Master Thesis

Hypercubes, Peak Patterns and Universal Positive Epistasis

Daniel Oros

Supervisor and First Examiner:
Prof. Dr. Joachim Krug
Second Examiner:
Prof. Dr. Kristina Crona

University of Cologne
Institute for Biological Physics
April 12, 2022

Abstract

Genes and their interactions with one another crucially affect reproductive success, commonly referred to as fitness. Biallelic models have been used in the past as a mathematical framework to model and explain these interactions. One approach is to represent L biallelic loci as hypercubic graphs, known as L-cubes. On these L-cubes, vertices model genotypes and the edges connect the vertices that differ by a single locus value. Assigning fitness values to genotypes gives edges a direction towards higher fitness. Local optimal genotypes, called peaks, then have a higher fitness than all their direct neighbors. Recently, researchers have introduced the notion of peak patterns, referring to the set of peaks that are unique up to relabeling of vertices. However, a complete characterization of all possible peak patterns has not yet been performed for $L \geq 4$. This work concerns itself with an analysis for $L = 4$ regarding peak patterns and all possible instances of sign and reciprocal sign epistasis, substantiating the importance of peak patterns. Additionally a lower bound $\propto 2^{2^{L-1}}$ is provided for the set containing all possible peak patterns for a given L. Informed by this, all peak patterns up to $L = 6$ are computed and joined with a variant of Fishers Geometric Model having a one dimensional phenotype. Moreover peak patterns are used to calculate the maximal number of peaks for the staircase triangulation up to $L = 8$.

Acknowledgement

First and foremost I would like to thank Prof. Dr. Joachim Krug. As my supervisor he guided me throughout my thesis, from a slow start in Anti-Corona-Measure times to more productive periods when restrictions started to ease. He always had the time, patience and advice needed to guide me throughout my research. I would like to thank him for providing me with the freedom and time needed to find and shape my topic of research.
Second I would like to thank all group members for our meetings, especially during times of Anti-Corona-Measures. Notably I would like to thank Benjamin for his inputs and thoughtful discussions, providing an outside look and pointing out directions of interest. For finally breaking the silence of being alone in an office, thank you Rotem.
For reading parts of my thesis prior to submission and providing helpful input, thank you Nina and Lukas.
Latifa for simply being there. For your constant support and love, thank you.
Last but not least I would like to thank my family and all the friends I have made during my years of study. The importance for having all of you around me can't be understated, thank you.

Contents

List of Figures

List of Tables

List of Tools

- The programming language has been julia
- All plots are made with matplotlib

1 Introduction

Genetic interaction is a central mechanism for the development of complex lifeforms. The question of how genes cooperate or interfere with each other and thereby affect an organisms phenotypes is a central topic in genetics. With the importance of understanding and predicting this mechanism being highlighted by the constant change and evolution of pathogens, such as the development of antibiotic resistant bacteria or the emergence of viruses and their variants. Both measurements and modeling of different genotypes aim to study their reproductive success, labeled as fitness. They play a crucial role in predicting evolution and aim to enable us to act with foresight to counter or prepare for these emerging problems.

The mathematical framework used in this thesis is composed of fitness landscapes and fitness graphs. The former maps genotypes to their respective fitness and the latter is the corresponding oriented acyclic graph. Genotypes are the vertices and edges connect genotypes which differ only by one mutation. In the case of all possible L biallelic loci, the graph corresponds to an L-cube, being oriented and acyclic. Former efforts to study the interplay between different alleles in this setting include the introduction of shapes in [1] and rank orders in [8]. Another approach to this problem are peak patterns, introduced in [6]. They focus on how many different configurations of peaks, meaning genotypes of higher fitness than all their nearest neighbors, the L-cube has up to renaming of vertices while keeping distances invariant. The shape approach relies on triangulations and rank orders infer interactions on the base of relative fitness values. Both have been studied for the 3- and 4-cube in [1] and [16] for shapes and in [8] and [7] for rank orders. Due to the exponential increase of genotypes with increasing number of loci, higher dimensional L-cubes become increasingly difficult to study. Peak patterns are aimed to study larger L-cubes than previously possible for the shape and rank order approaches. However, previous work on peak patterns is missing a formal definition and the calculation of all possible peaks patterns for $L \geq 4$. This thesis aims to close this gap by defining peak patterns for $L \geq 4$, calcu-

lating the set of all possible ones up to $L = 6$ and providing a lower bound on its size for arbitrary L. The latter can be used to make the case that it is not suitable to calculate the set of all peak patterns for $L \geq 7$ due to its enormous increase in size. Moreover, some special cases of peak patterns are observed and described as well as being combined with the shape approach and a version of Fishers Geometric Model, having a one dimensional phenotype and only individually beneficial mutations. A brief summary of the content is given next.

Sec.2 is concerned with the mathematical structure of the L-cube, defines peak patterns and develops the mathematical structure later needed to efficiently calculate them. In order to describe biological genetic systems sec.3 starts by giving an overview of the modeling terminology. Epistasis is defined next, which is used to quantify genetic interactions. It follows an introduction to a version of Fishers Geometric Model (FGM) as a genotype-phenotype-fitness map. A variant of FGM with a one dimensional phenotype and only individually beneficial mutations is defined next. Afterwards shapes are introduced and briefly explained in sec.4, focusing on the so called staircase triangulation. Results regarding peak patterns are combined in sec.5 and constitute the main result of this thesis. Starting with a proof for the maximal number of peaks for the variant of FGM. Subsequently it continues with the analysis of all possible 193270310 oriented acyclic 4-cubes, taking into account the set of all peak patterns, their corresponding partially ordered set by inclusion and two-way epistasis. Also included is the calculation of all possible peak patterns for the 5- and 6-cube. In order to explain why it might not be reasonable to go beyond $L = 6$, a lower bound for the size of the set containing all possible peak pattern for a given L is provided. Furthermore an analysis to check which peak patterns of the 5- and 6-cube are compatible with the defined variant of FGM, as well as an short analysis of possibly interesting peak pattern samples is conducted. The calculation of all compatible peak patterns with the staircase triangulation, by a necessary condition, up to $L = 8$ is performed in sec.6. Last but not least a summary of the results and an outlook is given in sec.7, followed by methods in sec.8.

2 The Cube Graph - Mathematical Background

A *graph* $\boldsymbol{G = (V, E)}$ is a collection of points and lines, referred to as *vertices* $\boldsymbol{V}$ and *edges* $\boldsymbol{E}$, with the latter connecting some subsets of the former. Edges can have a direction, marked as single-arrows using a graphical representation. If all edges have a direction, the graph is called a *directed graph*, while this definition does not exclude multiple edges connecting the same two vertices. If a graph has no multiple edges between the same two vertices, it is called an *oriented graph*.

The most important graph in this work is the hypercube, a generalized 3-cube. It is labeled by its dimension L, as *L-cube* or $\boldsymbol{Q_L}$. Familiar cubes are the 1- to 4-cube, being a line, square, cube and tesseract respectively.
The vertex coordinates of the unit L-cube can be directly formed by applying the graph cartesian product L times

$$\boldsymbol{Q_L = P_2 \square \ldots \square P_2} \tag{1}$$

to the path graph $\boldsymbol{P_2}$, which consists of two vertices connected by a single edge [10].
By labeling the vertices of $\boldsymbol{P_2}$ by 0 and 1, each vertex of an L-cube can be uniquely represented by an bit-String of length L, see fig.1. Hence vertices are labeled by a row vector $\vec{\sigma} \in \boldsymbol{G}$, with its entry $\sigma_i \in \{0, 1\}$ being the i-th coordinate. When assigning a direction to each edge, the L-cube becomes an oriented graph, denoted by $\boldsymbol{Q_L^O}$.

2.1 Paths in Oriented L-cubes

A sequence $\boldsymbol{p}$ of unique edges joining vertices is called a *path* and a sequence which minimizes the number of edges between two vertices is referred to as a *direct path*. A sequence of directed edges is called a *directed path*. Note that the distinction is important when referring to oriented graphs, as in those the direction of edges is neglected when referring to direct paths.

The structure of oriented graphs can still be taken into account for directed paths. If the direction of edges switches along a direct path, the vertices at which switching occurs are called *path maxima* if both edges are directed toward the vertex and *path minima* vice versa.

2.2 Hamming-Distance of the L-cube

Distances between vertices $\vec{\sigma}^A, \vec{\sigma}^B \in \boldsymbol{Q_L}$ within the unit L-cube are measured by their *Hamming-Distance*

$$\Delta(\vec{\sigma}^A, \vec{\sigma}^B) = \sum_{i=1}^{L} |\sigma_i^A - \sigma_i^B|. \tag{2}$$

Eq.(2) is a measure for the number of differing coordinates or the length of a direct path between them. As there are L coordinates needed to describe all vertices of an L-cube, the maximal distance between two vertices is L, being the *diameter*. When referring to the distance to a fixed vertex $\vec{\sigma}^A$, it's denoted as $\Delta_{\vec{\sigma}^A}(\vec{\sigma}^B) = \Delta(\vec{\sigma}^A, \vec{\sigma}^B)$. Two important Hamming-Distances are the ones from $\vec{\sigma} = \vec{0}$, $\Delta_0(\vec{\sigma}) \equiv \Delta_{\vec{0}}(\vec{\sigma})$, and $\vec{\sigma} = \vec{1}$, $\Delta_1(\vec{\sigma}) \equiv \Delta_{\vec{1}}(\vec{\sigma})$. Those count the number of zeros and ones respectively.

2.3 k-cubes of the L-cube

L-cubes contain lower-dimensional *k-faces*, which have the same cubic structure as the corresponding k-cube. The number of k-faces contained in an L-cube is

$$Q(k, L) = 2^{L-k} \frac{L!}{k!(L-k)!} \tag{3}$$

from [10]. The number of vertices and edges for $L = 2, 3$ and 4 can be found in table 1. As each edge has two directions to point in, the number of oriented graphs can be calculated from eq.(3), resulting in $2^{Q(1,L)}$.

2.4 Oriented and Acyclic L-cubes

Of particular interest are oriented L-cubes which do not contain any cycles. Meaning that by starting from a vertex $\vec{\sigma}$ it is impossible to return to $\vec{\sigma}$

L-Cube	Vertices	Edges	$\lvert \boldsymbol{B_L} \rvert$	Acyclic Graphs
2-Cube	4	4	8	14
3-Cube	8	12	48	1862
4-Cube	16	32	384	193270310

Table 1: Various properties of the 2-, 3- and 4-Cube.

by a directed path via other vertices. Those L-cubes are denoted as *acyclic L-cubes* or $\boldsymbol{Q_L^A}$.

The number of possible acyclic oriented L-cubes can be computed by using the L-cubes chromatic polynomial $\chi(\lambda)$ and evaluating it at $\lambda = -1$ [22]. The chromatic polynomials up to $L = 4$ are known from the OEIS A334159 [15]. Numbers of all possible acyclic oriented L-cubes for $L = 2, 3$ and 4 can be found in table 1.

Vertices which only have incoming (outgoing) edges are called *graph maxima* (*graph minima*) or simply *maxima* (*minima*), as they are local sinks (sources) with respect to the orientation of the L-cube. Maxima are throughout referred to as *peaks*. The set of all maxima is denoted by $\max(\boldsymbol{Q_L^O})$ or $\max(\boldsymbol{Q_L^A})$, with the same holding for minima being $\min(\boldsymbol{Q_L^O})$ or $\min(\boldsymbol{Q_L^A})$.

Oriented and acyclic L-cubes can have up to 2^{L-1} peaks, with the configuration having the largest number of peaks called the *Haldane Graph* [14].

Note that while path maxima are a necessary condition for a graph maxima, it is not a sufficient one.

2.5 Super and Subset

The super- and sub-set [9] of a vertex $\vec{\sigma} \in \boldsymbol{Q_L}$ are given by either all vertices containing the same coordinates being 1 for the former or 0 for the latter

$$\mathrm{super}(\vec{\sigma}) = \{\vec{\sigma}^s \in \boldsymbol{Q_L} : \text{if } \sigma_i = 1 \text{ then } \sigma_i^s = 1 \; \forall i \in \{1, \ldots, L\}\}, \tag{4}$$

$$\mathrm{sub}(\vec{\sigma}) = \{\vec{\sigma}^s \in \boldsymbol{Q_L} : \text{if } \sigma_i = 0 \text{ then } \sigma_i^s = 0 \; \forall i \in \{1, \ldots, L\}\}. \tag{5}$$

As these sets are sub-cubes of the L-cube, their sizes are

$$|\,\mathrm{super}(\vec{\sigma})| = 2^{|\vec{\sigma}|_1} = 2^{L-m}, \tag{6}$$

$$|\operatorname{sub}(\vec{\sigma})| = 2^{|\vec{\sigma}|_0} = 2^m \tag{7}$$

with $m = \Delta_0(\vec{\sigma})$ being the number of ones.
Hence, the size of the union of these two sets is

$$|\operatorname{super}(\vec{\sigma}_M) \cup \operatorname{sub}(\vec{\sigma}_M)| = 2^m + 2^{L-m} - 1 \tag{8}$$

Minimizing the size of the set in eq.(8) with regard to m yields $\left\lfloor \frac{L}{2} \right\rfloor$.

2.6 Peak Patterns

A *symmetry operation* of an L-cube is an isomorphism and defined by a map $s : \vec{\sigma}^s \mapsto \vec{\sigma}^d$ which maps each vertex $\vec{\sigma}^s$ to a unique vertex $\vec{\sigma}^d$, while preserving the edge-vertex connectivity of the graph. When representing vertices of an unit L-cube as $\vec{\sigma}$ from sec.2, all symmetry operations can be constructed by the composition of two maps. Either switching the values of bit pairs σ_i and σ_j, $\sigma_i \leftrightarrow \sigma_j$, or by negating sites σ_i value, $\sigma_i \to \overline{\sigma_i}\ \forall i, j \in \{1, \ldots, L\}$. As each site's value is a binary, negating a site results in $\overline{\sigma_i} = 1 - \sigma_i$, hence $\overline{0} \to 1$ and $\overline{1} \to 0$. This group is known as the *Hyperoctahedral Group* $\mathbf{B_L}$ [2]. Switching sides results in $L!$ and negating them in 2^L possibilities each. Hence the size of

$$|\mathbf{B_L}| = 2^L L!, \tag{9}$$

including the unit operation. The values up to $L = 4$ can be found in table 1 and a graphical representation in fig.1.

A *peak pattern (pp)* of an oriented L-cube is a symmetry class of the L-cubes peaks under the action of $\mathbf{B_L}$.
Let a set of vertices $\sigma^P = \{\vec{\sigma_1}, \ldots, \vec{\sigma_N}\}$ be all N peaks of an oriented L-cube, $\sigma_P = \max(\boldsymbol{Q_L^O})$. It's *orbit* is defined as

$$\mathbf{B_L}(\sigma^P) = \{s \cdot \sigma^P : s \in \mathbf{B_L}\} \tag{10}$$

where s acts on all $\vec{\sigma} \in \sigma^P$. Such an orbit is considered a peak pattern. Note that every peak pattern of $\boldsymbol{Q_L^O}$ is also valid for $\boldsymbol{Q_L^A}$. Without loss of

generality let $\vec{1}$ be a peak and the edge directions are chosen such that the direction is from vertices with m to $m+1$ ones. This configuration is acyclic and placing additional peaks doesn't alter this property. Hence, each peak pattern has at least one configuration of directed edges which is acyclic.
There are also many different realizations of the same peak pattern, when considering all possible configurations of $\boldsymbol{Q}_{\boldsymbol{L}}^{\boldsymbol{A}}$. E.g. all configurations having only $N = 1$ peak correspond to the same peak pattern. However, two configurations with $N = 2$ peaks each, and Hamming distance 2 and 3, can't correspond to the same peak pattern. This is due to the fact that $\mathbf{B_L}$ does conserve the vertex-edge connectivity and does not change distances. A graphical representation for two configurations with the same peak pattern of the 3- and 4-Cube can be found in fig.1.

