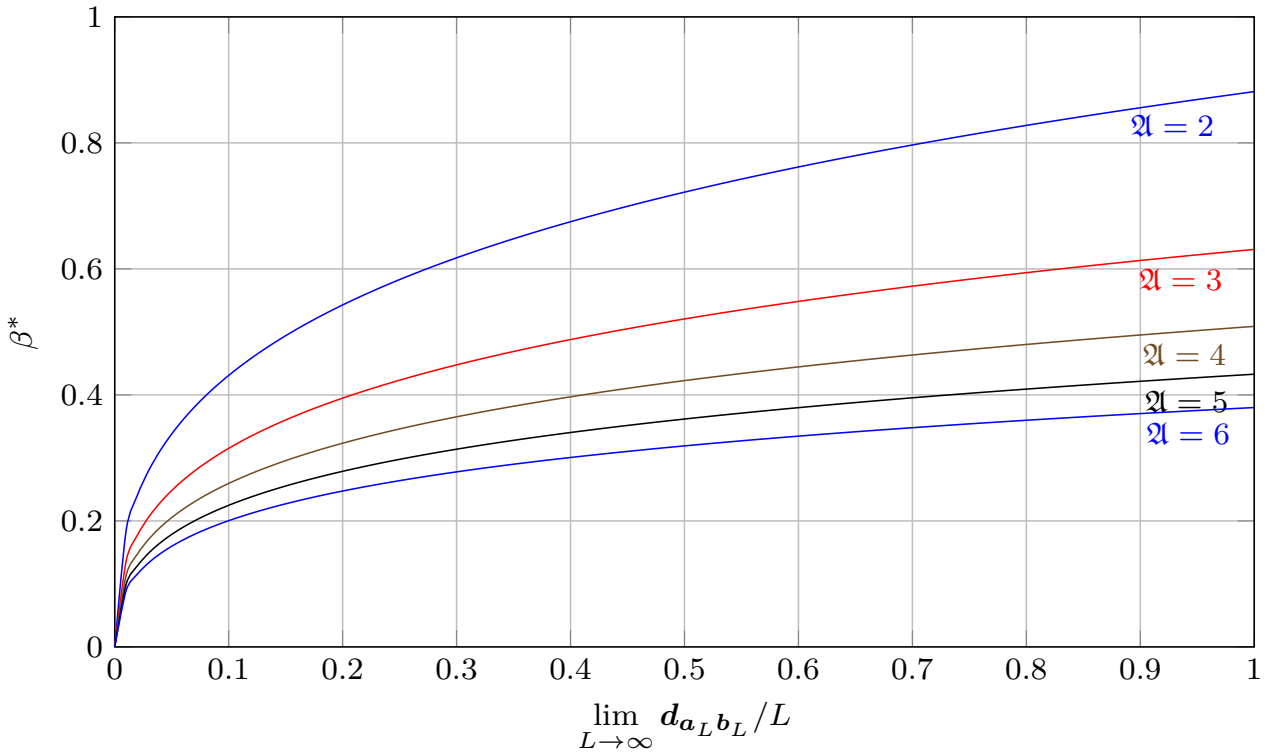
b) Smallest known mutation graph for which $\beta_c > \beta^*$ and $\beta^* < 1$

c) Path graph with $a = 3$: $\beta^*(\delta = 1) = \sqrt{2}^{-1} \ln(3 + 2\sqrt{2}) \approx 1.25 > 1$

The amino acid mutation graph ($a = 21$)



Accessibility threshold for the complete graph



- Accessibility threshold at full distance ($\delta = 1$) is

$$\beta_c(a) = \frac{\ln(a)}{a} + \frac{1+\ln(a)}{a^2} + \mathscr{O}\left(\frac{\ln(a)}{a^3}\right) \quad \text{for large } a$$

and the path length ℓ_c at the threshold is $\frac{\ell_c}{L} \approx \ln a + \frac{1+\ln a}{a}$

