Master's thesis

Benjamin Schmiegelt

Sign Epistasis Networks

July 18, 2016

Course of studies:	Physics (Master of Science)
Supervisor:	Prof. Dr. Joachim Krug
Second Examiner:	Prof. Dr. Johannes Berg
Institution:	University of Cologne, Institute for Theoretical Physics

Contents

1. Introduction

In this thesis I investigate the network structure of sign epistatic interactions for some typical stochastic fitness landscape models and its impact on dynamical properties of the evolutionary process under strong selection.

While proper definitions of the term evolution may be lacking, a common and important feature is "descent with modification"[20], i.e. a process of almost, but not fully, accurate replication. Abstracting away the biological details of this process one arrives at a large class of stochastic processes, which may vary in many details, but typically share the properties of reproduction with heritable traits, random mutations on these traits and (natural) selection favoring the reproductive success of individuals with certain sets of heritable traits. Other effects possible are recombination of heritable traits, either via sexual reproduction or horizontal DNA transfer, migration between populations and others.

In this thesis I consider only single populations reproducing asexually under higher selection pressure with mutation but without recombination, migration or any other such effect.

A population is generally made up of a finite number of individuals, which are distinguished from each other by a number of heritable traits, the union of which is called genotype of the individual. The genotype must be differentiated from the phenotype, which is the union of all expressed traits. Individuals in the population reproduce in such a way, that descendants inherit most of the ancestors genotype. In the biological context, inheritance of traits is assured via their encoding in the genome, which is chemically a massive polymer, the (desoxy-)ribonucleic acid (DNA/RNA), whose monomers, the nucleobases, come in four variants: cytosine, guanine, adenine and thymine (in case of RNA thymine is usually replaced by uracil). The string of these four bases encodes most of the heritable traits of an organism. The process of cell division with a complex chemical process of RNA/DNA synthesis and copy apparatus guarantees that daughter cells contain a copy of the parents genome. The genome is however generally not an exact copy and errors may occur due to stochasticity involved in the synthesis process as well as due to environmental degradation, e.g. through ionizing radiation. These mutations allow the population to change and to explore new traits over time.[1,34]

The third effect besides reproduction and mutation is selection. The number of viable and (fertile) offspring of an individual generally depends on its genotype. This is measured in the fitness value, meaning that an individual with a higher fitness will generally have more fertile offspring and its genotype will therefore prevail in the population over smaller fitness ones. There exists several different definitions of what exactly fitness is, however in this thesis most definitions will be considered equivalent and only the fact that higher fitness genotypes fixate in the population is of importance.[36]

The map assigning fitness values to genotypes, introduced by Sewall Wright[58], is called the fitness landscape and is the point of interest of this thesis.[16] This is not the most general and neither a biologically sufficient description of fitness in all settings, but will suffice for certain cases. Assignment of fitness values directly to genotypes cuts short genotypes-phenotype maps, which are considered intermediate functional layers instead.[46,51] Also inter-population interactions like non-linear dependence of the fitness on genotype frequencies, studied extensively in evolutionary game theory[23,52], and environmental effects changing the landscape are not incorporated directly, see e.g. phenotypic plasticity[9,48,57].

A "well-mixed" population (i.e. without significant spatial properties) is described as a map from genotypes to natural numbers counting the number of individuals with that genotype. If fitness differences are very large between any two genotypes, i.e. in the strong selection regime, then the smaller-fitness genotype has no chance of dominating the population ever.[19,38] Outside of the strong selection regime this is generally not true, because the fixation probability is related to the fitness differences, and for almost-neutral selection a lower-fitness genotype has a significant chance of fixating. The latter effect is also known as genetic drift and becomes stronger with smaller population sizes due to the inverse relation between population size and stochastic fluctuation magnitude.[5,18] How significant genetic drift is in evolutionary biology relative to natural selection has been a disputed issue for a long time.[37]

In the strong selection regime however the highest fitness genotype will always dominate in the long term and therefore it makes sense to declare it the population genotype. If also mutation is weak, then the fixation time of the dominant genotype is also shorter than the time between mutations and therefore the dominant

genotype will also be the only one existing in the population (except for the one previously dominant genotype). The dynamics arising in this limit case is that of an adaptive walk on the genotype space according to the fitness values assigned by the fitness landscape.[17,35,39]

In adaptive walk dynamics the population is a single dot on the fitness landscape moving step-wise on it towards higher fitness. This implies that an adaptive walk can only take a small subset of paths through the genotype space as determined by the fitness landscape. As soon as there is no genotype with higher fitness and reachable by mutation left, the dynamics stop and a terminal state is reached. Genotypes with this properties are local optima of the fitness landscape.

Since the adaptive walk dynamics are substantially more restricted in possible routes than in the general case, we are interested to know which parts of the landscape remain reachable for the population from given origin, which optima it will attain and which paths it will take to reach them. Restrictions on possible paths come about by sign-epistasis, the dependence of the sign of mutational fitness effects on the current background genotype, especially through reciprocal sign epistasis, that is mutations on two loci which are beneficial if applied together, but deleterious if applied one by one.[43,55,56] Sign epistasis is a extreme case of general epistasis, the dependence of mutational effect on background genotypes, which is considered to be of high importance of evolutionary dynamics.[49]

While experimental fitness landscapes become increasingly more available, they are often unsuitable to study (global) reciprocal sign epistasis due to selection bias of the chosen mutations and their small sizes.[11,47,50] Therefore I focus on theoretical models of landscapes, in particular a generalized version of the NK model. These models are defined on the hypercube, that is it is assumed that the genotype consist of a finite number of loci L, each of which can be found in two states/alleles. The only mutations allowed in one generation are those changing the alleles of one locus. The resulting mutation graph determining the adaptive walk transition graph (short of the monotone fitness requirement) is the L-dimensional binary hypercube, also known as Hamming graph $H(L, 2)$. In the NK model introduced by Kauffman and Weinberger[25,26] a number of small (partial) fitness landscapes over k loci each are added up to a total fitness value. The partial landscapes, they themselves considered high-dimensional random variables, are chosen i.i.d. and

may correspond to independent phenotypes or functional units. Several building blocks may share some of the same loci, thus constructing an explicit interaction network between loci, which is immediately related to the network of epistatic interactions between loci. The overlap in building blocks may represent functional inter-dependency or simultaneous (and potentially competing) effects of mutations on otherwise independent phenotypes, i.e. pleiotropy.[12,40,51] The parameter k determines the ruggedness of the model, which, scaling between 1 and L, indicates the amount of epistasis and the smoothness of the fitness landscape. In particular for $k = 1$, there is no epistasis and a simple linear fitness model is recovered, while higher k generally result in more complex landscape structures.

Kauffman and Weinberger suggested two choices of interaction schemes for the NK model and concluded that they behave very similar.[25,54] An additional interaction structure falling into the NK class has been considered by Perelson and Macken. In the block model (BN) loci are divided into disjoint sets of k loci and interactions are only present between loci of same sets.[41] However all three choices are similar in that, at large L and small k, they represent short-range interactions between loci. I will additionally consider an example of a diameter-2 interaction scheme, which will be seen to have qualitatively different properties in the $L \to \infty$ limit at constant k.

In this thesis I represent sign epistatic interactions as a directed graph over loci, with arrows determining whether a mutations on the source locus affect signs of mutations on the destination. I also add weights to the arrows determining the fraction of genotype backgrounds having this dependence.

This sign epistasis graph encodes a lot of information on limitations of adaptive walks. I will consider the expected weights of arrows on this graph, i.e. the strength of sign epistasis, as well as the global structure of the sign epistasis graph. Both may present limitations on adaptive paths on the fitness graph.

I will make use of some basics of (hyper-)graph theory and probability theory in this thesis. A short glossary with the precise definitions used can be found in the appendix.

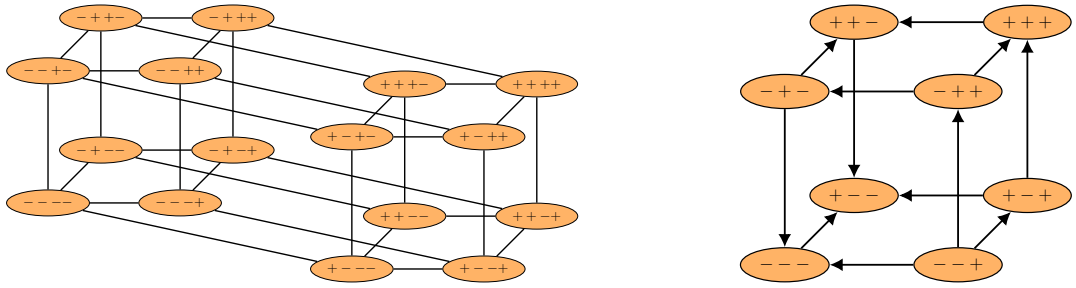

Figure 1: *Left: The hypercube representing mutations on the genotype space $\mathbb{H}$ for $L = 4$. Right: A fitness graph on the genotype space with $L = 3$. Arrows point towards higher fitness. This example has only one local maximum and minimum at $(+--)$ and $(--+)$ respectively. Nodes are labeled in both cases with the genotype sequence they represent. The hypercube projections onto the plane are also chosen such that edges/arrows on same loci are parallel.*

2. Mathematical framework

In this section I will introduce the fundamental mathematical formalism of evolution used in this thesis.

2.1. Genotype and mutation

The **genotype** consists of a finite number of **loci**, each of which may be in one of two states. The set of loci is denoted $\mathcal{L}$ and the **genome length** $L = |\mathcal{L}|$. I will always assume that the set of loci is totally ordered, i.e. $\mathcal{L} = \{l_1, \cdot, l_L\}$, where the assignment of indices to loci is fixed. Then a genotype g is any map a sequence of binary values $g = (g_{l_1}, \ldots, g_{l_L}) \in \{-1, 1\}^L$. The space of all genotypes is $\mathbb{H}_L = \{-1, 1\}^L$ or $\mathbb{H}_\mathcal{L} = \{-1, 1\}^\mathcal{L}$ if the labeling of loci should be relevant. I will write genotypes in the sequence representation as e.g. $(++-)$, meaning that $l_1 = +1$, $l_2 = +1$ and $l_3 = -1$.

Evolution proceeds by consecutive small mutations of the genome. Here I consider only basic point mutations at one locus. **Point mutations** are operators Δ_m for $m \in \mathcal{L}$ which switch the binary allele value of locus m, i.e. $\Delta_m g = (g_{l_1}, \ldots, -g_m, \ldots, g_{l_L})$. Note that Δ_m are their own inverses and with composition they generate a commutative group. The remaining elements of this group are multilocus mutation operators $\Delta_\mathcal{M}$ for $\mathcal{M} \subseteq \mathcal{L}$, such that $\Delta_\mathcal{M} = \prod_{m \in \mathcal{M}} \Delta_m$.

A natural metric describing the distance between genotypes in terms of minimum mutations needed to connect them is the **Hamming metric** $d_H(g,h) := \sum_{l\in\mathcal{L}}(1-\delta_{g_l h_l})$. This metric also induces a natural norm on the mutation group $\|\Delta_\mathcal{M}\| := |\mathcal{M}|$, such that $d_H(g,\Delta_\mathcal{M} g) = \|\Delta_\mathcal{M}\|$. The simple undirected graph over the genotype space with the Hamming metric as distance function is the L-dimensional **hypercube** $H(2,L)$.

The L-dimensional hypercube is the cartesian product of L copies of the complete graph on two vertices. Cartesian products of complete graphs are also known as Hamming graphs due to their relation to the Hamming metric. The notation of genotypes as sequences corresponds to the tuple representation of vertices in this cartesian product and loci are distinguishing labels of the cartesian factors. Elements of these sequences can be interpreted as coordinates and loci as the dimensions / labels of the coordinate axes.

2.2. Fitness landscape and fitness graph

A **fitness landscape**[16,58] is a mapping of fitness values to each genotype $F : \mathbb{H}_\mathcal{L} \to \mathbb{R}$. The space of all fitness landscapes over $\mathcal{L}$ will be denoted $\mathbb{F}_\mathcal{L}$, which can also be identified by the Euclidean space $\mathbb{R}^{2^L}$. For this choose any bijection $\phi : \mathbb{H}_\mathcal{L} \to \mathbb{N}_{\leq 2^L}$ and the derived bijection $\tilde{\phi} : \mathbb{F}_\mathcal{L} \to \mathbb{R}^{2^L}$ with $\tilde{\phi}(F) = \big(F(\phi^{-1}(1)), \ldots, F(\phi^{-1}(2^L))\big)$, i.e. each coordinate of the euclidean space corresponds to one fitness value of one genotype on the fitness landscape. I assume that the space of fitness landscapes inherits all geometrical and topological properties of the euclidean space via this bijection.

Thus the fitness values will also be written as $F_g := F(g)$, where the left-hand notation resembles the coordinate notation for vectors, only that the bases here are labeled by genotypes rather than natural numbers.

The effect of a mutation on the fitness value will be shortened by the notation $\Delta_l F(g) := F(\Delta_l g) - F(g)$. Here Δ_l can be interpreted as an operator on the space of fitness landscapes, i.e. $\Delta_l F$ is itself a fitness landscape in $\mathbb{F}_\mathcal{L}$ and can be thought of as the discrete analog to a partial derivative along locus/coordinate axis l.

For a given fitness landscape the **fitness graph**[6] is defined as the orientation of the hypercube graph constructed from genotype space, such that arrows point

towards increasing fitness. This is only well-defined as long as there are no neutral mutations, which will almost surely be true for all models I consider here by construction because I consider the strong selection limit and thus I will ignore any issues of defining the arrow direction for neutral mutations. The fitness graph encodes the difference signs of all possible mutations and so it bounds all adaptive walks.

In contrast to the number of possible fitness landscapes there are only finitely many fitness graphs on a given genotype space. Valid fitness graphs are exactly all acyclic orientations of the hypercube. Cycles may not appear, because that would imply that fitness increases after traversing it once. If there is however no cycle, matching fitness values can simply be assigned by traversing the graph in topological order. The asymptotic number of acyclic orientations of the hypercube is known to be $L^{\Theta(2^L)}$ with more accurate lower and upper bounds available.[24,30] This number scales superexponentially and the number of nodes in the graph is still exponential in L, so that the fitness graph is unsuitable as a visualization for anything but very small system sizes.

The **fitness rank landscape** of a fitness landscape is the map $R : \mathbb{H}_{\mathcal{L}} \to \mathbb{N}^{\leq 2^L}$, which maps to each genotype its fitness value rank. Since I explicitly disallow neutral mutations, this map is well-defined and always bijective. The properties under consideration in this thesis are mostly independent of fitness effect magnitudes and therefore the rank landscape contains all the information necessary. In fact the properties of the fitness graph were already sufficient for my purposes, but the rank landscape contains a superset of its information, though there are still only finitely many rank landscapes over any genotype space. Using that there are 2^L ranks to distribute onto 2^L genotypes in any possible permutation, there are $(2^L)!$ possible rank landscapes. The number of rank landscapes is significantly larger than the number of fitness graphs as can be seen by taking the logarithm of both values: $\ln(L^{\Theta(2^L)}) \sim \ln(L)\Theta(2^L)$ vs $\ln(2^L!) \sim 2^L L \ln 2$. Here the $\ln(L)$ factor is outweighted by the L factor.

2.3. Fourier transform

Being functions over a finite commutative group, fitness landscapes have Fourier transforms[33,53], here defined by:

$$F(g) = 2^{-\frac{L}{2}} \sum_{\tilde{g} \in \mathcal{P}(\mathcal{L})} \tilde{F}_{\tilde{g}} \prod_{l \in \tilde{g}} g_l$$

where $\tilde{F}_{\tilde{g}}$ are the real-valued Fourier components, which are indexed by the Fourier space elements, i.e. subsets of the locus set. The Fourier space is isomorphic to the genotype space and a subset of loci $\tilde{g}$ can be interpreted as a (Fourier) genotype by setting all alleles to ± 1 if $l \in \tilde{g}$ and ∓ 1 otherwise. The prefactor is chosen such that the inverse Fourier transformation is identical.

2.4. Projection and slicing

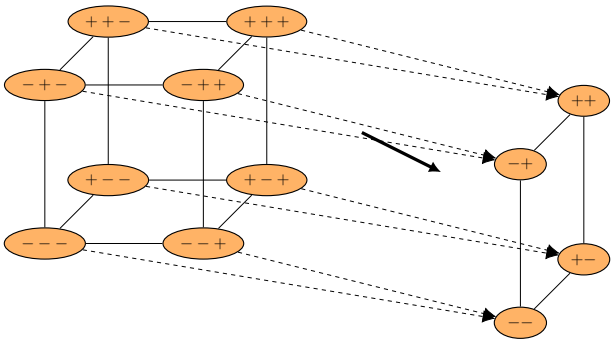

Figure 2: *Projection of the genotype hypercube for $L = 3$ onto $\mathcal{M} = \{l_1, l_2\}$.*

Given a subset of loci $\mathcal{M} \subseteq \mathcal{L}$, I denote the **projection** of a genotype g into the subspace $\mathbb{H}_{\mathcal{M}}$ by $\downarrow_{\mathcal{M}} g$. It is defined such that $(\downarrow_{\mathcal{M}} g)_l = g_l$ for all $l \in \mathcal{M}$. For example let $\mathcal{L} = \{l_1, l_2, l_3, l_4, l_5\}$, then the projection onto $\mathcal{M} = \{l_2, l_3, l_5\}$ in sequence representation is given by:

$$g = (g_{l_1}, g_{l_2}, g_{l_3}, g_{l_4}, g_{l_5}) \quad \mapsto \quad \downarrow_{\mathcal{M}} g = (g_{l_2}, g_{l_3}, g_{l_5}) \tag{1}$$

Here it is not obvious how loci in the projection are ordered. I will assume, that they retain their order from the original sequence.

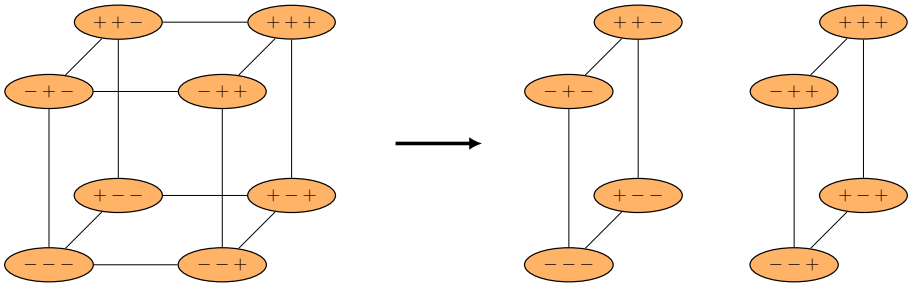

Figure 3: *The two (l_1, l_2)-slices with backgrounds $(l_3) = (\pm)$ of the genotype hypercube for $L = 3$.*

Given a subset $\mathcal{M} \subseteq \mathcal{L}$ and a genotype $h \in \mathbb{H}_{\mathcal{L}\setminus\mathcal{M}}$ the **$\mathcal{M}$-slice along** h is the subgraph of the genotype space $\mathbb{H}_\mathcal{L}$ induced by all $g \in \mathbb{H}_\mathcal{L}$ with $\downarrow_{\mathcal{L}\setminus\mathcal{M}} g = h$. h is called **background**.

For every $\mathcal{M} \subseteq \mathcal{L}$, there are $2^{|\mathcal{L}\setminus\mathcal{M}|} = 2^{L-|\mathcal{M}|}$ possible $\mathcal{M}$-slices. All $\mathcal{M}$-slices are disjoint and are themselves hypercubes of order $|\mathcal{M}|$. Consider $\mathbb{H}_\mathcal{L}$ with all edges in the $\mathcal{M}$-slices removed. The remaining graph has then $2^\mathcal{M}$ components, each one a hypercube of order $L - |\mathcal{M}|$. These are the $(\mathcal{L} \setminus \mathcal{M})$-slices of the genotype space. In some sense $\mathcal{M}$-slices and $(\mathcal{L} \setminus \mathcal{M})$-slices are therefore complementary and orthogonal. Choosing a background for the $\mathcal{M}$-slice and a background for the $(\mathcal{L} \setminus \mathcal{M})$-slice bijectively determines the genotype in $\mathbb{H}_\mathcal{L}$. All these properties are a consequence of the genotype space being a cartesian product of isomorphic graphs and not special for the hypercube in particular.