The distance between each pair of peaks can be mapped to a symmetric distance matrix $\{\Delta_{ij}\}$, with its trace values being zero. The entry Δ_{ij} corresponds to the Hamming-Distance between peaks i and j.
Note that while peak patterns are unique, their corresponding distance matrices are not. The peak pattern $\sigma^1 = \{0100, 0010, 0001, 0111\}$ of the 4-cube corresponds to the distance matrix $\Delta_{ij} = 2$ for $j \neq i$ and $\Delta_{ij} = 0$ for $j = i$. However $\sigma^2 = \{0110, 0101, 0011, 1111\}$ corresponds to the same distance matrix as σ^1, but they are not the same peak pattern. This relation can be checked computationally or by observing that no coordinate of every peak in σ^2 can be zero or one at the same time, as needed for σ^1.

The *average maxima distance* of a peak pattern σ^P with $N \geq 2$ is the mean of its upper triangular distance matrix, excluding the trace, given by

$$\bar{\Delta}_{\sigma^P} := \frac{2}{N(N-1)} \sum_{j>i} \Delta_{ij}. \tag{11}$$

The distance between peaks within an acyclic L-cube can only take real values between 2, as this is the minimal Hamming distance between two peaks, and L, being the L-cubes diameter, resulting in $\bar{\Delta}_{\sigma^P} \in [2, L]$. Note that fixing certain peaks and hence Δ_{ij} constrains the possible relative distance of other

entries within the distance matrix, as some configurations are not possible given a number of fixed peaks.
The *mean average distance* $\bar{\Delta}_N$ is the mean distance of n_N different peak patterns with N maxima, defined by

$$\bar{\Delta}_N = \frac{1}{n_N} \sum_{i=1}^{n_N} \bar{\Delta}_{p^i} \tag{12}$$

The *weighted mean average distance* $\bar{D}_N$ of peak patterns is the weighted arithmetic mean

$$\bar{D}_N = \frac{1}{\sum_{i=1}^{n_N} \omega_i} \sum_{i=1}^{n_N} \omega_i \bar{\Delta}_{p^i} \tag{13}$$

with ω_i as the number of all possible oriented acyclic L-cube configurations having peak pattern p^i.

2.7 Peak Pattern Normal Form

In order to check if two sets of peaks correspond to the same orbit it is possible to successively apply all elements of $\mathbf{B_L}$ to one set and comparing it to the other. This is however computationally expensive, as $\mathbf{B_L}$ has $2^L L!$ elements. Also checking set equality can be a bottleneck when comparing large numbers of peak patterns. Hence a unique representation for each element of a peak pattern can reduce the time to check if two sets of maxima belong to the same orbit enormously.
Arranging the bit-strings of N maxima as columns of a $L \times N$ matrix $\boldsymbol{M}$ leaves three degrees of freedom for the representation of an orbit. The first two are inherited from $\mathbf{B_L}$. First to permute bits, which corresponds to permuting rows, second to switch one or multiple bits, which is equivalent to switching all bits of one or multiple rows and third how the columns are ordered. The 0 and 1 in a columns bit-string can be thought of as a number in base 2. Define C_p for $p \in \{1, \ldots, N\}$ as the corresponding p-th column integer value in base 10. Then, all three degrees can be fixed by maximizing the column integer values and assuring that

$$C_i > C_j \text{ for } i > j. \tag{14}$$

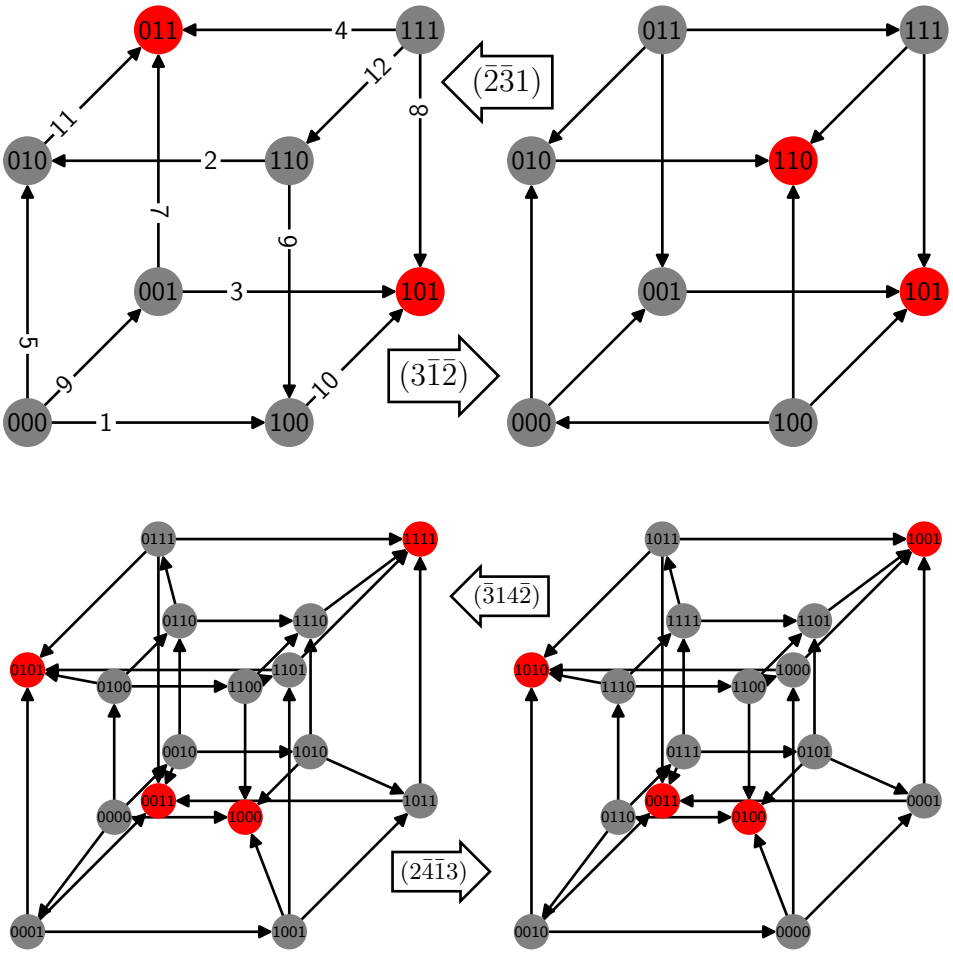


Figure 1: Peaks are colored in red and vertices are labeled by their bit string representation as in sec.2. The numbered edges of the upper left 3-cube are calculated in sec.8.1. Left and right cubes are mapped onto each other by the elements of $\mathbf{B_L}$, indicated by the corresponding arrows between the cubes. *top:* 3-Cube with two peaks and encodings 2218_{12} for the left and 3265_{12} for the right plot. Note that the cubes are rotations of each other, as both cubes have 000 as the lower left vertex. *bottom:* 4-cube with four peaks and encodings 37896800_{32} for the left and 619576898_{32} for the right plot. Spatial position of maxima are the same in both plots, as the vertices only got relabeled.

This translates to finding the optimal number of ones in a row. Fixing it then as the last row and repeating this iteratively for the remaining non-fixed rows, taking the already fixed rows into account. This scheme maximizes the number of ones in the lower right part of the matrix, and ensures that eq.(14) is true. One can see that this scheme also maximizes the integer row values R_l, being the l-th row bit-string in base 10, for $l \in \{1, \ldots, L\}$, with the ordering relation

$$R_i \geq R_j \text{ for } i > j. \tag{15}$$

Note that in eq.(14) C_i is strictly greater than C_j and in eq.(15) R_i and R_j are greater or equal. The first it due to the fact that C_i and C_j represent different vertices of the L-cube and hence can't be the same. R_i represents a row values and bit position i can be the same value for all peaks. The *peak pattern normal form* of $\boldsymbol{M}$ is then defined as the column vector

$$\vec{n}_M = \begin{pmatrix} R_1 \\ \vdots \\ R_L \end{pmatrix}. \tag{16}$$

A schematic structural representation of the corresponding binary matrix and their rows values can be seen in eq.(17). The last row has two domains, consisting of 0s and 1s. Within each domain, columns can be permuted without changing R_L. In the second last row, four domains exist, which have the same permuting rules and so on. By maximizing the number of ones in all the rightmost domains, from bottom to top, then the second most right and so on, the number of ones in each column and row are maximized.

$$\begin{array}{ccccccccccccccccc}
 & & & & & & & \vdots & & & & & & & & & \\
0 & | & 1 & | & 0 & | & 1 & | & 0 & | & 1 & | & 0 & | & 1 & \\
 & 0 & & | & & 1 & & | & & 0 & & | & & 1 & & \\
 & & & 0 & & & & | & & & & 1 & & & & \\
C_1 & & & \cdots & & & & \cdots & & & & \cdots & & & C_N &
\end{array}
\begin{array}{c} R_1 \\ \vdots \\ \vdots \\ R_L \end{array} \tag{17}$$

For example consider the set of four peaks, $\sigma^P = \{1010, 0100, 1001, 0011\}$ of the 4-cube with $N = 4$. This results in the 4×4 matrix

$$P = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}. \tag{18}$$

Transforming it into its peak pattern normal form yields

$$\begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix} \xrightarrow{(\bar{3}14\bar{2})} \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 \end{pmatrix} \xrightarrow[\text{columns}]{\text{sort}} \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 1 \end{pmatrix} \implies \vec{n}_P = \begin{pmatrix} 9 \\ 10 \\ 12 \\ 14 \end{pmatrix}.$$

The first step ensures that by using $s_1 = (\bar{3}14\bar{2}) \in \mathbf{B_L}$ the column integer values C_i are maximal, yet not ordered, hence violating eq.(14). Subsequent sorting of the column integer values results in an matrix form of eq.(17) and fulfills eq.(14). Note that s_1 is not unique, as either of $(2\bar{4}3\bar{1}), (\bar{1}\bar{4}23) \in \mathbf{B_L}$ would result in same peak pattern normal form after sorting the columns, while producing different second matrices. Additionally, P has the same peaks as the lower right 4-cube of fig.1 and the lower left plot the ones of its peak pattern normal form.

3 Biological Modeling

A *genotype* is a collection of genes. In this work *binary genotypes* are used, consisting of L biallelic loci. They are represented by a vector $\vec{\sigma} = (\sigma_1, \ldots, \sigma_L)^T$. If a genotypes site i is mutated, it is represented by $\sigma_i = 1$ and a non-mutated one by $\sigma_i = 0$. Hence, there are 2^L possible genotypes of a sequence of length L, with the *wild-type* $\vec{\sigma}_{WT} = \vec{0}$ and the *full-mutant* $\vec{\sigma}_{FM} = \vec{1}$. The set of all possible genotypes can be represented by an L-cube, with genotypes being connected by an edge if the Hamming distance between them is one.

Phenotypes are an organism's quantitative traits of arbitrary kind and influenced by its genotype and environment. A phenotype is represented as a collection of n real values in Cartesian space $\vec{z} = (z_1, \ldots, z_n)^T$. A common phenotype is e.g. the height of mammals. Specific labels for phenotypes are however not used in this work, as the phenotype space is only considered in an abstract form and will be defined when needed.

Fitness is the expected number of offspring, dependent on genotype and phenotype features. For bacteria the growth rate in a controlled environment is often used as a proxy for fitness. It can however become more difficult to measure fitness for organisms which e.g. reproduce sexually. Hence, fitness is not a global concept, as it needs to be interpreted in terms of the underlying setting. Nonetheless, fitness peaks are of certain interest, as they display genotypes which have a local optimal configuration in phenotype space. When only considering the relative fitness between genotypes of Hamming distance one and assuming that no two fitness values are exactly the same, each edge in the corresponding L-cube is directed from the genotype with lower to the one with higher fitness. The resulting oriented acyclic L-cube is referred to as *fitness graph* [5]. Fitness values are denoted by ω_g where g can refer to a genotypes vector $\vec{\sigma}_g$ entries or a set containing the loci of its 1-alleles.

3.1 Epistasis

Epistasis describes the effect of mutations behaving differently, depending on the genetic background, and thereby interacting with each other [7][8]. In this work only epistasis with regard to fitness will be considered.

Two-Loci Interaction: Let a two-locus system have fitness values ω_{00}, ω_{01}, ω_{10}, ω_{11}, with the subscript referring to the corresponding genotype. If there is no interaction between the two loci, fitness values are said to be additive, described by

$$\epsilon_2 = \omega_{00} + \omega_{11} - \omega_{01} - \omega_{10} \tag{19}$$

with $\epsilon_2 = 0$. If $\epsilon_2 \neq 0$ the loci are said to interact and exhibit epistasis.
When only considering the sign of ϵ_2, $\epsilon_2 > 0$ is said to have positive and $\epsilon_2 < 0$ negative epistasis. Combining the sign of ϵ_2 together with its corresponding fitness graph results in a more fine grained picture. For $L = 2$ there are $2^4 = 16$ possible graphs, two of which are cyclic and have no corresponding fitness landscape. The other 14 can be categorized into no epistasis, *sign epistasis (SE)* and *reciprocal sign epistasis (RSE)*, see fig.2. Note one can also define *magnitude epistasis (ME)*, which refers to the same fitness graph as no epistasis, with e.g. all arrows up. Therefore ϵ_2 and the corresponding fitness graph can be understood as incorporating complementary information. While in this case the computation of ϵ_2 is necessary to determine if there is no epistasis or ME, it does not contain information about SE or RSE beyond epistasis taking place or not. Checking for SE or RSE requires an analysis of the corresponding fitness graph.

Higher Loci Interaction: Higher order epistasis can be defined in a similar way to two-loci interaction. There is however more than one possible interaction for $L > 2$, as opposed to the two-loci case. For instance, in [1] these are defined and called *interaction coordinates* of which there are

$$C(L) = 2^L - L - 1. \tag{20}$$

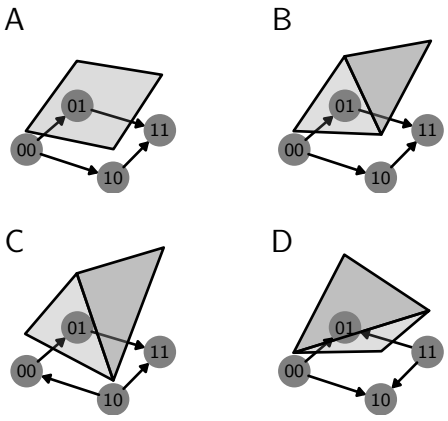


Figure 2: Schematic representation of two-way epistasis of a two-loci system and fitness being depicted as height of the fitness graphs genotypes. The gray shades depict the corresponding triangulation from sec.4.
A: $\epsilon_2 = 0$, no epistasis. **B:** $\epsilon_2 > 0$, magnitude epistasis. **C:** $\epsilon_2 > 0$, sign epistasis. **D:** $\epsilon_2 < 0$, reciprocal sign epistasis.