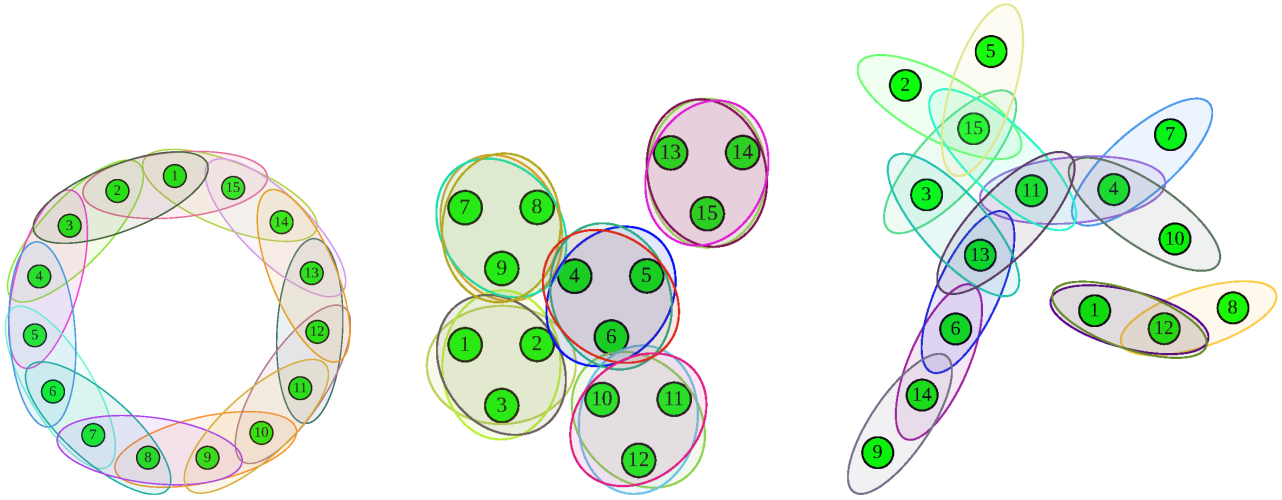
Evolutionary accessibility

of structured fitness landscapes

Kauffman's NK model

review: S. Hwang, B. Schmiegelt, L. Ferretti, JK, J. Stat. Phys. 2018

- Fitness is a sum of contributions, each of which is a random function of a subgroup of $k \leq L$ sites



adjacent NK block NK random NK

- Model interpolates between single peaked ($k = 1$) and random ($k = L$) landscapes

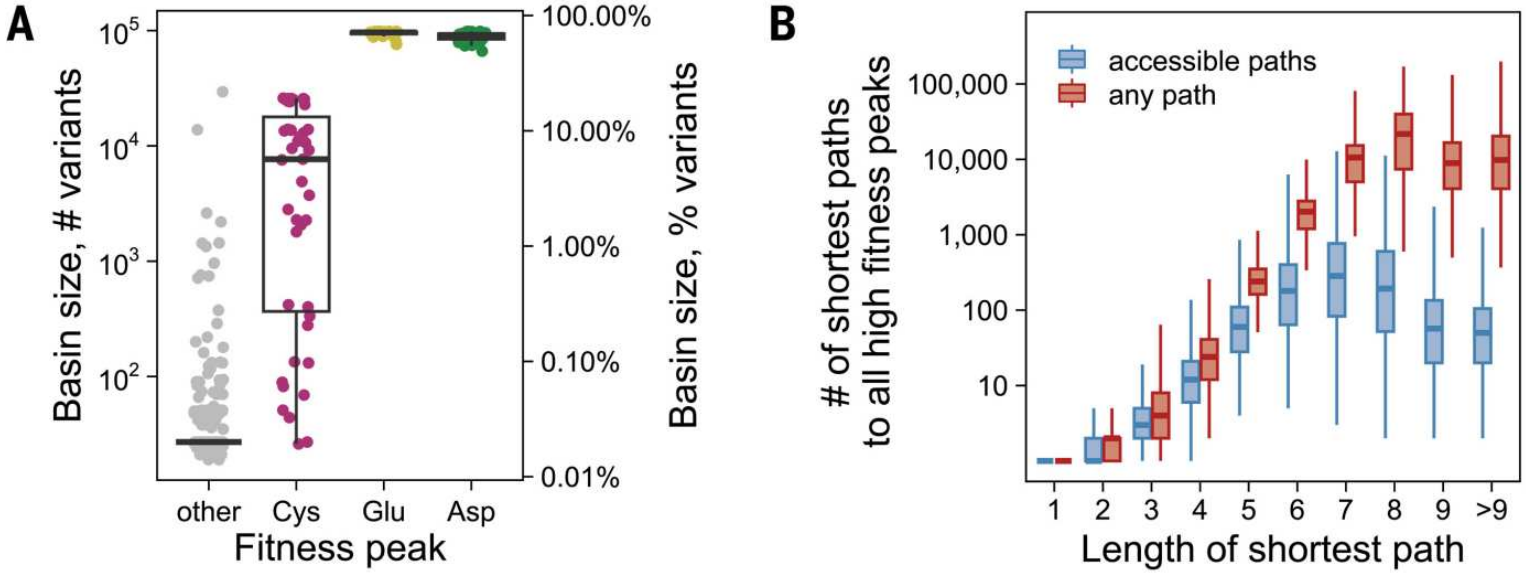
- Nevertheless the existence of accessible paths is exponentially unlikely (for $L \to \infty$) for any fixed $k > 1$

