

Epistasis, regular subdivisions and spanning trees

vorgelegt von
M. Sc.
Holger Eble

an der Fakultät II – Mathematik und Naturwissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
Dr. rer. nat.

genehmigte Dissertation

Promotionsausschuss:

Vorsitzender: Prof. Dr. Tobias Breiten
Gutachter: Prof. Dr. Michael Joswig
Gutachter: Prof. Dr. Svante Linusson

Tag der wissenschaftlichen Aussprache: 18. November 2022

Berlin 2022

Acknowledgements

I would like to thank my advisor and coauthor Michael Joswig for his scientific guidance, his professionalism, his support and his open-mindedness. Without him, this thesis wouldn't have been written. Further, I wish to thank my other coauthors Taylor Brysiewicz, Lukas Kühne, Lisa Lamberti and William Ludington. The scientific exchange with them had a great impact on this thesis and after all made it possible. I want to express my gratitude to Svante Linusson for serving as a referee and to Tobias Breiten for being the head of my Ph.D. committee. Last but not least, I want to thank Marta Panizzut for proofreading parts of this thesis and Benjamin Lorenz, Lars Kastner and Antje Schulz for incredible perpetual support regarding technical and administrative issues.

I wish to acknowledge the funding by SFB-TRR 195 “Symbolic Tools in Mathematics and their Application” of the German Research Foundation (DFG).

Abstract

In this thesis, we use techniques from polyhedral geometry and statistics in order to detect and quantify biological interactions within a system of genes or species described by a given data set. Our concept relies on the theory of regular subdivisions. A regular subdivision decomposes a space into convex cells and can be used to showcase some distinct aspects of the given data set in its cell structure. After all, one can implement and compute with regular subdivisions and this is an important feature of polyhedral and discrete geometry. The way these cells spatially relate to each other is exploited to determine a list of moderate length with potentially significant biological interactions. A statistical test allows us to diminish this list further and to point to few but statistically significant interactions. A major benefit of our method, and in a way this is reciprocal compared to other existing methods, is the concise extent of our findings which allows for communicating them in a comprehensive form, for instance in data tables or specifically developed bar diagrams.

We applied our methods to several experimentally obtained genetic and microbiome data sets. A central use case was the analysis of two instances of *Drosophila melanogaster* fly gut microbiome studies. The gut of these fruit flies has a microbiome with a small number of constituting species and can be manipulated in the laboratory by regulation of the food. The two *Drosophila* data sets we applied our methods on describe a microbiome system with five species and hence each of the two data sets can be related to some regular subdivision of the 5-dimensional 0/1-cube from where our method departs. We are able to point to significant higher dimensional interactions which are not perceived by other existing methods and, in particular, are not captured by looking at pairs of interacting species only.

Further, we reinterpret and analyze our method mathematically. It produces and is in some way equivalent to naming a network of shortest genetic distances, i.e. a minimum spanning tree of a certain fixed weighted graph with biological meaning. Tropical hypersurfaces, central objects of tropical geometry, which is an active field of research at the border of polyhedral, discrete and algebraic geometry, encapsulate these structures inside their 1-dimensional skeleton. We determine the parameter space of the minimum spanning trees arising this way. It turns out to be encoded by a collection of cones given by linear hyperplanes. For a few elected examples we computed an explicit representation of all occurring parameter cones. Yet, this rapidly reaches limits of complexity.

The rest of this thesis is about an achievement beyond these limits. Given a cell decomposition in some implicit form, one may not be able to recuperate the defining geometric data for every cell. But it still may be possible to enumerate them. We present a method for computing the number of chambers of a hyperplane arrangement in real euclidean space which uses purely combinatorial techniques and which makes use of the combinatorial symmetries of the given hyperplane arrangement. With this method, it was possible to compute the previously unknown number of chambers of the ninth resonance arrangement given by 511 hyperplanes in \mathbb{R}^9 .

Zusammenfassung

In dieser Arbeit werden Techniken aus der polyedrischen Geometrie und der Statistik vorgestellt, die benutzt werden können, um biologische Wechselwirkungen in einem durch Datensätze beschriebenen Gen- oder Speziensystem aufzufinden und zu quantifizieren.

Unser Konzept beruht auf der Theorie der regulären Unterteilungen. Eine reguläre Unterteilung zerlegt einen Raum in konvexe Zellen, die im vorliegenden Fall dazu dienen, ausgewiesene Eigenschaften der zugrundeliegenden Daten aufzuzeigen. Desweiteren lassen sich reguläre Unterteilungen in Computerprogrammen implementieren und berechnen, was allgemein einen wichtigen Aspekt der polyedrischen und diskreten Geometrie darstellt.

Der räumliche Bezug der Zellen zueinander wird hierbei benutzt, um eine Liste angemessener Länge mit potenziell signifikanten biologischen Wechselwirkungen zu erstellen. Desweiteren dient ein statistischer Signifikanztest zur weiteren Ausdünnung dieser Liste, die schließlich nur noch statistisch nachweisbar signifikante Wechselwirkungen enthält. Durch die Bündelung und Konzentration auf relevante Wechselwirkungen zeichnet sich unsere Methode wesentlich aus, da dies sich durchaus konträr zu den bereits existierenden Methoden verhält und eine stringente Kommunikation der Ergebnisse gestattet, beispielsweise in Form von Datentabellen oder eigens konzipierter Bardiagramme.

Wir haben unsere Methode auf mehrere Experimentaldatensätze mit Genetik- und Mikrobiombezug angewandt. Ein zentraler Anwendungsfall stellte dabei die Analyse zweier Datensätze dar, die das Mikrobiom des Magens der *Drosophila melanogaster* Fliege experimentell erfassen. Das Mikrobiom des Magens dieser Fruchtfliege hat die besondere Eigenschaft, durch eine geringe Anzahl von teilhabenden Spezien bestimmt und im Labor leicht manipulierbar zu sein, etwa durch Regulierung des Futters. Die zwei *Drosophila* Datensätze, die wir betrachteten, beschreiben jeweils ein Mikrobiomsystem mit fünf konstituierenden Spezien und können folglich mit regulären Unterteilungen des fünfdimensionalen 0/1 - Würfels assoziiert werden, welche die von uns entwickelte Methode verarbeitet. Es war uns möglich, höherdimensionale Wechselwirkungen zu finden, die von den bereits existierenden Methoden nicht gesehen werden und insbesondere vom Paarvergleich wechselwirkender Spezien übergangen werden.

Desweiteren interpretieren und analysieren wir unsere biologisch motivierte Methode innermathematisch. Im Einzelfall ist der Verlauf dieser äquivalent zur Konstruktion eines Spannbaums minimalen Gewichts in einem festgeschriebenen gewichteten Graph. Dieser Spannbaum lässt sich biologisch wiederum als Netzwerk kürzester genetischer Distanz interpretieren. Tropische Hyperflächen sind zentrale Objekte der tropischen Geometrie, eines eigens für sich aktiven Forschungsgebiets mit Anknüpfungspunkten zur polyedrischen, diskreten und algebraischen Geometrie. Diese Hyperflächen enthalten die betreffenden minimalen Spannbäume in ihrem eindimensionalen Skelett. Wir zeigen, dass der Parameterbereich dieser minimalen Spannbäume durch eine Sammlung polyedrischer Kegel gegeben ist. Für ausgewählte, kleine Beispiele gelingt es, eine explizite Darstellung für jeden einzelnen Parameterkegel zu berechnen. Dennoch stößt man dabei schnell auf unüberwindbare Komplexitätsschranken.

Der Rest dieser Arbeit beschäftigt sich mit einer Thematik, die jenseits dieser Komplexitätsschranken liegt. Zwar mag es für eine implizit gegebene Zellzerlegung mit den aktuellen Methoden unmöglich sein, für jede einzelne Zelle eine explizite geometrische Beschreibung zu errechnen, jedoch kann durchaus eine Abzählung der Zellen erfolgen. Wir präsentieren eine rein kombinatorische Methode zur Abzählung der Kammern eines reellen Hyperebenenarrangements, die wesentlich auf der Ausnutzung kombinatorischer Symmetrie fußt. Mit dieser Methode war es uns möglich, die zuvor unbekannte Kammeranzahl des neunten Resonanzarrangements zu bestimmen, das durch 511 Hyperebenen im \mathbb{R}^9 gegeben ist.

Contents

Introduction	1
0.1 Some notes on epistasis	3
0.2 Epistatic interactions and epistatic weights	4
0.3 Separating significant signal from data noise	5
0.4 Epistatic filtrations as minimum spanning trees	8
0.5 Counting chambers of hyperplane arrangements	8
Cluster partitions and fitness landscapes of the <i>Drosophila</i> fly microbiome	9
1.0 Abstract	9
1.1 Introduction	9
1.2 Mathematical background and terminology	10
1.3 Clusters in fitness landscapes	14
1.4 Significant cluster partitions	20
1.5 Epistasis, interaction coordinates and circuit interactions	24
1.6 Epistasis in <i>Drosophila melanogaster</i> fruit fly microbiomes	25
1.7 Discussion and outlook	31
1.8 Additional computations, figures and information for Chapter 1	32
Master regulators of evolution and the microbiome in higher dimensions	39
2.0 Abstract	39
2.1 Introduction	39
2.2 Results	40
2.3 Discussion and Conclusions	50
2.4 Acknowledgements	53
2.5 Materials and Methods	53
2.6 Terminology	54
2.7 Additional computations, figures and information for Chapter 2	61
The MST-fan of a regular subdivision	75
3.0 Abstract	75
3.1 Introduction	75
3.2 Regular subdivisions and tropical hypersurfaces	76
3.3 The MST-Fan $\mathcal{K}(h)$	79
3.4 An algorithm for computing $\mathcal{K}(T_{<})$	81
3.5 Permutation groups acting on tree orders	83
3.6 Matroids and Bergman fans	84
3.7 Application in population genetics: epistasis and spanning trees	85
3.8 Computations	86
Computing characteristic polynomials of hyperplane arrangements with symmetries	89
4.0 Abstract	89
4.1 Introduction	89
4.2 Hyperplane arrangements	90
4.3 A deletion-restriction algorithm	92
4.4 Automorphisms of hyperplane arrangements	94