For e.g. a fitness landscape of three loci and fitness values w_g, one of these interaction coordinates is

$$u_{111} = (\omega_{000} + \omega_{110} + \omega_{101} + \omega_{011}) - (\omega_{100} + \omega_{010} + \omega_{001} + \omega_{111}). \qquad (21)$$

As this work is only concerned with two-loci interaction, this paragraph is only included for reason of completeness and not further discussed.

L-Cubes and k-Loci epistasis: As described in sec.2.3, L-cubes have k-faces of lower dimension. These k-faces have the k-cubes structure, hence k-way epistasis can be measured for every k-face. This is of interest as it allows the computation and analysis of k-loci epistasis for systems with multiple loci under e.g. certain constraints or to check which kind of interactions are possible for multiple loci in general. Computing k-face epistasis is analogous to computing k-loci epistasis using interaction coordinates, just for every k-face. There will however only be two-loci epistasis of L-cubes be discussed and used throughout this work.

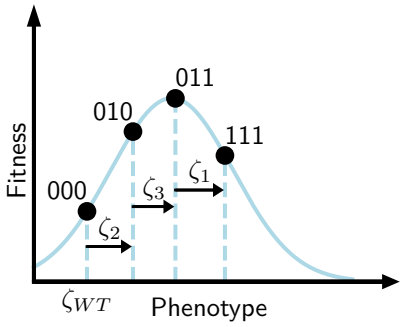


Figure 3: Schematic representation of a one dimensional phenotype-fitness function with a $L = 3$ biallelic loci system. To the position of the wild type ζ_{WT}, the mutation of locus 2 is added first, then locus 3 and last locus 1. The displacements are depicted by their corresponding ζ_i. This figure has been modified from [24].

3.2 Fishers Geometric Model

Fishers Geometric Model (FGM) has been first suggested by R. A. Fisher in [13]. The original idea has been extended and in this work a genotype-phenotype-fitness map is used, as described in [17]. The genotype-phenotype map $z : \boldsymbol{Q_L} \mapsto \mathbb{R}^n$ is additive and hence linear, while each mutation i has a fixed mutational displacement vector $\vec{\zeta_i} \in \mathbb{R}^n$. The position in phenotype space is given by

$$\vec{z}(\vec{\sigma}) = \vec{\zeta}_{WT} + \sum_{i=1}^{L} \sigma_i \vec{\zeta_i} \tag{22}$$

with $\vec{\zeta}_{WT}$ as the phenotype position of the wild type $\vec{\sigma}_{WT} = \vec{0}$.
Its phenotype-fitness function is a non-linear map $f : \mathbb{R}^n \mapsto \mathbb{R}$ of the phenotype. Note that the non-linearity is an important feature as it allows to incorporate gene interactions [23].
In the given setting of FGM, each genotype has a position in phenotype space, which in turn has a fitness value. In general, the graphical representation of the genotype-fitness map can be rough and have similarities with an actual landscape. Therefore it is also referred to as *fitness landscape*.

3.3 FGM - Construction and Constraints

In [24] the effect of beneficial mutations in TEM-1 β-lactamase that increase resistance of Escherichia coli to cefotaxime is studied. The correspond-

ing potential gene interactions regarding fitness are also described using a phenotype-fitness map. One focus is the dimension of the phenotype-fitness function used to model the individual effects of mutations on fitness under a certain genotypic background. With only beneficial mutations and a one dimensional phenotype-fitness function with a single maximum, an increase-decrease-increase pattern for different individually beneficial mutations is out ruled.

Building up on the work in [24] the phenotype-fitness function $f(\vec{z})$ used is considered to be smooth, with a single maximum at the origin, as in [17], and no other extreme points. This is due to the fact, that this setup turned out to be solvable as a simplified model and is compatible with some of the questions asked in this thesis. Note that in [24] it is laid out that for the experiment conducted a multi-dimensional phenotype is strongly hinted. Moreover throughout this work only individually beneficial mutations are considered for FGM, although their combination can be deleterious. Additionally, the phenotype dimension is set to be one dimensional, $n = 1$, resulting in $\vec{z}(\vec{\sigma})$, $\vec{\zeta}_{WT}$ and all $\vec{\zeta}_i$ to be scalars and not vectors. The vector symbol will be omitted from this point on. Furthermore, the wild type $\vec{\sigma}_{WT}$ and full mutant $\vec{\sigma}_{FM}$ are minima. This last condition ensures that while each individual mutation is beneficial, their combination eventually leads to a decrease in fitness. A schematic representation of this setup can be seen in fig.3. Note that the sign of ζ_i depends on the position of ζ_{WT} relative to the fitness optimum. If ζ_{WT} is on the left side of the fitness peak, it results in $\zeta_i > 0$ and if it's on the right side in $\zeta_i < 0$ $\forall i \in \{1, ..., L\}$. As each mutation is individually beneficial by construction, this leads to constraints on the displacements ζ_i. The fitness of any genotype with a single mutation is closer to the global maxima resulting in

$$f(\zeta_{WT} + \zeta_i) > f(\zeta_{WT}), \tag{23}$$

requiring $\zeta_i \neq 0$ for $\forall i \in \{1, ..., L\}$.

By choosing such a constrained setup, conditions can be obtained which enable to check if a certain oriented acyclic L-cube is compatible with the assumptions, as well as a necessary condition for the compatibility of peak

patterns with the given setup. Note that for this analysis no numerical information about the displacement vectors or the phenotype-fitness function are needed.
With the wild type and full mutant as graph minima, the number of an L-cubes edges that can change direction reduces by $2 \cdot L$. Using eq.(3), the number of edges which can change their direction reduces to 6 and 24 for the 3-cube and 4-cube respectively.
All possible oriented L-cubes will be numerically checked for $L = 4$. For $L = 5$ there are already 2^{80} different oriented configurations to check and 2^{70} when using the described variant of FGM. Hence, $L = 5$ is improbable to be checked in a brute-force manner when using standard computational equipment.

3.4 Path Condition

For the variant of FGM described in sec.3.3, each direct path from $\vec{\sigma}_{WT}$ to $\vec{\sigma}_{FM}$ can have only one path maxima for an arbitrary L. This condition is named the *path condition*.
There are $L!$ possible direct paths of length L from $\vec{\sigma}_{WT}$ to $\vec{\sigma}_{FM}$. A path maximum and a peak have only incoming edges, which results in the minimal Hamming distance of two between combinations of either path maxima or peaks. Hence, the largest possible number of path maxima within a direct path of length L is $\left\lfloor \frac{L}{2} \right\rfloor$. The maximal number of graph peaks are calculated in sec.5.1.
Let the path maxima of one of these paths be $\vec{\sigma}_m$, with m as the number of mutations. Then all $m!$ direct paths from $\vec{\sigma}_{WT}$ to $\vec{\sigma}_m$, as well as all $(L-m)!$ direct paths from $\vec{\sigma}_m$ to $\vec{\sigma}_{FM}$ can not contain any path maxima and hence no graph peaks. This results in the statement that super$(\vec{\sigma}_m)$ and sub$(\vec{\sigma}_m)$ can't contain any vertices that are peaks. Checking if for a set of peaks σ^P at least one realizations of $\mathbf{B_L}(\sigma^P)$ fulfills this statement, is therefore a necessary condition for the corresponding peak pattern of σ^P, to be compatible with the described variant of FGM.

4 Triangulation and Shapes Background

The shape approach is a method developed by N. Beerenwinkel et.al. in [1] and further clarified by K. Crona in [4]. It aims to understand gene-interactions between multiple loci by studying the geometry a fitness landscape imposes onto the convex hull of an L-cube, called the *genotope*. Two fitness landscapes are said to have the same *shape* if they impose the same *triangulation* onto the genotope.
Note that this method depends on a lot of mathematical background, which is not all covered in this introduction, but can be found in [1] and additionally in [18]. Nonetheless the main aspects are being explained in the next section, which summarizes the basics needed to get a understanding of the shape method. Therefore the ensuing subsection is a brief summary of [1], following its structure and only included for reason of completeness.

4.1 Triangulations and Shapes Introduction

Consider a finite alphabet Σ of size l, it can e.g. label different nucleotides or alleles of a gene. The biallelic case results in $l = 2$ and $\Sigma_2 = \{0, 1\}$. When considering a population of some sort, the individuals contain either the 0 or 1 allele. All probability distributions of this allele can be identified with the $l - 1$ dimensional standard simplex, for the biallelic case being

$$\Delta_{\Sigma_2} = \{(p_0, p_1) \in [0, 1]^2 : p_0 + p_1 = 1\} \tag{24}$$

and its general form

$$\Delta_{\Sigma} = \{(p_1, p_2, \cdots, p_l) \in [0, 1]^l : p_1 + p_2 + \cdots + p_l = 1\}. \tag{25}$$

When now creating a different alphabet, using the Σ_2 alphabet L times by the direct product, one gets Σ_2^L with $(\Delta_{\Sigma_2})^L \equiv \Delta_{\Sigma_2}^L$ being the L-cube having 2^L vertices. Note that not the whole set of possible genotypes in Σ_2^L needs to be considered for this approach to work, when e.g. some genotypes can't be realized biologically. $G \subseteq \Sigma_2^L$ is named the *genotype space* and its convex hull Π_G the genotope. In this work however, only the case $G \equiv \Sigma_2^L$ is considered,

which results in $\Pi_G \equiv \Pi_{\Sigma_2^L}$ as the convex hull of the L-cube $\Delta_{\Sigma_2}^L$. Now a point $v = (v_1, \ldots, v_L) \in \Pi_G$ is an L-tuple of allele frequencies, with the index of its components referring to the corresponding locus.
Rather than individual genotypes consider now a population of genotypes. Any probability distribution p on the set G can be considered such a population, with coordinates p_g representing the fraction of a population that is of genotype $g \in G$. Hence, a population is a point in the population simplex Δ_G. It is important to reemphasize the difference between Δ_G and Π_G. The former is used for distributions of populations and the latter for frequencies of alleles. They are related trough the marginalization map

$$\rho : \Delta_G \to \Pi_G, \quad (p_{\sigma_1 \ldots \sigma_L})_{\vec{\sigma} \in \Sigma^L} \mapsto \left(\left(\sum_{\vec{\sigma}: \sigma_i = \tau} p_{\sigma_1 \ldots \sigma_L} \right)_{\tau \in \Sigma} \right)_{i=1,\cdots,L}. \tag{26}$$

which maps a population p to its L-tuple of allele frequencies. Note that ρ is linear and all possible L-tuples, hence allele frequencies, can be realized by choosing the corresponding $v \in \Pi_G$. Notably every population realized by G can be mapped to the corresponding allele-frequency vector v via the marginalization map ρ.
The fiber ρ^{-1} then maps allele frequencies to populations, possibly being not unique. It is defined as

$$\rho^{-1}(v) = \{p \in \Delta_G : \rho(p) = v\}, \tag{27}$$

which is a polytope inside the population simplex.
A fitness landscape $\omega : G \to \mathbb{R}$ from sec.3.2 assigns a single fitness value to each genotype $g \in G$. But ω needs to be continuous in order to speak about the "shape" or "curvature" of the fitness landscape. Hence, consider populations $p \in \Delta_G$ instead of individuals. A populations fitness can be written as

$$\omega \cdot p = \sum_{g \in G} \omega_g \cdot p_g. \tag{28}$$

The corresponding continuous landscape $\tilde{\omega} : \Pi_G \to \mathbb{R}$ can be derived from ω by assigning every $v \in \Pi_G$ the maximum fitness among all populations p with allele frequencies v. Define

$$\tilde{\omega}(v) := \max\{p \cdot \omega : p \in \rho^{-1}(v)\} \qquad \forall v \in \Pi_G. \tag{29}$$

Note that a populations fitness $p \cdot \omega$ varies over the fiber ρ^{-1}. It does so because of the gene interactions underlying the fitness landscape ω.
The domains of linearity $\tilde{\omega}$ of are the cells in a regular polyhedral subdivision $\Pi_G[\omega]$ of the genotope Π_G. Those subdivisions $\Pi_G[\omega]$ are named the *shape* of the fitness landscape ω in [1]. For more information about polyhedral subdivisions see [18]. It is nonetheless important to note that the subdivision $\Pi_G[\omega]$ will be in most cases a *regular triangulation*, meaning a subdivision whose cells are simplices.
Now those simplices in $\Pi_G[\omega]$ have a interpretation in terms of alleles and populations: For any L-tuple of allele frequencies, $v \in \Pi_G$, a unique fittest population p exists. Hence a fittest population p fulfills $\rho(p) = v$, with the genotypes occurring in p being the vertices of the simplex $\Pi_G[\omega]$ that contains v. Concluding that either knowing all the fittest populations for a given ω or its shape is equivalent.

4.2 Triangulation of the 3-cube

For the 3-cube there are exactly 74 triangulations categorized into 6 *interaction types*, which differ only by labeling of the vertices. Hence, they are a symmetry class of the 3-cube. Moreover all triangulations are a combination of 4 different simplices that can be formed by the vertices of the 3-cube, again up to labeling of the vertices. For a graphical representation of the 4 simplices and the 6 interaction types, see fig.4.

4.3 The Staircase Triangulation and Peak Patterns

Combining peak patterns and triangulations has been done in [6]. An interesting finding is that not all peak patterns are compatible with any triangulation. The staircase triangulation number 6 in fig.4, composed only of the staircase simplex, can't have 3 or 4 peaks, leaving only three out of five peak patterns to be possibly compatible with it. The reason is that certain peak patterns, such as the one imposed by the Haldane graph, from sec.2.4, cut off genotypes. A genotype $\vec{\sigma}_g$ is said to be cut off, if there exists no other

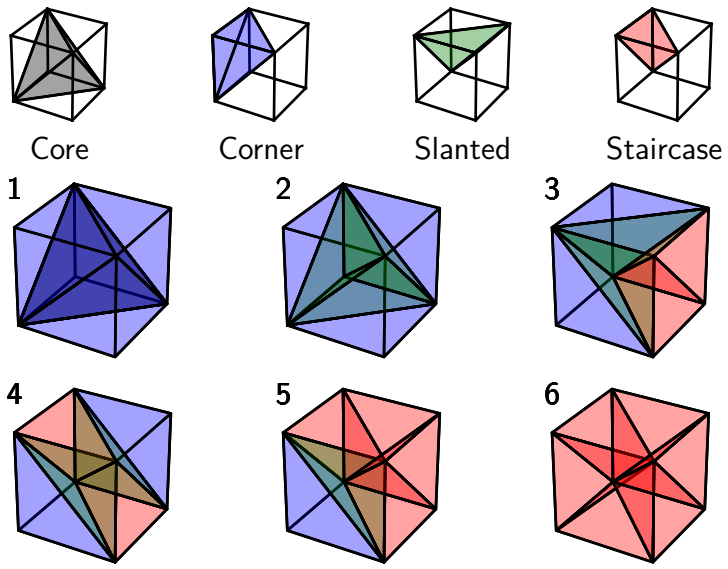


Figure 4: *First row:* All four possible simplices of the 3-cube with their respective names. *1-6:* All six possible interaction types up to cube-symmetry of the 3-cube. Number 6 is the staircase triangulation, which uses only the staircase simplex. Modified from [20].

genotype $\vec{\sigma}_{g'}$ that belongs to the same simplex and fulfills $\Delta(\vec{\sigma}_g, \vec{\sigma}_{g'}) = 2$. Meaning that there is no diagonal induced by the triangulation on one of the 2-faces that connects the cut off genotype $\vec{\sigma}_g$ to a another genotype.
The staircase triangulation is of particular relevance and generalized to L dimensions in [6]. Note that for $L = 3$ each simplex contains the $\vec{0}$ and $\vec{1}$ vertex. These are connected by the $L!$ direct paths between them and the $L + 1$ vertices of each path are the vertices of the individual staircase simplices. The generalization of the staircase triangulation of an L-cube is then given by the $L!$ simplices whose vertices are the ones of the $L!$ direct paths between $\vec{0}$ and $\vec{1}$. This results in a inequality for the fitness values in order for its triangulation to be the defined staircase one. Let g and g' be the set

representations of genotypes $\vec{\sigma}_g$ and $\vec{\sigma}_{g'}$ 1-allele loci,

$$\omega_{g\cup g'} + \omega_{g\cap g'} \geq \omega_g + \omega_{g'} \tag{30}$$

has been named *universal positive epistasis* in [6]. An interpretation is, that an uneven distribution of 1's in a pair of genotypes results in a greater or equal fitness than a more even one. While upper bounds for the maximal number of peaks of the staircase triangulation have been shown for $L = 4$ and $L = 5$ in [6], a compatibility check for which peak patterns are possible for $L \geq 4$ is however missing.