A rugged yet easily navigable fitness landscape

Papkou et al., Science 2023

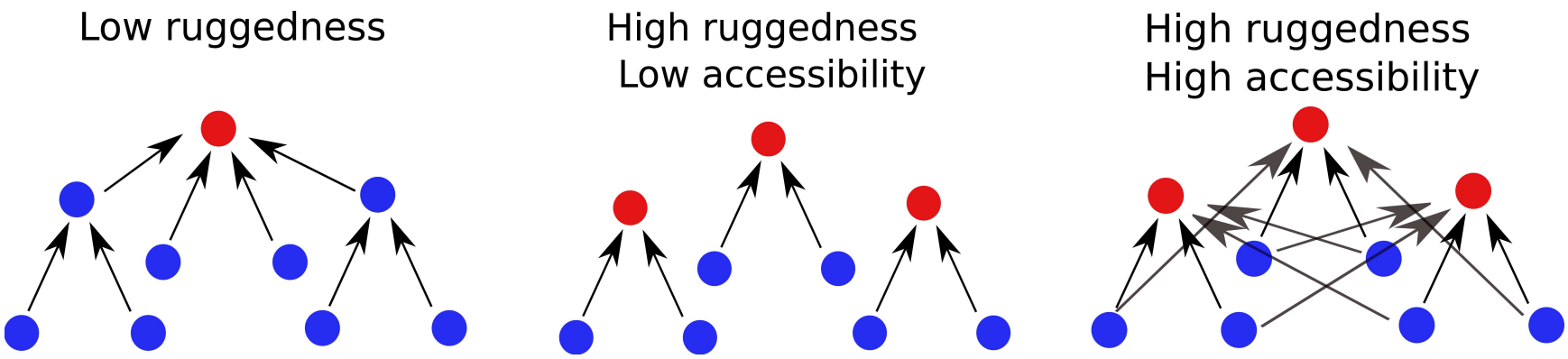
- $4^9 = 262,144$ combinations of nucleotides at 9 positions of the *fol A* gene in *E. coli* coding for dihydrofolate reductase (DHFR)

- Fitness measurements in trimethoprime yield 18,018 functional sequences



- 514 fitness peaks, 73 have high fitness and are highly accessible

Highly rugged yet highly accessible fitness landscapes



S.G. Das, S. Direito, B. Waclaw, R. Allen, JK, eLife 9:e55155 (2020)

The accessibility property

- Set notation: Identify a binary genotype σ with the subset of the locus set $\mathscr{L} = (1, 2, \ldots, L)$ at which $\sigma_i = 1$

- Example: $0000 = \emptyset$, $0001 = \{4\}$, $1010 = \{1, 3\}$, $1111 = \mathscr{L}$

- A fitness landscape has the subset-superset accessibility property if any peak is accessible from all its sub- and supersets along all direct paths