4.5	Enumeration algorithm with symmetry	95
4.6	Examples and integer sequences	98
4.7	Timings	100
4.8	Tables of Whitney numbers	101
References		105

Introduction

In data analysis, detecting dependencies and independencies within a given data set is an important task. Any sound method provided for a concrete investigation needs to reflect the sort of independency one is seeking for in its description in order to insure the interpretability of the gained results. This may happen in algebraic, geometric or statistical terms, for instance. In Chapter 1 of this thesis, we study a phenomenon from computational biology, called *epistasis*, where given data sets describe certain geometric situations provided with a local notation of independence, namely affine independence. More concretely, the genetic setup of an investigated organism is encoded as a point configuration \mathbf{A} of 0/1-vectors, i.e. any vector $a \in \mathbf{A}$ describes a distinct genotype of the organism indicating the presence or absence of each examined gene. A data set $D = \{D_a : a \in \mathbf{A}\}$ of phenotypes, drawn from an experiment for instance, might be interpreted as a genotype-phenotype map $h : \mathbf{A} \rightarrow \mathbb{R}$ where each value $h(a) = h(D_a)$ is obtained from the data row D_a via some prescribed arithmetic operation. Roughly speaking, epistasis is about detecting and measuring the degree of affine independences, called *epistatic interactions*, within the point set $\{(a, h(a)) : a \in \mathbf{A}\}$.

In this way, epistasis is naturally connected to the realm of polyhedral geometry whose theoretical and algorithmic body has already turned out useful to provide machinery for performing this specific data analytic task, cf. [9]. Based on that, we introduce in Chapter 1 a new method to study epistasis. One of its key features is the ability to deal with both the local nature of epistatic interactions and as well capture global information about the entire genotype-phenotype system (\mathbf{A}, h) via a clustering structure, called *epistatic filtration*, depicting a certain kind of epistatic behavior of the underlying data set D as a whole.

In general, a cell decomposition serves for splitting up a space into simpler pieces which are theoretically or computationally more accessible than the original space. Further, the combinatorial structure of this splitting may be used to encode some aspects of additional information which the original space might be equipped with. Each epistatic filtration is constructed from a cell decomposition of the convex hull of \mathbf{A} , namely a special kind of triangulation called *regular subdivision*, which is induced by the system (\mathbf{A}, h) and where every cell, i.e. a simplex in \mathbf{A} , has a biological meaning. Since all genotypes \mathbf{A} are considered simultaneously, both local and global notions of epistasis can be studied: An epistatic filtration features a list of epistatic interactions with distinct geometric properties and, as a whole, this collection showcases certain aspects of global epistatic behavior of the examined organism. In particular, this applies to higher dimensions where existing methods tend to struggle with the involved combinatorial complexity.

Further, we enriched our polyhedral techniques with a statistical component. Each perceived epistatic interaction receives a p -number from a p -test which serves for deciding about the statistical significance of the interaction. In this way, the actual statistics of the data set D is taken into account, i.e. importance is attributed to the distribution of the phenotype data points. Within the wide range of all epistatic interactions, our method usually points to very few relevant ones whose geometric significance is confirmed by the statistics of the underlying data set. Establishing this novel interplay between polyhedral geometry and statistics is one of the central achievements of Chapter 1.

Chapter 2 is devoted to presenting the material of Chapter 1 to the biology community. We applied our method profoundly to some authentic experimental data, for instance to two data sets examining the microbiome of the *Drosophila* fly gut, compare the outcomes and showcase how established procedures in the field of population genetics can be interpreted and investigated within our framework. An important role in our study played the concept of *context dependence*, where an epistatically interacting genotype system shares a common gene, called a *bystander*, which is then

simultaneously removed. Comparing the epistatic filtration of the resulting system with that of the original system describes the epistatic effect of adding the bystander gene, which is a relevant practice in biology.

Generally speaking, the focus of Chapter 2 lies on the study of data sets $D = \{D_a : a \in \mathbf{A}\}$ which describe phenotype data over higher-dimensional genotype sets \mathbf{A} , notably where \mathbf{A} is a hypercube of dimension bigger than three. These considerations are rare or even lacking to some degree in the biology literature, again due to dramatically increasing combinatorial complexity in higher dimensions where the 2-dimensional intuition turns out to be too simplistic or even wrong. In this practical investigation, we found geometrically and statistically significant higher-dimensional epistatic interactions which are overlooked by low-dimensional methods.

Chapter 3 takes a reversed approach and answers the following innermathematical questions which are motivated by the data analysis application from Chapter 1 and Chapter 2.

- (A) Which epistatic filtrations may arise?
- (B) Given a fixed epistatic filtration, what are the constraints on the data set D to retrieve it?
- (C) What is the parameter space of all feasible epistatic filtrations?

In Chapter 3, epistatic filtrations are reinterpreted in combinatorial terms, namely as minimum spanning trees of some weighted graph induced by the genotype-phenotype system (\mathbf{A}, h) . More precisely, the graph in question is the dual graph of the regular subdivision induced by the system (\mathbf{A}, h) and its edges encode the incidences of this cell decomposition. The edge weights indicate the strength of the epistatic interactions which are gathered in the filtration. Since this measure can be described as a linear functional in h , the parameter space of (C) lives in the linear world as well. It turns out to be a *fan*, i.e. a collection of polyhedral cones, and is hence given by another type of cell decomposition where again each cell carries specific information. A single epistatic filtration as in (B) is represented by exactly one such parameter cone. If the entire fan structure is computationally accessible, then it provides an answer to question (A). We also indicate computational bounds, which are reached very quickly when increasing the dimension of the point configuration \mathbf{A} . The reason is that questions (A) to (C) from above demand a concrete geometric description of every single parameter cone, which all live in $\mathbb{R}^{\mathbf{A}}$.

As is done in Chapter 3 with the parameter fan of epistatic filtrations, a cell decomposition of a space can be described by explicitly listing its cells in a concrete geometric encoding. For instance, the polyhedral cones of a fan can be given in half-space description. Alternatively, a cell decomposition sometimes might be described in a more implicit way. A real *hyperplane arrangement* is a collection of hyperplanes in some fixed \mathbb{R}^n . Each of its cells, also known as *chambers*, is a component of the complement of the arrangement and is indexed by a string of signs, plus or minus, describing its relative location with respect to each hyperplane of the arrangement. In this case, an explicit geometric cellular description is a priori missing and computing one efficiently is a separate task, cf. [89].

In Chapter 4 we present examples of real hyperplane arrangements for which implementations of currently known algorithms fail by far to provide an explicit description of the chambers due to complexity reasons. Instead, we limit our considerations to counting their number of chambers only. This attenuation allows the usage of coarser methods demanding much less time and memory. We present a chamber counting algorithm which evolved from the idea of systematically deconstructing a given hyperplane arrangement in depth-first manner along a binary tree and thereby keeping track of combinatorial invariants, the *Whitney numbers*. Our approach is purely combinatorial and avoids costly operations like computing convex hulls. The number of chambers can finally be reconstructed from the computed Whitney numbers. As a key modification of this rudimentary depth-first approach, we further make use of symmetry inherent to the arrangement to cut off branches of the binary tree which would result in redundant computations. This results in a significant speed-up of the algorithm and we were able to compute a chamber number which was formerly unknown.

The rest of the introduction presents the material covered in this thesis in more detail.

0.1 Some notes on epistasis

Epistasis is a subfield of genetics, which studies the effect of gene interactions on the underlying organism, cf. [121]. Historically, the term *epistasis* was first introduced by William Bateson in the early 1900 as a model for dominant or recessive effects in genetic systems, cf. [8]. However, in this thesis we stick to a slightly different paradigm established by R. A. Fisher in 1918 [53] who defined epistasis as deviation from linearity in the measured phenotypes. In the literature - both ancient and recent - this is often motivated with the following example: Given an organism with two genes or species A and B involved, the list \mathbf{A} of all possible genotypes consists of:

- $(0_A, 0_B)$, the *wild type*, meaning that neither species A nor species B is present,
- $(1_A, 0_B)$, meaning that species A is present but species B is absent,
- $(0_A, 1_B)$, meaning that species A is absent but species B is present and
- $(1_A, 1_B)$, meaning that both species A and B are present.

Now, assume that each genotype of \mathbf{A} receives a physical quantity via a genotype-phenotype map $h: \mathbf{A} \rightarrow \mathbb{R}$, say with $h(0_A, 0_B) = 0$. Then the single-mutation $(0_A, 0_B) \rightsquigarrow (1_A, 0_B)$ has the phenotypic effect $h(1_A, 0_B)$ and similarly for $(0_A, 0_B) \rightsquigarrow (0_A, 1_B)$. In Fisher's school, the system (\mathbf{A}, h) describes an *epistatic interaction* if the double-mutation $(0_A, 0_B) \rightsquigarrow (1_A, 1_B)$ cannot be additively deduced from the two single mutations, i.e. if $h(1_A, 1_B) \neq h(1_A, 0_B) + h(0_A, 1_B)$ holds.