The special case of the two dimensional hypercube slices generated by subsets with $|\mathcal{M}| = 2$ are called **squares**.

Slices were defined for the genotype space, but they naturally transfer to the fitness graph, which simply has directed edges instead. Taking the fitness values to be assignments to the genotype nodes in the graph, the slice of the fitness landscape is also defined as simple restriction of the function.

The fitness landscape restricted to a $\mathcal{M}$-slice along background h can also be Fourier transformed. Its order will be at most $|\mathcal{M}|$ instead of $\mathcal{L}$ for the full landscape. It is also easy to determine, because the $\mathcal{M}$-slice is a simple restriction, with the restriction being, that all alleles for loci in $\mathcal{L} \setminus \mathcal{M}$ are fixed by h. Simply fixing these alleles in the full Fourier transform one arrives at the Fourier coefficients of

the slice, here denoted by a tick:

$$\tilde{F}'_A = \sum_{A \subseteq B \subseteq \mathcal{L}} \tilde{F}_B \prod_{l \in B \setminus A} h_l \tag{2}$$

for all $A \subseteq \mathcal{M}$.

3. Adaptive walk and landscape properties

3.1. Adaptive walks

In the population, at any given time, there is only a subset of genotypes actually present. One of them has the highest fitness value and thus, in the strong selection limit, it is the only one that can successfully spread through the population. Thus it is sensible to consider this genotype the population genotype, assuming it does not go extinct immediately.

This reduces the state space of the stochastic process from all populations to just the genotype space. Because selection is strong it is also, except for very small populations and certain definitions of birth/death rate, impossible for a genotype with fitness lower than the majority genotype to fixate. Thus the resulting dynamic is that of an **adaptive walk**, i.e. a time-discrete Markovian stochastic process on genotype space with possible transitions determined by the fitness graph.[6,7,15,26]

Because the fitness graph has no cycles this implies that an adaptive walk will eventually halt at a genotype without outgoing arrows in the fitness graph, i.e. at a **local optimum** of the fitness landscape.

3.2. Local optima

A **local optimum** of a fitness landscape is a genotype g, such that for all point mutations Δ_l with $l \in \mathcal{L}$: $\Delta_l F(g) < 0$. An adaptive walker will always be stuck at a local optimum. These are the absorbing states of the process. Because there is typically more than one local optimum, the process is therefore non-ergodic.

Landscapes have at least one local optimum, the **global optimum**, which is the genotype with the the largest fitness value. I will write Ω for the global optimum.

3.3. Transition schemes

An adaptive walker has to move towards increasing fitness, however there may still be many such choices. The precise transition policy can be defined in several ways. A simple choice is the **random walk**, in which any of the available fitness-increasing steps at a given time point is chosen with equal uniform probability. Such an adaptive walk does not care for fitness magnitudes or global fitness ranks at all, but is fully determined by the fitness graph. The random walk arises in the strong selection limit combined with very low mutation rate, so that every new genotype arising by mutation is either fully fixed or completely lost before any new mutant can arise. This parameter region is also known as **strong selection weak mutation (SSWM) regime**.

Often however mutation rate is large enough for many mutants to spawn in one generation. Then there are potentially several mutants with higher fitness than the majority genotype present in transition phases and, due to the strong selection limit, only the one with the highest fitness will fixate. This is also known as **clonal interference**. Whether a mutation fixates is not solely determined by its individual fitness effect anymore. In the most extreme case every possible mutant arises immediately, so that the walker can choose the highest fitness value from all the possible fitness-increasing steps on the genotype space. This transition scheme is known as **greedy walk**. Here one has to be careful what mutants are considered allowed in one generation. Generally I consider here only one single point mutations as defined earlier per offspring, but realistically if the mutation rate is high enough to generate all mutants fast enough, there will also be a high chance of double-mutants or mutants with higher number of point mutations in one generation.

For the greedy walk one needs to retain global rank information of the fitness landscape. The fitness graph does not contain the necessary information to choose between two fitness-increasing mutations. But at least the fitness ranks are sufficient information and actual fitness values need not be considered. The greedy walk differs from the random walk in that it is fully deterministic. Given a fitness landscape and a starting point, the terminal state and the path to it are uniquely determined.

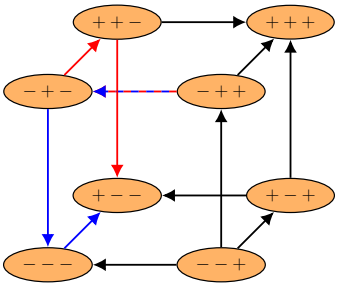

Figure 4: *Fitness graph with $L = 3$ and two local optima, at $(+++)$ and $(+--)$. The local optimum $(+++)$ can not be reached by any accessible path from its antipodal $(---)$. $(+--)$ can however be reached by its antipodal $(-++)$ by two accessible paths, marked red and blue. Both paths are short, i.e. of length 3 and without back mutation.*

The greedy walk increases fitness as fast as possible, often times getting stuck in local optima quickly. Conceptually it might be interesting to consider the opposing behavior, an adaptive walker which considers all fitness-increasing mutations, but chooses the one increasing fitness by the smallest positive amount possible. Such a walk is known as **reluctant walk** and one might expect it to have slow short-term fitness increase, but potentially by having longer walk lengths it might be able to increase fitness by a higher amount. Like the greedy walk, the reluctant walk is fully deterministic.

Between the random and greedy/reluctant walk interpolations have been used, e.g. by considering only a uniformly chosen fraction of fitness-increasing neighbors in every step or by choosing uniformly between the n top or bottom fitness ranks of fitness-increasing mutations.

My results will not depend on the scheme chosen. They will be general limiting statements on any adaptive walk, based on whether an adaptive walker can possibly take a path, independently of the probability it assigns to it.

3.4. Evolutionary accessible paths

Given a transition scheme and a starting point an adaptive walker will in finite time reach a local optimum and terminate. In addition to the question of which optimum it will reach, it is also interesting to consider the path it took to get there. A **(evolutionary/mutational) path** is a path in the graph theoretical

sense on the genotype hypercube, i.e. it is a finite non-empty sequence of genotypes $p = (g^{(1)}, \ldots, g^{(m)})$ without repeated genotypes such that two adjacent elements in the sequence can be reached via a single point mutation. The first element $g^{(1)}$ is called the **initial genotype** and the last one $g^{(m)}$ is called **final genotype**. The **(path) length** is the number of steps made, that is $m - 1$.

A path is called **accessible** if the fitness values along the sequence of genotypes is increasing monotonically. Because adaptive walkers may never make a transition decreasing fitness, accessible paths are the only possible trajectories they may take and testing accessibility can therefore indicate limitations on all kinds of adaptive walkers without having to consider dynamical properties.

One recent question of interest is the number of available paths to the global or local optima of the landscape. Here one usually considers path of maximal length. For every final genotype $g^{(\mathrm{fi})}$ there is exactly one in maximal distance L to it, the **antipodal genotype** $\Delta_{\mathcal{L}} g^{(\mathrm{fi})}$ having all loci in the opposite allele. Here mostly one considers the global optimum Ω as final genotype and its antipodal $\Delta_{\mathcal{L}}\Omega$ as initial genotype. Alternatively one may consider paths starting at a random genotype, although due to the topology of the hypercube, for large L, almost all genotype have about distance $\frac{L}{2}$ from any other genotype.

A helpful differentiation is that of short and long paths. A path is **short** if it is exactly as long as the Hamming distance between initial and final genotype. This is the minimal distance a path between to genotypes can have, by definition of the distance. A path is **long** if it is not short. Short paths on the hypercube are distinguishable in that they do not apply mutations to the same locus twice, i.e. only forward mutations happen, reversion of a mutation in a later step of the path does not happen.

Short paths are somewhat easier to handle mathematically, e.g. the number of short paths between two genotypes with Hamming distance d is exactly $d!$, because there are d loci in need to be mutated, while they may also only be mutated once and then only there order is left as choice. In particular the number of short paths between a genotype and its antipodal is $L!$.

A genotype is said to be **accessible**, if there is at least one accessible path to it from its antipodal.

3.5. Basin of attraction

Another value used to describe the structure of fitness landscapes is the **basing of attraction**.

The **(greedy) basin of attraction** of a local optimum g are all genotypes h, such that a greedy walk starting from h ends in g. The **reluctant basin of attraction** is similarly the set of genotypes h such that reluctant walks end in g. More generally the **weak basin of attraction** is the set of all genotypes h such that at least one adaptive walk can reach g, and the **strong basin of attraction** is the set of all genotypes h such that all adaptive walks will reach g.

Note that greedy and reluctant basin of attraction define an equivalence relation on genotypes, while weak basins of attraction for random adaptive walks can be overlapping for different local optima and strong ones will generally not cover the genotype space.

The relative sizes of basins of attraction suggest the distribution of adaptive walk outcomes. If sizes have low variance between local optima it is expected that all local optima are reached with similar probabilities, while for large variances some optima are preferred outcomes over others. The strong basin of attraction in particular determines the genotypes from which an adaptive walk has committed to one optimum. As far as an element of a strong basin of attraction is reached by any adaptive walk there is no outcome uncertainty left.

4. Specific models

The models for fitness landscapes used here are stochastic in nature, i.e. a fitness landscape model is a probability distribution over $\mathbb{F}_{\mathcal{L}}$ identified as Euclidean space $\mathbb{R}^{2^L}$. One hopes to recover general properties of real fitness landscapes from typical or average properties of these models. Due to the lack of sufficient empirical data and precise theory this seems to be one of the few approaches possible. Deduction of valid fitness landscape models from microscopic interactions is practically impossible due to the enormous complexity of even the simplest evolving biological systems (e.g. self-replicating RNA). For the same reason the models used are mostly based on theoretical guesstimates taking roughly into account known underlying principles

of theoretical biology and genetics.

4.1. House-of-Cards model

Probably the simplest stochastic models one may think of is that of completely random uncorrelated fitness. In relation to physics these kind of models are known as **random energy models**. For fitness landscapes specifically the name **House-of-Cards (HoC) model** has been established.[27] Given a base distribution of individual fitness values, the full landscapes is constructed by assigning each genotype an identically and independently distributed value from this distribution. The joint probability density of fitness values on the HoC landscape F is therefore just:

$$p_{\mathrm{HoC}}(F) = \prod_{g \in \mathbb{H}_{\mathcal{L}}} p_f(F_g)$$

where p_f is the base fitness value distribution's density, which I will here assume exists, i.e. the base distribution is supposed to be absolutely continuous on its support. This assures that for a finite number of genotypes almost surely no mutation has zero fitness effect and that no two fitness differences are exactly equal. For simplicity of the following calculations I also assume that the base fitness distribution has mean zero and that its variance exists.

In the HoC model all fitness values are by definition uncorrelated and even independent. This model is mathematically easy to handle, but might be considered very unlikely to describe actual empirical landscapes because given a somewhat well adapted genotype, a random mutation will result in complete loss of any adapted fitness, implying that the progress falls apart. Typically one would suspect however that single mutations are unlikely to drastically change the progress made so far. Nonetheless the House-of-Cards model is a good starting point to construct further models which incorporate some random contribution in addition to a more conservative fitness contribution.

The properties studied here in the strong selection regime do only depend on fitness ranks and the HoC model has the nice property that the actual base fitness distribution (as long as it is absolutely continuous on its support) does not matter for the distribution of the fitness ranks. Absolute continuity guarantees that ties

do not need to be handled and so the distribution of fitness ranks is simply uniform over all possible fitness rankings of the genotypes.

The HoC Fourier components are given by:

$$\tilde{F}_{\tilde{g}} = 2^{-\frac{L}{2}} \sum_{g \in \mathcal{P}(\mathcal{L})} F(g) \prod_{l \in g} \tilde{g}_l$$

Every fitness value appears exactly once in the sum, but with $\tilde{g}$-dependent sign. Because all $F(g)$ are i.i.d. and have mean zero, $\tilde{F}_{\tilde{g}}$ also has zero mean. If the variance of $F(g)$ is σ_f^2, then the variance of $\tilde{F}_{\tilde{g}}$ is also simply $2^{\frac{L}{2}}\sigma_f^2$. The Fourier transformation can be understood as an orthogonal operator on $\mathbb{F}_{\mathcal{L}}$ and therefore the off-diagonal elements of the covariance matrix of Fourier components are zero. This can be explicitly seen through the alternating signs of the product $\prod_{l \in g} \tilde{g}_l$. If the base fitness distribution is normal, then the joint distribution of Fourier coefficients is consequently also normal and they not only uncorrelated, but also independent. For other distribution, at least the marginal distributions of Fourier coefficients converge (up to the factor $2^{\frac{L}{2}}$) to a normal distribution with mean zero and variance σ_f^2 due to the central limit theorem.

Any slice of a HoC landscape is itself again a HoC landscape of smaller size, because the slicing is simply function restriction.

Many properties of the HoC landscape are known. By simple combinatorial arguments the mean number of optima is exactly $\frac{2^L}{L+1}$ and the mean number of accessible paths from the antipodal to the global optimum without backsteps is exactly 1.

The probability that at least one such path exists was shown to be asymptotically $\frac{\ln L}{L}$, decreasing slowly to zero, while the mean number of paths still stays 1.[22] Naively it would seem that it is difficult to find paths in the HoC model because mutational effects are so unpredictable and often destructive. This is shown by the small mean number of accessible paths in comparison to the number of possible paths ($L!$). However the rather slow decrease of the probability of zero accessible paths suggests that the high dimensionality of the problem still is high enough to not suppress the likelihood of long monotone fitness increases. As it will turn out there are much less rugged models with much smaller probability of accessible paths, showing that the high number of degrees of freedoms actually helps a lot in

finding an accessible path. If backsteps are allowed the probability even converges to a constant.[2] More detailed results on the number of accessible paths from the antipodal to the global optimum with or without backsteps were calculated by Hegarty and Martinsson[22,29] as well as Berestycki and Brunet[2,3].

Adaptive walks on HoC landscapes have asymptotically on average only a length of $e - 1$ steps in the greedy case and $\ln(L)$ steps in the random case.[31,35]

4.2. Linear model

On the opposite side of the spectrum of possible fitness landscape models one can find the **linear (additive/non-epistatic) model**, which is defined by a fitness contribution of every locus depending only on the state of said locus, i.e. the fitness can be written as

$$F(g) = \sum_{l \in \mathcal{L}} f_l g_l$$

where f_l are independently and identically distributed random variables. Again I will assume that the distribution of f_l is absolutely continuous with mean zero and finite variance. In this class of models long range correlations exist, as the correlation drops of linearly with distance.

The effect of every point mutation is always the same and they can effectively be handled as independently. The adaptive walk on such a landscape is actually decomposable into L short adaptive walks over single loci, which will generally only take zero or one step to the optimum, depending on whether they were already in the higher fitness state of the two alleles. It follows that the adaptive walk on such a landscape is equal in length to the initial distance from the one and only optimum and that every permutation of mutations is equally accessible. The number of accessible paths without backsteps from the antipodal to the global optimum is therefore the maximal value of $L!$.

The Fourier transform of this model is directly given by its definition, i.e. $\tilde{F}_{\{l\}} = \tilde{f}_l$ and $\tilde{F}_{\tilde{g}} = 0$ for $|g| \neq 1$, meaning that the highest order of interactions is 1, thus the naming of the model.

The linear model may be considered smooth in the sense that the effect of most mutations does not vary much (at all) for neighboring genotypes, which

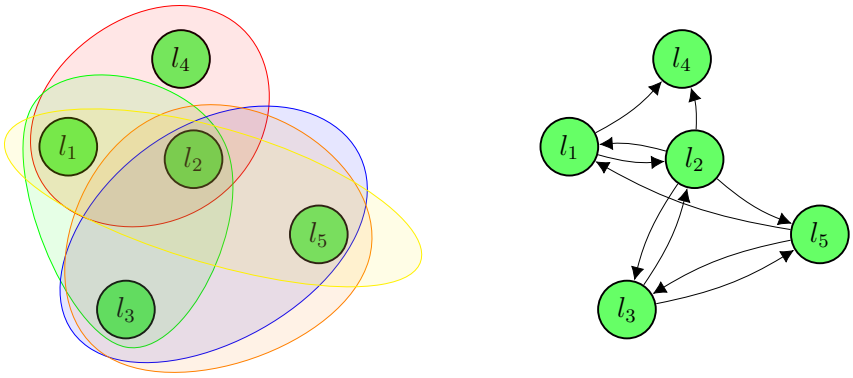

Figure 5: *Example of a classical NK structure as hypergraph on the left and in the simplified directed graph form for classical NK structures on the right. The parameters of this structure are $L = 5$ and $k = 3$, as well as $\#\mathbf{N} = L = 5$ as in all classical structures. Note that $|\mathbf{N}| = 4$, because one block has multiplicity 2 (blue and orange).*

is absolutely not the case for the HoC model, in which mutational effects on neighboring genotypes are almost always uncorrelated.

This model incorporates the property of consistent mutations which the HoC model lacks. One expects usually that a mutation which is beneficial in one individual is also beneficial in a individual with a slightly different genotype. Contrary to this toy model mutations in biological system are not always independent of the background genotype. The real world system is much more complex and would be too primitive to result in interesting behavior if it was purely linear. Rather one expects slight changes in mutation effects on slightly varying background genotypes. This effect moving away from the idealized linear model is known as epistasis and will be introduced in more detail later. Especially one expects certain combinations of mutations to be correlated stronger than others because they may e.g. be part of the same gene, functional unit or metabolic process.

4.3. Generalized NK model

Because both the House-of-Cards and the linear model are mostly toy models modeling two different aspects each only, it seems to be a good idea to find a model which can interpolate between the two. One choice for such a landscape model is

the NK model, introduced by Kauffman and Weinberger[25,26]. It is specified by a parameter k in addition to the number of loci L. The new parameter interpolates between 1 and L corresponding to the linear and HoC model respectively by summing L i.i.d. HoC landscapes over k-subsets of the overall locus set. Loci sharing one of the partial landscapes are then strongly correlated, while most pairs of loci are still uncorrelated on their own. Here I use a slightly generalized variant of the model than that of the original Kauffman and Weinberger papers. I also consider only the case of a fixed ruggedness parameter k while L goes to infinity. Other limits are also interesting and have been studied, in particular the limit of constant $\frac{k}{L}$ as $L \to \infty$.

A generalized NK fitness landscape model over a set of loci $\mathcal{L}$ is defined by an interaction network between loci, here called the NK structure, and a building block fitness landscape model. The NK structure is a k-uniform hypergraph $\mathbf{N}$ over the set of loci. Each edge $D \in \mathbf{N}$, here also called (NK) block, contains k loci, and corresponds to one partial landscape's fitness contribution, such that the total fitness of a genotype is:

$$F_L(g) = \sum_{D \in \mathbf{N}} \sum_{i=1}^{\mathcal{I}_{\mathbf{N}}(D)} f_{D,i}(\downarrow_D g)$$

where $f_{D,i}$ are i.i.d. (partial) fitness landscapes over k loci. Here I am referring to independence of the whole partial landscapes interpreted as elements of $\mathbb{R}^{2^k}$. I assume in this thesis that the partial landscapes are always HoC landscapes, although it is possible to generalize the main results to other cases as well. The HoC model fulfills certain properties simplifying things here. In particular it is invariant under permutation of loci and therefore it is not necessary to specify an order of loci in blocks. Also, with the HoC partial landscapes not only partial landscapes are as a whole independent, but so are their individual fitness values.

Due to linearity of the Fourier transform, the Fourier components of the NK model are simply the Fourier components of the individual partial landscapes summed. Because the Fourier components of the partial landscape behave like in the HoC model, this means that all Fourier components $\tilde{F}_{\tilde{g}}$ with $\tilde{g} \subseteq \mathcal{L}$ are unequal zero if and only if there is a NK block D, such that $\tilde{g} \subseteq D$. If they are unequal to

zero, they are, like in the HoC model marginally identically distributed, however their variance is multiplied by the number of NK blocks with $\tilde{g} \in D$.