  Das et al. 2020

- The accessibility property implies a lower bound

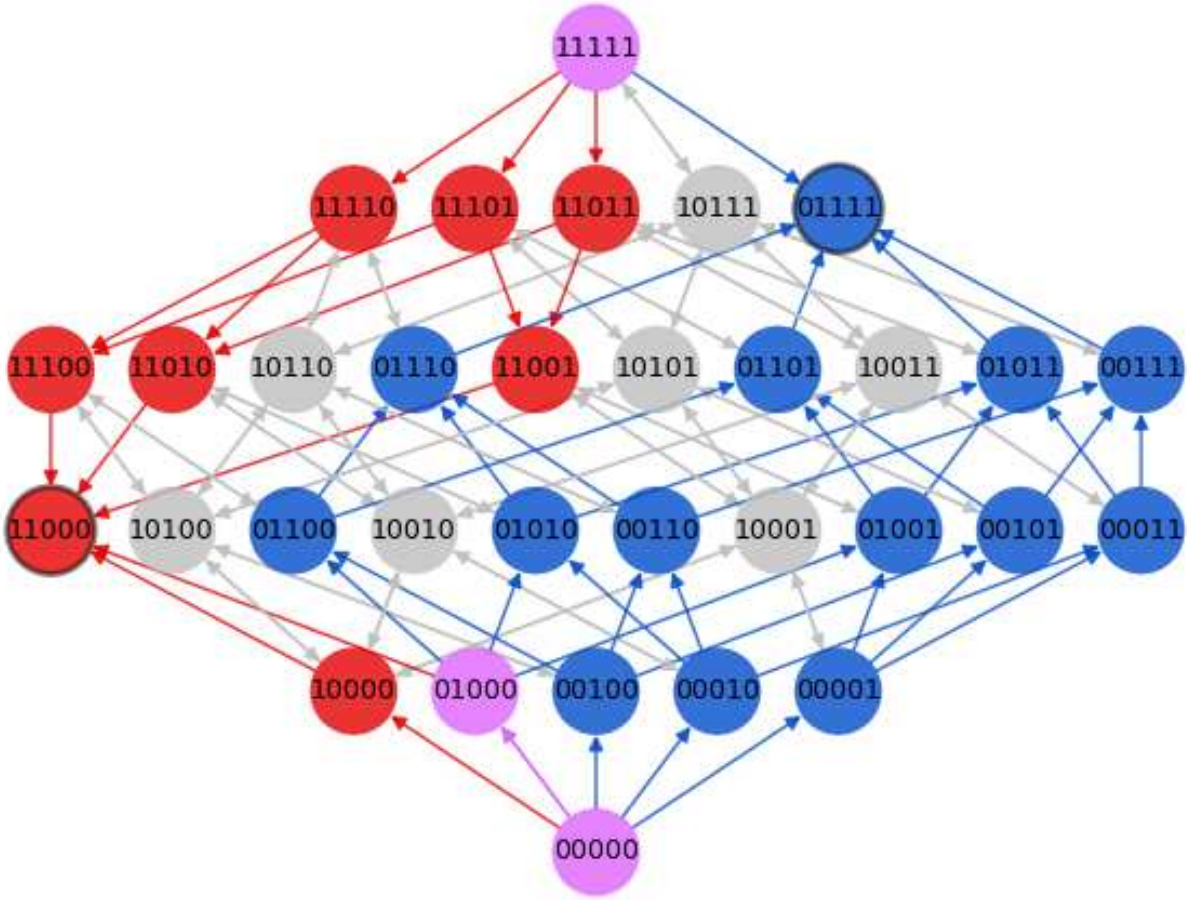
$$S_n \geq 2^n + 2^{L-n} - 2$$

  on the size S_n of the basin of attraction of a peak genotype with n mutations

- By Sperner's theorem, the property also implies an upper bound $N_{\text{max}} \leq \binom{L}{\lfloor L/2 \rfloor}$ on the number of fitness peaks

  Doros 2022

Illustration of the accessibility property for $L = 5$



red: sub/supersets of 11000

blue: sub/supersets of 01111

A sufficient condition for the accessibility property

- A fitness landscape displays universal negative epistasis (UNE), if for any two genotypes σ, σ' with $\sigma' \subset \sigma \subset \mathscr{L}$, and any subset $\tau \subseteq \mathscr{L} \setminus \sigma$

$$g_{\sigma \cup \tau} - g_{\sigma} \leq g_{\sigma' \cup \tau} - g_{\sigma'} \quad (*)$$

  i.e. the fitness effect of adding the mutations in τ is smaller in the background σ than in the background σ', if σ' is a subset of σ