5 Results for Peak Patterns

In this section all possible peak patterns are calculated for $L = 4, 5$ and 6 and further analyzed. They are also combined with the path condition from sec.3.4 for which the maximal number of peaks are calculated next.

5.1 Largest Possible Number of Peaks for the Path Condition and $n = 1$

Starting from the minimal size of the sub- and super-set union in eq.(8), one can choose all genotypes with distance $\left\lfloor \frac{L}{2} \right\rfloor$ from the wild type as peaks. This choice of peaks is valid, as none of the vertices is a sub- or super-set of each other, hence not violating the path condition from sec.3.4, and thus being a lower bound on the largest number of peaks

$$N_{max} \geq \binom{L}{\left\lfloor \frac{L}{2} \right\rfloor}. \tag{31}$$

An upper bound can be obtained by mapping this problem to set theory. By numbering each loci from 1 to L, the set containing all loci is defined as $X = \{1, \ldots, L\}$. It's power set, meaning the set of all possible combinations of the elements in X, is denoted by $\mathcal{P}(X)$ and is of size 2^L. Now one can construct a *partial set order*, which allows to compare two elements in $\mathcal{P}(X)$. The order will be determined by set inclusion which checks if a set is a subset of another. By constructing a *Hasse-Diagramm*, meaning to create a graph with its vertices being the sets of $\mathcal{P}(X)$ and connected by an edge if for $a, b, c \in \mathcal{P}(X)$ with $a \subset b$, there exists no c with $a \subset c \subset b$. For the case of the power set, the result is a graph which has the same structure as the L-cube, see fig.5 for the $L = 3$ case and is called a *partially ordered set* (poset). Now the maximal size of a set $S \subseteq \mathcal{P}(X)$ with no element in S being a subset of another, is the largest number of possible peaks $N_{max} = |S|$. The upper bound is known from *Sperner's Theorem* [11] to be

$$N \leq \binom{L}{\left\lfloor \frac{L}{2} \right\rfloor}. \tag{32}$$

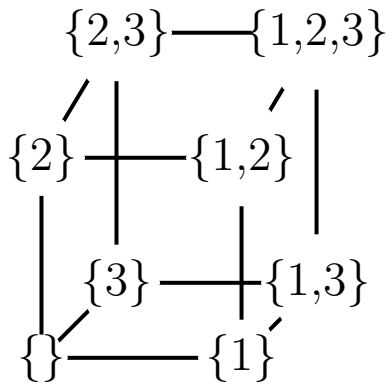


Figure 5: Hasse-Diagramm of the power set $\mathcal{P}(\{1,2,3\})$, which has the same structure as the 3-cube.

Combining eq.(31) and (32) results in in the largest number of possible maxima

$$N_{max} = \binom{L}{\lfloor \frac{L}{2} \rfloor}. \tag{33}$$

5.2 Full Analysis of the 4-Cube

All the following information can be found or directly computed from table 2 and it is referred to it if not stated otherwise. Also some values are explicitly calculated and pointed to if needed.

5.2.1 Distribution of Maxima and their Distance

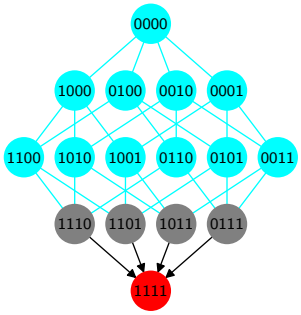
From [6] the number of peak patterns is known to be 5 for $L = 3$. The frequency of a peak pattern is the fraction of possible configurations of oriented acyclic L-cubes exhibiting this peak pattern and the number of realizations the overall number. All configurations for $L = 4$ are computed as in sec.8.2. This results in 20 possible peak patterns for acyclic oriented graphs, displayed as orthogonal graph plots in fig.6.
When including the additional path condition from sec.3.4, only 13 peak pattern remain possible. The left plot in fig.7 shows the number of peak patterns which are possible for the 4-Cube and the additional path condition, with its maximal number of peaks being 6, consistent with the proof in sec.5.1. Moreover already starting at 4 peaks not all possible peak patterns are compatible with the path condition. Hence when having sufficient data

available to compute the peak patterns, one can already determine if a one dimensional single peaked phenotype function is of use when fitting the data. Only taking into account the path condition for a one dimensional single phenotype function, the number of possible acyclic graphs reduces significantly, from 193270310 to 34572. The single most frequent configuration has two peaks with Hamming distance 4, accounting for more than half of all possible configurations. When including all possible acyclic oriented 4-cubes, this peak pattern is the least frequent one, see No.4 in table 2.
It's impact can be observed in fig.8 when computing the mean average distance of the possible peak patterns, especially when adding the corresponding number of possible acyclic graph realization as a weight. For two peaks, most acyclic graphs correspond to the peak pattern with both peaks being at distance two, having the smallest possible distance (▼). With the additional path condition, the observation reverses to the peak patterns with distance 4 (▲). Hence, the path condition has the tendency to increase the distance between maxima, which seems reasonable, as neither the sub- or super-set of a peak can contain another peak. This observation is however not true for the case with three peaks, but for any higher number, up to the maximum of six. The peak patterns which are possible for four peaks using the path condition (+) are the ones with a larger mean average distance when taking all possible ones (×) into account. This trend increases for five and six peaks.

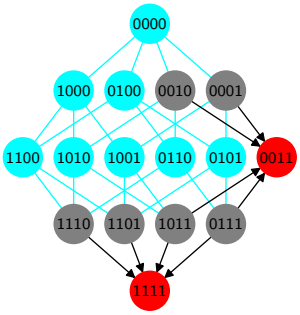
5.2.2 Partially Ordered Set of Peak Patterns

Another graph that is useful to take into account is the poset of peak patterns. If a peak pattern can be constructed by adding or removing a single peak, two nodes, each representing a peak pattern, are connected by an edge. The poset graph for the peak patterns of the 4-cube can be found in fig.9.
When measuring e.g. drug response curves one could measure how the peak patterns change dynamically. For example a dose response curve [9] of different mutants will have an effect on the corresponding oriented graph every time the fitness values of direct neighbors change their ordering. Measuring antibiotic concentration and the corresponding growth rate can result in a

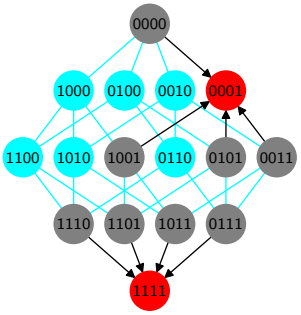
No.1 — N=1



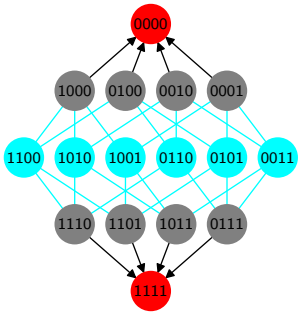
No.2 — N=2



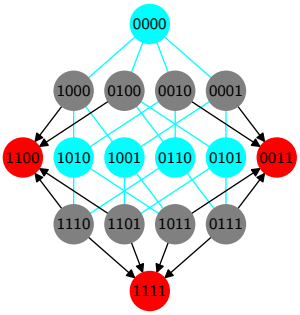
No.3 — N=2



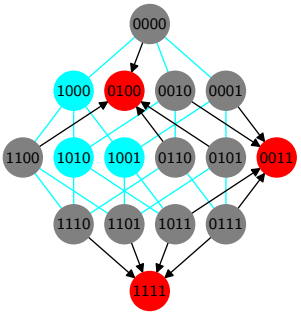
No.4 — N=2



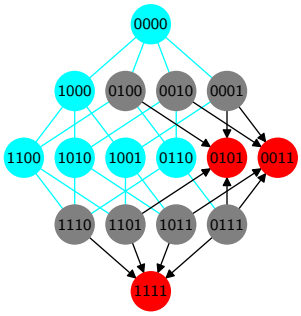
No.5 — N=3



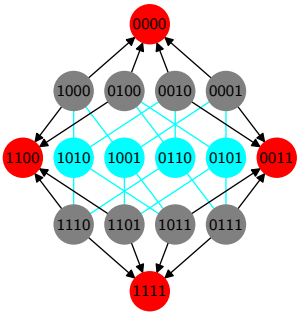
No.6 — N=3



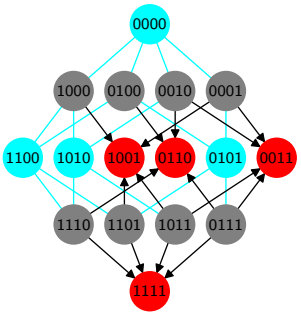
No.7 — N=3



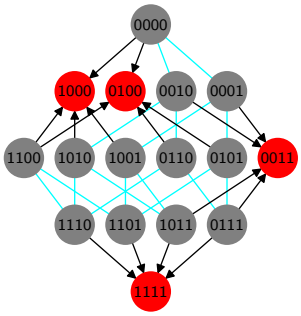
No.8 — N=4



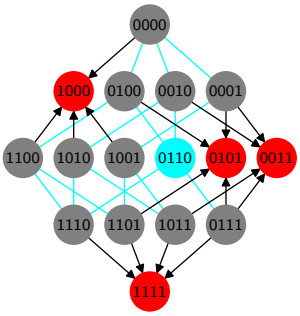
No.9 — N=4



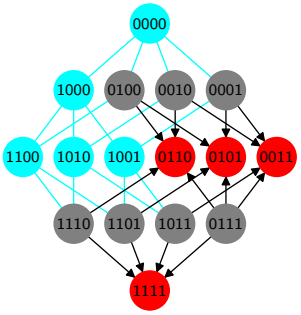
No.10 — N=4



No.11 — N=4



No.12 — N=4



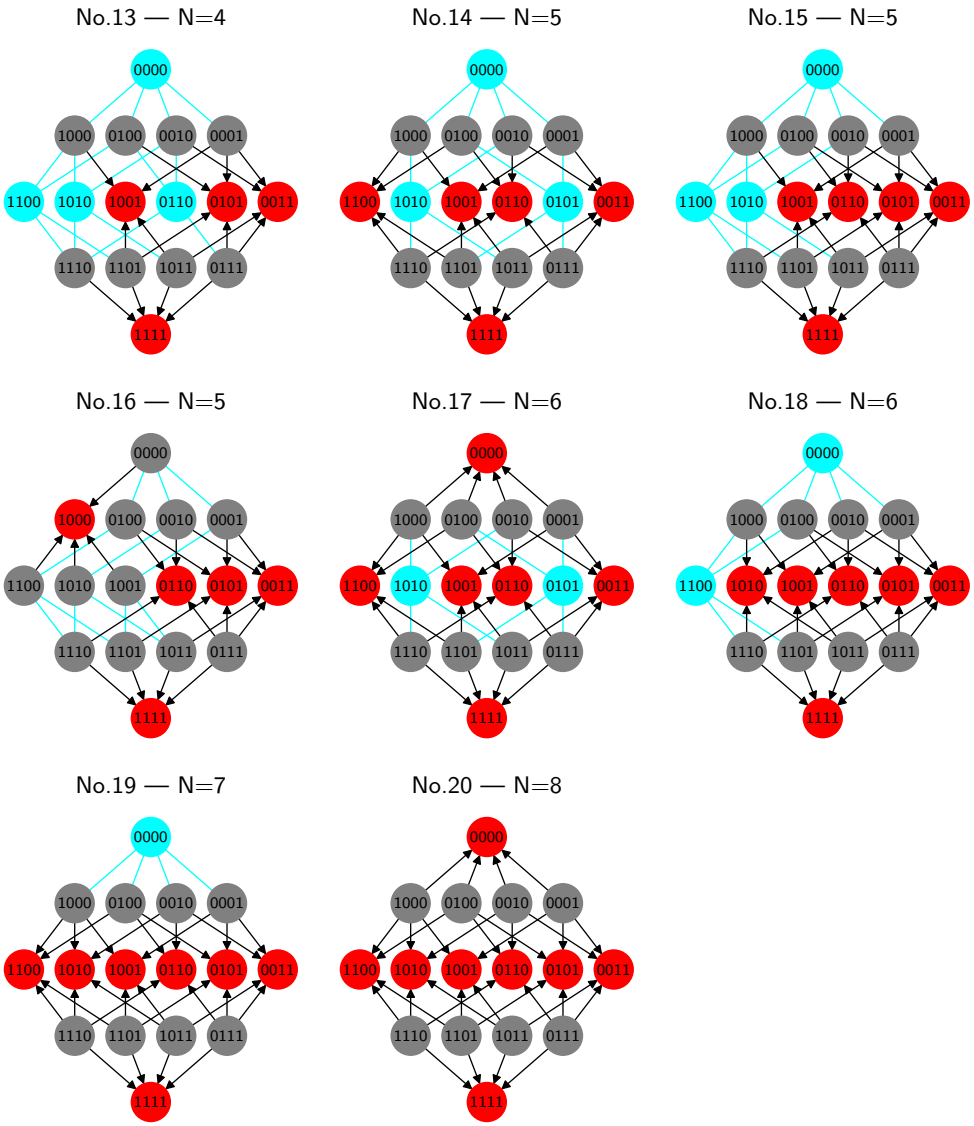


Figure 6: Orthographic graph plots for all possible peak patterns for $L = 4$ from tab.2. *Vertices:* Peaks are colored red, vertices constrained by nearest neighbor peaks grey and unconstrained vertices, by the same condition, cyan. *Edges:* Edges which are constrained by peaks have directed black-arrows and cyan lines are edges which are possibly not constrained by the peak pattern, up to the graphs acyclic condition.