I will discuss mainly properties in the limit $L \to \infty$. Therefore I consider not NK models for single parameter sets of $\mathbf{N}$ and k, but rather a sequence of such models with increasing L. A valid specification of the generalized NK model therefore needs to define a NK structure for every L. All variables are assumed to be implicitly functions of L, except when stated otherwise. Also I will allow the NK structure for each given L in the sequence to be a random variable, i.e. there does not need to be a fixed structure but a probability distribution over structures of the same size is sufficient. Specific choices for the NK structure will follow in later subsections.

The limits under consideration are usually under the assumption that k is fixed as $L \to \infty$. Other limits are also of interest, for example $L \to \infty$ with $\frac{L}{k} = \text{const.}$, but such cases will be mentioned explicitly. I also assume that the multiplicity of every element of $\mathbf{N}$ as $L \to \infty$ is bounded by a constant, such that over each finite subset of loci there are only finitely many possible NK structures. In order to avoid neutral mutations I will assume that for each L each locus appears in at least one block of $\mathbf{N}$. A side effect of these conditions is that the number of partial landscapes is at least linear in L but also at most $\mathcal{O}(L^k)$.

The NK model as introduced so far is more general than in the original. In order to recover the **classical NK structures** I require additionally that there is a bijection between loci and NK blocks, such that the NK block belonging to a locus contains the locus itself, i.e. the NK blocks can be indexed by loci, such that $l \in D_l$. This automatically fixes the number of NK blocks to exactly $\#\mathbf{N} = L$.

Structures of this kind can be alternatively represented as simple directed graphs instead of hypergraphs. The **simplified NK structure graph** is the simple directed graph over loci with arrows from l to m if $l \in D_m$. Usually I will represent NK structures in this way if possible. Note that, while the hypergraph representation is up to edge labels unique, the simplified representation is generally non-unique because there may be multiple bijections between loci and NK blocks.

4.3.1. Local boundedness

My results will be influenced by the topology of the NK structure hypergraph. In particular I need a property I call **local boundedness**. It roughly states that the number of close neighbors to nodes is not diverging to infinity with the number of loci.

Consider the ball or radius r around a locus l in the NK structure hypergraph and let its size, i.e. the number of loci it contains, be $B_r(l)$. Now suppose we choose a locus uniformly from all L possible choices. I then let B_r be the random variable giving the size of the r-ball around this random locus. I will denote the mean with respect to such a uniform locus choice $\mathbb{E}_{\text{loc}}[\cdot]$ and the mean with respect to a realization of the NK structure for given L by $\mathbb{E}_{\text{NK}}[\cdot]$. These have to be distinguished by the mean with respect to fitness value realizations $\mathbb{E}_{\text{f}}[\cdot]$. Similarly I distinguish probabilities with respect to each of these random choices by $\mathbb{P}_{\text{loc}}[\cdot]$, $\mathbb{P}_{\text{NK}}[\cdot]$ and $\mathbb{P}_{\text{f}}[\cdot]$. The mean/probability with respect to both structure and fitness choices will be simply denoted $\mathbb{E}[\cdot]$ and $\mathbb{P}[\cdot]$.

I say that a NK structure is **(almost surely) r-bounded everywhere** if there is a $n \in \mathbb{N}$, such that

$$\lim_{L\to\infty} \mathbb{P}_{\text{NK}}\left[\max_{l\in\mathcal{L}} B_r(l) \leq n\right] = 1$$

, i.e. if there are for sufficiently large L no loci with r-balls larger than n almost surely.

Conversely I say that the structure is **(almost surely) r-bounded nowhere** or **(almost surely) r-unbounded everywhere** if for all $n \in \mathbb{N}$:

$$\lim_{L\to\infty} \mathbb{P}_{\text{NK}}\left[\min_{l\in\mathcal{L}} B_r(l) > n\right] = 1$$

As a slightly weaker condition than the r-boundedness everywhere I say that a NK structure is **(almost surely) r-bounded in moments** if for all $s \in \mathbb{N}$, there are $c_{r,s} \in \mathbb{R}$, such that:

$$\lim_{L\to\infty} \mathbb{P}_{\text{NK}}\left[\mathbb{E}_{\text{loc}}\left[B_r^s\right] < c_{r,s}\right] = 1$$

for all $s \in \mathbb{N}$, i.e. if asymptotically all moments of the size distribution of r-balls

are bounded for almost all structure realizations.

Even weaker I say that the NK structure is **(almost surely) r-bounded in mean** if the condition above holds for $s = 1$ specifically.

If one of these properties holds for every $r \in \mathbb{N}$ I replace r-bounded by **∞-bounded**.

I will typically omit the phrase "almost surely". It is assumed implicitly.

Trivially every structure is surely 0-bounded everywhere. $B_1 - 1$ is simply the degree distribution in the primal graph of the NK structure hypergraph. Therefore the structure is 1-bounded in moments if the degree distribution is bounded in all moments and 1-bounded in mean if the mean degree is bounded. The latter is especially the case for classical NK structures at constant k, because each of the L NK block can induce at most $\binom{k}{2}$ edges in the primal graph, so that the mean degree must be bounded by $\binom{k}{2}$, too. 2-boundedness in mean does however not follow automatically and neither does 1-boundedness in moments. It is even possible to construct a fixed-k classical NK structure which is 1-bounded in mean but 2-bounded nowhere, as I will show later.

Nonetheless 1-boundedness everywhere implies ∞-boundedness everywhere, because if n is the largest 1-ball around any locus for large enough L the size or r-balls can be at most n^r and thus bounded in L.

Similarly it can be seen that 1-boundedness in moments implies ∞-boundedness in moments. I show this by induction over r. Consider the size of an $r+1$-ball around a uniformly chosen locus l. Its size is at most the total size of all r-balls of l's nearest neighbors, therefore:

$$\begin{aligned}(B_{r+1}(l))^s &\leq \left(\sum_{d(l,m)=1} B_r(m)\right)^s \leq (B_1(l))^{s-1} \sum_{d(l,m)=1} (B_r(m))^s \\ &\leq (B_1(l))^{2s-2} + \sum_{d(l,m)=1} (B_r(m))^{2s}\end{aligned}$$

and taking the mean with respect to loci over both sides

$$\mathbb{E}_{\text{loc}}\left[B_{r+1}^s\right] \leq \mathbb{E}_{\text{loc}}\left[B_1^{2s-2}\right] + \frac{1}{L}\sum_{l\in\mathcal{L}}\sum_{d(l,m)=1} (B_r(m))^{2s}$$

In the right-hand sum every $B_r(m)$ for any m appears exactly as often as the degree of m. Therefore:

$$\mathbb{E}_{\text{loc}}\left[B_{r+1}^s\right] \leq \mathbb{E}_{\text{loc}}\left[B_1^{2s-2}\right] + \mathbb{E}_{\text{loc}}\left[B_1 B_r^{2s}\right]$$

. By the Cauchy-Schwarz inequality for the mean:

$$\mathbb{E}_{\text{loc}}\left[B_{r+1}^s\right] \leq \mathbb{E}_{\text{loc}}\left[B_1^{2s-2}\right] + \sqrt{\mathbb{E}_{\text{loc}}\left[B_1^2\right]\mathbb{E}_{\text{loc}}\left[B_r^{4s}\right]} < c_{1,2s-2} + \sqrt{c_{1,s}c_{r,4s}} =: c_{r+1,s}$$

where the last inequality holds almost surely by induction assumption because all the means are bounded by the constants with probability 1. Thus almost surely $\mathbb{E}_{\text{loc}}\left[B_{r+1}^s\right]$ is bounded by $c_{r+1,s}$ completing the proof.

The degree distribution is therefore a good descriptor for local boundedness. Typical choices for the interaction scheme will all be ∞-bounded in moments at constant k. As contrast to these structures I will consider the star neighborhood, introduced later, which is 2-bounded nowhere.

Between these two extreme cases there are still some possible interpolations, however in order to make calculations simpler to follow I will restrict statements to these two cases.

4.3.2. Adjacent neighborhood (AN)

The adjacent neighborhood structure is defined by

$$D_{l_i} = \{l_{i+j \mod L} \mid j = 0 \ldots k-1\} \tag{3}$$

Loci are organized in a circular structure with k-nearest neighbor interactions. The corresponding structure graph looks like a circle with radius proportional to L and a "thickness" proportional to k. In the AN structure at constant k the number of degrees is identical for all loci and non-increasing in L. Therefore the AN structure is ∞-bounded in moments.

4.3.3. Block neighborhood (BN)

In the block neighborhood structure the set off loci is partitioned into $\frac{L}{k}$ disjoint subsets of size k and each of these subsets is used as an NK block with multiplicity k. The corresponding simplified structure graph consists of complete components over the blocks. By construction this model consists of several independent partial landscapes making it easier to study than the other interaction structures and many properties are already known.

The adaptive walk dynamics effectively decompose into independent walks on the blocks. The number of optima is just a multiple of the number of optima on each block and the number of accessible paths between any two genotypes is also a multiple of the number of (sub-)paths on each of the components times a combinatorial factor related to the independent order in which these subpaths are mutated. In particular it follows that, at fixed k, the mean number of accessible paths from the antipodal to the global optimum (with or without backsteps) is growing faster than exponential while the probability to find at least one such path is still decreasing exponentially.[41,45]

As for the AN model, the BN model at fixed k has fixed degree of $k-1$ for all loci for all L and therefore it is ∞-bounded in moments.

4.3.4. Random neighborhood (RN)

The random neighbor structure is chosen uniformly from all classical structures or equivalently for each D_l the chosen loci are l and $k-1$ uniformly chosen other ones.

The NK structure has a distinctly different look for $k=2$ and $k \geq 3$. For $k=2$ every node has to have exactly one in-coming neighbor and it is chosen uniformly from all other loci. This implies that starting from a random locus l_1 one can move along its in-degrees to obtain a reversed direction chain $l_1, l_2, l_3, \ldots$, which can only end by forming a cycle, i.e. by finding at some point a locus as in-degree that was already visited before. The cycle found in this way is certainly larger than constant size in L and has to be the only cycle in the component, because the remaining loci must be attached to the cycle via their only in-degree. Thus the remaining loci are attached to cycle nodes as trees directed towards their roots on the unique cycle.

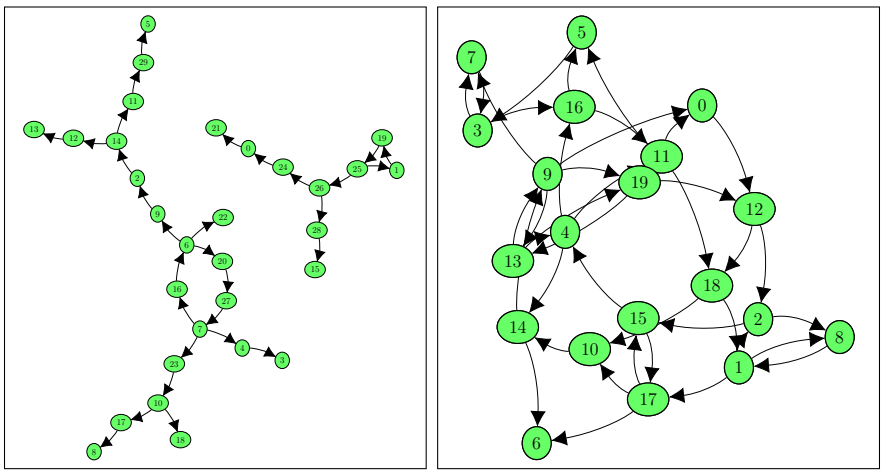

Figure 6: *Typical realizations of the RN simplified structure graph for $L = 40$,$k = 2$ (left) and $k = 3$ (right). For $k = 2$ multiple components are found, each containing one cycle at its center (5-cycle in the left one and 3-cycle in the right one). All other loci are attached as trees. For $k = 2$ the graph has no nice structure and is much stronger connected.*

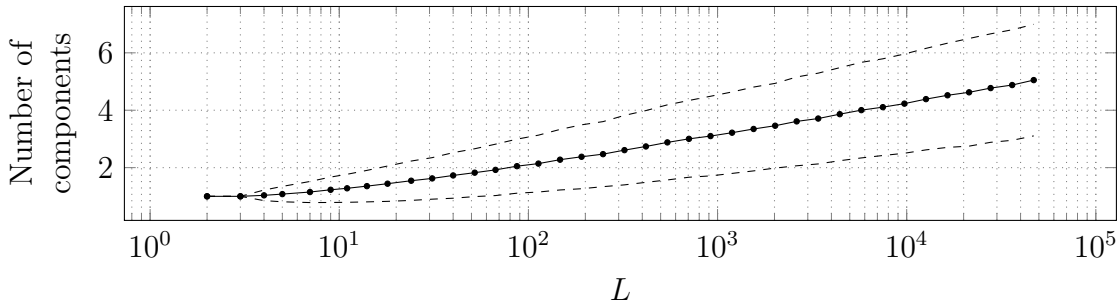

Figure 7: *Simulated mean number of components in the NK structure graph for the RN structure at $k = 2$ (10^5 realizations per data point) with standard deviation (dashed lines). $k = 2$ (and trivially $k = 1$) are the only values of k, for which the number of components is increasing. In fact there is asymptotically exactly one component for every larger k and I was unable to find counterexamples by random sampling, except for L close to $2k$, the minimal at which 2 components can exist.*

For $k \geq 3$ however there may be multiple cycles per component because branching along the in-degrees is possible. Consequently the NK structure is much stronger connected and it can be seen that simulation results indicate that there is actually only one giant component spanning the whole graph.

For any k, the limit distribution of the in- and out-degrees for a locus $l \in \mathcal{L}$ in the simplified structure graph can be calculated. The in-degree is by definition $k-1$. The out-degree depends on the choices of members for all NK blocks D_m with $m \neq l$. Each of these $L-1$ blocks contains l with a probability $\frac{k-1}{L-1}$ Therefore the out-degree is binomial distributed with $L-1$ trials with trial probability $\frac{k-1}{L-1}$ or in the limit of large L Poisson distributed with mean $k-1$. The total degree in the simplified structure graph is therefore k plus a Poisson distributed random variable with mean $k-1$, i.e. a shifted Poisson distribution. The degree in the primal graph of the full structure hypergraph is larger, because every out-degree, belonging to a $m \in \mathcal{L}$ with $l \in D_m$ implies additional $k-2$ degrees. Note that the probability that there is a $m' \in D_m$, such that also $l \in D_{m'}$ or $m' \in D_l$, i.e. that m' is also a neighbor of l is tending to zero as $\mathcal{O}(L^{-1})$ and therefore the additional degrees relative to the simplified structure graph are all asymptotically unique, implying that the degree distribution in the primal graph of the structure hypergraph is asymptotically $(k-1)(\xi_{k-1}+1)$, where ξ_{k-1} is a Poisson random variable with mean $k-1$. Consequently $B_1(l)$ is distributed like $k+(k-1)\xi_{k-1}$ for every locus $l \in \mathcal{L}$ individually.

The degree distributions of any (randomly chosen) constant-size set of loci are asymptotically independent, because given the degree distribution of a finite set of loci, only a finite set of NK blocks are restricted in any way, which can not affect an additional locus degree choice in the $L \to \infty$ limit.

This is also true if the set of loci is constraint to be neighbors in some way, as long as the degrees induced in this way are taken into account. This implies that the local topology of the structure graph is that of a tree with branching number (along in- and out-degrees) $k-2+\xi_{k-1}$. Note that the branching number has mean $2k-3$. For $k=2$ this is 1, i.e. seeing this as a branching process the extinction probability is 1 and therefore the size of trees converges in distribution as $L \to \infty$. For $k \geq 3$ the branching number is larger than 1 and every locus has at least 1 branch. Therefore the extinction probability is zero and the tree, and therefore the

component, will span an asymptotically non-zero fraction of the structure graph, i.e. there will be a giant component.

The RN structure is ∞-bounded in moments: If it was not ∞-bounded in moments, then there would be a $s \in \mathbb{N}$ and a function $t(L) = \omega(1)$, such that

$$\limsup_{L\to\infty} \mathbb{P}_{\mathrm{NK}} \left[\mathbb{E}_{\mathrm{loc}} \left[B_1^s\right] > t(L)\right] = \alpha > 0$$

Then also, because B_1 is non-negative:

$$\limsup_{L\to\infty} \mathbb{E}_{\mathrm{NK}} \left[\mathbb{E}_{\mathrm{loc}} \left[B_1^s\right]\right] \geq \alpha t(L) = \omega(L)$$

i.e. $\mathbb{E}_{\mathrm{NK}} \left[\mathbb{E}_{\mathrm{loc}} \left[B_1^s\right]\right]$ would need to diverge at least in limit superior. However this is not the case, because:

$$\mathbb{E}_{\mathrm{NK}} \left[\mathbb{E}_{\mathrm{loc}} \left[B_1^s\right]\right] = \mathbb{E}_{\mathrm{loc}} \left[\mathbb{E}_{\mathrm{NK}} \left[B_1^s\right]\right]$$

As shown $\mathbb{E}_{\mathrm{NK}} \left[B_1^s\right]$ converges to the s-th moment of the shifted Poisson distribution and is therefore asymptotically bounded for every locus individually and so is then $\mathbb{E}_{\mathrm{NK}} \left[\mathbb{E}_{\mathrm{loc}} \left[B_1^s\right]\right]$. Therefore the structure is ∞-bounded in moments.

Also note that strictly speaking the RN structure does not satisfy the condition of bounded multiplicity of NK blocks. However the probability for any pair of loci to appear together in more than two partial landscapes decreases to zero in L, so that the condition is still satisfied almost surely.

4.3.5. Star neighborhood (SN)

All previous NK structures shared the property of ∞-boundedness in moments, while the following by construction does not. I construct the star neighborhood as a neighborhood structure which is explicitly r-unbounded everywhere for every $r \geq 2$, in drastic contrast to the other three structures introduced so far. Mark $k-1$ loci as **center loci** $c_1, \ldots, c_{k-1}$ and all other loci as **ray loci**. Then define the NK blocks by $D_l = \{l, c_1, \ldots, c_{k-1}\}$. Note that for the center loci this set is too small and therefore, only for center loci, another uniformly chosen locus is added (or the same locus is used for all center-associated building blocks, it doesn't really

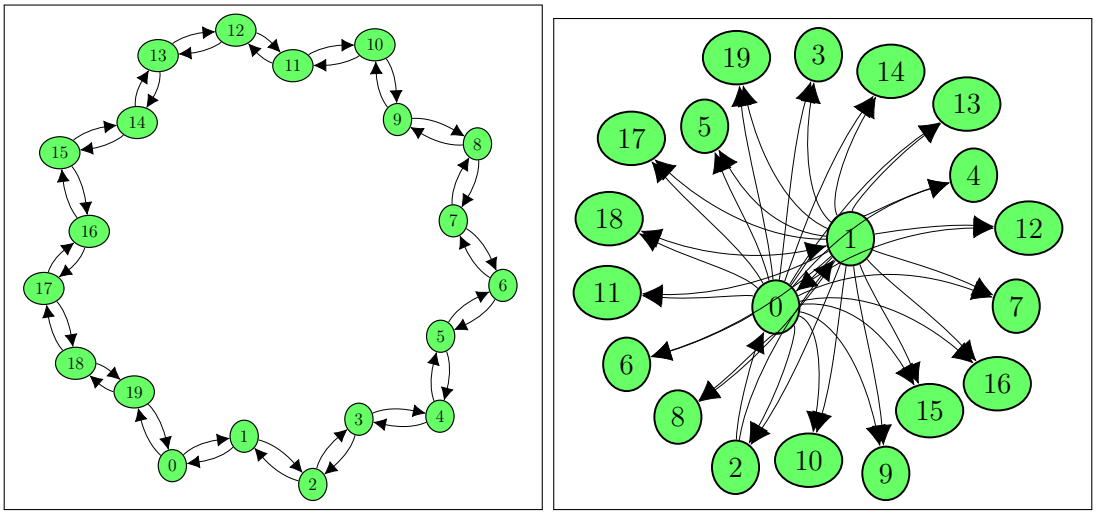

Figure 8: *Simplified structure graph for the AN (left) and SN (right) structures at $L = 20$ and $k = 3$. Loci in the AN structure are far away from each other, while it is always possible to reach all loci via two steps over the center in the SN model.*

matter).

All loci in the star neighborhood are reachable from each other by one step through the center (assuming $k \geq 2$) and so every locus is contained in 2-balls. Consequently the SN structure is 2-bounded nowhere. It is still 1-bounded in mean though, as all classical NK structures with finite k have to be. Only the constant number of center loci does not have bounded 1-balls.