  K. Crona, JK, M. Srivastava, J. Math. Biol. 2023

- For any peak genotype σ

$$g_{\sigma \cup \{i\}} - g_{\sigma} < 0 \quad \text{and} \quad g_{\sigma} - g_{\sigma \setminus \{j\}} > 0$$

  for all $j \in \sigma$, $i \in \mathscr{L} \setminus \sigma$

- Together with $(*)$ this immediately proves the accessibility property

Constructing landscapes with UNE

- Fisher's geometric model (FGM) generates rugged fitness landscapes by composing a linear genotype-phenotype map with a non-monotonic, single-peaked phenotype-fitness map Φ: Park et al., J. Phys. A 2020

$$\sigma \quad \rightarrow \quad z(\sigma) = \sum_{i=1}^{L} a_i \sigma_i \quad \rightarrow \quad g_\sigma = \Phi\left(\sum_{i=1}^{L} a_i \sigma_i \right)$$

- The expected number of fitness peaks in FGM grows exponentially in L

- FGM displays UNE if Φ is concave and $a_i > 0$:

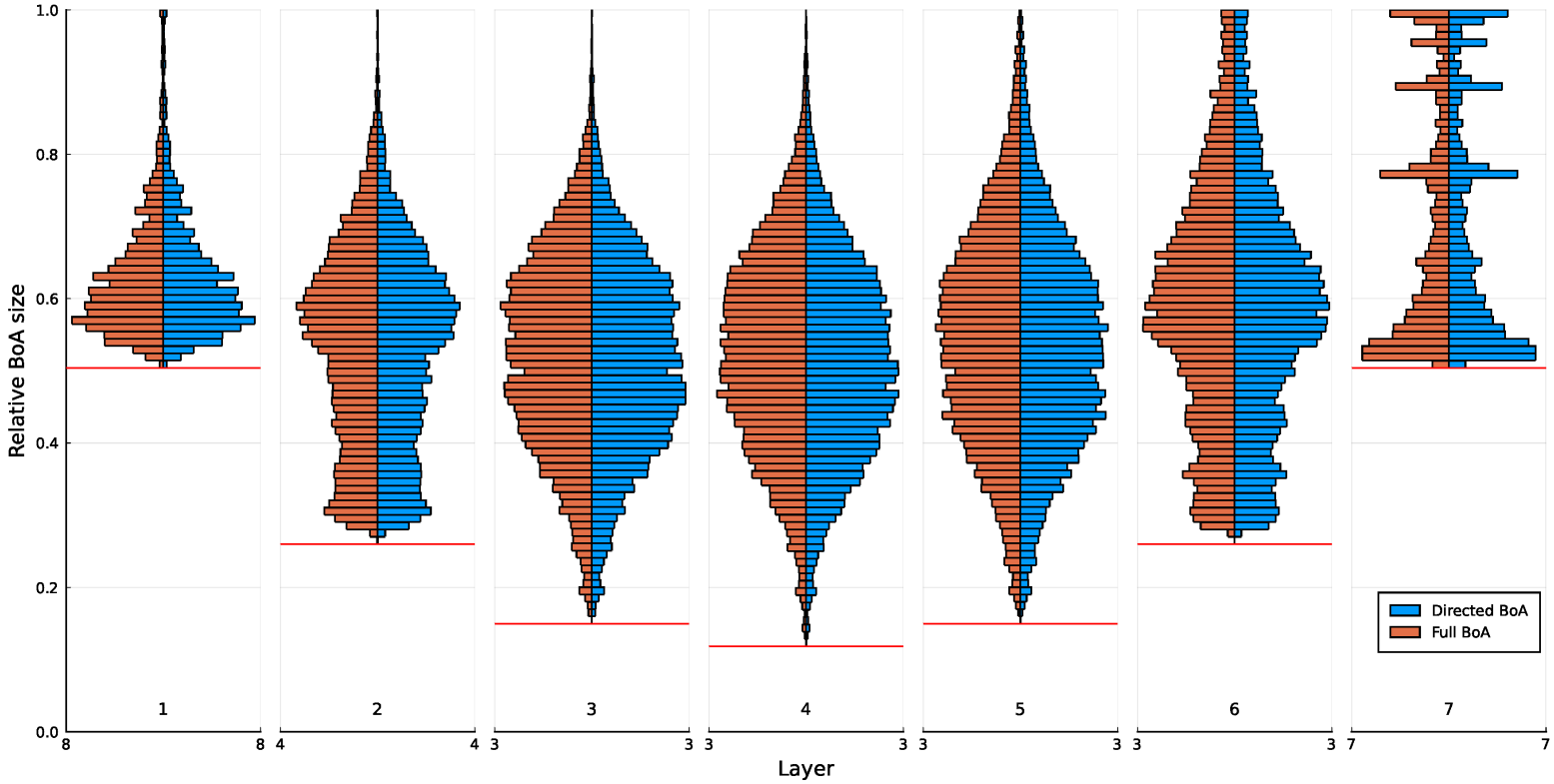
$$g(\sigma \cup \tau) - g(\sigma) = \Phi[z(\sigma) + z(\tau)] - \Phi[z(\sigma)] <$$

$$< \Phi[z(\sigma') + z(\tau)] - \Phi[z(\sigma')] = g(\sigma' \cup \tau) - g(\sigma')$$