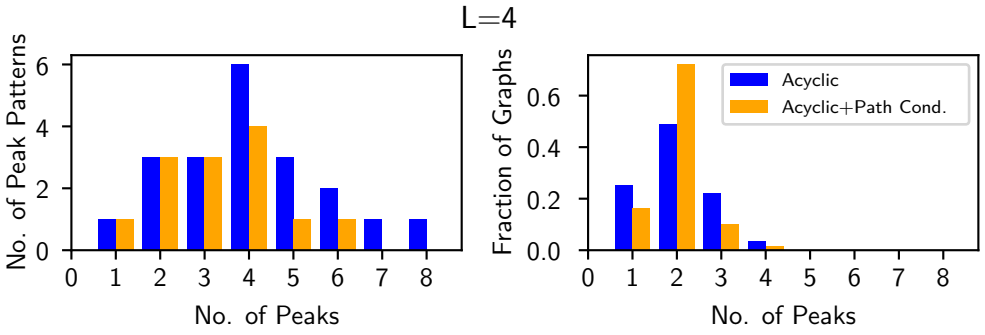


Figure 7: Results for the 4-cube. *Left*: Bar plot of the total number of possible peak patterns. *Right*: The fraction of possible graph configurations for both cases. The total number of graphs for the acyclic case is 193270310, see table 1, and when including the additional path condition 34572.

large amount of different oriented acyclic L-cubes. A course grained analysis can therefore be analyzing the change of peak patterns. Note that from the outlined viewpoint only a single edge changes in every step, hence the number of peaks can only increase or decrease by one or not change at all.
Note that each peak pattern with a fixed number of maxima N constrains the same number of edges, being $N \cdot L$, through the next neighbor condition. Due to the graph structure the same is not true for vertices, compare e.g. No.10 and No.12 in fig.6. Both have 4 peaks, but the former has 0 and the latter 5 unconstrained vertices. Of particular interest are the peak patterns with only incoming edges from the left, but no outgoing edges on the right, as they directly constrain possible fitness values for all vertices. Those are discussed in more detail sec.5.6.
When considering the path condition of sec.3.4 one sees that there are certain paths of peak patterns which can no longer be realized, e.g. from no. $13, 15, 18, 19$ to 20, implied by transparent edges and nodes. This is due to the fact, that if one peak pattern is not possible, all other ones which might be constructed from these aren't possible either. While there are only 20 peak patterns for the 4-cube, the path conditions effect can be observed more clearly for $L = 5, 6$ in sec.5.5.

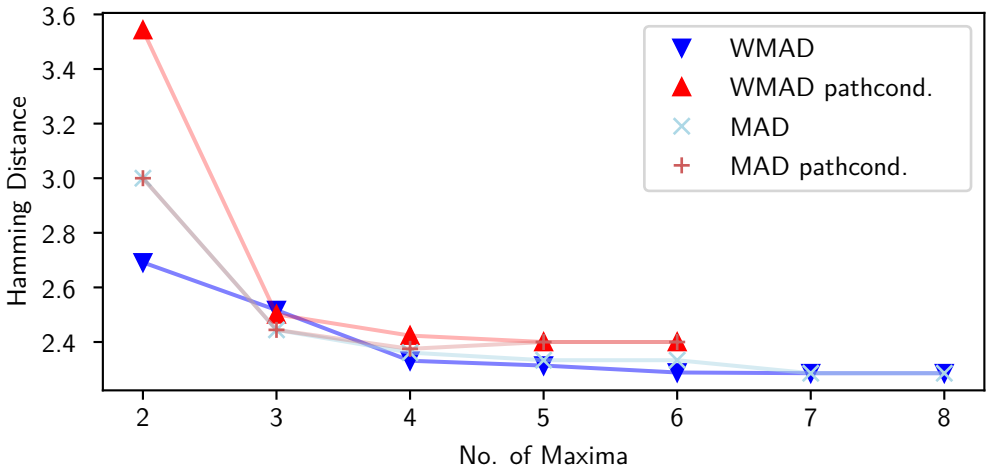


Figure 8: Scatter plot of the mean (weighted) average distance of peak patterns maxima (WMAD) MAD for a given number of maxima. The weight is given by the number of possible graph realizations for each peak pattern. $N = 1$ is not displayed, as it is of no relevance in this setup.

5.2.3 Two-Loci Epistasis

The possible interplay of epistasis and peak patterns is of interest as only looking at the number of peaks might not be enough in order to distinguish potential interactions of mutations. Fig.10 shows every possible combination of peak patterns and SE/RSE of the 4-cube. Each 4-cube has 24 2-faces of which a certain number express SE and RSE. Iterating over all possible oriented acyclic 4-cubes and counting the number of SE/RSE shows an interesting behavior. The left heat map, showing instances of SE, seems more smooth than the right one representing RSE. First, the structure of both heat maps in terms of SE and RSE are discussed, followed by a short note on the overall number of realizations.

SE: Two immediate observations from the left heat map are that the number of 2-faces with SE decreases with an increase in peaks. Also, while being not frequent, there are instances for every peak pattern, having no SE at all, seen in the most left bluish column, but none with one or two 2-faces,

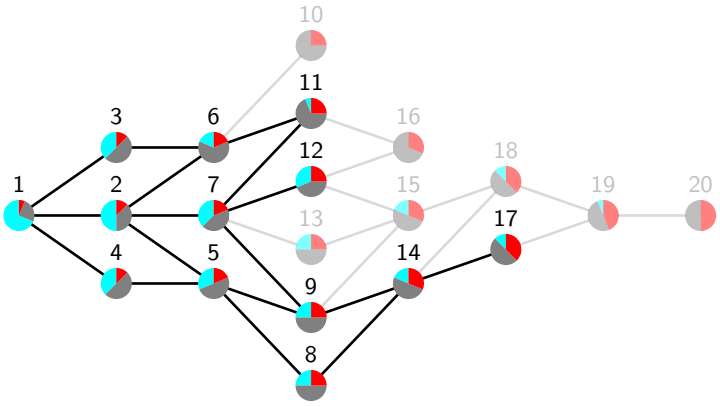


Figure 9: Partially ordered set of the peak patterns from table 2, including the corresponding labels. If two nodes are connected, the peak patterns are the same, up to removing or adding a single peak. The pie chart for each node represents the fraction of maxima of a certain kind. The coloring is the same as in fig.6, with the *cyan* and *gray* nodes representing unconstrained and constrained vertices by the next neighbor condition and peaks being *red*. Transparent nodes are not possible by the path condition laid out in sec.3.4.

seen as two white columns on the left. Additionally there are gaps with a certain number of 2-faces with no SE, while other 4-cube realizations can have more or less 2-faces with SE. For peak pattern No.1, containing only a single peak, there are no instances of 22 and 23 2-faces showing SE, while there are for 21 and 24. Also for six peaks, there is a gap for peak pattern No.17, showing no possible 4-cubes with 5 2-faces displaying SE. This is all the more interesting as both six peaks peak pattern No.17 and 18 constrain the same number of vertices, see fig.9. Further, for a number of peaks with more than one peak pattern, there are different kind of boundaries. On one hand, for 2, 5 and 6 peaks the lower bound, when excluding no SE, is two, and the upper bound 21, 12 and 8 respectively, being sharp. On the other hand, the structure is more complex for 3 and 4 peaks, with the lower bound being the same, but the upper one being not constant. For the three peak

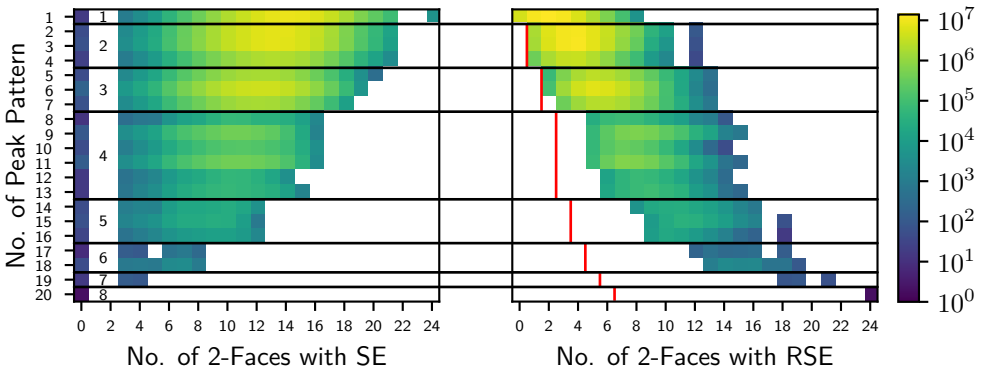


Figure 10: Logarithmic heat map of a 20×25 matrix, with the y-axis being the 20 peak patterns from table 2 in the same order. The x-axis corresponds to the number of 2-faces of the 4-cube showing sign epistasis (SE) or reciprocal sign epistasis (RSE), from 0 to 24. Displayed on the *left* is the count of SE and on the *right* for RSE of the 2-faces for each of the 193270310 oriented acyclic 4-cubes. White spots correspond to zero realizations of SE/RSE. Black horizontal lines separate the peak patterns with a different number of peaks N, which are shown between those lines in the left plot. Red vertical lines on the right plot are the lower bound $N-1$ for the number of 2-faces showing RSE given the number of peaks from [3]. Note that the right plot is discussed and some of it's features partly proven in [3], while the left one is not.

patterns with 3 peaks, No.5, 6, and 7, the upper bound is 20, 19 and 18 respectively. For the six peak patterns having 4 peaks, No.8, 9 and 10 have each an upper bound of 16 2-face epistasis and No.12 and 13 have 14 and 15 respectively. This shows that the cubic configuration of maxima has to be taken into account and not only the number of peaks or their distance when being interested in SE.

RSE: The right heat map shows an increase in RSE when increasing the number of 2-faces. As known from [21] a necessary condition for multiple peaks is RSE, which can be directly observed for the 4-cube. Although the lower bound of $N-1$ 2-faces showing RSE from [3], indicated by red vertical lines, underestimates the occurrences of RSE for $N \geq 4$. This underestimation increases with increasing number of peaks N. A more rough structure than for the left SE heat map can also be observed, with the lower and upper bound of the number of 2-faces showing RSE being less constant. For 3, 4, 5 and 6 peaks the lower bound is not constant, being 2-3, 5-6, 8-9 and 12-13 respectively. The right bound is rough, containing e.g. gaps. Notably for peak patterns No.2, 3 and 4 with 2 peaks, there are no oriented acyclic 4-cubes with 11 2-faces showing RSE, but with 12. A similar behavior is observable for peak pattern no.15 and 16 having 5 peaks and not a single instance of 17 2-faces showing RSE. Again there is a peak pattern disrupting this scheme, as No.14 has at most 16 RSE 2-faces, missing the gap. Peak patterns No.17 and 18 with 6 peaks differ by not showing 17 and 19 for the former and 12 2-face RSE for the latter, but do vice versa. In total, it can be concluded that when observing RSE, one should take into account the underlying peak patterns in order to understand which interactions are theoretically possible and which not.

Realizations: A more detailed figure in regard to the number of instances showing SE or RSE can be found in fig.17 from the appendix. Note that it differs from fig.10, as the y-axis is linear. On one side this makes it more difficult to find if and when SE or RSE occurs for a given number of 2-faces. On the other side it shows clearer that the acyclic instances of the 4-cube showing SE or RSE have an reciprocal behavior with regard to the increase or decrease in number of peaks. Note that each row in both plots of fig.10 and each subplot of fig.17 sums up to the number of realizations in table 2. For peak patterns No.1-16, the bar plot of RSE is more sharply peaked than its SE counterpart, resulting in a higher maxima. For some subplots both bar plots seem to have a bell shape, it is often tilted, and in other cases absent.

5.3 Lower Bound for the Number of Peak Patterns of an L-cube

A lower bound for the number of possible peak patterns $|PP_L|$ can be obtained by comparing the growth of possible peak patterns and of the Hyperoctahedral group $\mathbf{B_L}$.

One can start by calculating a lower bound for peak patterns of an L-cube with N peaks, $|PP_{L,N}|$. Using the known configuration of the Haldane graph, one can simply choose N peaks to get a configuration σ_P. Dividing the number of all possible sets σ_P by the size of $\mathbf{B_L}(\sigma_P)$, which is $2^L L!$ from eq.(9), and taking the ceiling value, results in a lower bound

$$|PP_{L,N}| \geq \left\lceil \frac{1}{2^L L!} \binom{2^{L-1}}{N} \right\rceil \tag{34}$$

for the number of possible peak pattern with N peaks.

Note that $|PP_{L,N}|$ can only be a natural number. If both sides in eq.(34) are equal, the right side is a natural number and if the right side is not a natural number, the left side needs to be larger. Therefore to simplify calculations the ceiling function can be neglected. Taking the sum over all possible numbers of peaks is then a lower bound for $|PP_L|$ and one obtains

$$|PP_L| \geq \sum_{N=1}^{2^{L-1}} \frac{1}{2^L L!} \binom{2^{L-1}}{N} = \frac{2^{2^{L-1}-L}}{L!}. \tag{35}$$

Eq.(35) results in $|PP_5| \geq 17$, $|PP_6| \geq 93206$ and $|PP_7| \geq 2.8 \cdot 10^{13}$. Note that the choice to start with the Haldane graph leaves out configurations of peaks which can not be created from it, e.g. peak patterns which only have incoming edges from the left in fig.9.

An algorithm storing only the bit strings of N peaks, would need at least $N \cdot L$ bits of memory storage per peak pattern, resulting in $\approx 8 \cdot 10^5$ GB as a lower bound for $L = 7$. Concluding that calculating all possible peak patterns is only reasonable up to $L = 6$.

5.4 Peak Pattern Algorithm for Arbitrary L

Calculating all possible peak patterns in a brute force manner by trying out every possible configuration of oriented acyclic L-cubes, as in sec.5.2 for $L = 4$, is not suitable for e.g. $L \geq 5$. Already the corresponding oriented graph for $L = 5$ has 80 edges, resulting in 2^{80} configurations to check when applying brute force methods. Peak patterns can however be calculated when only focusing on the configuration of peaks.

Of particular importance are the following properties in order:

1. Due to the acyclic condition, every graph has at least one peak and one sink. Fixing without loss of generality the WT as peak, and choosing the other edges such that vertices with more 1's have a lower fitness, meaning all arrows down, the graph is acyclic.

2. When now placing peaks on the graph and keeping in mind that two peaks have at least Hamming distance 2, arbitrary peak patterns can be constructed, as adding sinks does not alter the acyclic property.

3. One can see that there is only one peak pattern with one peak and $L - 1$ with two peaks. The latter ones are having Hamming distance $2, 3, \ldots, L$, as two peaks have so to speak no angles between them, which would prohibit the transformation of one configuration to another. Additionally there is only one peak pattern with 2^{L-1} peaks, the Haldane graph.

4. There are in general three types of vertices, see for example fig.6. Peaks (red), constrained vertices with a peak as next neighbor (grey) and unconstrained vertices which can potentially be either of the former (cyan).

5. Having a peak pattern with $N + 1$ peaks and removing one results in a peak pattern with N peaks, which is a sub-structure of the former one. This procedure also works in the opposite way, allowing to create peak patterns with $N + 1$ peaks from peak patterns N peaks.

An algorithm to use above properties to calculate all peak patterns for a given L-cube is as follows:

I Starting from all possible peak patterns p_{max} with N peaks, e.g. the known ones for $N = 1$ or $N = 2$ from prop.3. Those seed peak patterns can be seen as sub-structures of peak patterns with $N+1$ maxima, with one maxima being removed. Create a list $\mathbf{a}$ of peak patterns with $N+1$ maxima, which is empty at first.

II Choose the p'th (1 at start for a given N) peak pattern with N maxima. Compute a list of all u_{max} unconstrained vertices $\mathbf{u}$.

III Add a maxima to the u'th (1 at start for a given p) unconstrained vertex, resulting in a valid peak pattern with $N + 1$ maxima.

IV Compute its normal form from sec.2.7 and compare it to all known ones in $\mathbf{a}$. Add to $\mathbf{a}$ when unknown, discard if known.

V If $u < u_{max}$ increase u by one and go back to step III.