One could interpret the star neighborhood as the extreme case of a regulatory site (center) affecting the expression of a large group of other proteins (rays).

4.3.6. Previous results on the NK model

Local optima have been studied rigorously for the AN structure. Asymptotically at constant k the mean number of optima will be exponential in the number of loci with the exponential constant of proportionality depending on k and the fitness distribution.[13] The same holds for the block model, where a genotype is a local optimum if and only if it is a local optimum on all projections onto NK blocks, and therefore the mean number of local optima is exactly an exponential in L with the base given by the number of optima in the HoC model with k loci.

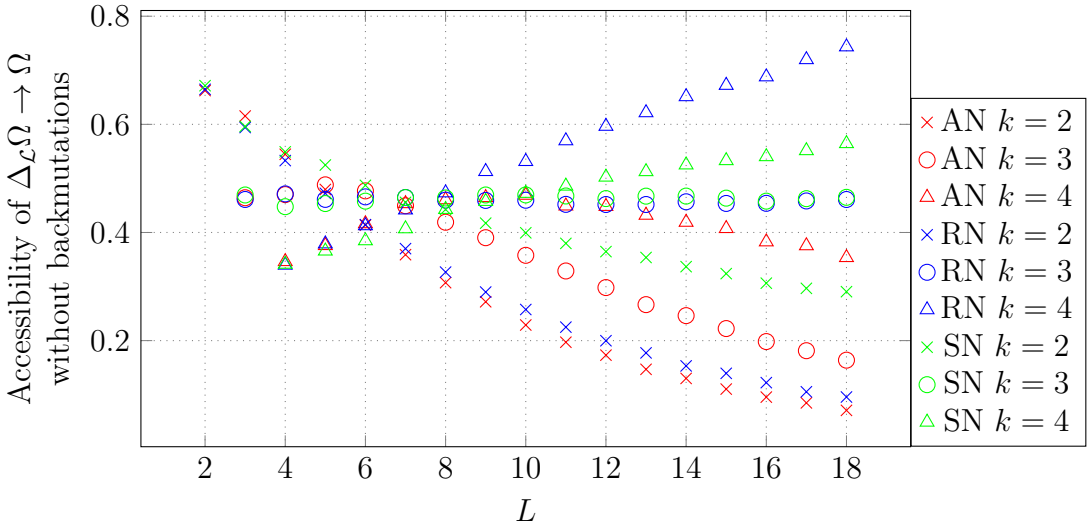

Figure 9: *Simulated accessibility of the global optimum from its antipodal without backmutations in the RN, AN and SN structure NK models. (Gaussian fitness and* 10^5 *realizations per data point.) For* $k = 2$ *accessibility drops with* L *for all three structures, for* $k = 3$ *however the RN and SN accessibility accessibility seem to be rather constant. For* $k = 4$ *both even seem to be increasing in* L*, while for the AN structure there is still a (albeit slower) decrease. I will show later that actually for* $L \to \infty$*, all curves for AN and RN will exponentially decrease to zero, while for the SN structure the probability will converge to a non-zero value for all* $k \geq 2$*.*

Accessible paths from the antipodal to the global optimum, mainly without backsteps, have been analyzed via simulation for constant k and constant $\frac{k}{L}$.[15,45] These simulation results suggest that, while the mean number of paths without backsteps increases faster than exponential at constant $k \geq 2$ for AN, RN and BN, the probability to find at least one path behaves distinctively different. For the AN and RN case with $k = 2$ it seems to be decreasing in L, while for $k = 3$ in the RN case it is almost constant in the simulated range of L. For $k \geq 4$ in both the RN and AN case it seems to be increasing in L. For the BN model in contrast the decrease is always exponential to zero, because the probability to find any path is the product of the probabilities to find any path on each independent NK block.[45]

4.3.7. Lower bound on accessibility in the NK model

An exponentially decreasing lower bound on the accessibility of genotypes in maximal distance in the NK model can be derived for all fitness distributions and NK structures. Suppose there are $\dim \mathbf{N}$ different NK blocks and a random genotype is chosen as starting point with a random permutation of loci defining a path without backsteps from the starting genotype to its antipodal. Each partial landscape has then a probability of $\frac{1}{(k+1)!}$ to have monotone fitness increase along this path because they are of HoC-type and only k of the steps in the path modify it. If all partial landscapes share this property, then surely the path is also accessible on the whole landscape. Therefore a lower bound on the probability that a random path without backsteps of length L and starting from a random genotype is accessible is $e^{-\dim \mathbf{N} \ln((k+1)!)}$ or for classical NK structures $e^{-L \ln((k+1)!)}$. It seems reasonable to assume that this is also a bound if the destination of the path is conditioned to be the global optimum, because the only change on the level of partial landscapes would be a slight bias for the path's projection on the partial landscape to also end in a higher fitness state. In fact this result also implies that the mean number of paths without backsteps for a classical NK structure will be at least $L! e^{-L \ln((k+1)!}$ and therefore growing superexponentially like $e^{L \ln(L) + \mathcal{O}(L)}$ at constant k. The leading order of the exponent is unchanged, even if k grows slowly, not faster than $(\ln(L))^{1-\epsilon}$ for some $\epsilon > 0$.

4.4. Rough-Mount-Fuji (RMF) model

Another possible way to interpolate between the linear and the HoC model is given by the Rough-Mount-Fuji (RMF) model.[32] It is conceptually a bit simpler than the NK model and does have a smaller parameter space, only depending on the ruggedness interpolating parameter $c \in \mathbb{R}^+$. The RMF landscape is the sum of a HoC landscape with fitness variance of 1 and a linear model with all its fitness effects set to $\tilde{f}_l = c$. The model is by construction an interpolation between the HoC and linear model. For $c = 0$ the linear model does not contribute and the resulting landscape is pure HoC. For $c \to \infty$, the mutational effects of the linear part will surely eventually outscore the HoC contributions of fixed distribution.

In the Fourier space this model is, due to linearity of the Fourier transform, also

simply a sum of the HoC and linear contributions. This implies that all properties of the HoC Fourier components apply, only that the order one (linear) terms will be increased by the constant c.

Note that often the model is defined slightly differently. Usually it is assumed that of the additive component is rotated such that its global optimum coincides with the global optimum of the HoC contribution. The sign epistasis properties considered later will not dependent on this distinction.

5. (Sign) epistasis

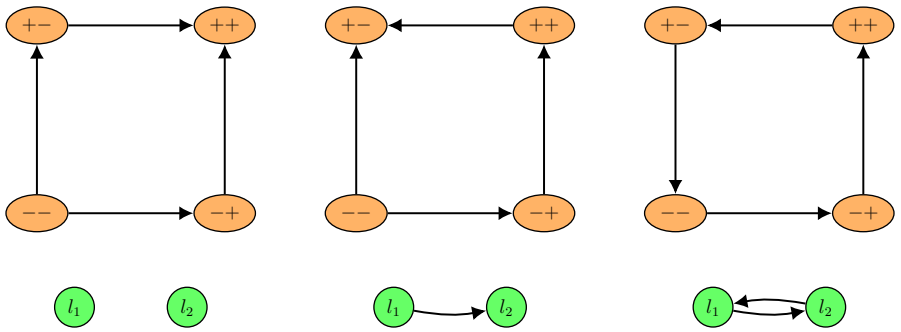

Figure 10: *Top: Example $\{l_1, l_2\}$-squares with, from left to right, no sign epistasis, sign epistasis dependence only of l_2 on l_1, but not the other way around, and reciprocal sign epistasis. Bottom: Corresponding subgraph of the sign epistasis graph induced by $\{l_1, l_2\}$.*

In the linear model each locus has a well-defined contribution to the overall fitness. If the effect of a mutation at one genotype is known, then it is also known on all other background genotypes, corresponding to the L degrees of freedom of a realization of the model. The HoC model however has many more degrees of freedom to realize, namely the maximum possible, 2^L. Therefore the effect of a single mutation cannot be fully described by just the effect on one background, but rather the effect on most backgrounds must be taken into account explicitly.

Epistasis is the effect of a locus' fitness contribution depending not only on the state of the locus itself, but on the rest of the background genotype as well. A particular simple form of this epistatic effect is described by **pairwise epistasis**. Here we use the following definition for the epistatic effect of a locus l's mutation

on the fitness change due to a point mutation on l:

$$\tilde{E}_{lm}(g) = \Delta_l F(\Delta_m g) - \Delta_l F(g)$$

If the value of $\tilde{E}_{lm}(g)$ is 0, a mutation on m will not change the fitness effect of an immediately following mutation on l. If it is larger than 0, then the effect of Δ_l will be larger if m is mutated immediately beforehand, while a value smaller than 0 indicates the opposite. The following symmetry for the object $\tilde{E}$ holds

$$\tilde{E}_{lm}(g) = \tilde{E}_{ml}(g)$$

Also note that

$$\tilde{E}_{lm} = \Delta_m \Delta_l F$$

, i.e. it is the discrete analog to the Hessian.

If a locus l affects a locus m on some genotype we may expect it to also affect the locus on other genotypes. Especially in the models we consider here this is almost surely true and thus we consider the reduction of $\tilde{E}_{lm}(g)$ to a value E_{lm} independent of the background genotype, which is 1 if there is at least one g with $\tilde{E}_{lm}(g) \neq 0$ and 0 otherwise.

E_{lm} can be viewed as a binary-valued symmetric matrix and as such it can also be treated as the adjacency matrix of an undirected graph, which will be referred to as the **epistasis graph**. Since we assume that there are no neutral mutations, $E_{ll} = 1$ for all loci l. These trivial loops will be ignored in the plots of the graph, making it simple.

In terms of the Fourier transform $E_{lm} = 1$ if there is at least one $\tilde{g} \subseteq \mathcal{L}$ with $l \in \tilde{g}$ and $m \in \tilde{g}$, such that $\tilde{F}_{\tilde{g}} \neq 0$.

In the HoC and RMF models, almost surely, the epistasis graph is complete, because the probability of two independent fitness differences to be equal is zero for absolutely continuously distributed fitness values.

In the generalized NK model, the epistasis graph is limited by the NK structure. Two loci l and m can only be epistatic if there is a NK block with $\{l, m\} \subseteq D$, because otherwise all Fourier coefficients resulting in epistasis are zero. The epistasis graph is then the primal graph of the NK structure hypergraph.

Every undirected simple graph is a valid epistasis graph as can be seen easily by defining a NK model with $k = 2$ and NK blocks corresponding to the edges of the epistasis graph.

While there are many interesting effects due to epistasis[8,42], a certain subset of epistatic interactions is particularly interesting here. These are so-called **sign epistatic interactions**.[14,55] Interactions which do not only let the magnitude of a mutations fitness effect be dependent on other loci, but also the sign of fitness. Sign epistasis is important, because given strong enough selection pressure populations are very unlikely to fixate mutations resulting in a fitness decrease, while positive mutations have a high chance of fixation. In the limit case of very large selection pressure movement towards deleterious mutations becomes impossible. As in the previous subsection, we will use a simple notion of pairwise sign epistasis defined as[1]:

$$\tilde{S}_{lm}(g) = \mathrm{sgn}\Delta_l F(\Delta_m g) - \mathrm{sgn}\Delta_l F(g)$$

If $\tilde{S}_{lm}(g)$ is zero, although the magnitude of a mutation on l might be changed by one on m, a beneficial mutation will stay beneficial and a deleterious will stay deleterious. A value of $\tilde{S}_{lm}(g) = 2$ indicates that the mutation Δ_l is deleterious at g, but beneficial at $\Delta_m g$, while a value of -2 indicates the opposite. Analogues to the process in the previous subsection we introduce a reduced version of $\tilde{S}_{lm}(g)$, i.e. S_{lm}, independent of g, such that $S_{lm} = 1$ if $\tilde{S}_{lm}(g) \neq 0$ for at least one g and 0 otherwise. The property S_{lm} is again a binary matrix, however not a symmetric one as can be seen from the definition. Thus a locus l might depend on the state of locus m sign epistatically, while m does not depend sign epistatically on l. Anyway if $S_{lm} = 1$ then also $E_{lm} = 1$, i.e. sign epistasis is a subset of general epistasis. This implies by symmetry of E_{lm}, that even if $S_{lm} = 1$ but $S_{ml} = 0$, then still $E_{ml} = 1$.

The appropriate representation for the asymmetric binary matrix is a directed graph, again if the trivial loops are ignored a simple one. This **sign epistasis graph** reduces the information content of the complex fitness graph to a small-scale representation which, given that there is a clear underlying genetic structure on the genome, conveys the interaction between loci much more clearly. It is also more

[1] We leave the zero case of the sign function undefined as the models used almost never have neutral mutations.

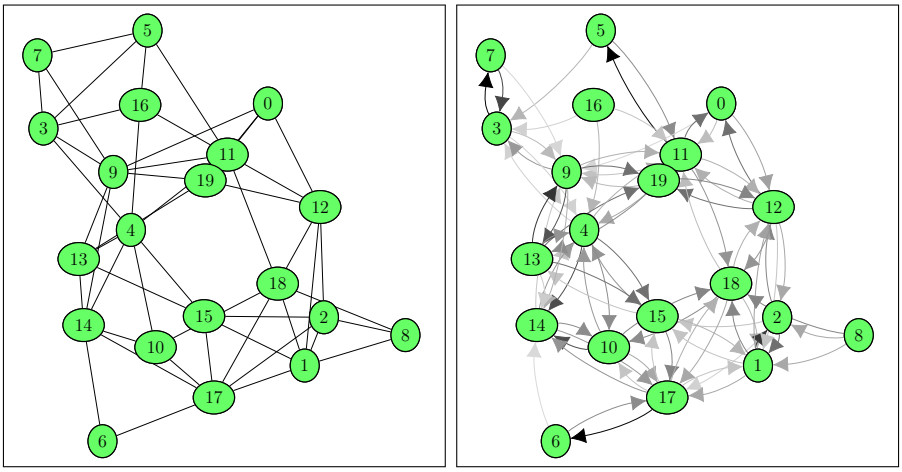

Figure 11: *Example of epistasis graph (left) and (weighted) sign epistasis graph for the NK model with RN structure at* $L = 20$ *and* $k = 3$*. Every edge in the epistasis graph is a triangle, because this is a* $k = 3$ *NK model. The sign epistasis graph is almost the same as the epistasis graph with bidirectional arrows, however once in a while arrows are missing. By chance of the chosen fitness values, certain loci's mutation signs are independent of loci which are otherwise sharing fitness values partially. For example there is no sign epistasis between loci 7 and 5 although they are epistatic. The sign epistasis weight is indicated by the grayscale of arrows. Most sign epistasis is present on less than half the backgrounds, but some are even present on all backgrounds, e.g.* $11 \to 5$ *and* $17 \to 6$ *and* $3 \to 7$*, but not* $7 \to 3$*.*

relevant than the epistasis graph, because low amplitude noise does not usually contribute to it as much.

The simple undirected graph underlying the sign epistasis graph is a subgraph of the epistasis graph, because sign epistasis is a special kind of epistasis as seen.

Defining the sign epistasis graph to have arrows even if there is sign epistasis only on one background genotype removes a lot of information one might consider important. Also empirical landscapes will generally not be perfectly measured and spurious sign epistasis may be the result of measurement errors. Therefore a more general definition for the sign epistasis network would be that of a weighted sign epistasis graph, in which each arrow is additionally assigned a rational number between 0 and 1, determining the ratio of background genotypes at which the sign epistasis is present. I call this the **weighted sign epistasis graph**. Arrows with weight 0 are assumed to not exist.

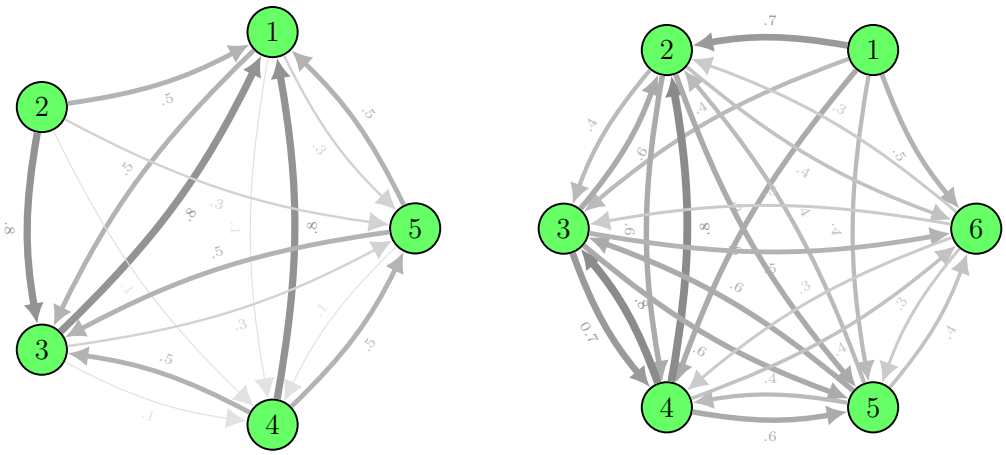

Figure 12: *Two examples of the weighted sign epistasis graphs for empirical landscapes. Left: Based on data set of five point mutations in β-lactamase jointly increasing resistance to antibiotics drastically by Weinreich et al.[56] Right: Based on data of six individually deleterious mutations in different pathways of* Saccharomyces cerevisiae *by Hall et al.[21] Arrow thickness and opacity is determined by the arrow weights. The weights are also given rounded to one decimal as arrow label. No loci are sign epistatic everywhere, but some combinations are never sign epistatic. The graphs do neither look like House-of-Cards models, which are almost completely uniformly weighted* 0.5*, nor RMF models, which are almost completely uniformly weighted* < 0.5*, nor NK models with* $k = 1$ *or* $k = 2$*, which would have much less arrows and larger variance in arrow weights. A better model for these empirical graphs seem to be mixtures of HoC or RMF models with NK models.*

5.1. Reciprocal sign epistasis

Sign epistasis is generally not symmetric. If however at a background genotype g mutations on l are sign epistatically dependent on m and the other way around, then one speaks of **reciprocal sign epistasis**. Reciprocal sign epistasis is known to be a necessary condition for emergence of multiple local optima and thus seems to be important for the ruggedness of landscapes.[28,44] It also imposes a limitation on accessible paths, because it is never possible to mutate both loci immediately one after another with monotone fitness increase on the background they are reciprocal on. This can be seen by looking at the orientation of the fitness graph $\{m, l\}$-slice along g. In this square both pairs of parallel arrows need to be oriented opposite to one another, leaving no way to cross from any corner to the antipodal one. However (local) reciprocal sign epistasis still leaves accessible paths switching both loci with intermediate mutations on other loci.

5.2. Global sign epistasis

If a locus l is sign epistatically dependent on m everywhere, i.e. for every background genotype, I speak of **global sign epistasis**. This is equal to the a weight of 1 in the weighted sign epistasis graph. A special kind of global sign epistasis is **sole (global) sign epistasis**, which I define as global sign epistasis such that mutation signs on l depend on all backgrounds on m, but only on m, i.e. there is no other locus affecting the mutation signs on l. These two variants are different in that there are two possible orientations of sign epistasis on a square (either $\Delta_l F(g)$ is positive or it is negative). For simple global sign epistasis the orientation may dependent on the background, but sole global sign epistasis explicitly excludes this by assuming that no other locus except m affects $\mathrm{sgn}\Delta_l F$ at all.

Global sign epistasis implies that mutations on l and m immediately one after another are always only accessible in one order at most. Sole global sign epistasis takes this limitation further, such that this order is the same at all backgrounds and independent of additional mutations in between. It implies a strict order in accessible paths and can be recognized by sign epistasis arrows with weight 1 and no incoming arrows to l except that from m.