- The positivity condition on the a_i can be relaxed

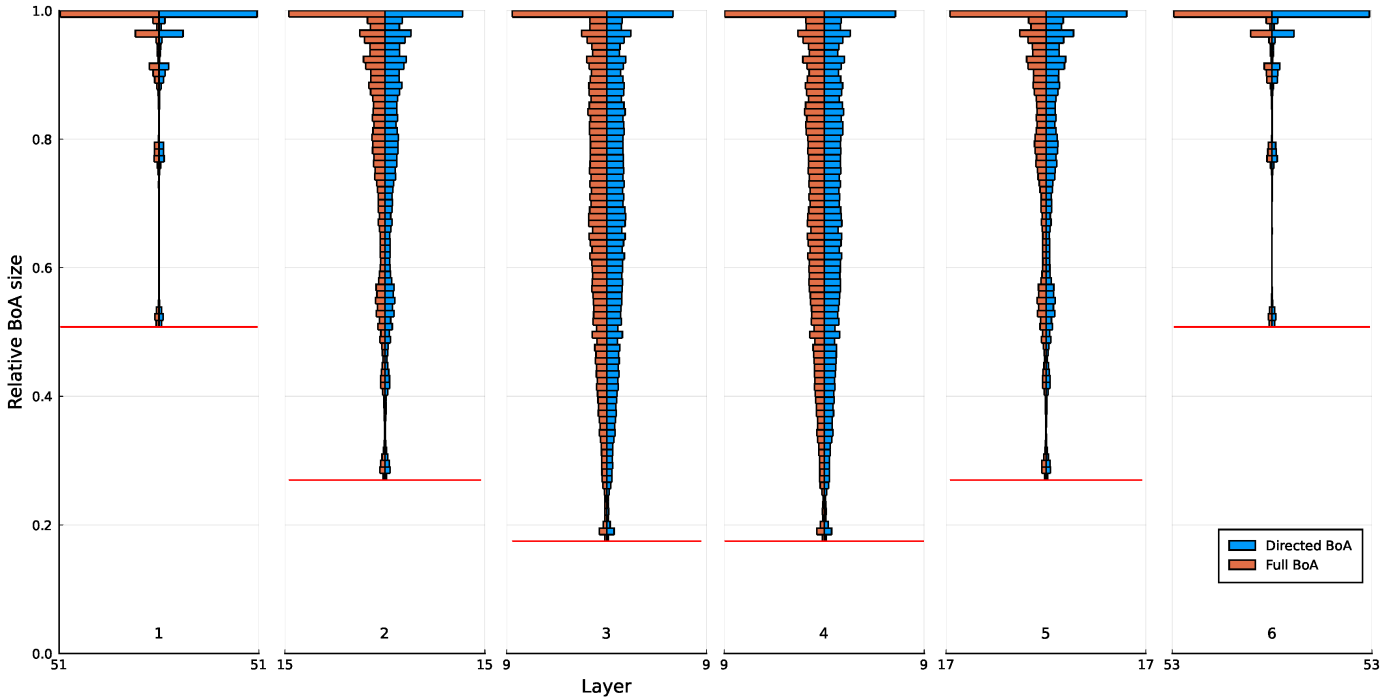
One-dimensional FGM with positive coefficients

D. Oros



- FGM with $L = 8$, Exp(1) coefficients a_i and $\Phi(z) = -(z-4)^2$
- Figure shows the distribution of the sizes of basins of attraction of peaks with $n = 1, \ldots, 7$ mutations