VI ($u = u_{max}$, checked all peak patterns which can be created by adding a maxima to peak pattern p) If $p < p_{max}$ increase p by one and go back to step II.

VII ($p = p_{max}$, checked all peak patterns with N maxima) If $N < 2^{L-1}$ go to step I and increase N by one.

VIII All possible peak patterns of the L-cube have been computed.

5.5 Peak Patterns for the 5- and 6-cube

Using the algorithm from sec.5.4 and taking into account the lower bound on the number of peak patterns from sec.5.3, it is feasible to calculate all possible peak patterns up to $L = 6$.
Logarithmic bar plots for the distribution of peak patterns given the number of peaks can be found in fig.11.

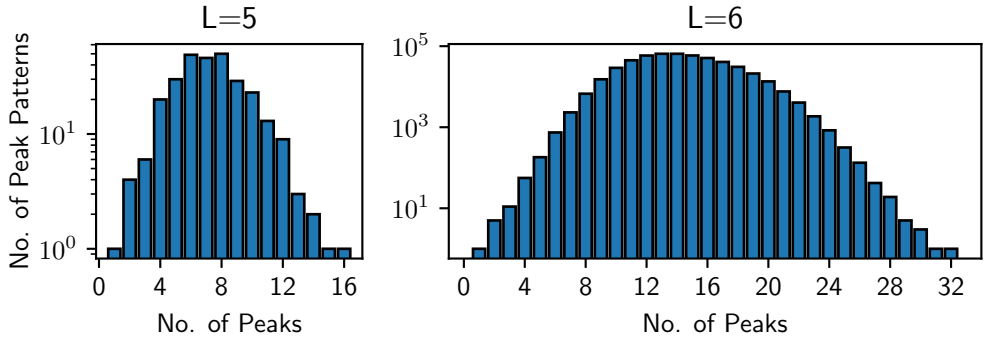


Figure 11: *Left:* Bar plot for peak patterns of the 5-cube, with the total number of peak patterns being 287. The maximal number of peak patterns for a given number of peaks is at $N = 8$. *Right:* Bar plot for peak patterns of the 6-cube, with the total number of peak patterns being 519194. The maximal number of peak patterns for a given number of peaks is at $N = 13$.

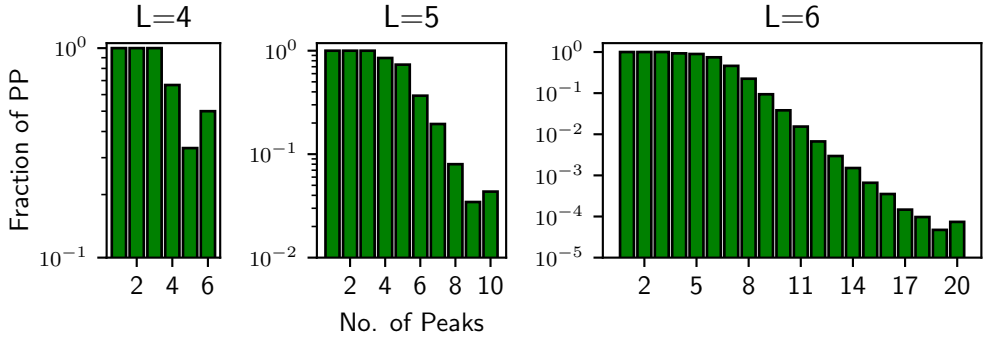


Figure 12: *From left to right:* Logarithmic bar plots displaying the fraction of peak patterns for the 4-, 5- and 6-cube which are compatible with the path condition from sec.3.4, namely 13 (65%), 83 ($\approx 28.92\%$) and 7348 ($\approx 0.01\%$) respectively. The maximal number of peaks is given by eq.(33).

5-Cube: For the 5-cube there are in total 287 peak patterns. The distribution given the number of maxima N seems slightly tilted to the left. This is to be expected, as with every additional peak the graph has more constraints for vertices becoming peaks themselves due to the next neighbor condition. What stands out is a sudden drop for $N = 7$, hinting that the next neighbor condition creates a complex interaction between maxima.

6-Cube: There are 519194 peak patterns for the 6-cube in total. As for $L = 5$ the distribution is slightly tilted for the same reason and has its maximum at $N = 13$. It appears also smoother and has a single peak in comparison to $L = 5$.

When taking into account the path condition from sec.3.4 the number of possible peak patterns changes significantly. The fraction of these peak patterns, given the number of peaks, is shown in fig.12.
One notices that the shape of all three plots is similar, as they each show an initial constant period, followed by and exponential decrease in the fraction of suitable peak patterns with N_{max} being an exception. Note that for each of the 4-,5- and 6-cube, all peak patterns having $N = 1, 2$ or 3 peaks are compatible with the path condition. Also observable in all three plots is the sudden increase for the maximum number of possible peaks $N_{max} = \binom{L}{\lfloor L/2 \rfloor}$, see eq.(33). This is due to the fact, that the peak pattern with $N = N_{max}$, of which there is only one, has only peaks with the same number m of ones. When removing a single peak it is not possible to move all remaining $N_{max}-1$ peaks to a set of vertices with either $m - 1$ or $m + 1$ ones, because as there is no such set. The next largest set of vertices with the same number of ones in it has $\binom{L}{L/2\pm1}$ vertices for an even L and $\binom{L}{\lfloor L/2 \rfloor -1}$ or $\binom{L}{\lceil L/2 \rceil +1}$ vertices for an uneven L, of which none has $N_{max} - 1$ vertices. Hence there is only one suitable peak pattern with $N_{max} - 1$ peaks. This peak pattern is the same up to a single peak as the single one with N_{max} peaks. But as known from fig.7 and fig.11, the number of overall peak patterns decreases from $N_{max} - 1$ to N_{max}, resulting in a smaller fraction.
Overall all three plots in fig.12 show interesting similarities, while the num-

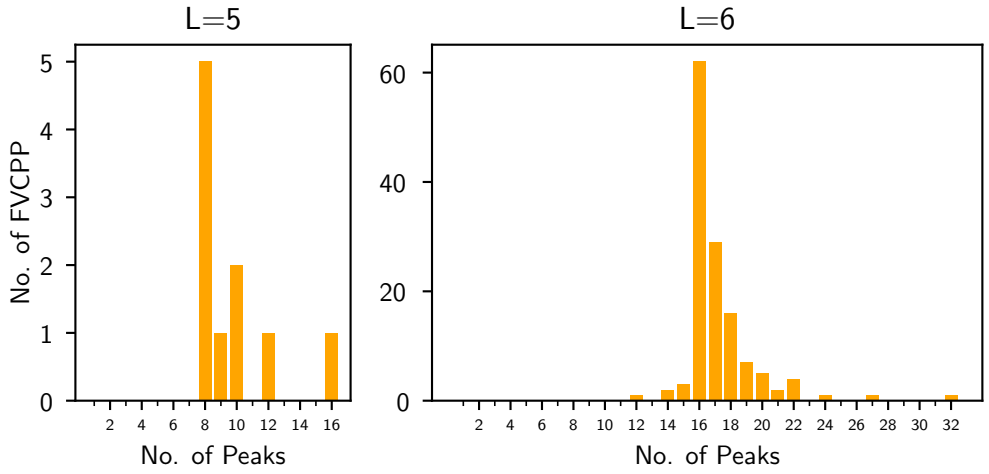


Figure 13: Bar plots for the number of fully vertex constrained peak patterns (fvcpp), *left* being the 5-cube, totaling 10 fvcpp, and *right* the 6-cube, having 134 fvcpp.

ber of vertices increases exponentially and of peak patterns at least double-exponentially, see eq.(35).

5.6 Fully Vertex Constrained Peak Patterns

A subset of peak patterns that stands out are the ones in which every vertex is either a peak or can't be one due to the next neighbor condition, referred to as *fully vertex constrained peak patterns (fvcpp)*. While the Haldane graph is a trivial example of such, with every vertex being either a peak or minima, peak patterns with less peaks are not.
In table 2 they have zero unconstrained vertices and can be found as leaves in fig.9. For the 4-cube there are three of these peak patterns (15%) out of the 20 possible ones, No.10, 16 and 20 with 4, 5 and 8 peaks respectively.
For the 5- and 6-cube there are 10 and 134 fvcpp, $\approx 3.48\%$ and $\approx 0.02\%$ of all possible peak patterns respectively. Bar plots matching the number of peaks and fvcpp are shown in fig.13. Both bar plots have a clear maximum for $8 = 2^3$ and $16 = 2^4$ peaks. A general structure has not been observed,

although this result can also be written as 2^{L-2} for $L = 5$ and $L = 6$. A lower and upper bound for the minimal number of peaks for an L-cube can be calculated.

Lower bound: As each peak has L neighbors, this results in a lower bound of $\left\lfloor \frac{2^L}{L+1} \right\rfloor$, while not taking into account the L-cubes graph structure, hence not guarantying existence. Note that this result is also known as the mean value of peaks for the *House of Cards* model [19].

Upper bound: As described in eq.(1), constructing an L-cube out of two $(L-1)$-cubes can be done by connecting them via the graph cartesian product. A more visual way of the graph cartesian product acting on two $(L-1)$-cubes is as follows: Add an additional coordinate for each vertex and set it 0 for one and 1 for the other $(L-1)$-cube. Afterwards connect each pair of vertices which are the same up to the added coordinate by an edge. Applying this procedure and constructing the 4-cube out of two 3-cubes can be seen in fig.1. The lower left plots inner 3-cube has 0 for its last bit coordinate, while for the outer 3-cube it is 1.
The peak pattern with 2 maxima in distance 3 of the 3-cube is a fvcpp and has its peaks at some of the 3-cubes diagonal opposite vertices. Taking now a similar one, with its peaks being at a different diagonal allows one to create a fvcpp with 4 peaks. By this scheme one can get higher dimensional fvcpp, resulting in an upper bound of 2^{L-2} peaks for the minimal number of peaks for an fvcpp of an L-cube.

It is interesting that there are no fully constrained peak patterns for a small number of peaks and only a few for the opposite side with a lot of peaks. The first case is reasonable due to the lower bound above. The latter case has at least two factors involved. First, fig.11 shows that there is a several magnitude difference in the number of peak patterns having e.g. 16 and 28 peaks. Hence, leaving less peak patterns to possibly be a fvcpp. Second, once an L-cube has a high number of peaks, it leaves few options for more peaks. This is due to the next neighbor condition, which constrains more vertices

the more peaks are present and in turn leaves few vertices which have a lot of unconstrained vertices as neighbors.
Standing out is the fvcpp having the smallest number of peaks for the 6-cube, namely 12. A plot of this fvcpp can be found in the appendix fig.18.

6 The Staircase Triangulation, Universal Positive Epistasis and Peak Patterns

As laid out in sec.4.3 upper bounds for the maximal number of peaks for the staircase triangulation are known from [6] for $L = 4$ and $L = 5$. This section combines the peak pattern approach and the staircase triangulation. It starts with an algorithm of how to check if a peak pattern is compatible with eq.(30) and subsequently calculates all compatible peak patterns for $L = 4$ to $L = 8$.

6.1 Algorithm for Staircase Triangulation Compatible Peak Patterns

Eq.(30) can be used as a necessary condition to check which peak patterns are compatible. It is not a sufficient condition, as peak patterns encode only the direction of $N \cdot L$ edges, up to the acyclic condition, and no numerical fitness values to compute the triangulation.
A peak pattern σ^P is said to be compatible in the sense of eq.(30), if there exists an element in its orbit, $\sigma^F \in \mathbf{B_L}(\sigma^P)$, see eq.(10), for which none of the $\frac{N(N-1)}{2}$ pairs of peaks $(\vec{\sigma}_g, \vec{\sigma}_{g'})$ Hamming distances

$$\begin{aligned} d_{g,\cup} &= \Delta(\vec{\sigma}_g, \vec{\sigma}_{g\cup g'}) \\ d_{g',\cup} &= \Delta(\vec{\sigma}_{g'}, \vec{\sigma}_{g\cup g'}) \\ d_{g,\cap} &= \Delta(\vec{\sigma}_g, \vec{\sigma}_{g\cap g'}) \\ d_{g',\cap} &= \Delta(\vec{\sigma}_{g'}, \vec{\sigma}_{g\cap g'}) \end{aligned} \tag{36}$$

are allowed to fulfill

$$(d_{g,\cup} = 1 \wedge d_{g',\cap} = 1) \vee (d_{g',\cup} = 1 \wedge d_{g,\cap} = 1). \tag{37}$$

If a pair does fulfill eq.(37), it means that the sum of the fitness values of $\vec{\sigma}_{g\cup g'}$ and $\vec{\sigma}_{g\cap g'}$, having distance one to the peaks $\vec{\sigma}_g$ and $\vec{\sigma}_{g'}$, are greater than the peaks fitness sum. Hence eq.(30) can't be true, as peaks must have a higher fitness than their neighbors. This is a generalization of an idea in [6].

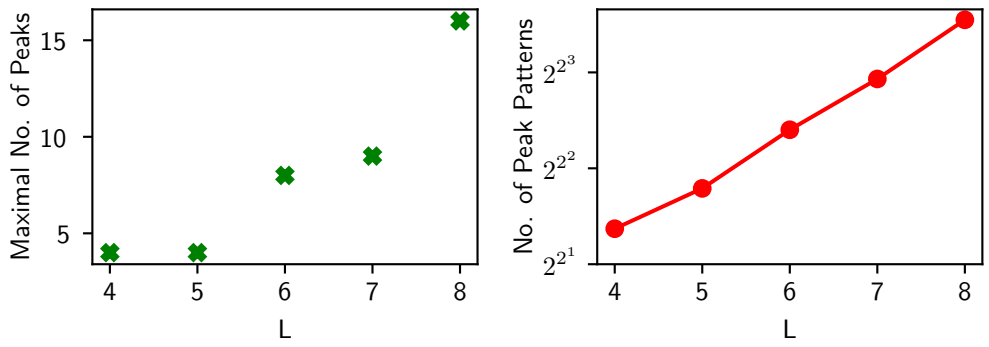


Figure 14: Peak Pattern that are compatible with the staircase triangulation of the corresponding L-cube. *Left:* Maximal number of peaks of the peak patterns. *Right:* Number of peak patterns, note the double exponential y-scale.

From sec.5.5 the peak patterns up to $L = 6$ are known and can be checked individually if they are compatible with the staircase triangulation or not. While eq.(35) gives a lower bound when calculating all possible peak patterns, it is not relevant setting. Note that for each pair of peaks eq.(37) needs to be checked, therefore only adding a peak to a peak pattern that is compatible can lead to another compatible peak pattern. This is due to the fact, that for each pair g and g' eq.(37) needs to be checked and if a pair does not fulfill this property, adding another vertex does not change this. Resulting in an enormous reduction in the computation of peak patterns with $N + 1$ peaks from peak patterns with N peaks. Additionally, checking if any $\sigma^F \in \mathbf{B_L}(\sigma^P)$ does fulfill the conditions above can be a bottleneck, as checking all elements in $\mathbf{B_L}(\sigma^P)$ can result in a complexity of $\mathcal{O}(L!2^L)$, see eq.(9). However, the permutations part of $\mathbf{B_L}$, being of complexity $\mathcal{O}(L!)$ doesn't need to be checked. This is due to the fact that the compatibility check of eq.(30) is only based on the number of 1-alleles, which is not changed by permutation.