Global reciprocal sign epistasis (GRSE) is reciprocal sign epistasis present

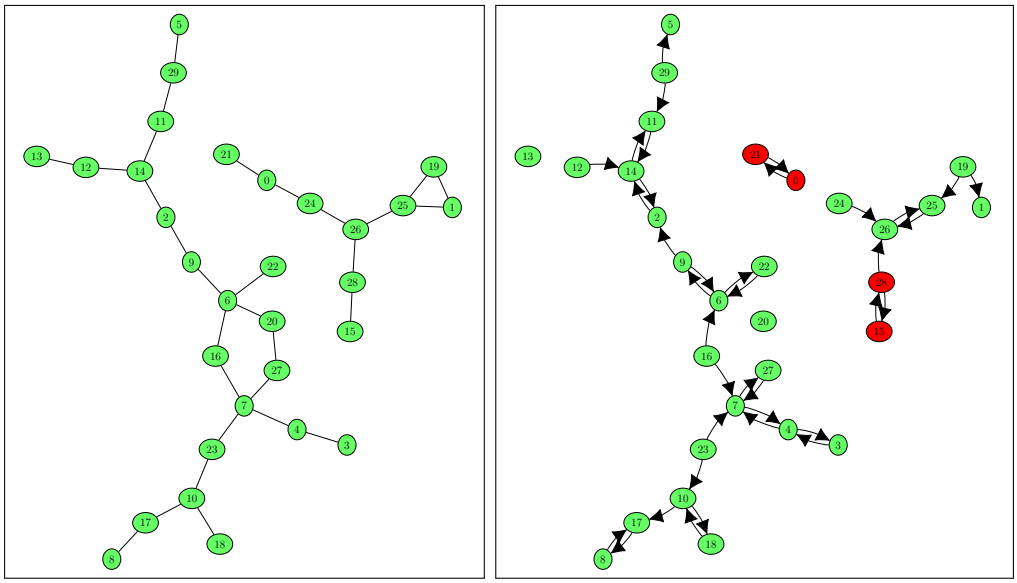

Figure 13: *Example of epistasis graph (left) and (unweighted) sign epistasis graph (right) for the NK model with RN structure at $L = 30$, $k = 2$ and gaussian fitness. Pairs of loci with global reciprocal sign epistasis are marked in red. They can be identified by bidirectional arrows between them, but no other incoming arrows. Further outgoing arrows are allowed, as can be seen here in the 15/28 pair. In the case of the 0/21 pair we have separable global sign epistasis, i.e. the global reciprocal sign epistasis interaction also forms a weak component.*

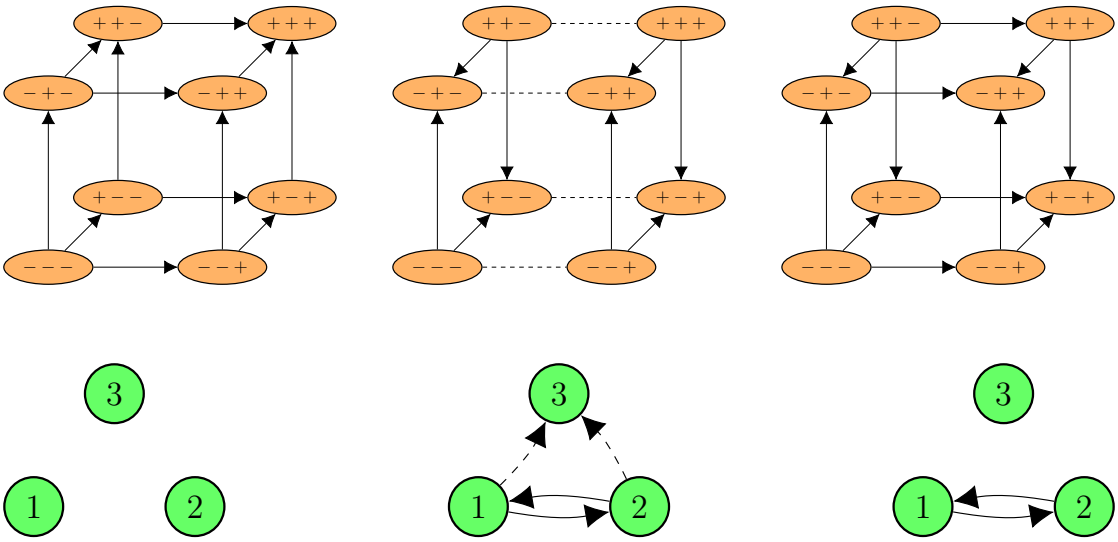

Figure 14: *Upper row: From left to right the fitness graphs associated with a non-epistatic, a globally reciprocal and a separable reciprocal landscape of three loci. The orientation of fitness increases in the non-epistatic model does not depend on the position in the other dimensions for any locus. In the middle mutations Δ_1 and Δ_2 are reciprocal in the foreground, as well as the background (i.e. for both $\sigma_3 = 0$ and $\sigma_3 = 1$). The orientation of mutations Δ_3 is not constrained by locus l_1 and l_2 being globally reciprocal, as indicated by the missing arrow heads. In the case of separable reciprocal epistasis (right) however the orientation of Δ_3 arrows needs to be the same for all positions on the hypercube, as it is completely separated from Δ_1 and Δ_3. Lower row: Sign epistasis graphs associated with the fitness graphs. Dashed arrows may or may not be present.*

at all backgrounds. Here again two orientations of reciprocal sign epistasis on a square are possible (again fixed by choosing the orientation of any one edge). But I directly assume GRSE is sole reciprocal sign epistasis, i.e. that there are no other loci affecting mutation signs on Δ_l and Δ_m. In the sign epistasis graph GRSE can be recognized as bidirectional arrows with weight 1 and no further incoming arrows to the two loci. GRSE poses much stronger restrictions than the other variants mentioned. It implies that loci l and m may not be mutated both without the path becoming inaccessible. In contrast to local reciprocal sign epistasis this holds strictly. Even with other intermediate mutations it is impossible to cross both l and m. This is not only true for short, but also long paths with back-mutations.

Therefore there are no accessible paths of maximal length on landscapes with at least one locus pair with GRSE, neither to the global optimum, nor any other genotype. The maximal distance crossed by accessible paths becomes $L - 2Z_{\mathrm{GRSE}}$, where Z_{GRSE} is the number of GRSE interactions. Note that each locus can only be present in one GRSE interaction, because it is by definition of GRSE only sign epistatic with its unique interaction partner. Furthermore, under those distance restrictions, any walker starting at any landscape point can only explore at most a fraction $\frac{3}{4}^{Z_{\mathrm{GRSE}}}$ of the genotype space. This is in particular a restriction on the size of the basins of attraction of any kind for local optima.

A further special variant of global reciprocal sign epistasis I call **separable global sign epistasis (SGRSE)**. It is global reciprocal sign epistasis, such that not only l and m are not sign epistatically dependent on any other third locus, but that no third locus is sign epistatically dependent on l or m. In the sign epistasis graph this corresponds to two loci with bidirectional arrows but no further arrows in- or out-going. The two loci then form a weak component of the sign epistasis graph.

Separable global sign epistasis has further effects on the landscape structure. Suppose g is a local optimum of the landscape. Then by mutating both l and m, one arrives at a new local optimum, because each one of the mutations switches the sign of mutations on l and m. Switching the sign twice again gives a state in which all mutations would have negative effect, i.e. a local optimum. Other loci's mutation signs are unaffected. Distance 2 is also the smallest distance that two local optima can be separated. Therefore local optima come in clusters of at least

$2^{Z_{\text{SGRSE}}}$ local optima connected by distance-2 jumps.

5.3. Probability of sign epistasis

Suppose l and m are different loci and h is a genotype with $l, m \notin h$ and consider the $\{l, m\}$-square along some background h in the fitness graph. These squares are the elementary setting to consider sign epistasis. On the square arrows, corresponding to the direction of fitness increase, may be set in three distinct ways. Either both pairs of parallel arrows are each oriented in the same direction, or only one is, or both are. The first case is the one without sign epistasis between l and m at this background, while the second one corresponds to sign epistatic dependence of either l on m or reversed, but not both. The last option is known as **reciprocal sign epistasis**, because l depends sign epistatically on m as well as the reverse.

In real fitness space the choice between these forms is determined by the sign of fitness differences. It is however also possible to look at the Fourier space situation of the square.

The square's Fourier expansion consists of only four terms:

$$\tilde{F}_{\{\}} + \tilde{F}_{\{l\}} g_l + \tilde{F}_{\{m\}} g_m + \tilde{F}_{\{l,m\}} g_l g_m$$

The individual Fourier components are retrieved from the full landscape's one by setting the background h and summing all terms of equal l and m alleles remaining.

A mutation on l has then an effect proportional to:

$$\tilde{F}_{\{l\}} g_l + \tilde{F}_{\{l,m\}} g_l g_m$$

The first term does not depend on g_m and the second term switches signs depending on g_m. Therefore l depends sign epistatically on m at this locus if and only if $|\tilde{F}_{\{l\}}| < |\tilde{F}_{\{l,m\}}|$.

Using this property and some combinatorial arguments one can calculate the probability of sign epistasis at a random genotype between two random mutations. This probability is also the mean weight of the sign epistasis arrow between the two loci.

5.3.1. HoC model

In the HoC model mutations at different genotypes are (except for backmutations) independent and therefore the probability for the sign of a mutation to change by application of a different mutation is just $\frac{1}{2}$. This implies that the mean weight of any arrow in the weighted sign epistasis graph is also $\frac{1}{2}$.

Furthermore, there are 2^{L-2} $\{m, l\}$-squares of the hypercube and all of their fitness values are independent. Consequently the weight of an sign epistasis arrow is the scaled sum of 2^{L-2} i.i.d. symmetric Bernoulli random variables and therefore the marginal distribution of any arrow weight is a scaled symmetric Binomial distribution over 2^{L-2} trials and in the limit of $L \to \infty$, the marginal distribution of sign epistasis edges converges to the deterministic distribution with value $\frac{1}{2}$.

Covariances between arrow weights are also negligible in the large L limit because they have to also converge to zero if the weights' variances converge to zero. Correlation coefficients however are scaled to the variance and therefore they might not vanish. Correlation coefficients between disjoint pairs of mutations must be uncorrelated because slices of the first pair and slices of the second pair share at most one genotype, the fitness of which can be chosen arbitrarily without affecting the likelihood of sign epistasis. The remaining non-disjoint pairs of arrows may generally have finite correlation coefficients and can be calculated from combinatorial arguments by averaging over all rank orders of fitness values on two squares joined in one edge, although the results will not be presented here.

Under these considerations the HoC sign epistasis graph will almost surely be the complete graph for $L \to \infty$ and the weighted sign epistasis graph will in the same limit have weights of exclusively $\frac{1}{2}$ with vanishing variation. Simulation shows that this limit is reached quite fast (as expected due to the exponential increase in terms relevant for the central limit theorem) and already for small system sizes empty edges cannot be found.

In fact the probability of **global sign epistasis**, i.e. sign epistasis between two loci l and m at all backgrounds follows to be $2^{-2^{L-2}}$. The same holds for the probability that a sign epistasis arrow does not exist at all.

The probability of global reciprocal sign epistasis is of course even smaller than that, but can be exactly calculated. For a given square of the fitness graph it can

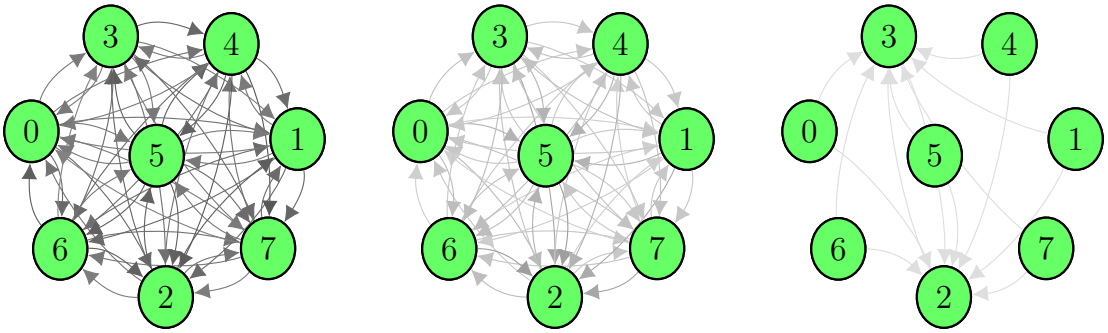

Figure 15: *Weighted sign epistasis in the RMF model with $L = 8$ and standard normal fitness and (from left to right) $c = 0, 2, 4$. For $c = 0$ this is the HoC model, while for large c it approaches the linear model. The arrow weight is very uniform with mean weight $\frac{1}{2}$ for the HoC case and decreasing for larger c. For $c = 4$ already arrows vanish and ultimately for large enough c no sign epistasis is left.*

be seen easily by counting rank orders, that there is a $\frac{1}{3}$ probability of reciprocal sign epistasis if all fitness values are i.i.d. and there are two possible, equally likely, orientations of arrows corresponding to reciprocal sign epistasis. The probability, that the one initially chosen remains on all other $2^{L-2} - 1$ backgrounds, which have completely i.i.d. fitness values, is then $\frac{1}{6}\frac{1}{3}^{2^{L-2}-1}$. Multiplying by the number of locus pairs $\frac{L(L-1)}{2}$ one gets the mean number of sign epistasis interactions $L(L-1)6^{-2^{L-2}}$ and by Markov's inequality this is also an upper bound on the probability that there is at least one GRSE interaction.

Thus I can conclude that global sign epistasis or even GRSE is practically impossible in the HoC model for about $L \geq 6$.

5.3.2. RMF model

Consider a $\{l, m\}$-square at background g. In the RMF model we are starting out with the pure HoC model at $c = 0$. As c is increased only the linear terms in the Fourier expansion of the square increase deterministically. However sign epistasis is present if the magnitude of the quadratic term is larger than the magnitude of the linear terms. Consequently, as the linear contribution increases with c, the probability of sign epistasis y decreases, initially from $\frac{1}{2}$ to 0. If the support of the HoC fitness distribution is bounded, eventually the linear term will become

large enough so that it can never be overcome by the randomized contribution and the sign epistasis graph will surely be empty. For unbounded support this cannot happen for any finite c, but the probability of sign epistasis becomes smaller and smaller. The sign epistasis graph, initially starting from the HoC case being uniformly weighted with $\frac{1}{2}$ first decreases uniformly in mean weight as c increases until edges vanish completely and ultimately only the empty graph is left.

The remaining results from the HoC analysis carry over, because independence between squares on different backgrounds still holds and thus the probability of global sign epistasis will be $y^{2^{L-2}}$ and the probability of reciprocal sign epistasis between any pair of loci will be at most $(L-1)Ly'^{2^{L-2}-1}$ for some $y' < \frac{1}{2}$ monotonically decreasing in c.

Global reciprocal sign epistasis is even considerately less likely than in the HoC model, even for moderate values of c and L.

5.3.3. Generalized NK model

For the generalized NK model I again consider a random $\{m, l\}$-square and its Fourier expansion.

Sign epistasis dependence of l on m is present if $|\tilde{F}_{\{l,m\}}| > |\tilde{F}_{\{l\}}|$. Only partial landscapes on blocks D with $l \in D$ contribute to either (because otherwise they have no non-zero $\tilde{f}_{\{l,m\}}$ or $\tilde{f}_{\{l\}}$) and only D with $\{l, m\} \subseteq D$ contribute to $\tilde{F}_{\{l,m\}}$, let the number of NK blocks only l be α and the number of NK blocks with both l and m be β.

The two relevant Fourier components are obtained from the four fitness values on the square. In particular:

$$\tilde{F}_{\{l,m\}} = \frac{1}{2}\left(F(++) - F(-+) - F(+-) + F(--)\right)$$

and

$$\tilde{F}_{\{l\}} = \frac{1}{2}\left(F(++) - F(-+) + F(+-) - F(--)\right)$$

Contributions to each of these four real space fitness values of different partial landscapes are independent. For the β partial landscapes with $l \in D$ and $m \in D$, the contributions to the four values are i.i.d., but for the α ones with only l, the

contributions to $F(++)$ are equal to $F(+-)$ and the contributions to $F(-+)$ equal to those of $F(--)$. However all contributions are identically distributed with the base fitness distribution with mean 0 and variance σ_f^2.

Therefore, with $Q_{i,j}$ being i.i.d. fitness contributions, the Fourier components can be written:

$$\tilde{F}_{\{l,m\}} = \frac{1}{2}\sum_{i=1}^{\beta}(Q_{1,i} - Q_{2,i} - Q_{3,i} + Q_{4,i})$$

and

$$\tilde{F}_{\{l\}} = \frac{1}{2}\sum_{i=1}^{\beta}(Q_{1,i} - Q_{2,i} + Q_{3,i} - Q_{4,i}) + \frac{1}{2}\sum_{i=1}^{\alpha}(2Q_{5,i} - 2Q_{6,i})$$

If $\alpha = 0$, then the probability that $|\tilde{F}_{\{l,m\}}| > |\tilde{F}_{\{l\}}|$ is $\frac{1}{2}$ as expected because the square is then essentially following the HoC model. For $\alpha \geq 1$ there are additional contributions to $\tilde{F}_{\{l\}}$ which do have mean zero, but increase in the variance of $\tilde{F}_{\{l\}}$ relative to $\tilde{F}_{\{l,m\}}$. In fact one can see that the probability for $|\tilde{F}_{\{l,m\}}| > |\tilde{F}_{\{l\}}|$ decreases monotonically in α. Suppose for $\alpha = 0$ the Fourier components are given and suppose we add the r.v. $\sum_{i=1}^{\alpha}(Q_{5,i} - Q_{6,i})$ to $\tilde{F}_{\{l\}}$, then:

$$\begin{aligned}
&\mathbb{P}\left[\left|\tilde{F}_{\{l,m\}}\right| > \left|\tilde{F}_{\{l\}} + \sum_{i=1}^{\alpha}(Q_{5,i} - Q_{6,i})\right|\right] \\
=&\mathbb{P}\left[\left|\tilde{F}_{\{l,m\}}\right| - \left|\tilde{F}_{\{l\}} + \sum_{i=1}^{\alpha}(Q_{5,i} - Q_{6,i})\right| > 0\right] \\
\leq&\mathbb{P}\left[\left|\tilde{F}_{\{l,m\}}\right| - \left|\tilde{F}_{\{l\}}\right| - \left|\sum_{i=1}^{\alpha}(Q_{5,i} - Q_{6,i})\right| > 0\right] \\
\leq&\mathbb{P}\left[\left|\tilde{F}_{\{l,m\}}\right| - \left|\tilde{F}_{\{l\}}\right| > 0\right]
\end{aligned}$$

Therefore the probability of sign epistasis is at most $\frac{1}{2}$ at $\alpha = 0$, but it is monotonically decreasing in α. For $\beta = 0$, or $\alpha \to \infty$ at constant β, the probability is even 0. If the fitness distribution has finite variance then the asymptotic as $\alpha \to \infty$ at constant β is $p(\alpha, \beta) \propto \frac{1}{\sqrt{\alpha}}$, as can be seen by expression of the probability in the characteristic function of the fitness distribution, i.e. in Fourier transformed space. There addition of i.i.d. random variables, i.e. convolution of probability

distributions, become simple products, in particular a α-power of the characteristic function appears. For large α the inverse Fourier transform may then be calculated by saddle-point approximation giving the factor $\frac{1}{\sqrt{\alpha}}$. This works out naively only if the second moment exists, because otherwise the characteristic function may not be twice differentiable at zero, where its value is maximal and equal to one.

For stable distributions, in particular the normal distribution, as fitness distribution, the values of $p(\alpha, \beta)$ are only dependent on the fraction $\frac{\alpha}{\beta}$, while generally, especially for small β, there might be explicit dependence on both parameters.

For epistatic loci always $\beta \geq 1$ and the interpolation between $\frac{\alpha}{\beta} = 0$ with mean sign epistasis weight $\frac{1}{2}$ and $\frac{\alpha}{\beta} \to \infty$ with mean sign epistasis weight 0 should be visible in the AN, RN and SN structures.

In the AN structure the number of partial landscapes sharing two loci depends negatively on their index distance, i.e. two loci l and $l+1$ share $k-1$ partial landscapes, but only one landscape contains l, but not $l+1$, i.e. $\beta = 1$ while $\alpha = k-1$. Loci in maximum correlated distance on the other hand share only one partial landscape and $k-1$ contain only each of the loci individually. Of course for more distant loci, there is no shared block and therefore $\beta = 0$ and the probability of sign epistasis is zero.

In the RN structure the probability of sign epistasis simplifies, because it is very unlikely for two loci to be contained together in the more than partial landscape. Therefore mostly $\beta = 1$, while α is Poisson distributed with mean $k-1$. Therefore the mean sign epistasis weight is generally larger on low degree loci and lower on high degree loci.