Basins of attraction in tradeoff-induced landscapes

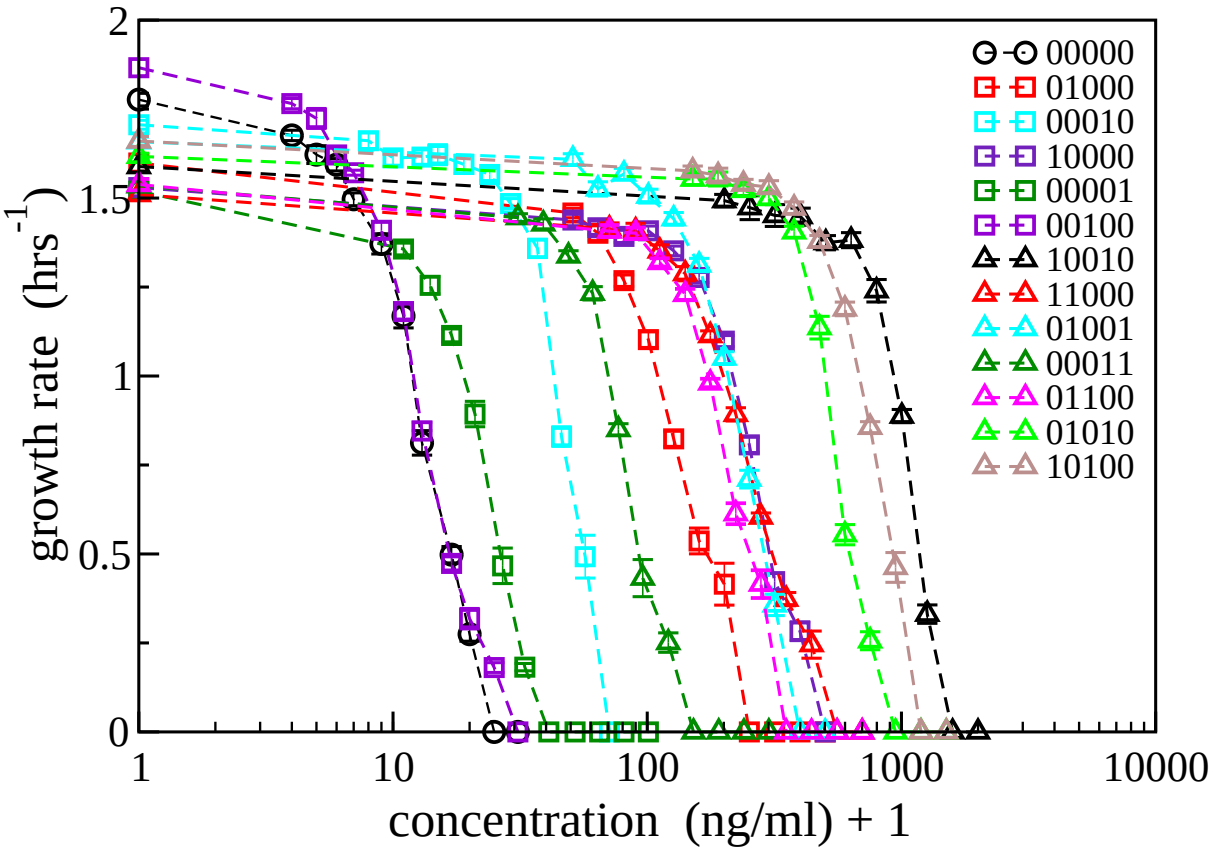


- The accessibility property was first found in a model of antibiotic resistance evolution at varying drug concentrations with adaptational tradeoffs

  Das et al. 2020

- Figure shows the distribution of the maximal sizes (across concentrations) of basins of attraction of peaks with $n = 1, \ldots, 6$ mutations

Dose-response curves with tradeoffs



- Growth rate as function of the concentration of ciprofloxacin for resistance mutants of *E. coli*

S. Direito, B. Waclaw, R. Allen (Edinburgh)

The tradeoff-induced landscape model (TIL)

- L mutations $i = 1, \ldots, L$ characterized by null-fitness r_i and resistance m_i relative to the "wild type" $(0, 0, \ldots, 0)$