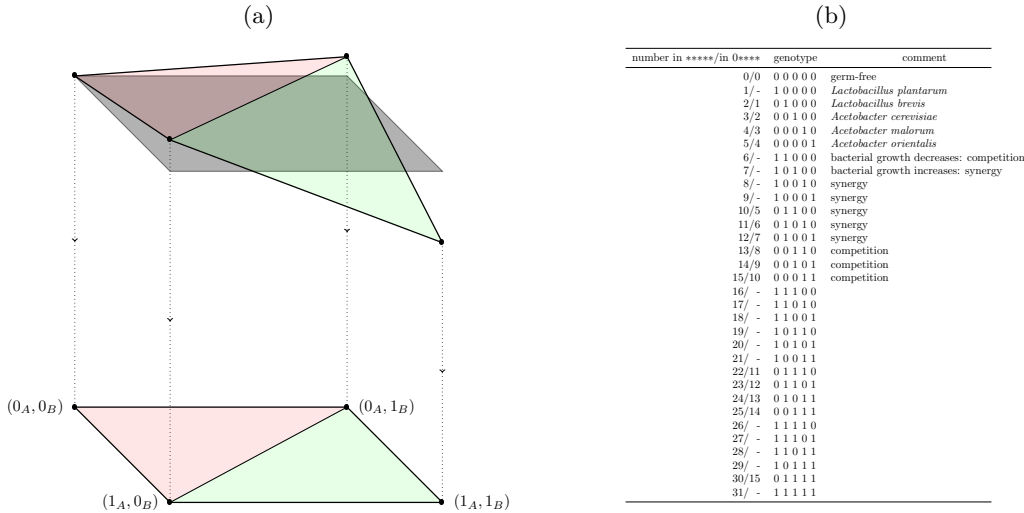


Figure 1: (a) A genotype-phenotype map $h: \mathbf{A} \rightarrow \mathbb{R}$ for which the system (\mathbf{A}, h) describes an epistatic interaction. The value $h(1_A, 1_B)$ is lower than expected and cannot be linearly deduced from the values $h(0_A, 0_B)$, $h(1_A, 0_B)$ and $h(0_A, 1_B)$. The reference hyperplane for the wild type is pictured in gray. (b) An encoding of the genotypes involved in an experimental investigation of the *Drosophila* fly gut microbiome featuring two types of bacteria, the lactobacilli and the acetobacters. This *Eble data set*, bi-allelic on five loci, is analyzed in Chapter 2. The idea of interpreting the genotypes as 0/1 - vectors and lifting them one dimension higher via h dates back to 1932, cf. [156].

The system (\mathbf{A}, h) is an epistatic interaction if and only if the quantity

$$\overline{E} := h(0_A, 0_B) + h(1_A, 1_B) - h(1_A, 0_B) - h(0_A, 1_B) ,$$

which is omnipresent in the epistasis literature, does not vanish. Further, if \overline{E} changes its sign from minus to plus, then the *combinatorial* situation switches as well: The diagonal $\{(1_A, 0_B), (0_A, 1_B)\}$ of the square at the bottom of Figure 1(a) is then replaced by $\{(0_A, 0_B), (1_A, 1_B)\}$. But this perception is based on a bias towards choosing the wild type as point of reference: In general, the phenotype $h(0_A, 0_B)$ of the wild type receives the positive coefficient 1 in \overline{E} but the equation $(-\overline{E})$ would equally serve for decisions on epistasis.

In this thesis, we adapt these ideas with regard to detecting and studying specific epistatic interactions, but focus on higher dimensional systems where signs of equations like \bar{E} do not have an unbiased meaning and occur in a vast number due to much more complicated combinatorics.

0.2 Epistatic interactions and epistatic weights

In Section 1.3 we provide a general framework for the detection and the study of a certain kind of epistatic interactions in any *fitness landscape* (\mathbf{A}, h) , where $\mathbf{A} \subset \mathbb{R}^n$ is some point configuration and $h: \mathbf{A} \rightarrow \mathbb{R}$ is any height function on \mathbf{A} . There we introduce a non-negative quantity e_h , which is designed for measuring the strength of epistatic interactions, called *epistatic weight*. Our method exploits the cell structure of the regular subdivision $\Sigma(h)$ of \mathbf{A} induced by h . For a generic height function h , each maximal cell of $\Sigma(h)$ is an n -dimensional simplex and arises as the projection of an upper facet of the polytope $P_h := \text{conv}\{(a, h(a)): a \in \mathbf{A}\} \subset \mathbb{R}^{n+1}$, cf. [40, Definition 2.2.10.].

A bipyramid in \mathbf{A} is a pair (s, t) of adjacent n -dimensional simplices s and t , both spanned by points of \mathbf{A} , sharing a common $(n - 1)$ -dimensional face. An epistatic interaction in \mathbf{A} is a bipyramid in \mathbf{A} such that the polytope $(s, t)_h := \text{conv}\{(a, h(a)): a \in \text{vert}(s \cup t)\} \subset \mathbb{R}^{n+1}$ is not contained in a hyperplane. The epistatic weight $e_h(s, t)$ of an epistatic interaction (s, t) is defined to be a dimensionally scaled adaption of the normalized volume of the lifted bipyramid $(s, t)_h$ and hence measures how far the lifted vertices $\{(x, h(x)): x \in \text{vert}(s \cup t)\}$ are away from living on a hyperplane. Note that for a generic height function $h: \mathbf{A} \rightarrow \mathbb{R}$, e.g. for a reasonable outcome h of some real-world experiment, the lifted bipyramid $(s, t)_h$ almost never lies on a hyperplane. Hence, we may refer to epistatic interactions in \mathbf{A} simply as the bipyramids in \mathbf{A} and the number of these is big, even for small point configurations \mathbf{A} .

From a data analysis point of view, the foremost purpose while developing our techniques was to be able to point to few but specific and significant epistatic interactions. We first restricted our consideration to epistatic interactions of $\Sigma(h)$, i.e. to pairs (s, t) of adjacent maximal simplices of the regular subdivision $\Sigma(h)$. In the biological application we had in mind, these maximal cells encode fittest populations with respect to the height function h and varying allele frequencies, cf. Section 1.2.2. The specific geometric context related to $\Sigma(h)$ can be described as follows. Each maximal cell s of $\Sigma(h)$ gives rise to a unique affine map $\tilde{h}_s: \text{conv}(s) \rightarrow \text{conv}(s)_h \subset P_h$, which identifies the cell s with the upper facet of P_h it came from and which affinely extends to a map $\text{aff}[\tilde{h}_s]: \text{conv}(\mathbf{A}) \rightarrow \mathbb{R}^{n+1}$. Now, according to [23, Section 1.F], the piecewise linear and concave map $\tilde{h}: \text{conv}(\mathbf{A}) \rightarrow \mathbb{R}^{n+1}$, patched together from the various \tilde{h}_s , takes for any $x \in \text{conv}(\mathbf{A})$ the value

$$\tilde{h}(x) = \min \{ \text{aff}[\tilde{h}_s](x) : s \text{ is a maximal cell of } \Sigma(h) \} \quad (1)$$

and serves as support function for $\Sigma(h)$: Its regions of linearity are precisely the maximal cells of $\Sigma(h)$. Now, an epistatic interaction $b = (s, t)$ of $\Sigma(h)$ with adjacent maximal simplices s and t of $\Sigma(h)$ features the two distinct satellite vertices, or satellite genotypes, $b_1 = s \setminus t$ of s and $b_2 = t \setminus s$ of t . From equation (1) one sees that in this situation, the epistatic weight $e_h(s, t)$ measures the magnitude to which the two *satellite data points* $(b_1, h(b_1))$ resp. $(b_2, h(b_2))$ are lower than expected with respect to a given reference system inherent to (s, t) , namely the hyperplane spanned by the data points $\{(x, h(x)): x \in \text{vert}(t)\}$ resp. $\{(x, h(x)): x \in \text{vert}(s)\}$.

In the bi-allelic n -loci case, a fitness landscapes (\mathbf{A}, h) has the n -dimensional 0/1-cube as underlying point configuration \mathbf{A} . In the two-dimensional example from Section 0.1 above, the height function h is hence defined over the unit square and the epistatic weight

$$e_h(\{(0, 0), (1, 0), (0, 1)\}, \{(1, 0), (0, 1), (1, 1)\})$$

coincides with the absolute value of the quantity \bar{E} , upto a dimensionally related factor of $\sqrt{2}$.

The regular triangulations of \mathbf{A} are parametrized by full-dimensional cones in $\mathbb{R}^{\mathbf{A}}$, called *secondary cones*, which fit together nicely to form the *secondary fan* of \mathbf{A} . More precisely, any choice of height functions from the interior of a fixed secondary cone yields the same triangulation of \mathbf{A} . Consequently, small perturbations h' of h occurring under the influence of data noise tend to produce the same triangulation $\Sigma(h') = \Sigma(h)$. The separation of data noise from relevant epistatic signal hence cannot be achieved by solely looking at the combinatorics of $\Sigma(h)$. In Section 0.3 we briefly present the metric and statistical techniques we developed to do so.

0.3 Separating significant signal from data noise

In order to get closer to our actual use cases in the field of population genetics, let $\mathbf{A} \subset \mathbb{R}^n$ be a point configuration representing a microbiome, i.e. \mathbf{A} is the vertex set of the n -dimensional 0/1-cube, in the bi-allelic case, or more generally of any n -dimensional product of simplices, and let $h: \mathbf{A} \rightarrow \mathbb{R}$ be a height function on \mathbf{A} , cf. [10]. Assume that h is sufficiently generic, i.e. h lies in the interior of a secondary cone and hence the regular subdivision $\Sigma(h)$ of \mathbf{A} is a triangulation. Since big epistatic weights describe big deviations from linearity, we first compile a list L of some length m which comprises all epistatic interactions (s_i, t_i) of (\mathbf{A}, h) , i.e. of $\Sigma(h)$, ascendingly ordered by their epistatic weight $e_h(s_i, t_i)$. The driving idea is to seek for some subset $I \subset [m]$, ideally oriented towards the tail of L , such that an interaction $(s_i, t_i) \in L$ can verifiably be considered significant if $i \in I$ holds. This is achieved in two levels.