The SN model takes this to the extreme. Two ray loci are not epistatic at all. The mean sign epistasis weight from the ray to the center converges to zero, because there is only one partial landscape sharing both, but $L-1$ other ones containing only the center locus. On the other hand the weight from the center to the ray is $\frac{1}{2}$ because there is exactly one shared partial landscape and no other partial landscape containing only the ray locus. Finally two center loci share all L partial landscapes and therefore also have probability $\frac{1}{2}$.

The probability of global sign epistasis in the NK model can also be bounded. Let α and β be defined as before and the probability of sign epistasis at a random genotype given by $p(\alpha, \beta)$. Note that we have been considering only two of the k loci

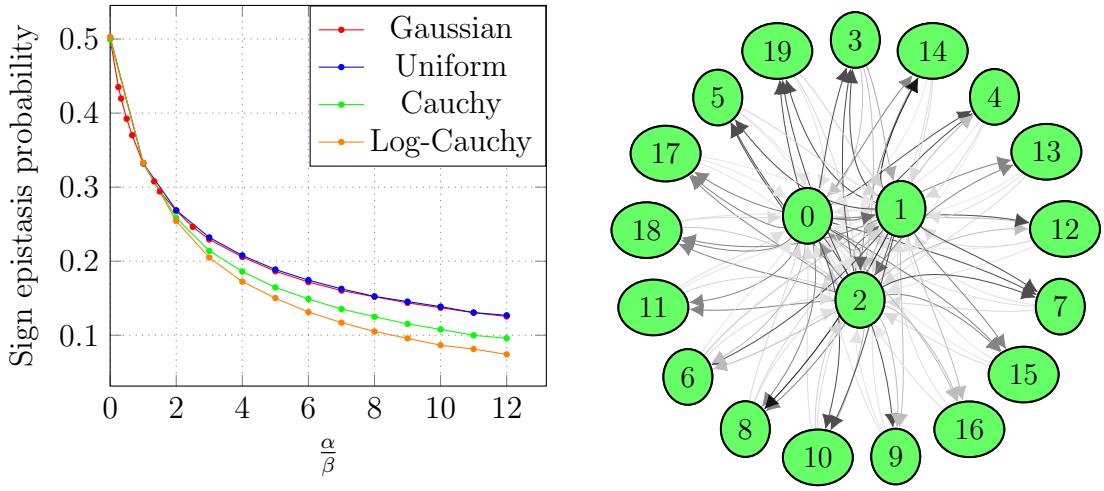

Figure 16: *Left: Probability of sign epistasis between two loci at random backgrounds depending on the fraction of shared partial landscapes $\frac{\alpha}{\beta}$ with Gaussian, uniform and Cauchy distributed base fitness values. For the latter two the probability may be dependent on β, for them here $\beta = 1$. (10^5 realizations per data point) For increasing $\frac{\alpha}{\beta}$, i.e. increase fraction of unshared partial landscapes, sign epistasis becomes uncommon. The fitness distribution has a minor impact, but at least the tail behavior seems to make a significant difference. Note that even though the Gaussian and uniform case seem to be identical, e.g. for $\frac{\alpha}{\beta} = 2$ they differ by approximately 0.05% with a p-value of 10^{-4}.*

Right: Example of weighted sign epistasis graph for the star neighborhood with $L = 20$ and $k = 4$. Ray nodes share $\beta = 1$ NK block with center loci and are contained in $\alpha = 0$ other NK blocks. Therefore arrows from center to ray loci have on average $\frac{1}{2}$ weight. Loci in the center however are contained in additional $\alpha = 19$ other NK blocks, therefore arrows from ray to center loci are of low mean weight. Note that most of these arrows actually have the smallest possible gray-scale value. Loci in the center share all partial landscapes and therefore also have mean arrow weight $\frac{1}{2}$ between each other.

in each partial landscape so far. If we modify any of the remaining $k-2$ ones, the partial landscape's contributions to the square are swapped with i.i.d. ones. If we do this for every partial landscape contributing to the square, then all contributing $Q_{i,j}$ will have been replaced by i.i.d. ones. Because every partial landscape has 2^{k-2} to choose these backgrounds, there are therefore at least 2^{k-2} independent backgrounds for the property of sign epistasis and therefore the probability of global sign epistasis is at most $p(\alpha,\beta)^{2^{k-2}} \leq 2^{-2^{k-2}}$. Therefore the sign epistasis becomes quickly improbable with increasing k. Note in particular that $p(\alpha,\beta)$ may additionally be decreasing k, especially in the RN structure typically α increases linearly with k, possibly sending $p(\alpha,\beta)$ to zero quickly, too. Additionally there are for non-zero α additional non-independent backgrounds on which sign epistasis needs to hold. This number becomes quickly large, too, especially if the α NK blocks do not overlap except for l.

The limit on sole global sign epistasis can with the same methods be bounded above by $2\left(\frac{p(\alpha,\beta)}{2}\right)^{2^{k-2}} \leq 2\cdot 4^{-2^{k-2}}$. The additional factors 2 and $\frac{1}{2}$ are due to the necessity to retain one of the two possible sign epistasis orientations, which are both equally likely to occur on any independent background.

For (global) reciprocal sign epistasis of course the limits above also apply, but they may be sharpened. Local reciprocal sign epistasis on a square dependents additionally on $\tilde{F}_{\{}m\}$ and so the α contributions need to be split into α_1 and α_2. α_1 is then the number of NK blocks containing only l and α_2 the number of NK blocks containing only m. The probability of reciprocal sign epistasis between l and m at a random background is then $\bar{p}(\alpha_1,\alpha_2,\beta)$ dependent on all three numbers. With the same inequalities used for sign epistasis one can see, that at constant β, $\bar{p}$ is decreasing monotonically in both α_1 and α_2. Also at $\alpha=0$, the probability of reciprocal sign epistasis on a square is, like in the HoC model, $\frac{1}{3}$. Therefore the probability of (local) reciprocal sign epistasis is always smaller or equal to $\frac{1}{3}$. For large α_1 and α_2 and fitness distributions with finite variance, the probability will also decrease for the same reasons as previously with $\frac{1}{\sqrt{\alpha_1\alpha_2}}$.

The probability of global reciprocal sign epistasis between l and m is then, again by the argument of 2^{k-2} independent backgrounds and two equally likely orientations, at most $2\left(\frac{\bar{p}(\alpha_1,\alpha_2,\beta)}{2}\right)^{2^{k-2}} \leq 2\cdot 6^{-2^{k-2}}$.

The probability to find GRSE between any pair of loci at all is therefore at most $L(L-1)6^{-2^{k-2}}$, because there are only $\frac{L(L-1)}{2}$ pairs of loci, or for classical NK structures at most $k(k-1)L6^{-2^{k-2}}$. These values are at the same time upper bounds on the mean number of GRSE interactions.

This gives a limit on the probability to find GRSE as a function of k in the $L \to \infty$ limit. Even if k is only growing faster than $\log_2 \log_6(L)$, the probability to find GRSE between any pair of loci has to decrease to zero for every structure. Practically I find that it is already difficult to find GRSE in simulated landscapes for $k = 4$, but basically impossible for $k = 5$. An estimation of the orders of magnitude can be made from the bounds derived. For $k = 5$ the upper bound on the probability of global reciprocal sign epistasis between two loci is already about 10^{-6}, for $k = 7$ it is already around 10^{-24}, smaller than one over Avogrado's constant. In my simulations however I see GRSE for $k = 4$ only appear at about $L \approx 10^6$, it seems as if k is of by one in the estimate. This can be heuristically explained because for GRSE there are not only 2^{k-2} backgrounds but also two typically independent Fourier components $\tilde{F}_{\{l\}}$ and $\tilde{F}_{\{m\}}$ each of which has independent 2^{k-2} backgrounds, at least if α_1 and α_2 blocks are usually non-overlapping, as is the case for the AN and RN case, but not the BN case. Then each of the $2 \cdot 2^{k-2}$ backgrounds may destroy reciprocal epistasis, effectively shifting k by one.

Nonetheless in the following sections I will show that for structures ∞-bounded in moments the probability to find GRSE increases to 1 exponentially as $L \to \infty$ for every constant k and that the mean indeed increases linearly in L as the bound above suggests.

This is not true for the star neighborhood. GRSE edges can not be adjacent in the NK structure primal graph because a locus may only be part of one GRSE interaction. In the SN structure however it is only possible to find k non-adjacent edges (because each edge contains a center locus). Therefore the number of GRSE interactions is limited to k in the SN structure. Simulation also shows that the probability to find GRSE quickly decreases for every k in L. This can be partially be understood because all locus pairs in the SN structure have at least one of α_1 or α_2 increasing linearly in L and therefore $\bar{p}(\alpha_1, \alpha_2, \beta)$ decreasing to zero in L.

6. Local NK properties

I define a **local NK property** as a set of uniformly bounded random variables X_l defined for every locus l, such that X_l is a deterministic function of only the fitness value realizations up to distance some constant distance d. Constant distance here means, that d is not asymptotically bounded in L and uniformly bounded means that all X_l share an upper and lower bound.

Many properties, such as GRSE are local NK properties. Typically these properties will be binary, however here they can be arbitrary.

Let $Z = \sum_{l \in \mathcal{L}} X_l$ or with the mean over loci $Z = L\mathbb{E}_{\text{loc}}[X_l]$. If the property is binary Z counts the number of appearances of a local attribute (like GRSE).

For the mean of Z over structure and fitness realizations we then have

$$\mathbb{E}[Z] = L\mathbb{E}[\mathbb{E}_{\text{loc}}[X_l]] = L\mathbb{E}_{\text{loc}}[\mathbb{E}[X_l]]$$

Because the X_l are uniformly bounded, $\mathbb{E}_{\text{loc}}[\mathbb{E}[X_l]]$ is also bounded and therefore $\mathbb{E}[Z] = \mathcal{O}(L)$.

Also:

$$\mathbb{E}\left[Z^2\right] = L\mathbb{E}\left[\mathbb{E}_{\text{loc}}\left[X_l^2\right]\right] = L\mathbb{E}_{\text{loc}}\left[\mathbb{E}\left[X_l^2\right]\right]$$

is $\mathcal{O}(L)$ for the same reason. This already indicates that the increase of the variance in slow enough, such that a law of large numbers should hold.

Let furthermore $\tilde{Z} = \frac{Z - \mathbb{E}[Z]}{\sqrt{L}}$.

6.1. Central limit theorem

I claim the following variant of the central limit theorem: If $\mathbb{E}[\mathbb{E}_{\text{loc}}[X_l]]$ and $\mathbb{E}[\mathbb{E}_{\text{loc}}[X_l^2]]$ converge as $L \to \infty$ and the NK structure is ∞-bounded in moments, then $\tilde{Z}$ converges in distribution to either a centered normal distribution or the constant 0 as $L \to \infty$.

I show this by calculating all cumulants of Z. The first cumulant of Z is $L\mathbb{E}[\mathbb{E}_{\text{loc}}[X_l]]$. Higher order cumulants can be calculated using linearity of the joint cumulant:

$$\kappa_n(Z) = \sum_{(l_1,\ldots,l_n) \in \mathcal{L}^n} \kappa(X_{l_1}, \ldots, X_{l_n})$$

The joint cumulant vanishes, if $\{X_{l_1} \dots X_{l_n}\}$ can be separated in two disjoint sets, such that every element in the first subset is independent of every element of the second set. X_l is independent of X_m if they are in distance larger than $2d$ in the NK structure, because each one is a deterministic function of its neighbors up to distance d in the NK structure. Therefore the primal graph of the NK structure with additional edges between loci up to distance $2d$ acts as a dependency graph of the random variables X_l. Thus, if $\{l_1, \dots, l_n\}$ is not connected in this contraction, then $\kappa(X_{l_1}, \dots, X_{l_n}) = 0$. The remaining joint cumulants can be expressed as a polynomials in mixed moments and because the random variables are uniformly bounded, all remaining joint cumulants are also absolutely bounded by some constants c_n.

Suppose one wants to count the non-separable sets of $(l_1, \dots, l_n) \in \mathcal{L}^n$ in the dependency graph. If the set is non-separable, then the maximal span of the set in the dependency graph is n, because otherwise there is no chance to connect the maximal distant loci with $n-1$ steps. In the primal structure graph this translates to a maximum span of $2dn$. Consequently all elements of the non-separable set need to be included in a $2dn$-ball around l_1 and therefore the total number of non-separable sets is at most $\sum_{l \in \mathcal{L}} (B_{2dn}(l))^{n-1}$.

With this the cumulant can be bounded.

$$|\kappa_n(Z)| \leq c_n \sum_{l \in \mathcal{L}} (B_{2dn}(l))^{n-1} = c_n L \mathbb{E}_{\text{loc}} \left[B_{2dn}^{n-1} \right]$$

Due to ∞-boundedness in moments $\mathbb{E}_{\text{loc}} \left[B_{2dn}^{n-1} \right]$ is almost surely asymptotically bounded and thus:

$$|\kappa_n(Z)| = \mathcal{O}(L)$$

.

The cumulant generating function of Z is thus at most growing linearly in L [2] and for $\tilde{Z}$ all cumulants except the second vanish for $L \to \infty$. The first cumulant of $\tilde{Z}$ vanishes by construction, the second converges and the higher ones are of order $\mathcal{O}(L^{-1})$, because $\kappa_n \left(\frac{Z}{\sqrt{L}} \right) = L^{-n} \kappa_n(Z)$. Then by Lévy's continuity theorem

[2] Note that this is was shown only for each cumulant separately almost surely, however the countable intersection of almost sure events is also almost sure.

$\tilde{Z}$ converges in distribution to a centered normal distribution or the constant 0 if the second cumulant is converges to 0.

6.2. Binomial bound

Suppose now that additionally X_l are binary, like for example the GRSE property. I am then interested in the likelihood that there $Z = 0$, however the central limit theorem cannot provide good bounds on this probability.

Using a similar approach as before it is however possible to find an exponential upper bound on this probability.

Suppose again that the NK structure is ∞-bounded in moments and let there be a asymptotically not vanishing fraction $\alpha > 0$ of loci with $\mathbb{E}[X_l] \geq \mu_0 > 0$.

For some $n \in \mathbb{N}$ there exists then out of this α-fraction of loci a non-vanishing fraction $0 < \beta \leq \alpha$ of loci which additionally have $B_{2d} < n$ asymptotically. This must be the case because otherwise there would be a fraction of loci with B_{2d} diverging, implying that $\mathbb{E}[B_{2d}]$ also diverges contradicting the assumption of ∞-boundedness in moments. Out of this β-fraction I can surely choose $\frac{\beta L}{n}$ loci which are pairwise separated by a distance larger than $2d$. The set of corresponding X_l is then mutually independent and, having $\mathbb{E}[X_l] \geq \mu_0$. It follows that Z is at least as large as a binomial distributed random variable over $L\gamma = L\frac{\beta}{n}$ trials with success probability μ_0. In particular therefore:

$$\mathbb{P}[Z = 0] \leq \mu_0^{\gamma L}$$

with $\gamma > 0$.

This implies that the probability for none of the binary random variables to be 1 decreases at least exponentially in L.

7. Local NK properties of the sign epistasis graph

7.1. Weak sign epistasis components

The limits shown above for local NK properties on locally bounded NK structures can be used to give some interesting qualitative results on these NK landscapes.

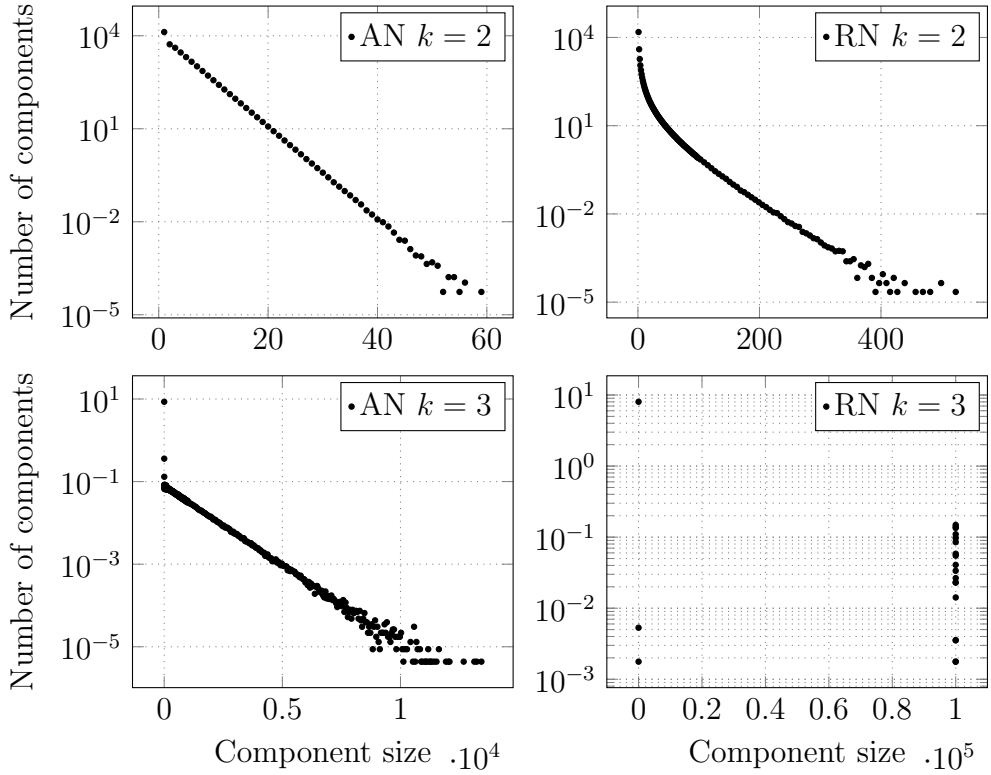

Figure 17: *Simulated mean distribution of component sizes at $L = 10^5$ with Gaussian fitness and varying number of realizations. For the AN structure at all k and the RN structure at $k = 2$, the distribution seems to have an exponential tail in accordance with quasi-one-dimensionality. Also the largest components found via simulation are much smaller than L, showing that there is no giant component. This is different for the RN model at $k \geq 3$. It has only few very small components and one giant component left.*

In particular in terms of the sign epistasis graph, properties related to the local structure around loci is of interest. The definition of local NK property assumes that the radius of influence for the property in the structure graph is limited by a constant, therefore the global structure can not be studied this way, but the local structure may very well.

Here I will consider small components. Weak components in the sign epistasis graph are interesting because they practically decouple the strong selection dynamics, i.e. two weak sign epistasis components behave in this limit as if the two components were not dependent at all. An adaptive walk over such a landscape decouples in two independent adaptive walks over each component, in the same way as it happens in the BN structure by construction. This allows one to calculated several properties of the landscape easier, like the number of global optima or the number of accessible paths.

So how many components are there? Of course the number of components is at least as large as the number of components in the NK structure. As shown in the introduction to the typical structure choices, the BN model has by definition only small components, while the AN and SN model has by definition only one giant component. The RN model however has potentially many components, although this doesn't seem to be the case for $k \geq 3$.

But as one can see even for the AN and RN model the sign epistasis graph will have many (i.e. linearly growing) number of small weak sign epistasis components. To see this I consider the following local NK property: Let s be the size of the component in question and X_l the binary property of locus l being part of a weak sign epistasis component of size s. The weak component may have at most diameter $s-1$ and therefore whether or not l is part of a component of size s is then at most dependent on mutational effects in loci up to constant distance s and consequently this is a local NK property.

The number of weak sign epistasis components is $\frac{Z}{s}$ where $Z = \sum_{l \in \mathcal{L}} X_l$ is the sum of the NK property defined.

The probability that a locus is part of a weak sign epistasis component of size s is certainly non-zero if the size of the structure component is at least s and 0 otherwise, because there is always a non-zero chance of epistasis edges not realizing any sign epistasis thereby reducing the structure component size to s in the sign

epistasis graph. The probability, if it is non-zero, is also bounded from below for structures ∞-bounded in moments because there are only finitely many NK structures of bounded size with a diameter of s.

Due to the previous central limit theorem we therefore know that asymptotically as $L \to \infty$ the number of components of size s will be distributed normally with linearly increasing mean and variance if the NK structure is ∞-bounded in moments and at least a non-vanishing fraction of loci are part of NK structure components of size at least s.