- Fitness of a mutant $\sigma = (\sigma_1, \ldots, \sigma_L)$ at concentration x is

$$g_\sigma(x) = r_\sigma f(x/m_\sigma)$$

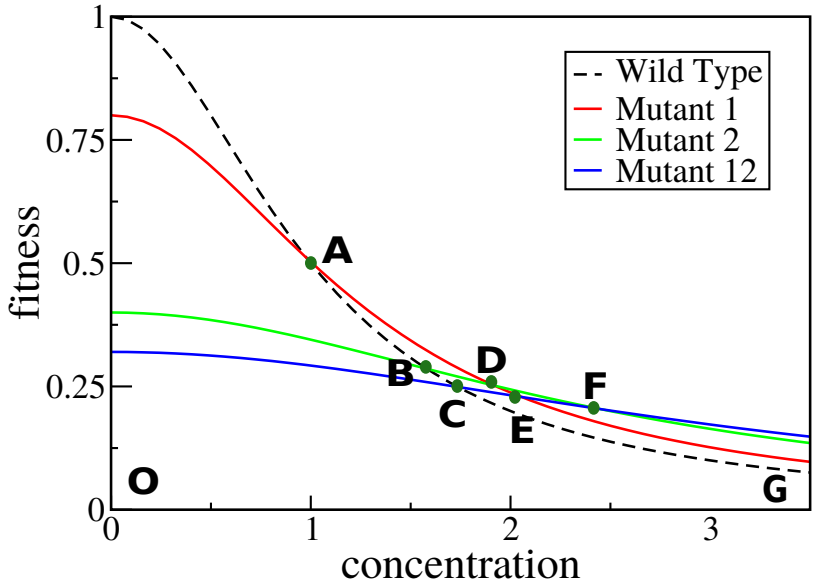
  with a single, monotonically decreasing shape function $f(x)$

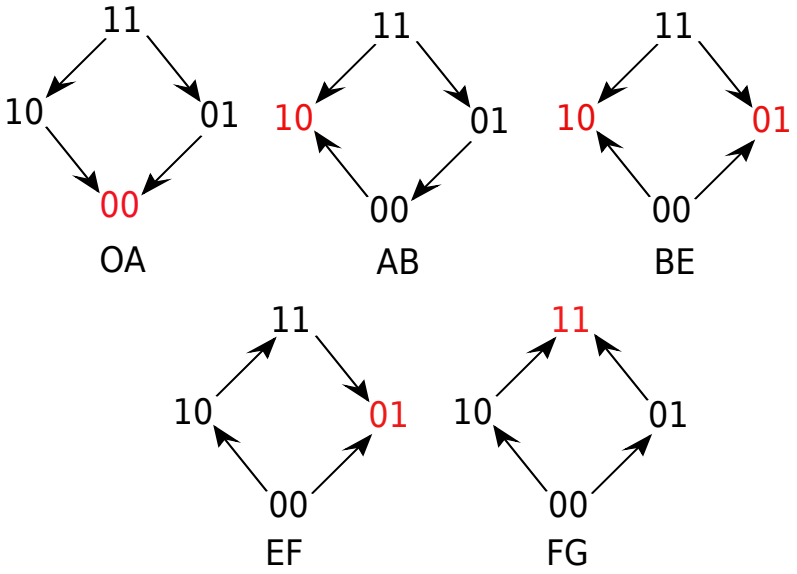
- The scaling parameters combine multiplicatively as

$$r_\sigma = \prod_{i=1}^{L} (r_i)^{\sigma_i} \quad \text{and} \quad m_\sigma = \prod_{j=1}^{L} (m_j)^{\sigma_j}$$

- Tradeoff: Every additional mutation increases resistance $(m_i > 1)$ and decreases growth rate $(r_i < 1)$

The tradeoff-induced landscape model (TIL)





- Crossing of dose-response curves flips arrows in the fitness graph
- The accessibility property follows from the ordering of crossing points
- Number of peaks at intermediate concentrations is exponential in L

S.G. Das, JK, M. Mungan, PRX 2022

Summary

- Spectacular advances in the empirical exploration of fitness landscapes have rekindled the interest in the underlying mathematical structures
- Structured landscapes can be less or more accessible than random ones
- Beyond accessible paths, the organization of the basins of attraction of fitness peaks is of interest empirically and theoretically

Summary

- Spectacular advances in the empirical exploration of fitness landscapes have rekindled the interest in the underlying mathematical structures
- Structured landscapes can be less or more accessible than random ones
- Beyond accessible paths, the organization of the basins of attraction of fitness peaks is of interest empirically and theoretically

Thanks to

- Kristina Crona, Suman G. Das, Daniel Oros, Jasper Franke, Muhittin Mungan, Stefan Nowak, Benjamin Schmiegelt
- Arjan de Visser and lab (Wageningen University)