0.3.1 Geometric level: The epistatic filtration process

In order to extract a global statement about relevant epistatic information, we established a procedure called *epistatic filtration*, which successively keeps track of the epistatic interactions of L and which records how they relate to each other, going from lowest to highest epistatic weight and avoiding redundant geometric information. The rough idea on which the filtration process is based is that a single epistatic interaction (s, t) of (\mathbf{A}, h) is recorded by merging the cells s and t to form the cluster $(s \cup t) \subset \text{conv}(\mathbf{A})$, which is the projection of the lifted bipyramid $(s, t)_h$ the epistatic weight $e_h(s, t)$ measures the volume of. Processing along the list L yields the following evolution of cell decompositions, also called *cluster partitions* in Section 1.3.2, of $\text{conv}(\mathbf{A})$:

- (A₀) Start with the cell decomposition C_0 of $\text{conv}(\mathbf{A})$ given by the maximal simplices of $\Sigma(h)$.
- (A_k) Assume that the first $k - 1$ epistatic interactions of L are processed and $\text{conv}(\mathbf{A})$ has a cell decomposition C_{k-1} into connected sets c_i . Then the k -th interaction (s_k, t_k) of L determines unique cells c_{s_k}, c_{t_k} of C_{k-1} with $\text{conv}(s_k) \subset c_{s_k}$ and $\text{conv}(t_k) \subset c_{t_k}$. If merging s_k and t_k is geometrically redundant, i.e. if $c_{s_k} = c_{t_k}$ holds, then the interaction (s_k, t_k) is called *uncritical*. In this case, step (A_{k+1}) is performed with the cell decomposition $C_k := C_{k-1}$ and the next interaction $(s_{k+1}, t_{k+1}) \in L$ if there is any. Otherwise, we have $c_{s_k} \neq c_{t_k}$. In this case, we call the interaction (s_k, t_k) *critical* and we merge the cells c_{s_k} and c_{t_k} . Step (A_{k+1}) is hence performed with the cell decomposition $C_k := (C_{k-1} \setminus \{c_{s_k}, c_{t_k}\}) \cup \{c_{s_k} \cup c_{t_k}\}$ of $\text{conv}(\mathbf{A})$ and the next epistatic interaction $(s_{k+1}, t_{k+1}) \in L$ if there is any.

The ordered list of tuples $\mathcal{F}(h) := (C_k, e_h(s_k, t_k))$, where (s_k, t_k) runs through the critical epistatic interactions only, is called *epistatic filtration* of (\mathbf{A}, h) . The concept of merging cells in this way is motivated by the following theorem.

Theorem (*Undo Theorem*, Theorem 1.3.3). *Let $\Sigma(h')$ be a regular subdivision of \mathbf{A} which is not a triangulation. Then there is an $\varepsilon > 0$ such that for any choice of $h \in B_\varepsilon(h')$ the epistatic filtration $\mathcal{F}(h)$ features some cluster partition C_k which coincides with the cell decomposition $\Sigma(h')$ of $\text{conv}(\mathbf{A})$.*

The *Undo Theorem* has the following interpretation in terms of data analysis. Assume there is a comprehensive data collection D of experimental measured data, which associates to every $a \in \mathbf{A}$ some data row D_a . Now, let the height function h be an arithmetic outcome of D , e.g. the value $h(a)$ is the sample mean of D_a . Then the function h is almost always generic. Yet, there may exist epistatic interactions of (\mathbf{A}, h) whose non-negative epistatic weights are derived from inaccuracies in the measuring process but may tend to vanish in an ideal experimental environment. Phrased differently, some affine dependencies are concealed by data noise inherent to D . By the Undo Theorem, these dependencies are then revealed at some stage of the epistatic filtration $\mathcal{F}(h)$.

0.3.2 Statistical level: A significance test for epistatic interactions

Still we assume that there is a data collection $D = (D_a)$ we want to analyze. In order to enrich regular subdivisions with a statistics, we consider a vector of random variables $\bar{X} = (\bar{X}_a)_{a \in \mathbf{A}}$ where each \bar{X}_a is normally distributed with the first two moments given by the sample mean and the standard error of D_a , cf. Section 1.4. Taking $h: \mathbf{A} \rightarrow \mathbb{R}$ to be the sample means of D yields a regular triangulation $\Sigma(h)$ of \mathbf{A} . Now, for every epistatic interaction (s, t) of (\mathbf{A}, h) our significance test pursues the question if there is statistical evidence that the de facto positive epistatic weight $e_h(s, t)$ is not based on data noise of D . This is done by establishing a null hypothesis via a certain dummy random variable Z with expected value zero. According to the practice of p -tests, the null hypothesis for the epistatic interaction (s, t) of (\mathbf{A}, h) is refuted if the p -value $p(s, t) := \mathbb{P}(Z \geq e_h(s, t))$ is sufficiently small. Here, the p -value bound 0.05 seems to be an engraved law, i.e. the epistatic interaction (s, t) is *significant* if $p(s, t) < 0.05$ holds, but we decided to install a second class of *semi-significant* interactions, namely those which fulfill $0.05 \leq p(s, t) < 0.1$.

Intuitively, critical significant epistatic interactions of (\mathbf{A}, h) should always be big, i.e. appear in advanced stages of the filtration $\mathcal{F}(h)$. But also interactions with small epistatic weight have the chance to be significant since the second moment of the random variable Z depends on the standard errors of D . However, on all computations with real data we run, significant epistatic interactions only occurred at the very end of the filtrations, often preceded by a remarkable gap between the epistatic weights of the last non-significant and the first significant critical interaction.

0.3.3 Visualization

Since our primal goal was to be able to select few but significant information out of a big collection of epistatic interactions, namely *all* bipyramids in \mathbf{A} , the development of a concise visualization of our findings became central and we came up with bar diagrams as in Figure 2(a). There, we showcase the epistatic filtration of the 4-dimensional face $\mathbf{A} := \{0, 1\}^4 \xrightarrow{0****} \{0, 1\}^5$ of the Eble data set analyzing the *Drosophila* gut microbiome with the first (*Lactobacillus plantarum*) of five species required to be absent, cf. Section 2.2.6 for instance.

0.3.4 Application: Parallel transport

Context dependent epistasis is an important subfield of study, which is concerned about the effect of changing bystanders of epistatic interactions. More precisely, assume we have a fixed bi-allelic fitness landscape $h: \mathbf{A} \rightarrow \mathbb{R}$ on n loci, i.e. $\mathbf{A} = \{0, 1\}^n$ is the n -dimensional 0/1-cube. For instance, if we want to consider epistatic interactions of genotypes where the first species is always present, then we may pass to the $(n - 1)$ -dimensional subcube $\mathbf{A}' := 1**\dots*$ of \mathbf{A} , restrict h to $h': \mathbf{A}' \rightarrow \mathbb{R}$ and compute the epistatic filtration $\mathcal{F}(h')$. Each epistatic interaction of (\mathbf{A}', h') now has the first species as a bystander. In order to study the context dependence of $\mathcal{F}(h')$, we pass to its parallel epistatic filtration: Parallely translating $1**\dots*$ onto the facet $0**\dots*$ of the n -cube associates to each critical interaction (s, t) of $\mathcal{F}(h')$ a parallel epistatic interaction (s_0, t_0) , i.e. a bipyramid in the point configuration $0**\dots*$, for which the first species is always absent. The parallel epistatic interactions equally allow for the computation of epistatic weights and its p -values. Comparing these quantities of the original filtration $\mathcal{F}(h')$ with their parallel counterparts accounts for a method which contributes to the study of context dependence, cf. Section 1.6.6.

For this technique, an important situation arises when the occurrences of all but two species are prescribed. Then the point configuration \mathbf{A}' is the vertex set of a square, which has exactly two triangulations. Passing to any parallel square $\mathbf{B}' \subset \{0, 1\}^n$ by reverting the constraint for precisely one fixed species yields the sign $+$ if the triangulations of \mathbf{A}' and \mathbf{B}' induced by h parallely translate into each other, and the sign $-$ if the triangulations are reversed. Note that with the introduction of non-negative epistatic weights we omitted signs as the direct outcomes of numerical functions evaluating the height function h . In our opinion, this is necessary to accomodate the complicated combinatorics in higher dimensions.

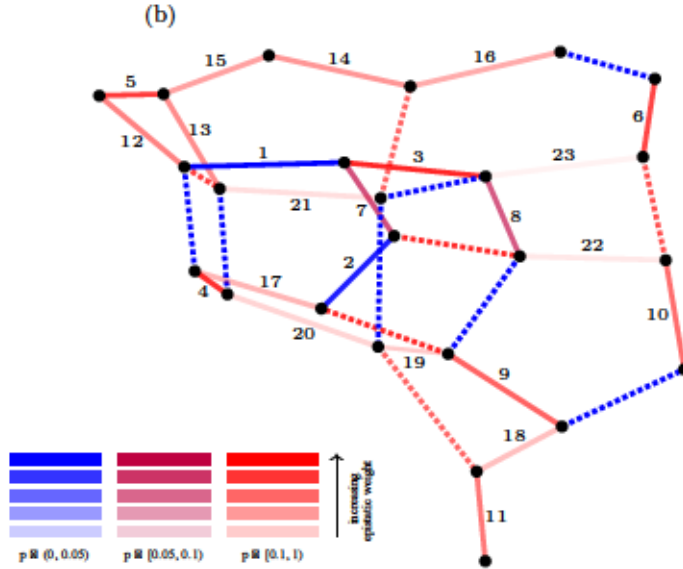
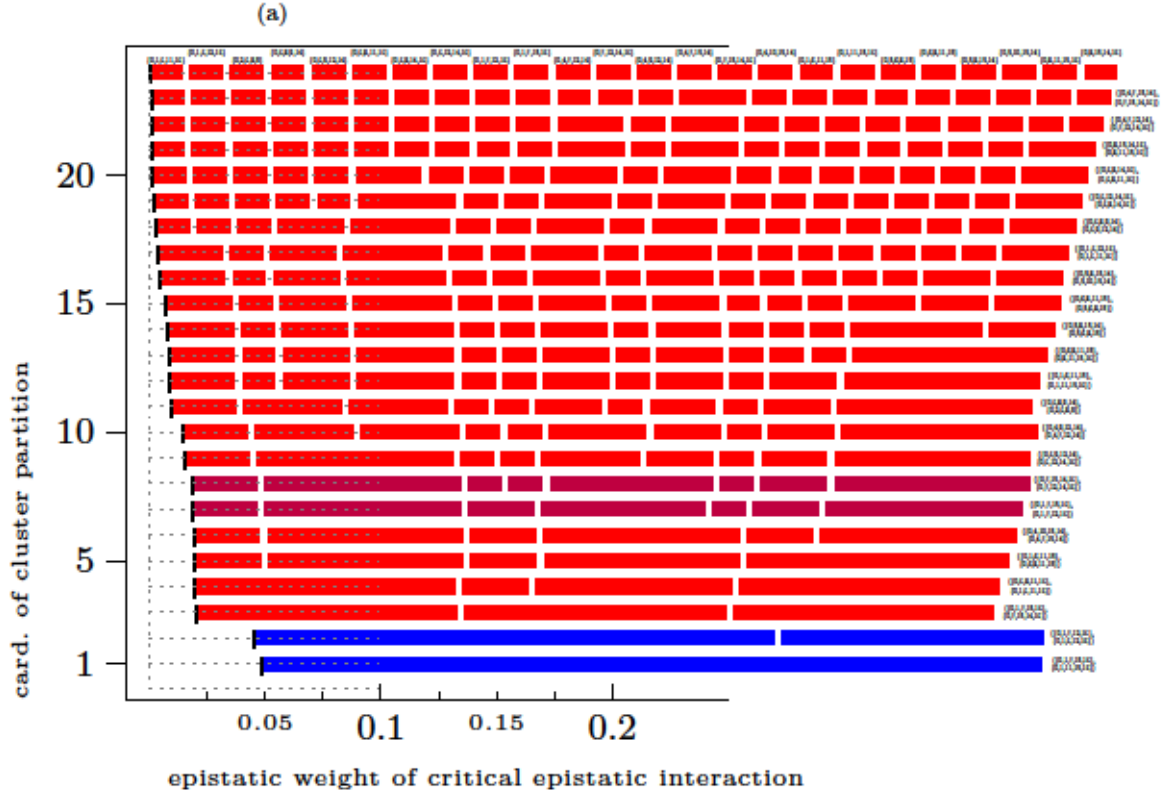