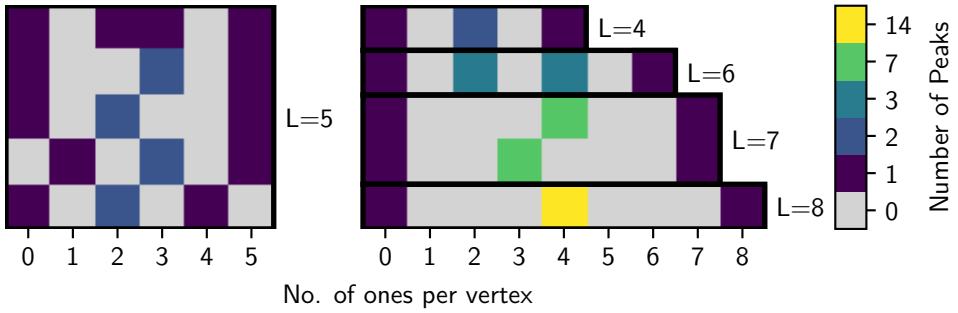


Figure 15: All possible different layer configurations of peak patterns compatible with the staircase triangulation and having the maximal number of peaks. Shown are the cases for $L = 4$ up to $L = 8$.

6.2 Compatible Peak Patterns up to $L = 8$

Using as a seed only peak patterns which are compatible for the algorithm described in sec.5.4, all compatible peak patterns from $L = 4$ up to $L = 8$ have been computed. The maximal number of peaks and the overall number of those peak patterns can be seen in fig.14. The right plot has a double exponential scale, which seems in line with eq.(35). There are however not enough data-points to substantiate this statement. If it is true though, it limits substantially the number of staircase triangulation peak patterns that seem reasonable to calculate with the outlined algorithm for higher L. The left plot, showing the maximal number of peaks, does not seem to show a pattern, but the corresponding peak patterns do.

Let σ_S be a set of peaks that is compatible with the staircase triangulation as outlined before, then at least one element in $\mathbf{B_L}(\sigma^S)$ is compatible. As the difficulty of interpreting peak pattern plots using the full L-cube increases with higher L, omitting some of the information comes at hand. Instead of plotting every single vertex and edge, clustering all peaks with an equal number of ones into the same layer is performed. There are $L + 1$ layers, labeled by their distance m from $\vec{0}$, $m = \Delta_0(\vec{\sigma})$. All possible unique layer configurations for $L = 4$ up to $L = 8$, having the maximal number of peaks, are shown in fig.15. There are $1, 5, 1, 2, 1$ possible configurations for the 4- to

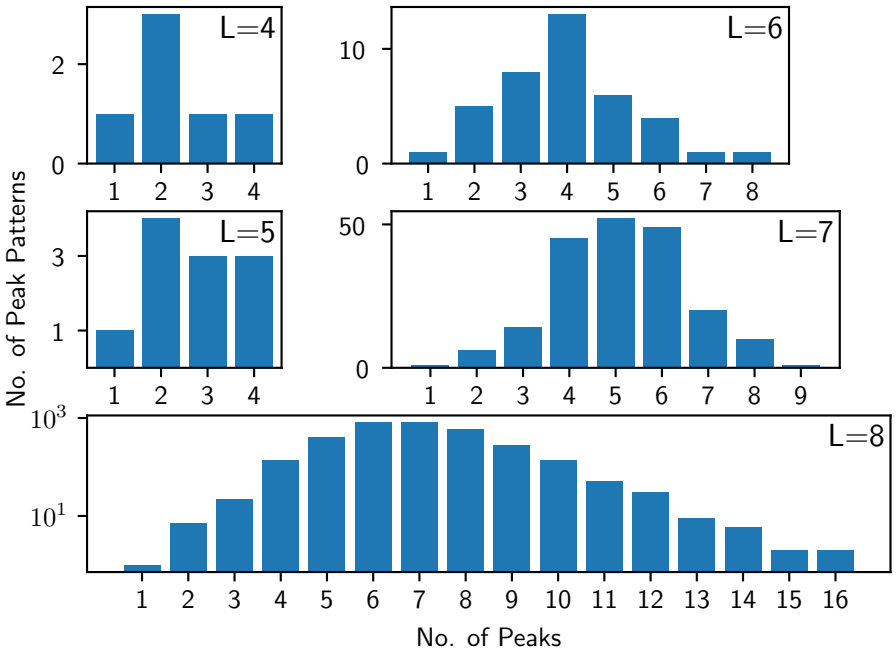


Figure 16: Bar plot for the number of peak patterns that are compatible with the staircase triangulation for the 4− to 8-cube. Note the logarithmic y-scale for $L = 8$.

8-cube respectively. For $L = 4, 7$ and 8, peaks are only at $m = 0, L$ and at layers having the most vertices, $m = \frac{L}{2}$ for $L = 4$ and 8 and at $m = \left\lfloor \frac{L}{2} \right\rfloor, \left\lceil \frac{L}{2} \right\rceil$ for $L = 7$. With its 5 different realizations, the $L = 5$ case stands out, where every layer has at least one peak in at least one pattern. This is likely due to the fact, that there are only 4 peaks in total, being the same as for $L = 4$, but with double the number of vertices to choose from. With only one realization and peaks at $m = 0, 6, \frac{L}{2} \pm 1$, the 6-cube has again a different pattern.
Bar plots of peak patterns given the number of peaks N are shown in fig.16. As outlined in [6], there are at maximum 4 peaks for $L = 4$ and $L = 5$ and the three different possible configurations for $L = 5$ are also given. However there seems to be no immediate structure between the number of peak patterns having the maximal number of peaks and the number of their

layer configurations from fig.15. There are $1, 3, 1, 1, 2$ peak patterns for the 4- to 8-cube respectively, while having $1, 5, 1, 2, 1$ layer patterns.

7 Results and Outlook

This section is a summary of this thesis and points out difficulties, solutions and possible further research questions and directions.

Computation of Peak Patterns: The approach of checking all oriented acyclic 4-cubes, and thereby encountering each possible peak pattern at least once, is a straightforward method for calculating all 4-cube peak patterns from fig.6. It is however a slow one and practically not viable for $L \geq 5$, as for e.g. $L = 5$ this would require checking 2^{80} oriented 5-cubes, see sec.2.3. Using the peak pattern normal from sec.2.7 in combination with the insights and corresponding algorithm from sec.5.4 turned out to be a fast and efficient method for calculating all peak patterns from $L = 4$ to $L = 6$. Additionally knowing from sec.5.3 that for $L = 7$ there are at least $2.8 \cdot 10^{13}$ possible peak patterns led to the conclusion that calculating all the peak patterns for a given $L \geq 7$ is not a viable approach. However, the calculation of a peak patterns normal form turned out to be the computationally most expensive part for $L \leq 6$. As the number of peak patterns with N peaks is at least $\left\lceil \frac{1}{2^L L!} \binom{2^{L-1}}{N} \right\rceil$ from eq.(34), comparing a peak patterns normal form to all other already found ones, while storing them in a simple list as in the algorithm used from sec.5.4, also becomes computationally expensive. Using e.g. eq.(16) to cluster the known peak patterns into batches would provide a speedup, as each new peak pattern would only get compared to a subset of the already found ones. The resulting batches can be compared to each other after all peak patterns with $N+1$ peaks have been computed from the ones with N peaks. Optimizing both parts might allow one to calculate the minimal number of peaks for a fully vertex constrained peak patterns (fvcpp) for $L = 7$ and all peak patterns compatible with the staircase triangulation for $L = 9$.

Peak Patterns: The number of peak patterns for $L = 1, 2$ and 3 are $1, 2$ and 5 respectively from [6] and 20 for $L = 4$ from fig.6, as well as 287 for $L = 5$ and 519194 for $L = 6$ from fig.11. This results in the integer sequence

for the number of peak patterns given L to be $1, 2, 5, 20, 287, 519194$ for $L = 1$ to $L = 6$. At the time of submission of this thesis, this sequence was absent in the *OEIS - On-Line Encyclopedia of Integer Sequences*[1], a website storing integer sequences from various fields. Moreover peak patterns turned out to have an interesting structure, as for the same number of peaks the number of constrained vertices can vary. This led to the definition of fvcpp in sec.5.6. The result that for a fvcpp of the 6-cube the minimal number of peaks is $N = 12$, see fig.18, came as a surprise. With the next possible fvcpp having $N = 14$ peaks, it highlights the range of patterns that can be formed by peaks constraining the fitness of their next neighbors.

Peak patterns have not been applied to empirical data in this work, but fig.10 shows that it might not be sufficient in general to look at the number of peaks alone, but rather at their peak patterns. Already for $L = 4$ some peak patterns, with the same number of peaks, show interesting behavior for sign epistasis (SE) and reciprocal sign epistasis (RSE). While for some peak patterns some instances of SE or RSE are possible, they are not possible for others having the same number of peaks. Nonetheless applying the idea of peak patterns to empirical data is an important step to potentially show their relevance for real biological systems. In order to develop the peak pattern approach further, additional steps are necessary. For a biallelic system empirical data might not be available as a full set of all 2^L fitness values due to experimental or biological reasons. Going further, the system of interest could have not only biallelic loci. Hence further work is needed, including the generalization of peak patterns for subsets of the L-cube, as well as for other than collections of biallelic loci. Regarding the partially ordered set (poset) of peak patterns by inclusion from sec.5.2.2 no analysis has been performed for the $L = 5$ and $L = 6$ cases.

Potentially leaving the weak mutation strong selection regime, that is usually considered when using the L-cube with edges between genotypes of Hamming distance one, might also be an option. When connecting e.g. genotypes with Hamming distance two, there may also be a peak pattern-like structure to

[1] `https://oeis.org/`

observe.

Fishers Geometric Model: The exponential decrease in peak patterns fulfilling the path condition in fig.12 has similarities between the 4-, 5- and 6-cube, while having vastly different numbers of peak patterns for the respective L-cube. All peak patterns with $N = 1, 2$ and 3 peaks notably do fit the variant of FGM. However an analysis for a higher dimensional phenotype space, $n \geq 2$, has not been performed, but is achievable using the calculated peak patterns when having a path condition for these cases at hand. It might also be worth generalizing the poset from fig.9 for the 4-cube to the 5- and 6-cube. Observing if models of real evolution experiments follow the allowed paths on the poset, could give insights that potentially point to the necessity of adapting the model used. Additionally an increase in distance between peaks when including the path condition in fig.8 has been observed, with three peaks being an exception.

Epistasis: One of the most notable plots is fig.10, as it shows that taking into account the spatial configuration of peak patterns and not only their number is possibly a necessary step to find a lower bound for the number of instances of RSE. Moreover it might still be feasible to compute the $L = 5$ case in a similar manner, possibly showing an even more diverse structure due to its 287 different peak patterns. Also looking at subsets of peak patterns such as the fvcpp could provide insight into how the distance between peaks influence SE or RSE. With only 10 fvcpp for the 5-cube, the 6-cube might be more significant in this manner. For its 134 fvcpp certain edges are already fixed du to neighboring peaks. One could choose the remaining ones randomly in order to create the full fitness graph, check if they are acyclic and calculate their instances of SE or RSE. Note that it becomes less likely to get an acyclic graph when choosing edge directions by e.g. a 50% chance for each of both possible directions. With the 5- and 6-cube already having 80 and 192 edges, see eq.(3), choosing a peak pattern preliminary instead of creating all edges at random and checking the peak pattern later on, the number of random edges gets reduced by $N \cdot L$.

Staircase Triangulation: An application for peak patterns is to find the maximal number of peaks for the staircase triangulation. The algorithm from sec.5.4 in combination with the insights of looking at peak patterns as a poset allowed the extension of the former limit, being at $L = 5$, to higher L. Calculations up to $L = 8$ are included in sec.6.2. While the inequality provided by universal positive epistasis from eq.(30) was only checked for pairs of peaks, hence being only a necessary condition rather than a sufficient one, it still led to interesting insights. Fig.14 shows the maximal number of peaks for $L = 4$ up to $L = 8$, as well as the overall number of compatible peak patterns. The latter notably shows a potentially double exponential behavior, just as the upper bound for peak patterns in eq.(35). As the staircase triangulation is highly symmetric, due to only using one type of simplex, the methods to check if a peak pattern is compatible with a triangulation could in principle be extended to other triangulations. This however depends on the inequalities defining them, which would need to allow to make use of peaks and their next neighbors with lower fitness. The switching from a peak pattern standpoint to a layered one in sec.6.2 and shown fig.15 marks an important last step in this thesis. It shows that it is suitable for larger L to switch to a different representation when trying to find structure in the L-cubes peak patterns, as it becomes increasingly difficult to interpret plots of higher dimensional L-cubes. Hence the final remark is that peak patterns are suitable for some range of L, depending on the purpose, but further reduction of the information stored in the full oriented acyclic L-cube is needed for larger systems.

8 Methods

8.1 Encoding of Oriented L-cubes

Each oriented L-cube has $N_e = Q(1, L)$ edges and needs N_e-bits to be stored. A unique map between those edges can be constructed, up to choosing an origin which corresponds to a vertex which will be labeled by only zeros in the L-cubes vertex bit representation. Reading the N_e-bit word c from left to right, like all following bit strings, the n-th bit encodes the direction of an edge between vertices v_l and v_h. Within the vertex bit representation bit number k is 0 for v_l and 1 for v_h. Removing bit number k results in an $L-1$ bit string with integer value r. The n-th bit position on the N_e-bit string is then calculated by

$$n = 2^{L-1}(k-1) + r + 1, \tag{38}$$

while the bits value is 0 for an directed edge from $v_l \to v_h$ and 1 for $v_l \leftarrow v_h$. This idea from the GitHub repository of Devin Greene, found in [8].
The N_e-bit string is translated from base 2 into an unsigned integer x of base 10, while the number of edges is represented in the subscript, e.g. x_{N_e}. Fig.1 shows examples for oriented acyclic 3- and 4-cubes. If a graph has fixed edges due to boundary conditions, it is represented by an N_e-bit string nonetheless. For example consider the upper left plot from fig.1. First calculate n for each edge, see (39), and subsequently the values of c_i with $i \in \{1, \cdots, 12\}$. The resulting bit string is $c = 010101010001$. Using base 10 instead of 2 for c results in $\sum_{n=1}^{12} c_n 2^{n-1} = 2218_{12}$.

$$
\begin{aligned}
000 \to 100: &\quad k=1, r=0 \implies n=1, &\quad c_1 = 0 \\
000 \to 010: &\quad k=2, r=0 \implies n=5, &\quad c_5 = 0 \\
000 \to 001: &\quad k=3, r=0 \implies n=9, &\quad c_9 = 0 \\
100 \leftarrow 110: &\quad k=2, r=1 \implies n=6, &\quad c_6 = 1 \\
100 \to 101: &\quad k=3, r=1 \implies n=10, &\quad c_{10} = 0 \\
010 \leftarrow 110: &\quad k=1, r=1 \implies n=2, &\quad c_2 = 1 \\
010 \to 011: &\quad k=3, r=2 \implies n=11, &\quad c_{11} = 0 \\
110 \leftarrow 111: &\quad k=3, r=3 \implies n=12, &\quad c_{12} = 1 \\
001 \to 101: &\quad k=1, r=2 \implies n=3, &\quad c_3 = 0 \\
001 \to 011: &\quad k=2, r=2 \implies n=7, &\quad c_7 = 0 \\
101 \leftarrow 111: &\quad k=2, r=3 \implies n=8, &\quad c_8 = 1 \\
011 \leftarrow 111: &\quad k=1, r=3 \implies n=4, &\quad c_4 = 1
\end{aligned}
\tag{39}
$$

8.2 Checking Graph Properties

Using the encoding from sec.8.1 and the number of oriented L-cubes from sec.2.3, each possible oriented 3-cube, iterating from 0_{12} to 4095_{12} $(2^{12}-1)$, and oriented 4-cube, from 0_{32} to 4294967295_{32} $(2^{32}-1)$, is being checked for three properties:

P.1 Is it acyclic?