This is a rather powerful statement, because it implies that in the $L \to \infty$ limit sign epistasis components of every size (assuming the size is not already limited by the NK structure) will appear in linearly growing number, i.e. a non-vanishing fraction of loci will belong to components of size 1, 2, 3, etc.

All of the introduced NK structures except the SN structure are ∞-bounded almost everywhere and so the result applies. Their sign epistasis graphs will have a linearly increasing number of components, however it is not obvious from the previous statements whether the whole graph will fall into small components or whether there may still be a giant component left (assuming one existed previously).

7.2. Quasi one dimensional structure

In general my results cannot prove whether a giant component is left in the sign epistasis graph, but for the AN model a more detailed result can be given.

I say that a NK structure is **quasi one dimensional**, if, in the limit $L \to \infty$, the size of d-shells around all loci are bounded by a constant in d, i.e. when the number of loci in increasing radii are not increasing. In a usual geometrical sense this would correspond to the surface not growing with the volume, which is only true for the one-dimensional geometry.

The AN structure at constant k satisfies this property, because the size of d-shells is always $2(k-1)$.

With this property, there cannot be any giant component in the sign epistasis graph. This can be seen by considering any locus potentially belonging to the giant component. For every $r \in \mathbb{N}$ consider the union of $3r$- and $4r$-shell around a locus. These shell unions are mutually independent for different r because their loci do not

share NK edges between shells in distance ≥ 2. For each r there is also a non-zero probability for the edges between $3r$- and $4r$-shell to result in no sign epistasis, because as shown every NK structure generates every directed graph compatible with epistasis with finite probability. But this happening is independent for every r and there are only finitely many such probability values (because the size of shells is bound due to quasi one dimensionality), it follows that the probability of this event not happening for any r smaller than r' decreases exponentially in r'. But this event marks the separation of the shells interior, i.e. the $3r$-ball from the outer structure in the sign epistasis, i.e. it forms a component with bounded size.

Thus there cannot be any giant component, because it would require r to grow to infinity while the separation probability does not decrease faster than linear. Furthermore it implies that the tail of the size distribution of sign epistasis components is at least exponential.

This behavior of the AN structure has been used before to derive properties in the $L \to \infty$ limit, for example by Durrett and Limic to show that the scaled height of optima is asymptotically normal distributed and that the number of local optima is asymptotically log-normal distributed.[13]

7.3. Independent and isolated loci

The smallest component size to consider is $s = 1$. I call such components **isolated loci**.

A slightly weaker case are **independent loci**. They are not weak components but simply loci without incoming sign epistasis arrows, i.e. mutation signs on them are independent of the states of all other loci.

Independent loci are special in that they are present in linearly growing number in all NK structures at constant k, not only the ∞-bounded ones. This is because whether a locus is independent or not is solely a function of the fitness values for partial landscapes it is part of. Each partial landscape contains however only a constant number of loci and therefore it must be possible to choose a linear subset of loci not sharing any partial landscapes. These loci are then mutually independent in their probability to become independent loci and thus there is asymptotically a normal distributed number of independent loci with increasing mean for all NK

landscapes.

The same is not true for isolated loci, however because AN, BN and RN are ∞-bounded in moments the number of isolated loci will also be asymptotically normal distributed with linearly increasing mean.

Given the linearly growing number of independent loci on classical NK structures, it follows that adaptive walks from random starting points take at least linearly many steps, because there will be a linear number of independent loci in sub-optimal state initially and their mutation effects can only change by stepping along it.

It also implies that a non-vanishing fraction of loci are in identical state over all optima of the landscape, i.e. this fraction of a locally optimal genotype is deterministic. This is not generally the case. The probability for all optima to share one locus in the same state would decrease exponentially in the number of optima if they were distributed uniformly.

The probability that a certain locus is independent has a similar behavior as the probability of global sign epistasis. For a locus to be independent mutation on it must always have the same sign. There are at least 2^{k-1} independent background choices for every locus, due to them being part of partial landscapes of size k and therefore the probability for a locus to be independent is at most $2^{1-2^{k-1}}$, giving an upper limit of $L2^{1-2^{k-1}}$ for the mean number of loci. This is a fast decreasing in k and the bound seems to be rather bad concerning walk lengths.

7.4. Global reciprocal sign epistasis

Global reciprocal sign epistasis is also a NK local property, because whether there is (S)GRSE or not is only dependent on the partial landscapes in immediate surrounding of the focal loci. It is also in principle possible for every epistatic locus pair to turn into GRSE with non-zero probability. Therefore the central limit theorem applies and there will for structures ∞-bounded in moments be a normal distributed number of (S)GRSE interactions with linearly increasing mean. By the binomial bound additionally the probability that there is no (S)GRSE at all will decrease to zero exponentially.

This implies that the maximal distance accessible by an adaptive walker on these NK landscapes is asymptotically only a fraction of the full distance, i.e. the

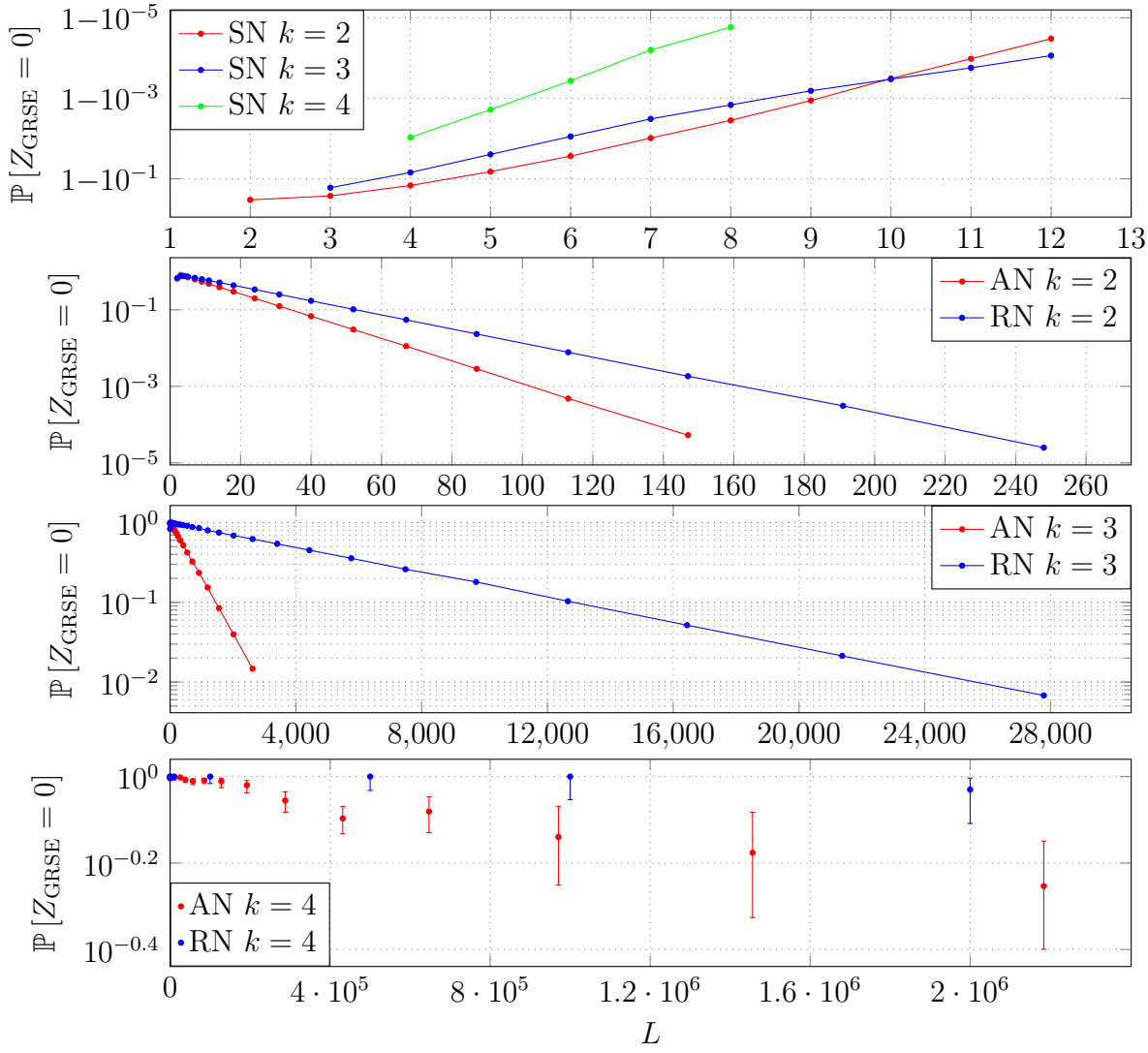

Figure 18: *Simulated probability not to find any GRSE in some NK variants. (Gaussian fitness and varying number of realizations per data point.) Error bars indicate 95% confidence intervals. Note the different scales of both x- and y-axes. In the SN structure the probability quickly goes to* 1*. For the AN and RN it does however decrease exponentially towards zero, but with vastly different speeds depending primarily on k, but also varying between AN and RN. These probabilities are also upper bounds on the accessibility of maximal distant genotypes.*

difference $L - Z_{\mathrm{GRSE}}$, implying that the fraction of the landscape reachable is exponentially decreasing. The probability that there is at least one accessible path of full length L is then also decreasing exponentially, implying exponential decrease in the accessibility of the global optimum from its antipodal.

SGRSE implies that local optima cluster in these landscapes in clusters of exponentially increasing size in L.

8. Star neighborhood

Most of the results so far do not apply to the SN structure. It is specially chosen to contrast the properties of ∞-bounded NK structures. Here I will show some of the qualitative differences.

First consider the number of local optima. Given any fixed state of the center loci, there can only be one local optimum in this subspace, because each ray locus can be optimized independently of the other ray loci. Therefore the SN model has at most 2^k local optima, in contrast to the ∞-bounded structures which had clusters of local optima increasing exponentially in size and so especially have exponentially many local optima in L.

The low number of local optima implies that there are not many ways for an adaptive walker to be trapped and accessibility of local optima will therefore be generally high.

Consider a reversed adaptive walker starting at the global optimum. Initially all mutations are deleterious and mutations of ray loci do not affect other ray loci's mutation signs. Therefore it is surely possible to mutate all ray loci in an accessible way in any order. After doing so all partial landscapes will have been modified and the only steps left are the center loci. The center loci taken for themselves form a HoC landscape on any fixed background. Therefore there is a non-zero probability depending on k that they may also be traversed by the reverse walker. The fitness values on this HoC slice of the center loci may be slightly biased because the background was not chosen randomly, however this bias does not extend to the rank landscape. Thus the probability that the there is an accessible path to the global optimum as constructed above is non-zero and independent of L. The accessibility of the global optimum at constant k is therefore not converging to

zero, in contrast to ∞-bounded structures, which will eventually for large enough L be inaccessible.

The low number of optima also implies that basins of attraction must be large on average, i.e. spanning a non-vanishing fraction of the landscape, which is not true for ∞-bounded structures.

9. Generalizations

Most results of this thesis can be further generalized in principle straight forward, although sometimes more complex in technicalities. In this section I will shortly mention some of these possibilities.

9.1. More alleles per locus

First of all the restriction to two-locus alleles is actually unnecessary, in general the notion can be extended to arbitrary, but bounded in L, number of alleles per locus and even arbitrary mutational structures on each locus. The notion can be formalized from the observation that the hypercube genotype space is actually the cartesian product of L copies of the complete graph on two vertices. Loci are simply these factors in the cartesian product. The generalization would then be to allow arbitrary graphs as factors in this product as long as their size is bounded for $L \to \infty$. The cartesian product naturally defines the projections onto locus subsets and each genotype can still be written as a sequence of locus states. Mutations on different loci are then still associative and even commutative (on the genotype space, not with regards to fitness). The HoC model's definition would still be valid, assigning each genotype an i.i.d. random fitness value and the definition of NK structures is unaffected. The definition of epistasis and sign epistasis graph would have to be amended to handle multiple mutations on the same locus, but the actual probabilities of sign epistasis are not radically different. Local NK properties still would satisfy the central limit theorem under the same conditions and consequently for the arguments on components, GRSE, etc. would not be affected except for the unspecified constants. However GRSE would need to be defined as GRSE between any pair of mutational transitions on the two loci.

One limitation on locus graphs should however be, that they are connected and at least of size 2.

9.2. Non-HoC partial landscapes

Partial landscapes are not really required to be of HoC type. This is however what was used in the original definition of the NK model. In principle any fitness landscapes on k loci is sufficient as long as it is invariant under permutation of loci. This requirement is necessary to avoid having to decide the mapping of loci in NK blocks to inputs of the partial landscape. The definition of NK structures is not affected by this and the central limit theorem on local NK properties would still hold. Only the actual probabilities of sign epistasis, the realizable sign epistasis graphs etc. might become limited, if the chosen fitness landscape model is not a continuous probability distribution in the space of possible distributions or it is not supported on the whole landscape. If however the probability for the property does not become zero by this change, then the previous results will still hold, i.e. either GRSE vanishes completely in the AN, RN and BN structures or it still grows linearly, but nothing in between. A trivial example without GRSE would be choosing the partial landscapes to be linear fitness landscapes. Then the resulting NK model will be also a linear model independently of the structure chosen, also implying that the number of size-1 components in the sign epistasis graph will grow linearly, while all other are impossible. If there is however even a slight non-zero probability of GRSE in the partial landscape model, then the GRSE count for ∞-bounded NK structures will still grow linearly.

9.3. Non-uniform NK structure

The NK edges do not generally need to be chosen of equal size. Arbitrary hypergraphs may be allowed, however to conserve the properties for the case of constant k in L, the size of edges should at least be bounded by a constant $\tilde{k}$ in the limit $L \to \infty$. Given such a bound, the NK structure will essentially behave like a constant k structure, because all arguments actually only required k to be bounded. Some calculated properties originally depending on k might end up depending either on the largest NK block size or the smallest one or anything in-between,

so care has to be taken there. However the result on the central limit theorem is unaffected and so is the probability of GRSE, the number of components, etc.

9.4. Correlation of partial landscapes

Generally the independence of partial landscapes is one of the main important properties of the NK model, which allows to treat it as I did with the central limit theorem, however some correlations can be tolerated. In particular if one assumes some arbitrary correlation of partial landscapes, as long as this correlation is only dependent on the distance between them, while still leaving partial landscapes in larger distances mutually independent, then effectively nothing changes, except that all d-boundedness properties need to be scaled to the incorporate the additional correlated distance, i.e. if the correlation distance is r, then all boundedness conditions in theorems would need to be replaced from d-boundedness to $d + r$-boundedness. With this adjustment all theorems on the central limit theorem would still apply, but the probabilities of sign epistasis might change significantly. Potentially there is also some additional limitation if there are strong correlation (e.g. correlation coefficients of 1).

I also suspect that correlations of unbounded distance may be introduced without invalidating the central limit theorem, as long as the correlation strength falls of quick enough with distance, however I have not considered such a case in detail.

9.5. Including spin glass models

Given the generalizations mentioned above (which also can be in principle combined), several other common stochastic models are included in the class of generalized NK models. For example the Edwards-Anderson model on a d-dimensional lattice with usual spins may be viewed as an NK model with $k = 2$, NK blocks corresponding to edges of the d-dimensional lattice and a partial landscape distribution with first order Fourier component corresponding to the external magnetic field and second order component corresponding to the spin interactions, both distributed in some given way. At least if the interaction distribution is continuous, the created NK model can still generate arbitrarily sized sign epistasis components and GRSE. The NK structure of the d-dimensional lattice is ∞-finite everywhere for every

fixed d and the therefore the limit results on the number of small components, GRSE, the structure of optima, etc. apply like in the AN case. One-dimensional lattices are also quasi one-dimensional and so the sign epistasis graph would fall into components of constant sizes.

Constant size sign epistasis components in this case should be interpreted as the spin glass (or a fraction thereof) behaving as if it were many small spin glasses in the low-temperature limit (corresponding to the strong selection limit).

The Sherrington-Kirkpatrick model would correspond to a $k = 2$ NK structure edges between every pair of loci. Obviously this structure is 1-bounded nowhere and therefore none of my results would apply to it. The same would be true for higher-order variants of the SK model.

Potts spins would correspond to loci with higher number of alleles.

10. Summary

After defining a generalized NK type model and introducing the (weighted) sign epistasis graph as representation of mutation interactions in the strong selection regime, I have demonstrated how these graphs look in the $L \to \infty$ limit for the House-of-Cards, the Rough-Mount-Fuji and the generalized NK model at constant ruggedness parameter with locally bounded interaction structure.

In particular HoC and RMF sign epistasis graphs do not contain any structural information on the landscape, because these models do not differ between loci in their definition anyway. Their sign epistasis graphs are essentially complete graphs with uniform arrow weights.

The NK model however defines a structure on loci explicitly, the NK structure. I have shown that the probability of sign epistatic dependence usually decreases with the number of NK blocks specific to one considered locus, while increasing with the number of NK blocks shared by both loci.

For sufficiently local bounded NK structures, in particular the AN, RN and BN structures, I have proven the validity of a central limit theorem for random variables on loci depending only on local neighborhoods. As examples of such properties I showed that in these models the sign epistasis graph will have linearly growing numbers of components of constant sizes and linearly growing number of

GRSE pairs. The latter results in an exponential decrease in the probability to find accessible paths, contrasting previous simulation results, exponential size clustering of local optima and exponential decrease in the reachable fraction of the landscape.

With the star neighborhood, which is not locally bounded, I showed that not all (classical) NK structures fulfill these asymptotic properties at all. It has instead a converging mean number of optima and a non-zero converging probability to find accessible paths spanning the landscape to the global optimum. This shows that although it is often assumed to be insignificant, the choice of underlying structure for the NK model can in fact influence global properties qualitatively. Nonetheless to see this difference rather special NK structures with highly variable locus degrees have to be used. All the common choices are asymptotically bounded in arbitrary distance. Highly variable degrees may however be adequate descriptions of genetic interactions, because there are often genes/proteins controlling the expression of large number of other genes, thereby acting as a high-degree node versus genes/proteins which do not affect expression at all.

For all classical NK structures, including SN, I showed the linearly growing number of independent loci resulting in linearly growing minimal length of adaptive walks, linearly growing number of frozen loci across local optima, and factorial increasing mean number of accessible paths.

A. Mathematical prerequisites

A.1. Landau notation

For real-valued functions $f(x)$ and $g(x)$ I use the usual definition for the Landau notation:

$$\begin{aligned} f \in o(g) &:\Leftrightarrow \lim_{x\to\infty} \left|\frac{f(x)}{g(x)}\right| = 0 \\ f \in \mathcal{O}(g) &:\Leftrightarrow \limsup_{x\to\infty} \left|\frac{f(x)}{g(x)}\right| < \infty \\ f \in \Omega(g) &:\Leftrightarrow g \in \mathcal{O}(f) \\ f \in \omega(g) &:\Leftrightarrow g \in o(f) \\ f \in \Theta(g) &:\Leftrightarrow f \in \mathcal{O}(g) \wedge g \in \mathcal{O}(f) \\ f \sim g &:\Leftrightarrow \lim_{x\to\infty} \left|\frac{f(x)}{g(x)}\right| = 1 \end{aligned}$$

As is commonly done, I also use this class notation in place of a representative, e.g. $f(x) = 3+o(x)$ means there exists a function $g(x) \in o(x)$, such that $f(x) = 3+g(x)$. I say that f grows **sub-linearly** if $f = o(x)$, **linearly** if $f = \Theta(x)$, **super-linearly** if $f = \omega(x)$, **sub-exponentially** if $f = e^{o(x)}$, **exponentially** if $f = e^{\Theta(x)}$ and **super-exponentially** if $f = e^{\omega(x)}$. I say that f grows **sub-polynomially** if $\forall_{\epsilon>0} : f = o(x^\epsilon)$, **polynomial** if there are $\epsilon_1 > 0$ and $\epsilon_2 > 0$, such that $f = \mathcal{O}(x^{\epsilon_1})$ and $f = \Omega(x^{\epsilon_2})$ and **super-polynomially** if $\forall_{\epsilon>0} : f = \omega(x^\epsilon)$. I use the same terminology for functions falling to zero with respect to their reciprocals, e.g. a function $f(x)$ is falling exponentially if $\frac{1}{f(x)}$ is growing exponentially.