Figure 2: (a) Epistatic filtration of the 4-dimensional face $0****$ of the bi-allelic Eble data set on five loci. The top row showcases the cluster partition C_0 with 24 labeled maximal simplices. Every row below arises from a merging step (A_k) from Section 0.3.1 for a critical epistatic interaction which is indicated at the right end of that row. Altogether, this epistatic filtration has $23 = 24 - 1$ critical epistatic interactions, which is not a coincidence, cf. Section 0.4. The two significant interactions $(\{0, 1, 7, 12, 15\}, \{0, 1, 5, 12, 15\})$ and $(\{0, 1, 7, 13, 15\}, \{0, 1, 11, 13, 15\})$ are colored in blue, the two semi-significant interactions are purple. These labels refer to a specific encoding of the point configuration \mathbf{A} , cf. Figure 1(b). The first of these two significant interactions has satellite genotypes $5 = (0, 1, 1, 0, 0)$ and $7 = (0, 1, 0, 0, 1)$, both describing synergetic effects since both types of bacteria are involved. The bar diagram is reconstructed from a unique rooted binary tree with 24 labeled leaves, cf. Section 1.3.3. (b) The minimum spanning tree (bold) of the weighted dual graph $\Gamma(h)$, cf. Section 0.4, corresponding to the filtration of (a). The two significant interactions are close to each other in the shortest path metric.

0.4 Epistatic filtrations as minimum spanning trees

In fact, the epistatic filtration process as described in Section 0.3.1 can be reformulated in terms of a well-known optimization algorithm on a certain weighted graph, the dual graph $\Gamma(h)$ of $\Sigma(h)$. The nodes of the graph $\Gamma(h)$ correspond to the maximal simplices of $\Sigma(h)$ and there is an edge connecting two nodes of $\Gamma(h)$ precisely if the corresponding maximal simplices share a common codimension one face. To every edge $\{s, t\} \in E(\Gamma(h))$ we assign the epistatic edge weight $e_h(s, t) = e_h(t, s)$. In this terminology, the procedure described in the merging steps (A_k) of Section 0.3.1 produces a spanning tree of $\Gamma(h)$ with minimum total weight. In terms of biology, we may interpret this minimum spanning tree as a network of shortest genetic distances. The parameter space of all minimum spanning trees of $\Gamma(h)$, i.e. of all epistatic filtrations which can be realized over the fixed point configuration \mathbf{A} for height functions h varying in its secondary cone, is a fan.

Theorem (Theorem 3.4.3). *Let $sc(h)$ be the secondary cone of the regular triangulation $\Sigma(h)$. The cone $sc(h)$ splits into relatively disjoint full-dimensional parameter cones K_T , such that every choice $g \in \text{int}(K_T)$ produces the same minimum spanning tree $T(g) = T$ of $\Gamma(h)$ during the epistatic filtration process. The parameter cones K_T form a fan.*

In biology terms, the cones K_T are precisely the shapes of bar diagrams as in Figure 2, which can arise for a fixed regular subdivision $\Sigma(h)$, i.e. for a fixed cell decomposition of $\text{conv}(\mathbf{A})$ into fittest populations. In Section 3.4 we give an algorithm which computes the underlying fan structure. Further, we relate our results to tropical geometry, where the dual graph $\Gamma(h)$ is the bounded 1-skeleton of a certain tropical hypersurface associated to the height function h .

0.5 Counting chambers of hyperplane arrangements

The secondary fan of point configurations \mathbf{A} which are Lawrence polytopes describe a class of hyperplane arrangements in \mathbb{R}^A , cf. [40, Section 5.5.3]. In this case, the *chambers* of the hyperplane arrangement are precisely the interiors of the maximal cones of the secondary fan. In general, from an applied point of view one might primarily be interested in an explicit representation of every single maximal cone of the fan, in half-space description for instance. In Chapter 4, we discuss examples of hyperplane arrangements for which this is computationally not feasible with currently available algorithms and hardware. Still, some relevant information can be extracted using purely combinatorial techniques: Given a hyperplane arrangement \mathbf{H} in \mathbb{R}^n , the number of chambers of \mathbf{H} is in fact determined by its *characteristic polynomial* $\chi_{\mathbf{H}}(t)$, whose coefficients, the *Whitney numbers*, sum up in absolute value to the number of chambers of \mathbf{H} , cf. Section 4.2.1. Further, any hyperplane arrangement allows for two operations, *deletion* and *restriction*, which produce two new hyperplane arrangements, temporarily called artefacts of \mathbf{H} here in the introduction, and which are compatible with taking characteristic polynomials, cf. Lemma 4.3.1. This observation results in the elementary depth-first binary tree Algorithm 2. It produces a left child for every deletion and a right child for every restriction step. Globally taking a vertical perspective of the course of this algorithm, the idea of Algorithm 4 is now to horizontally identify artefacts of \mathbf{H} which coincide upto combinatorial symmetry of \mathbf{H} . Thus, Algorithm 4 is a breadth-first style algorithm whose strength lies in avoiding computations of redundant combinatorial information. This modification allowed us to compute the number of chambers of the ninth resonance arrangement given by 511 hyperplanes in \mathbb{R}^9 . This number reads 1,955,230,985,997,140 and the computation took roughly 11 days, cf. Section 4.7.

Cluster partitions and fitness landscapes of the *Drosophila* fly microbiome

This chapter is based on the published article “Cluster partitions and fitness landscapes of the *Drosophila* fly microbiome” [47] by Holger Eble, Michael Joswig, Lisa Lamberti and William Ludington. This version of the article has been accepted for publication, after peer review and is subject to Springer Nature’s AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: <https://doi.org/10.1007/s00285-019-01381-0>. The article appeared in the Journal of Mathematical Biology volume 79.3, pp. 861–899.

1.0 Abstract

The concept of genetic epistasis defines an interaction between two genetic loci as the degree of non-additivity in their phenotypes. A fitness landscape describes the phenotypes over many genetic loci, and the shape of this landscape can be used to predict evolutionary trajectories. Epistasis in a fitness landscape makes prediction of evolutionary trajectories more complex because the interactions between loci can produce local fitness peaks or troughs, which changes the likelihood of different paths. While various mathematical frameworks have been proposed to calculate the shapes of fitness landscapes, Beerenwinkel, Pachter and Sturmfels (2007) suggested studying regular subdivisions of convex polytopes. In this sense, each locus provides one dimension, so that the genotypes form a cube with the number of dimensions equal to the number of genetic loci considered. The fitness landscape is a height function on the coordinates of the cube. Here, we propose cluster partitions and cluster filtrations of fitness landscapes as a new mathematical tool, which provides a concise combinatorial way of processing metric information from epistatic interactions. Furthermore, we extend the calculation of genetic interactions to consider interactions between microbial taxa in the gut microbiome of *Drosophila* fruit flies. We demonstrate similarities with and differences to the previous approach. As one outcome we locate interesting epistatic information on the fitness landscape where the previous approach is less conclusive.

1.1 Introduction

In evolutionary biology, the concept of a fitness landscape plays a prominent role in the study of genetic mutations, evolutionary trajectories, and the consequences for organismal health and disease [41]. These landscapes were introduced by Sewall Wright in [156] and typically arise as high dimensional discrete or continuous genotype–phenotype mappings. The underlying coordinates in these mappings encode alleles at n genetic loci of interest and are called *genotypes*. The convex hull of these genotypes, which might be viewed as points in \mathbb{R}^n , is a polytope called the *genotope*. In non-technical terms, we can think of each genetic locus as a separate dimension. The collection of all combinations of these loci (e.g. of double mutants, triple mutants, etc.) forms a cube where the dimension n is the number of loci considered. Thus, for a two locus, bi-allelic set, the genotope is a square (i.e., two-dimensional cube) with vertices (i.e., corners) representing the wild type, each single mutant, and the double mutant.