P.2 Which peak pattern does it correspond to?

P.3 Does it fulfill the path condition from sec.3.3?

If **P.1** is computed to be false, the next oriented graph is checked. The function *is_cyclic* is used, with its algorithm using *depth-first search*, from [12]. As the number and configuration of peak patterns is not known from the beginning for **P.2**, each oriented graph gets checked and if a formerly unknown peak pattern arises it gets saved to a list. Then each following graphs peak pattern gets compared to the peak patterns within that list.
To check **P.3**, all possible $L!$ paths from the WT to the FM for the path

condition are being checked.
Note that distributed computing is being used for $L = 4$ in order to reduce computation time. The 2^{32} possible combinations are split into batches, computed individually and joined for subsequent computations and analysis.

References

[1] Niko Beerenwinkel, Lior Pachter, and Bernd Sturmfels. "EPISTASIS AND SHAPES OF FITNESS LANDSCAPES". In: *Statistica Sinica* 17.4 (2007), pp. 1317–1342. ISSN: 10170405, 19968507. URL: http://www.jstor.org/stable/24307677.

[2] William Y.C. Chen and Richard P. Stanley. "Derangements on the n-cube". In: *Discrete Mathematics* 115.1-3 (May 1993), pp. 65–75. DOI: 10.1016/0012-365x(93)90479-d. URL: https://doi.org/10.1016/0012-365x(93)90479-d.

[3] Nate Chenette et al. *OCCURRENCES OF RECIPROCAL SIGN EPISTASIS IN SINGLE AND MULTI-PEAKED THEORETICAL FITNESS LANDSCAPES*. Preprint on webpage at https://faculty.valpo.edu/lpudwell/papers/RSEs.pdf, last accessed 06.04.2022. Jan. 2022.

[4] Kristina Crona. "Polytopes, Graphs and Fitness Landscapes". In: *Recent Advances in the Theory and Application of Fitness Landscapes*. Springer Berlin Heidelberg, 2014, pp. 177–205. DOI: 10.1007/978-3-642-41888-4_7. URL: https://doi.org/10.1007/978-3-642-41888-4_7.

[5] Kristina Crona, Devin Greene, and Miriam Barlow. "The peaks and geometry of fitness landscapes". In: *Journal of Theoretical Biology* 317 (Jan. 2013), pp. 1–10. DOI: 10.1016/j.jtbi.2012.09.028. URL: https://doi.org/10.1016/j.jtbi.2012.09.028.

[6] Kristina Crona, Joachim Krug, and Malvika Srivastava. "Geometry of fitness landscapes: Peaks, shapes and universal positive epistasis". In: (2021). DOI: 10.48550/ARXIV.2105.08469. URL: https://arxiv.org/abs/2105.08469.

[7] Kristina Crona, Mengming Luo, and Devin Greene. "An uncertainty law for microbial evolution". In: *Journal of Theoretical Biology* 489 (Mar. 2020), p. 110155. DOI: 10.1016/j.jtbi.2020.110155. URL: https://doi.org/10.1016/j.jtbi.2020.110155.

[8] Kristina Crona et al. “Inferring genetic interactions from comparative fitness data”. In: *eLife* 6 (Dec. 2017). DOI: 10.7554/elife.28629. URL: https://doi.org/10.7554/elife.28629.

[9] Suman G Das et al. “Predictable properties of fitness landscapes induced by adaptational tradeoffs”. In: *eLife* 9 (May 2020). DOI: 10.7554/elife.55155. URL: https://doi.org/10.7554/elife.55155.

[10] Anthony Delgado et al. “Subcubes of hypercubes”. In: *Congressus Numerantium* 189 (Jan. 2008), pp. 25–32.

[11] Konrad Engel. *Sperner Theory*. Cambridge University Press, Jan. 1997, p. 1. DOI: 10.1017/cbo9780511574719. URL: https://doi.org/10.1017/cbo9780511574719.

[12] James Fairbanks et al. *JuliaGraphs/Graphs.jl: an optimized graphs package for the Julia programming language*. 2021. URL: https://github.com/JuliaGraphs/Graphs.jl/.

[13] Ronald Aylmer Fisher. *The genetical theory of natural selection*. Clarendon Press, 1930. DOI: 10.5962/bhl.title.27468. URL: https://doi.org/10.5962/bhl.title.27468.

[14] J. B. S. Haldane. “A Mathematical Theory of Natural Selection. Part VIII. Metastable Populations”. In: *Mathematical Proceedings of the Cambridge Philosophical Society* 27.1 (Jan. 1931), pp. 137–142. DOI: 10.1017/s0305004100009439. URL: https://doi.org/10.1017/s0305004100009439.

[15] Andrew Howroyd. *Irregular triangle read by rows*. Entry A334159 in The On-Line Encyclopedia of Integer Sequences. Apr. 2020. URL: https://oeis.org/A334159.

[16] Peter Huggins et al. “The hyperdeterminant and triangulations of the 4-cube”. In: *Mathematics of Computation* 77.263 (Sept. 2008), pp. 1653–1679. DOI: 10.1090/s0025-5718-08-02073-5. URL: https://doi.org/10.1090/s0025-5718-08-02073-5.

[17] Sungmin Hwang, Su-Chan Park, and Joachim Krug. "Genotypic Complexity of Fisher's Geometric Model". In: *Genetics* 206.2 (June 2017), pp. 1049–1079. DOI: 10.1534/genetics.116.199497. URL: https://doi.org/10.1534/genetics.116.199497.

[18] Jesús A. De Loera, Jörg Rambau, and Francisco Santos. *Triangulations*. Springer Berlin Heidelberg, 2010. DOI: 10.1007/978-3-642-12971-1. URL: https://doi.org/10.1007/978-3-642-12971-1.

[19] C A Macken and A S Perelson. "Protein evolution on rugged landscapes." In: *Proceedings of the National Academy of Sciences* 86.16 (Aug. 1989), pp. 6191–6195. DOI: 10.1073/pnas.86.16.6191. URL: https://doi.org/10.1073/pnas.86.16.6191.

[20] Jeanne Pellerin, Kilian Verhetsel, and Jean-FranÇois Remacle. "There Are 174 Subdivisions of the Hexahedron into Tetrahedra". In: *ACM Trans. Graph.* 37.6 (Dec. 2018). ISSN: 0730-0301. DOI: 10.1145/3272127.3275037. URL: https://doi.org/10.1145/3272127.3275037.

[21] Frank J. Poelwijk et al. "Reciprocal sign epistasis is a necessary condition for multi-peaked fitness landscapes". In: *Journal of Theoretical Biology* 272.1 (Mar. 2011), pp. 141–144. DOI: 10.1016/j.jtbi.2010.12.015. URL: https://doi.org/10.1016/j.jtbi.2010.12.015.

[22] Richard P. Stanley. "Acyclic orientations of graphs". In: *Discrete Mathematics* 5.2 (1973), pp. 171–178. DOI: 10.1016/0012-365x(73)90108-8. URL: https://doi.org/10.1016%2F0012-365x%2873%2990108-8.

[23] J. Arjan G.M. de Visser and Joachim Krug. "Empirical fitness landscapes and the predictability of evolution". In: *Nature Reviews Genetics* 15.7 (June 2014), pp. 480–490. DOI: 10.1038/nrg3744. URL: https://doi.org/10.1038/nrg3744.

[24] Mark P. Zwart et al. "Unraveling the causes of adaptive benefits of synonymous mutations in TEM-1 β-lactamase". In: *Heredity* 121.5 (July 2018), pp. 406–421. DOI: 10.1038/s41437-018-0104-z. URL: https://doi.org/10.1038/s41437-018-0104-z.

9 Appendix

No.	Encoding	Peak Pattern Normal Form	Maxima	Uncons-trained Vertices	Realizations (Single Peak)
1	0_{32}	$1,1,1,1$	1	11	(5654)48649200
2	16448_{32}	$2,2,3,3$	2	8	(4356)41536560
3	1052688_{32}	$2,2,2,3$	2	6	(2688)40995968
4	16843009_{32}	$2,2,2,2$	2	6	(17879)12296152
5	134758464_{32}	$5,5,6,6$	3	5	(1536)9214656
6	67387458_{32}	$4,5,6,6$	3	3	(1056)24193728
7	4210784_{32}	$4,5,6,7$	3	6	(856)9663408
8	151601473_{32}	$10,10,12,12$	4	4	(108)1050384
9	1075863624_{32}	$10,9,13,14$	4	4	(336)323088
10	101073474_{32}	$9,10,12,12$	4	0	242832
11	37896800_{32}	$9,10,12,14$	4	1	(64)2300160
12	1077952616_{32}	$8,11,13,14$	4	5	(2)2170464
13	6316128_{32}	$9,10,12,15$	4	4	311604
14	1210605640_{32}	$21,19,26,28$	5	3	(36)57728
15	1080057960_{32}	$18,21,25,30$	5	3	134080
16	1111638632_{32}	$17,22,26,28$	5	0	118992
17	1227448649_{32}	$42,38,52,56$	6	2	(1)1800
18	1616930920_{32}	$37,42,51,60$	6	2	9264
19	1751672936_{32}	$75,85,102,120$	7	1	240
20	1768515945_{32}	$150,170,204,240$	8	0	2

Table 2: Calculated peak patterns and corresponding properties for $L = 4$, which results in i) 20 possible peak patterns for all acyclic oriented graphs and ii) 13 including the additional path condition from sec.3.4. Peak patterns which do not have a single instance fullfilling the path condition are marked gray and the possibly differing values for ii) are given in brackets. The encoding given corresponds to the peak pattern normal form with all arrows up.

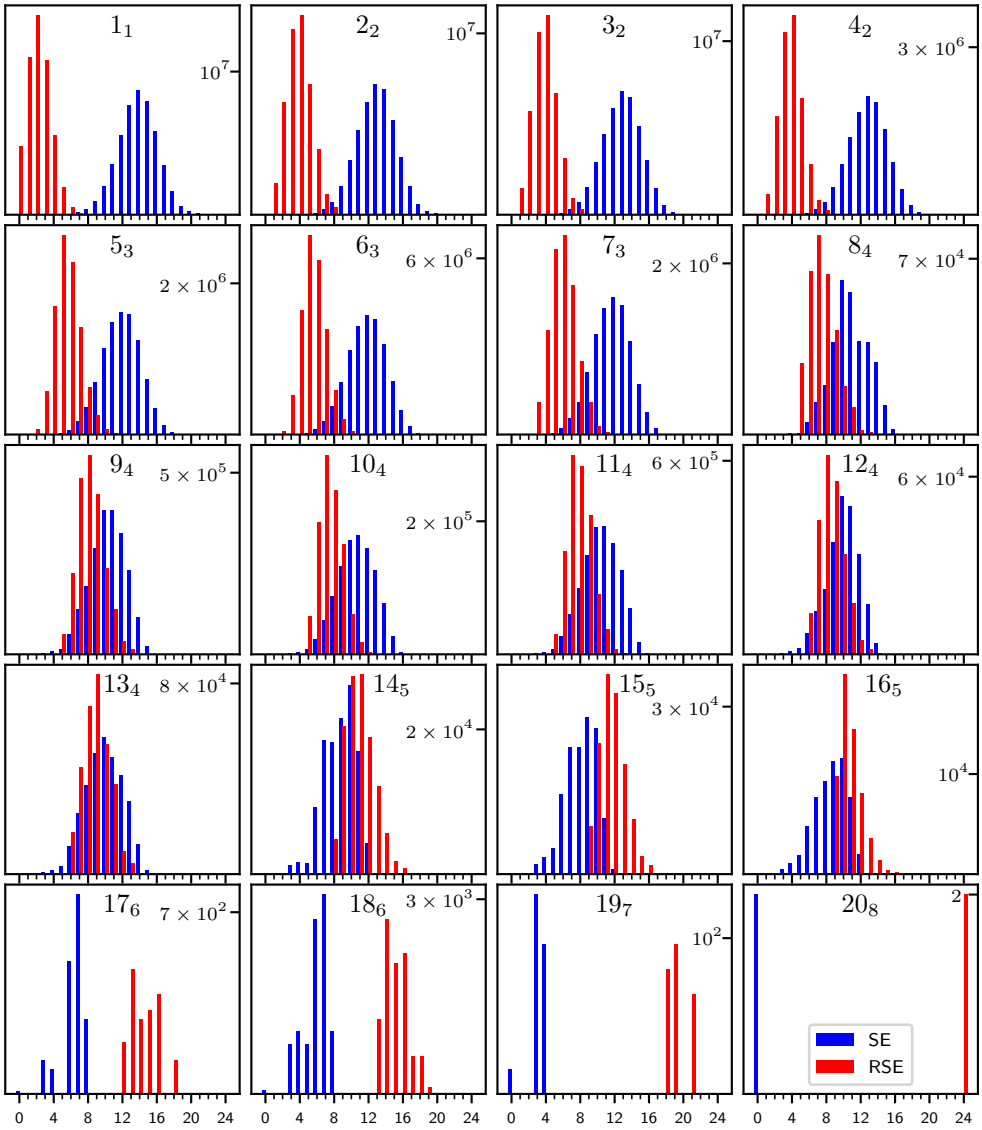


Figure 17: Grouped bar plots, with linear scales, for the same data shown in fig.10, with the central upper number in each subplot referring to the peak pattern no. in table 2 and the subscript to the number of peaks.

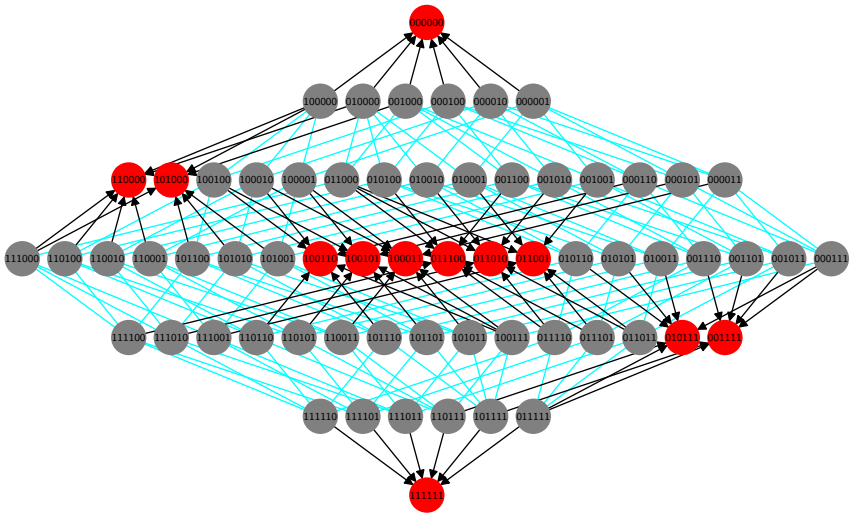


Figure 18: Fully vertex constrained peak pattern with 12 peaks, being the minimal number of peaks to fully constrains all vertices of the 6-cube. The coloring is the same as in fig.6.

Erklärung

Hiermit versichere ich an Eides statt, 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. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer anderen Prüfung noch nicht vorgelegt worden. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Köln den 12.04.2022
Daniel Oros