A.2. Probability theory

The **r-th moment** of a real valued random variable X is defined as the expectation of the r-th power of X, i.e. $m_r = \mathbb{E}[X^r]$.

The **moment generating function** of a real valued random variable X is defined as $M(t) = \mathbb{E}\left[e^{tX}\right]$.

The **characteristic function** of a real valued random variable X is defined as $\chi(t) = \mathbb{E}\left[e^{itX}\right]$.

The characteristic function does always exist, while the moment generating function does not necessarily. If however both exist, then $\chi(t) = M(it)$.
If all moments of X exist, then $\chi(t)$ is analytic and $\chi(t) = \sum_{r=1}^{\infty} \frac{m_r (it)^r}{r!}$.
The **cumulant generating function** of X is defined as $K(t) = \log \chi(t) = \log \mathbb{E}\left[e^{itX}\right]$.
The **r-th cumulant** of X is defined as $\kappa_r = (-i)^r K^{(r)}(0)$.
If all moments exist, then also all cumulants exist and the cumulant generating function is analytic with $K(t) = \sum_{r=1}^{\infty} \frac{\kappa_r (it)^r}{r!}$.
Given a vector of real-valued random variables $\vec{Y} = (Y_1, \ldots, Y_n)$ their **joint moment generating function** is defined as $M(\vec{t}) = \mathbb{E}\left[e^{\vec{t}\vec{Y}}\right]$, their **joint characteristic function** as $\chi(\vec{t}) = \mathbb{E}\left[e^{i\vec{t}\vec{Y}}\right]$ and their **joint cumulant generating function** as $K(\vec{t}) = \log \chi(\vec{t})$.
Again the joint characteristic function always exists while the joint moment generating function might not, but if they do then $\chi(\vec{t}) = M(i\vec{t})$.
Every list of indices $i_1 \ldots i_m$ defines a **mixed/joint moment** by $\mathbb{E}\left[Y_{i_1} \ldots Y_{i_m}\right] = \frac{\partial^m}{\partial t_{i_1} \ldots t_{i_m}} M(\vec{t})\Big|_{\vec{t}=\vec{0}}$.
Similarly the **mixed/joint cumulant** or **connected correlation function** is defined by $\kappa(Y_{i_1}, \ldots, Y_{i_m}) = (-i)^m \frac{\partial^m}{\partial t_{i_1} \ldots t_{i_m}} K(\vec{t})\Big|_{\vec{t}=\vec{0}}$.
Both the mixed moments and cumulants are linear in their arguments, i.e. $\mathbb{E}\left[(Y_1 + Y_2)Y_3\right] = \mathbb{E}\left[Y_1 Y_3\right] + \mathbb{E}\left[Y_2 Y_3\right]$ and $\kappa(Y_1 + Y_2, Y_3) = \kappa(Y_1, Y_3) + \kappa(Y_2, Y_3)$.
If the random variables $Y_{i_1} \ldots Y_{i_m}$ are separable into two disjoint non-empty sets such that elements of the first set are mutually independent of elements of the second set, then the joint cumulant is zero. This is also known as **linked-cluster theorem**.

A.3. Multisets

The **power set** of a simple set S is the simple set $\mathcal{P}(S)$ containing all subsets of S.
A (finite) **multiset M** over a finite base set A is defined by the **multiplicity function** $\mathcal{I}_{\mathbf{M}} : A \to \mathbb{N}_0$.
Multisets will be written in bold font, while simple sets will be written using normal font weight.
The base set will usually be implied by context and not explicitly mentioned.

For $x \in A$ I write $x \in \mathbf{M}$ iff $\mathcal{I}_{\mathbf{M}}(x) \geq 1$.
The **support** of $\mathbf{M}$ is the simple set $S = \mathrm{supp}\,(\mathbf{M}) \subseteq A$, such that $x \in S \Leftrightarrow x \in \mathbf{M}$.
The **size** of $\mathbf{M}$ is $\#\mathbf{M} := \sum_{x \in A} \mathcal{I}_{\mathbf{M}}(x)$.
The **dimension** of $\mathbf{M}$ is the size of its support and denoted $|\mathbf{M}| = |\mathrm{supp}\,(\mathbf{M})|$.
A multiset $\mathbf{N}$ is a **(multi-)subset** of $\mathbf{M}$, written $\mathbf{N} \subseteq \mathbf{M}$, iff $\mathcal{I}_{\mathbf{N}}(x) \leq \mathcal{I}_{\mathbf{M}}(x)$ for all $x \in A$.
If in any of the binary set operations base sets are not equal, they are implied to be extended to their union.

A.4. Graph theory

The definitions and statements in this section are partially based on the introductory books *Hypergraph Theory* by Alain Bretto[4] and *Graph Theory* by Reinhard Diestel[10].

A **(simple undirected finite) graph** is a tuple $(V, \mathbb{E}\,[])$ of a finite vertex set V and an edge set $\mathbb{E}\,[\subseteq]\,\{e \in \mathcal{P}(V) \mid |e| = 2\}$.
A **(non-simple finite) directed graph** is a tuple $(V, \mathbb{E}\,[])$ of a finite vertex set V and an edge (arrow) set $\mathbb{E}\,[\subseteq]\,V^2$.
A **(multi-)hypergraph** is a tuple $(V, \mathbf{E})$ of a finite vertex set V and an (hyper-)edge multiset $\mathbf{E}$ with support $\mathcal{P}(V)$.
A hypergraph is k**-uniform** iff $\forall e \in \mathbf{E} : |e| = k$.
Note that simple undirected graphs are a special case of hypergraphs and all following definitions for hypergraphs apply in that sense also to them.
An **isomorphism** between two hypergraphs (directed graphs) $G = (V, \mathbf{E})$ and $G' = (V', \mathbf{E}')$ is a bijection $\phi : V \to V'$, such that $\mathcal{I}_{\mathbf{E}}(A) = \mathcal{I}_{\mathbf{E}'}(\phi(A))$ for all $A \in \mathcal{P}(V)$ $(A \in V^2)$.
G and G' are said to be isomorphic iff there is an isomorphism between them.
Isomorphism is an equivalence relation and its quotient space elements are called **isomorphism classes**.
A **subgraph** of a hypergraph (directed graph) $G = (V, \mathbf{E})$ is a hypergraph (directed graph) $G' = (V', \mathbf{E}')$, such that $V' \subseteq V$ and $\mathbf{E}' \subseteq \mathbf{E}$.
The **subgraph induced** by a vertex set $V' \subseteq V$ is the unique subgraph $G' =$

$(V', \mathbf{E}')$ with maximal sized $\mathbf{E}'$.
The **primal graph** of a hypergraph $G = (V, \mathbf{E})$ is the simple undirected graph $G' = (V, E')$, such that $e' \in E' \Leftrightarrow \exists e \in \mathbf{E} : e' \subseteq e$.
The **underlying graph** of a directed graph $G = (V, E)$ is the undirected graph $G' = (V, E')$, such that $(u, v) \in E \Leftrightarrow \{u, v\} \in E' \vee u = v$.
An **orientation** of an undirected graph (V, E) is a directed graph (V, E'), such that $\{u, v\} \in E \Leftrightarrow (u, v) \in E' \vee (v, u) \in E'$ and $(v, u) \in E' \Rightarrow (u, v) \notin E'$.
Two different vertices v and u are **adjacent** iff there is an edge e, such that $v \in e$ and $u \in e$.
A vertex v is **incident** on an edge e, iff $v \in e$.
Two edges e and f are **incident** iff $e \cap f \neq \emptyset$.
An edge (v, v) in a non-simple (directed) graph is called a **loop**.
A **path** of length m in a hypergraph with boundary is a sequence $(v_1, \ldots, v_m) \in V^m$, such that v_i and v_{i+1} are adjacent for $i = 1, \ldots, m-1$.
A **(undirected) path** of length m in a directed graph with boundary is a path of length m in its underlying graph.
A **directed path** of length m in a directed graph with boundary is a sequence $(v_1, \ldots, v_m) \in V^m$, such that $(v_i, v_{i+1}) \in E$ for $i = 1, \ldots, m-1$.
The **distance** between two vertices v and u is the length of the shortest (undirected) path between u and v.
The **directed distance** from vertex v to vertex u is the length of the shortest directed path from u to v.
The **distance** between a vertex v and an edge e is the minimal distance between v and any vertex in e.
The r**-ball** $B_r(v)$ of radius r around a vertex $v \in V$ is the subgraph induced by all vertices in distance at most r to v.
The r**-ball** $B_r(A)$ of radius r around a vertex set $A \subseteq V$ is the subgraph induced by all vertices in distance at most r to any element of A.
The r**-shell** $S_r(v)$ of radius r around a vertex $v \in V$ is the subgraph induced by all vertices in distance exactly r to v.
The r**-shell** $S_r(A)$ of radius r around a vertex set $A \subseteq V$ is the subgraph induced by all vertices in distance exactly r to one element of A, but not in smaller distance to any other element of A.

The (r, s)**-shell** $S_{(r,s)}(v)$ of radius r around a vertex $v \in V$ is the subgraph induced by all vertices in distance between r and s to v.
The (r, s)**-shell** $S_{(r,s)}(A)$ of a vertex set $A \subseteq V$ is the subgraph induced by all vertices in distance between r and s to one element of A, but not in distance smaller than s to any other element.
Two vertices are **(directed-)connected** iff there is a (directed) path between them.
The equivalence classes of connected loci are called **connected components** in undirected graphs, **weakly connected components** for undirected paths in directed graphs and **strongly connected components** for directed paths in directed graphs.
Given a series of graphs of increasing size, if the size of largest component grows asymptotically linearly with the graph size, this component is called a **giant component**.
Given two (un-)directed graphs (V, E) and (V', E') their **cartesian product** is the graph (V'', E'') with $V'' = V \times V'$ and $\{(v, v'), (u, u')\} \in E''$ iff $u = v \wedge \{u', v'\} \in E'$ or $u' = v' \wedge \{u, v\} \in E$.

References

[1] Bruce Alberts et al. *Molecular Biology of the Cell, Fourth Edition*. 4th ed. Garland Science, 2002.

[2] Julien Berestycki, Éric Brunet, and Zhan Shi. "Accessibility percolation with backsteps." In: *arXiv preprint arXiv:1401.6894* (2014).

[3] Julien Berestycki, Éric Brunet, Zhan Shi, et al. "The number of accessible paths in the hypercube." In: *Bernoulli* 22.2 (2016), pp. 653–680.

[4] Alain Bretto. *Hypergraph Theory*. Mathematical Engineering. Heidelberg: Springer International Publishing, 2013.

[5] Brian Charlesworth. "Effective population size and patterns of molecular evolution and variation." In: *Nature Reviews Genetics* 10.3 (2009), pp. 195–205.

[6] Kristina Crona, Devin Greene, and Miriam Barlow. "The peaks and geometry of fitness landscapes." In: *Journal of theoretical biology* 317 (2013), pp. 1–10.

[7] J. Arjan G. M. de Visser and Joachim Krug. "Empirical fitness landscapes and the predictability of evolution." In: *Nature Reviews Genetics* 15.7 (July 2014), pp. 480–490.

[8] J. Arjan GM de Visser, Tim F. Cooper, and Santiago F. Elena. "The causes of epistasis." In: *Proceedings of the Royal Society of London B: Biological Sciences* 278.1725 (2011), pp. 3617–3624.

[9] Thomas J. DeWitt, Andrew Sih, and David Sloan Wilson. "Costs and limits of phenotypic plasticity." In: *Trends in ecology & evolution* 13.2 (1998), pp. 77–81.

[10] Reinhard Diestel. *Graph Theory*. Springer Science & Business Media, Jan. 2006. 436 pp.

[11] Jeremy A. Draghi and Joshua B. Plotkin. "Selection biases the prevalence and type of epistasis along adaptive trajectories." In: *Evolution* 67.11 (2013), pp. 3120–3131.

[12] Aimée Marie Dudley et al. “A global view of pleiotropy and phenotypically derived gene function in yeast.” In: *Molecular Systems Biology* 1.1 (Mar. 29, 2005), E1–E11.

[13] Richard Durrett and Vlada Limic. “Rigorous results for the NK model.” In: *Annals of probability* (2003), pp. 1713–1753.

[14] Luca Ferretti et al. “Measuring epistasis in fitness landscapes: The correlation of fitness effects of mutations.” In: *Journal of Theoretical Biology* 396 (May 7, 2016), pp. 132–143.

[15] Jasper Franke et al. “Evolutionary Accessibility of Mutational Pathways.” In: *PLOS Comput Biol* 7.8 (Aug. 18, 2011), e1002134.

[16] Sergey Gavrilets. *Fitness landscapes and the origin of species (MPB-41)*. Princeton University Press Princeton, NJ, 2004.

[17] John H. Gillespie. “A simple stochastic gene substitution model.” In: *Theoretical Population Biology* 23.2 (Apr. 1, 1983), pp. 202–215.

[18] John H. Gillespie. “Is the population size of a species relevant to its evolution?” In: *Evolution* 55.11 (2001), pp. 2161–2169.

[19] John H. Gillespie. “Some properties of finite populations experiencing strong selection and weak mutation.” In: *American Naturalist* (1983), pp. 691–708.

[20] Brian Hall et al. *Strickberger’s evolution*. Jones & Bartlett Learning, 2008.

[21] David W. Hall, Matthew Agan, and Sara C. Pope. “Fitness Epistasis among 6 Biosynthetic Loci in the Budding Yeast Saccharomyces cerevisiae.” In: *Journal of Heredity* (Jan. 1, 2010), esq007.

[22] Peter Hegarty, Anders Martinsson, et al. “On the existence of accessible paths in various models of fitness landscapes.” In: *The Annals of Applied Probability* 24.4 (2014), pp. 1375–1395.

[23] Josef Hofbauer and Karl Sigmund. *Evolutionary games and population dynamics*. Cambridge university press, 1998.

[24] Nabil Kahale and Leonard J Schulman. “Bounds on the chromatic polynomial and on the number of acyclic orientations of a graph.” In: *Combinatorica* 16.3 (1996), pp. 383–397.

[25] Stuart A Kauffman and Edward D Weinberger. “The NK model of rugged fitness landscapes and its application to maturation of the immune response.” In: *Journal of theoretical biology* 141.2 (1989), pp. 211–245.

[26] Stuart Kauffman and Simon Levin. “Towards a general theory of adaptive walks on rugged landscapes.” In: *Journal of theoretical Biology* 128.1 (1987), pp. 11–45.

[27] John FC Kingman. “A simple model for the balance between selection and mutation.” In: *Journal of Applied Probability* (1978), pp. 1–12.

[28] Daniel J. Kvitek and Gavin Sherlock. “Reciprocal Sign Epistasis between Frequently Experimentally Evolved Adaptive Mutations Causes a Rugged Fitness Landscape.” In: *PLOS Genet* 7.4 (Apr. 28, 2011), e1002056.

[29] Anders Martinsson. “Accessibility percolation and first-passage site percolation on the unoriented binary hypercube.” In: *arXiv preprint arXiv:1501.02206* (2015).

[30] Jiři Matoušek. “The number of Unique-Sink Orientations of the hypercube*.” In: *Combinatorica* 26.1 (2006), pp. 91–99.

[31] Johannes Neidhart and Joachim Krug. “Adaptive walks and extreme value theory.” In: *Physical review letters* 107.17 (2011), p. 178102.

[32] Johannes Neidhart, Ivan G. Szendro, and Joachim Krug. “Adaptation in Tunably Rugged Fitness Landscapes: The Rough Mount Fuji Model.” In: *Genetics* 198.2 (Oct. 1, 2014), pp. 699–721.

[33] Johannes Neidhart, Ivan G. Szendro, and Joachim Krug. “Exact results for amplitude spectra of fitness landscapes.” In: *Journal of Theoretical Biology* 332 (Sept. 7, 2013), pp. 218–227.

[34] Martin A. Nowak. *Evolutionary dynamics*. Harvard University Press, 2006.

[35] H. Allen Orr. “A minimum on the mean number of steps taken in adaptive walks.” In: *Journal of Theoretical Biology* 220.2 (2003), pp. 241–247.

[36] H. Allen Orr. “Fitness and its role in evolutionary genetics.” In: *Nature Reviews Genetics* 10.8 (Aug. 2009), pp. 531–539.

[37] H Allen Orr. “The genetic theory of adaptation: a brief history.” In: *Nature Reviews Genetics* 6.2 (2005), pp. 119–127.

[38] H. Allen Orr. “The Population Genetics of Adaptation: The Adaptation of Dna Sequences.” In: *Evolution* 56.7 (July 1, 2002), pp. 1317–1330.

[39] H. Allen Orr. “The population genetics of adaptation: the distribution of factors fixed during adaptive evolution.” In: *Evolution* (1998), pp. 935–949.

[40] Sarah P. Otto. “Two steps forward, one step back: the pleiotropic effects of favoured alleles.” In: *Proceedings of the Royal Society of London B: Biological Sciences* 271.1540 (Apr. 7, 2004), pp. 705–714.

[41] A. S. Perelson and C. A. Macken. “Protein evolution on partially correlated landscapes.” In: *Proceedings of the National Academy of Sciences* 92.21 (Oct. 10, 1995), pp. 9657–9661.

[42] Patrick C. Phillips. “Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems.” In: *Nature Reviews Genetics* 9.11 (2008), pp. 855–867.

[43] Frank J. Poelwijk et al. “Empirical fitness landscapes reveal accessible evolutionary paths.” In: *Nature* 445.7126 (Jan. 25, 2007), pp. 383–386.

[44] Frank J. Poelwijk et al. “Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes.” In: *Journal of theoretical biology* 272.1 (2011), pp. 141–144.

[45] B. Schmiegelt and J. Krug. “Evolutionary Accessibility of Modular Fitness Landscapes.” In: *Journal of Statistical Physics* 154.1-2 (Oct. 29, 2013), pp. 334–355.

[46] Peter F. Stadler and Bärbel MR Stadler. “Genotype-phenotype maps.” In: *Biological Theory* 1.3 (2006), pp. 268–279.

[47] Ivan G Szendro et al. “Quantitative analyses of empirical fitness landscapes.” In: *Journal of Statistical Mechanics: Theory and Experiment* 2013.01 (2013), P01005.

[48] Sara Via et al. “Adaptive phenotypic plasticity: consensus and controversy.” In: *Trends in Ecology & Evolution* 10.5 (1995), pp. 212–217.

[49] J. Arjan G. M. de Visser, Tim F. Cooper, and Santiago F. Elena. "The causes of epistasis." In: *Proceedings of the Royal Society of London B: Biological Sciences* 278.1725 (Dec. 22, 2011), pp. 3617–3624.

[50] J Arjan GM de Visser and Joachim Krug. "Empirical fitness landscapes and the predictability of evolution." In: *Nature Reviews Genetics* 15.7 (2014), pp. 480–490.

[51] Günter P. Wagner and Jianzhi Zhang. "The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms." In: *Nature Reviews Genetics* 12.3 (2011), pp. 204–213.

[52] Jörgen W. Weibull. *Evolutionary game theory*. MIT press, 1997.

[53] E. D. Weinberger. "Fourier and Taylor series on fitness landscapes." In: *Biological Cybernetics* 65.5 (Sept. 1991), pp. 321–330.

[54] Edward D. Weinberger. "Local properties of Kauffman's \textit{N} - \textit{k} model: A tunably rugged energy landscape." In: *Physical Review A* 44.10 (Nov. 1, 1991), pp. 6399–6413.

[55] Daniel M. Weinreich, Richard A. Watson, and Lin Chao. "Perspective: sign epistasis and genetic costraint on evolutionary trajectories." In: *Evolution* 59.6 (2005), pp. 1165–1174.

[56] Daniel M. Weinreich et al. "Darwinian Evolution Can Follow Only Very Few Mutational Paths to Fitter Proteins." In: *Science* 312.5770 (Apr. 7, 2006), pp. 111–114.

[57] Mary Jane West-Eberhard. "Phenotypic plasticity and the origins of diversity." In: *Annual review of Ecology and Systematics* (1989), pp. 249–278.

[58] Sewall Wright. *The roles of mutation, inbreeding, crossbreeding, and selection in evolution*. Vol. 1. na, 1932.

Eidesstattliche Erklärung gemäß Prüfungsordnung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche kenntlich gemacht.