Fitness landscapes then arise by mapping the reproductive success (or other fitness traits) of each genotype to its corresponding vertex on the cube. The concept of genetic epistasis defines an interaction between two genetic loci as the degree of non-additivity in their phenotypes. Shapes of fitness landscapes reveal the epistasis and are regular subdivisions of the genotope induced by genotype–phenotype mappings, meaning that the degree to which a set of vertices interact has a

physical shape that we can measure (as studied in Beerenwinkel et al. [10, 11]). Such subdivisions play a key role in determining interaction patterns among altered genes and pathways on the genotype, and they shed light on the possible orders in which genetic mutations might occur. For instance, where an epistatic interaction lowers fitness, evolutionary paths traversing that vertex would have lower odds of occurring. The study of interaction patterns is a general one that applies also to economies, social networks, and food webs in ecology.

Recently the gut microbiome has arisen in biology as a major factor shaping the genotype–phenotype mappings in animals [94]. The microbiome itself is an ecological interaction network of microbial species. Resolving the structure of the microbiome interaction patterns and their impacts on genotype–phenotype mappings in the host is a major unsolved problem in biology. However, the number of possible interactions is massive in general, as the number of species in the microbiome is on the order of hundreds to thousands. Thus, there is a great need to develop methods capable of detecting interactions without drowning in data. Here, we focus on a naturally simple gut microbiome, the *Drosophila* fruit fly's gut, with only five species of bacteria. For the purposes of the present paper, we keep the notations and terminology of genetic interactions, noting that they apply also to microbiome interactions.

In this work, we build on the approach developed by Beerenwinkel et al. [10, 11] and propose a new method to process metric interaction information, also known as epistasis, usually arising from interactions among altered genes. The main idea we bring to the theory of interactions are cluster partitions and cluster filtrations. These deal with connected components, called *clusters*, of some subgraph of the dual graph associated to a regular subdivision that is induced, e.g., by a genotype–phenotype mapping. To build these clusters, we first associate positive weights to pairs of maximal and adjacent simplices in the regular subdivision. Similar to an angle, these weights measure the deviation of the two adjacent simplices from being affinely dependent. Using these weights we form progressively bigger clusters by gluing the ones with lower weight together in a continuous process, which we call *cluster filtration*. The important new aspect in this algorithm is that the filtration process is designed to statistically distinguish an essential biological signal, encoded in the positive weight, from noise. In this way, filtrations enable us to handle and interpret the usually vast interaction information, which is currently only fully characterized for double and triple mutants; cf. [10, 11].

To describe the strength of our approach, we consider fitness landscapes for microbiome-modified *Drosophila* flies. We then describe similarities and differences to previous approaches. More precisely, the data set we inspect in this work consists of *Drosophila* flies prepared with up to five different bacterial species in their gut. We then view each bacterial combination as a genotype. In this way, the genotype is given by a 5-dimensional cube. The fitness landscapes we consider are defined by daily fecundity (referred to as *fec*), time to death (*ttd*) and development time (*dev*) mappings. For each such fitness landscape we study the induced regular subdivision and describe their properties in the language of clusters and cluster filtrations. To compare the epistatic information between fitness landscapes we use cluster partitions. Our results show that cluster filtrations detect interactions when these are present in the sense of [10, 11]. Additionally, we locate statistically relevant and previously undetected epistatic information. Comparing our findings with previous studies confirms that cluster filtrations can also be used to strengthen existing analysis and prompt new possible conclusions in interaction networks. Cluster partitions and their filtrations are a new mathematical idea, and we hope that this will find applications also beyond *Drosophila* microbiome data.

1.2 Mathematical background and terminology

Our approach relies on the theoretical framework for revealing epistatic interactions in genetic systems given in [10, 11], and we use the same terminology. The theory of regular subdivisions is developed in the 2010 monograph by DeLoera, Rambau, and Santos [39].

1.2.1 Genotypes and their regular subdivisions

We consider a fixed n -dimensional convex polytope P in \mathbb{R}^n . That is, P is the convex hull of finitely many points and we will assume that P affinely spans the entire space. A point $v \in P$ for which there exists an affine hyperplane which meets P only in v is called a *vertex* of P . The set of

vertices, denoted as V , forms the unique minimal set which generates P as the convex hull. Our second ingredient is a *height function* on the vertices, which is any function $h : V \rightarrow \mathbb{R}$ that assigns a real number to each vertex of P .

We will be particularly interested in the case where $V = \{0,1\}^n$, and $P = [0,1]^n$ is the n -dimensional unit cube. Following the approach of [10, 11] we call $[0,1]^n$ the *genotope* of an n -biallelic system. The vertices in $\{0,1\}^n$ are identified with binary strings of length n called *genotypes*. In the biological applications we have in mind, the points in the genotope correspond to the allele frequencies in a population; cf. Section 1.2.2 below. Height functions then correspond to traits, such as reproductive fitness of an organism or other experimental measurements —also called *phenotypes*— on the genotypes of the n -biallelic system.

The set of lifted points

$$V(h) := \{(v, h(v)) \mid v \in V\}$$

generates a polytope $P(h) := \text{conv}V(h)$ in \mathbb{R}^{n+1} . We will assume that the height function h is *nontrivial* in the sense that the lifted polytope $P(h)$ has full dimension $n+1$. By construction, the points in $V(h)$ are precisely the vertices of $P(h)$. In general, there are three types of facets of $P(h)$: if ν is an outward normal vector on the facet F , then F is called an *upper/vertical/lower facet* of $P(h)$ if the $(n+1)$ -st coordinate of ν is positive/zero/negative. It may happen that there are no vertical facets, but there are always upper and lower facets. The upper facets form a polyhedral ball sitting in the boundary complex of the lifted polytope $P(h)$. Projecting them back to P , by omitting the last coordinate, yields a polyhedral subdivision $\mathcal{S} = \mathcal{S}(V, h)$ of P . A *polyhedral subdivision* of P is a finite family of polyhedra, whose elements we call the *cells* of the subdivision, such that each face of a cell is a cell, and the intersection of any two cells is a (possibly empty) cell. Those subdivisions which are induced by a height function are called *regular*; cf. Definition 2.3.1 and Lemma 2.3.11 in [39]. A polyhedral subdivision for which all cells are simplices is called a *triangulation*. Triangulations induced by a height function are generic in the following sense. If each value of the height function is chosen at random (e.g., uniformly in a fixed interval) then the induced regular subdivision is a triangulation almost surely.

The height function h is called a *fitness landscape* in [10, §3]. The genotope subdivision $\mathcal{S}(V, h)$ is known as the *shape* of the fitness landscape h . Geometric properties of these shapes of fitness reflect interactions among organisms or genotypes. In the biological setting we have in mind, height functions are given by certain fittings of replicated measurements. Depending on the nature of the data, these fittings can, for instance, be means, medians or modes of the observed measurements, as well as expected values of nonparametric density estimators, described as in [58, Chapter 9]. As such we can assume that the height functions are generic, so they induce regular triangulations of the n -cube $[0,1]^n$. In the concrete computations below, we consider means of replicated measurements.

Before we continue with our exposition, let us consider an example which we will revisit later. This is about the smallest nontrivial case which arises.

Example 1.2.1. *We consider a 2-biallelic system, and so the genotope is the unit square $P = [0,1]^2$. Its four vertices 00, 10, 01 and 11 form the set V ; here, 01 is shorthand notation for the point with coordinates $(0,1)$. Assume that some measurement gives the height function h which reads*

$$h(00) = 53.25 \quad h(10) = 46.65 \quad h(01) = 43.16 \quad h(11) = 43.48 \quad . \quad (1.2)$$

The lifted polytope $P(h)$ is a 3-dimensional simplex (a.k.a. tetrahedron). In this case there are two upper and two lower facets; no vertical ones. Figure 1.3 shows the upper facets of $P(h)$ and the resulting genotope subdivision $\mathcal{S}(V, h)$. The latter is a triangulation with two maximal cells, indicated in green and red.

Remark 1.2.2. *Most polytopes admit triangulations which are not regular, i.e., not induced by any height function; cf. [39, Theorem 6.3.11] for examples of non-regular triangulations of $[0,1]^n$ for $n \geq 4$. While the triangulations of interest here will always be regular, the algorithmic methods discussed below generalize to the nonregular setting.*

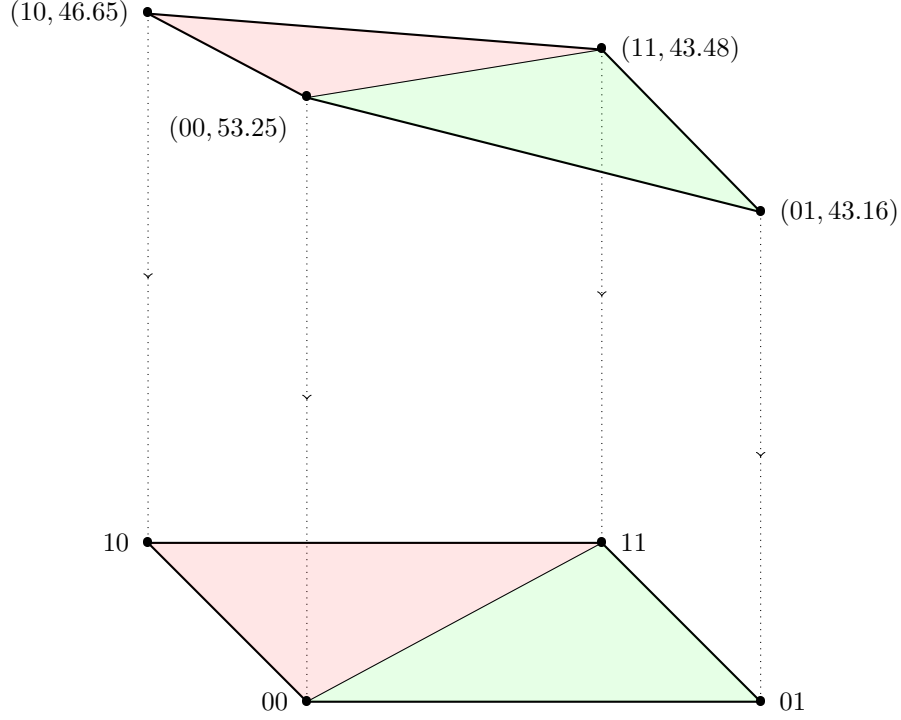


Figure 1.3: Upper facets of the lifted polytope $[0, 1]^2(h)$ and the induced regular triangulation $\mathcal{S}(\{0, 1\}^2, h)$.

1.2.2 Fittest populations

Again we consider the genotypes $V = \{0, 1\}^n$ of an n -biallelic system. A map $p : V \rightarrow \mathbb{R}$ is a (*relative*) *population* if it attains only nonnegative values which sum to one. This yields a point

$$\rho = \rho(p) := \sum_{v \in V} p(v)v$$

in the genotype $[0, 1]^n$, which is the *allele frequency vector*. Its coordinate ρ_i describes the probability for the population p to have allele 1 in its i -th locus. The set of all relative populations, denoted as Δ_V , is a simplex of dimension $2^n - 1$.

Now we add the height function $h : V \rightarrow \mathbb{R}$ to the picture. For a fixed allele frequency vector $w \in [0, 1]^n$ this gives rise to the linear program

$$\begin{aligned} & \text{maximize} && h \cdot p \\ & \text{subject to} && p \in \Delta_V \text{ and } \rho(p) = w . \end{aligned} \quad (\text{LP}(h, w))$$

The coordinates of the vector p are the variables to be determined. If h and w both are generic then $\text{LP}(h, w)$ has a unique optimal solution, the *fittest population* $p^* = p^*(h, w)$, and this a vertex of the polytope

$$\Delta_{V,w} := \{p \in \Delta_V \mid \rho(p) = w\} ,$$

which is contained in the 2^n -dimensional vector space \mathbb{R}^V . The condition $\rho(p) = w$ is equivalent to n linear equations, one for each coordinate of the allele frequency vector w . It follows that the fittest population p^* is the convex combination of at most $n + 1$ vertices of Δ_V , and the projection $\rho(p^*) = w$ gives rise to a representation of $w = \lambda_1 v_1 + \dots + \lambda_{n+1} v_{n+1}$, with $\lambda_i \geq 0$ and $\sum \lambda_i = 1$, as the convex combination of at most $n + 1$ genotypes $v_1, \dots, v_{n+1} \in V$. These genotypes are precisely the vertices of the unique simplex s of $\mathcal{S}(V, h)$ which contains w . The genericity of h implies that $\mathcal{S}(V, h)$ is a triangulation, while the genericity of w implies that the simplex s is unique. The optimal value of $\text{LP}(h, w)$ is $h \cdot p^* = \sum \lambda_i (h(v_i))$. In this way we obtain the piecewise linear function

$$\begin{aligned} h^* : [0, 1]^n &\longrightarrow \mathbb{R} \\ w &\longmapsto h \cdot p^*(h, w) \end{aligned}$$

on the genotype. Now the regions of linearity of h^* coincide with the maximal cells of the regular triangulation $\mathcal{S}(V, h)$.

Applying our methods to measurement data will almost always establish generic height functions, and thus the relevant polyhedral subdivisions will almost always be triangulations.

Remark 1.2.3. *For every fitness landscape h there are two shapes. One shape is induced by the upper facets of the convex hull $P(h)$, the other by the lower facets. Considering simultaneously both shapes might be advisable. However, previous approaches [10, 11] adopted the convention of considering upper facets in the definition of a regular subdivision, see also [64]. This convention is consistent with modelling the fitness of a population via the linear program $\text{LP}(h, w)$, which is a maximization problem.*

With a few straightforward technical adjustments, such as phrasing fitness in terms of a minimization problem, all our results also hold using lower facets.

1.2.3 The dual graph of a subdivision and its complexity

Let \mathcal{S} be a polyhedral subdivision of some n -dimensional polytope P . Two maximal cells of \mathcal{S} are *adjacent* if they share a common $(n-1)$ -dimensional cell. This adjacency relation induces a graph structure as follows: the nodes are the maximal cells (of dimension n), and an edge connects two nodes if the two cells are adjacent. This is known as the *dual graph* of \mathcal{S} , and we denote it by $\Gamma(\mathcal{S})$. Notice that $\Gamma(\mathcal{S})$ is always connected. The edges of $\Gamma(\mathcal{S})$ are the *dual edges* of \mathcal{S} .

The lemma below gives essential complexity bounds in the case of most interest to us. Let us denote the minimal number of maximal cells of any triangulation of $[0, 1]^n$ by $k_*(n)$.

Lemma 1.2.4. *Let \mathcal{S} be a triangulation of the unit cube $[0, 1]^n$, and let k be the number of nodes of $\Gamma(\mathcal{S})$. Then*

$$2^n - n \leq k_*(n) \leq k \leq n! . \quad (1.3)$$

The lower bound is attained if and only if $\Gamma(\mathcal{S})$ has no cycles. Moreover, the number of dual edges is at most $k(n+1)/2 - n \cdot k_(n-1) < (n+1)!$.*

Proof. The first part of the claim is a special case of [39, Theorem 2.6.1]. For the second part, observe that each n -simplex is adjacent to at most $n+1$ other simplices as this is the number of its facets. Yet \mathcal{S} induces a triangulation on each of the $2n$ facets of $[0, 1]^n$. Therefore, at least $2n \cdot k_*(n-1)$ of the $k(n+1)$ cells of dimension $n-1$ in \mathcal{S} lie in the boundary of $[0, 1]^n$. These $(n-1)$ -cells are contained in a unique maximal one. We arrive at the estimate of at most

$$k(n+1) - 2n \cdot k_*(n-1) \leq (n+1)! - 2n(2^{n-1} - n + 1)$$

incident pairs of nodes and edges of $\Gamma(\mathcal{S})$. Dividing by two gives the upper bound on the number of dual edges. \square

For $n = 3$ we get $2^3 - 3 = 5$ as the lower bound in (1.3), and $3! = 6$ as the upper bound. Both bounds are tight; i.e., there are triangulations of $[0, 1]^3$ with five and six facets, respectively. While, for $n = 4$, the lower bound in (1.3) is only $2^4 - 4 = 12$ we have $k_*(4) = 16$; cf. [39, Example 6.3.14]. To determine the exact lower bound $k_*(n)$ for $n \geq 8$ is a difficult open problem; cf. [39, §6.3.3] and Table 1.1 for an overview.

Table 1.1: The minimal number $k_*(n)$ of maximal cells of a triangulation of $[0, 1]^n$.

n	1	2	3	4	5	6	7	8
$k_*(n)$	1	2	5	16	67	308	1493	$\leq 11\,944$
$2^n - n$	1	2	5	12	27	58	121	248
$\frac{2^n n!}{(n+1)^{(n+1)/2}}$	1	1.54	3	6.87	17.78	50.78	157.5	524.41

The proof of Lemma 1.2.4 is based on the interplay between the size of a triangulation and the volumes of its maximal cells. This can be carried further to derive bounds which are better asymptotically. The key ingredient is Hadamard's famous problem of giving an upper bound for the determinant of a matrix with given entries; cf. [21] for a survey. In our context this yields the following.

Lemma 1.2.5. *The normalized volume of any simplex spanned by vertices of $[0, 1]^n$ is bounded by*

$$\frac{(n+1)^{(n+1)/2}}{2^n}.$$

This can be employed to derive a lower bound for $k_*(n)$. The even better bound

$$\frac{2^n n!}{(n+1)^{(n+1)/2}} \leq k_*(n) \quad (1.4)$$

arises from the Hadamard inequality for matrices with ± 1 coefficients; cf. [39, §6.3.3]. Starting from $n \geq 7$ the bound (1.4) is better than the more naive lower bound $2^n - n$ from Lemma 1.2.4.

1.3 Clusters in fitness landscapes

Our next goal is to show how epistatic information can be extracted from the dual graph associated to the triangulation of the genotype $[0, 1]^n$ by a given height function. To do this we first associate a positive weight with each dual edge, we then relate this information to interaction coordinates and epistasis. Later, we introduce cluster and cluster filtrations as a new tool to filter, summarize and analyze epistatic information.

1.3.1 Epistatic weight

As before, let P be an arbitrary n -polytope in \mathbb{R}^n , equipped with a generic height function h . This induces a regular triangulation $\mathcal{S} = \mathcal{S}(V, h)$, where V is the vertex set of P . Let s and t be two adjacent n -simplices in \mathcal{S} . Then there are altogether $n+2$ vertices v_1, v_2, \dots, v_{n+2} of P such that

$$s = \text{conv}\{v_1, v_2, \dots, v_{n+1}\} \quad \text{and} \quad t = \text{conv}\{v_2, v_3, \dots, v_{n+2}\}.$$

We consider the $(n+2) \times (n+2)$ -matrix

$$E_h(s, t) := \begin{pmatrix} 1 & v_{1,1} & v_{1,2} & \dots & v_{1,n} & h(v_1) \\ 1 & v_{2,1} & v_{2,2} & \dots & v_{2,n} & h(v_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & v_{n+2,1} & v_{n+2,2} & \dots & v_{n+2,n} & h(v_{n+2}) \end{pmatrix}, \quad (1.5)$$

where $v_{i1}, v_{i2}, \dots, v_{in}$ are the coordinates of $v_i \in \mathbb{R}^n$. The *epistatic weight* of the dual edge connecting s and t is then defined as

$$e_h(s, t) := |\det E_h(s, t)| \cdot \frac{\text{nvol}(s \cap t)}{\text{nvol } s \cdot \text{nvol } t}, \quad (1.6)$$

where $\text{nvol } s$ is the *normalized volume* of s , i.e., the determinant of the submatrix $N(s)$ obtained from $E_h(s, t)$ by omitting the last column and the row corresponding to the vertex v_{n+2} not lying in s . Similarly, $N(t)$ is the submatrix of $E_h(s, t)$ obtained by omitting the last column and the row corresponding to the vertex v_1 not lying in t . The determinant of $E_h(s, t)$ is the volume of the convex hull of the $(n+2)$ vertices of s and t lifted to \mathbb{R}^{n+1} by the height function h . The intersection $s \cap t$ is spanned by v_2, v_3, \dots, v_{n+1} and separates the two *satellite vertices* v_1 and v_{n+2} . Its normalized $(n-1)$ -dimensional volume $\text{nvol}(s \cap t)$ coincides with the n -dimensional normalized volume of a pyramid over $s \cap t$ with height 1.

The epistatic weight $e_h(s, t)$ vanishes if the lifted point configuration of $s \cup t$ with respect to h lies in a hyperplane and is positive otherwise. If the denominator of (1.6) is one, we say that the simplices s and t are *unimodular*. One then says that a *triangulation is unimodular*, if all its simplices are unimodular.

For $\lambda > 0$, the scaled height function λh over the scaled genotype λP provides the same combinatorial data, i.e., the same labeled maximal cells. The factor $\text{nvol}(s \cap t) / (\text{nvol } s \cdot \text{nvol } t)$ in (1.6) makes the epistatic weight $e_{\lambda h}(s, t) = e_h(s, t)$ invariant under the scaling by λ .

If h' is non-trivial and not generic, then $\mathcal{S}' = \mathcal{S}(V, h')$ is a non-trivial regular subdivision. By Lemma 2.3.4 in [39], all maximal cells of \mathcal{S}' are full-dimensional, but not necessarily simplices. Let

C and D be two maximal cells in \mathcal{S}' which are adjacent, i.e., the intersection $C \cap D$ is a common face of codimension one. A maximal collection of points v_1, v_2, \dots, v_{n+2} is called a *bipyramid* in the dual edge (C, D) of \mathcal{S}' if v_1, \dots, v_{n+1} are affinely independent vertices of C and v_2, \dots, v_{n+2} are affinely independent vertices of D . In this way a bipyramid spans a pair (s, t) of adjacent n -simplices contained in the union of C and D , and $s \cap t \subseteq C \cap D$. For each such pair (s, t) one can form the matrix $E_{h'}(s, t)$ as in (1.5), and find the corresponding epistatic value $e_{h'}(s, t)$ via (1.6). To define $e_h(C, D)$, we take the mean over all $e_{h'}(s, t)$ where (s, t) runs through the bipyramids in the dual edge (C, D) .

Example 1.3.1. We continue Example 1.2.1 where $P = [0, 1]^2$, and h is given by (1.2). The induced triangulation has precisely two maximal cells. The dual graph has a single edge connecting these two cells. We compute the epistatic weight

$$e_h(\{00, 10, 11\}, \{00, 01, 11\}) = \left| \det \begin{pmatrix} 1 & 0 & 0 & 53.25 \\ 1 & 1 & 0 & 46.65 \\ 1 & 0 & 1 & 43.16 \\ 1 & 1 & 1 & 43.48 \end{pmatrix} \right| \cdot \sqrt{2} \approx 9.786 . \quad (1.7)$$

Notice that the denominator in (1.6) is one. The factor $\sqrt{2}$ is the one-dimensional volume of the shared face $\text{conv}(\{00, 11\})$. For this 2-locus system, the computation (1.7) thus agrees with the usual epistasis formula of [10, Example 3.7] up to the factor $\sqrt{2}$, which does not depend on h :

$$\varepsilon(00, 11, 10, 01) := h(00) + h(11) - h(10) - h(01) .$$

Biologically, the non-vanishing of the epistatic weight means that the additive epistatic assumption is violated: the fitness of the double mutant is higher than what one would expect by knowing the fitness of the single mutants and the wild type.

Remark 1.3.2. Since we did not fix orderings of the vertices of s and t , the matrix $E_h(s, t)$ is only defined up to row reordering. However, our approach solely rests on the epistatic weights from (1.6); taking absolute values here makes those values independent of any ordering.

Let us now summarize the biological information encoded in the dual graph valued by the epistatic weight. First, each node in $\Gamma(\mathcal{S})$ corresponds to an $(n + 1)$ -tuple of genotypes that can be realized as the support of a fittest population, i.e., of an optimal solution of $\text{LP}(h, w)$ for some w . In this sense, we can state (by a slight abuse of language) that each edge of $\Gamma(\mathcal{S})$ describes the union of genotypes occurring in two fittest populations. This union consists of $n + 2$ genotypes with n genotypes shared by the two fittest populations and where the remaining two are satellites. The edges incident to a given node s in $\Gamma(\mathcal{S})$ thus encode exactly those genotypes which together with n genotypes of s form a fittest population in the above sense.

Finally, the epistatic weight associated with an edge $e = \{s, t\}$ of $\Gamma(\mathcal{S})$ measures how far the supporting genotypes of the two adjacent fittest populations s and t are away from being affinely dependent. In this sense, $e_h(s, t)$ can be seen as a deformation of the usual statistical correlation notion, see the discussion in Section 1.4. The non-vanishing of the epistatic weight of e thus means that knowing the fitness of the genotypes supported by the fittest population s does not allow us to deduce the fitness of the satellite genotype not in s .

1.3.2 Cluster partitions and epistatic filtrations

Let Γ' be a spanning subgraph of the dual graph $\Gamma = \Gamma(\mathcal{S})$ of the subdivision \mathcal{S} . *Spanning* means that Γ' has all the nodes of Γ but some dual edges may be missing. We call a connected component of Γ' a Γ' -*cluster*. Further, we call the partition of the nodes of Γ into Γ' -clusters the Γ' -*cluster partition* of \mathcal{S} . That is, a cluster partition is an additional combinatorial structure imposed on \mathcal{S} by the choice of the spanning subgraph Γ' .

Our next goal is to define a filtration process on \mathcal{S} by a sequence of nested cluster partitions. For this, consider a *threshold value* θ , where $\theta \geq 0$. The specification of a threshold value defines a not necessarily connected subgraph, $\Gamma(\theta)$, of Γ by deleting those dual edges whose normalized epistatic weight exceeds θ . The $\Gamma(\theta)$ -clusters and the $\Gamma(\theta)$ -cluster partition are shortened to θ -*clusters* and the θ -*cluster partition*, respectively.

The intuition behind these concepts comes from the following. Consider two height functions, h' and h'' , on the vertices of P such that h' is not generic and such that h'' is a small perturbation. More precisely, we assume that there is a positive number ε such that $|h''(v)| < \varepsilon$ for all $v \in V$. Our assumption on h' means that $\mathcal{S}(V, h')$ is not a triangulation. We call h'' *sufficiently generic* for h' if $h' + h''$ is generic, i.e., $\mathcal{S}(V, h' + h'')$ is a regular triangulation. Note that it does not suffice to require h'' to be generic: e.g., if h' is generic, then $-h'$ is generic, too, but their sum $h' + (-h') \equiv 0$ is not. Now θ -cluster partitions can detect the perturbation by h'' in the following sense.

Theorem 1.3.3. *For every height function h' there are positive numbers ε and θ such that the following holds: If h'' is sufficiently generic for h' and additionally satisfies $|h''(v)| < \varepsilon$ for all $v \in V$, then each maximal cell of $\mathcal{S}(V, h')$ corresponds to exactly one θ -cluster of the triangulation $\mathcal{S}(V, h' + h'')$.*

Proof. First assume that h' is generic, and thus the partition $\mathcal{S} = \mathcal{S}(V, h')$ induced by h' is a triangulation. Pick $\varepsilon > 0$ sufficiently small such that for $|h''(v)| < \varepsilon$ we have $\mathcal{S}(V, h' + h'') = \mathcal{S}(V, h')$. That is, h' and $h' + h''$ lie in the same secondary cone, defined as in [39, Def. 5.2.1]. Such an ε can always be found since the secondary cone of a triangulation is an open subset of \mathbb{R}^{m-n-1} , where m is the cardinality of V ; cf. [39, §5.2.1]. Then the claim in the generic case becomes trivial, e.g., with $\theta = 0$. The maximal cells of $\mathcal{S}(V, h')$ are precisely the 0-clusters of $\mathcal{S}(V, h' + h'')$.

A second case arises when $\mathcal{S}(V, h')$ has only one maximal cell given by the entire polytope $\text{conv}(V)$. Then we can pick $\varepsilon > 0$ arbitrary and make θ sufficiently large such that the θ -cluster partition of $\mathcal{S}(V, h' + h'')$ comprises a single cluster.

Now we consider the only interesting case where h' is not generic and $\mathcal{S} := \mathcal{S}(V, h')$ is a non-trivial regular decomposition of V induced by h' which is not necessarily a triangulation. Here we pick $\varepsilon_1 > 0$ small enough such that for all sufficiently generic h'' with $|h''(v)| < \varepsilon_1$ the subdivision $\mathcal{S}(V, h' + h'')$ is a triangulation which refines \mathcal{S} . Such an ε always exists as shown for instance in Lemma 2.3.15 in [39]. This claim can equivalently be expressed by saying that h' lies in the boundary of the (full-dimensional) secondary cone of $\mathcal{S}' := \mathcal{S}(V, h' + h'')$.

Let C and D be two maximal cells in \mathcal{S} which are adjacent. Let (s, t) be the adjacent n -simplices spanned by a bipyramid contained in the union of C and D , and $s \cap t \subseteq C \cap D$. Let $e_{h'}(s, t)$ be the corresponding epistatic value given by (1.6). Now we let

$$\theta := \frac{1}{2} \cdot \min \{e_{h'}(s, t) \mid s \cup t \text{ bipyramid in some dual edge of } \Gamma(\mathcal{S})\} , \quad (1.8)$$

which is the minimum taken over a finite set of non-zero positive real numbers and thus $\theta > 0$. Let (s', t') be adjacent n -simplices in the triangulation \mathcal{S}' . We call the dual edge (s', t') *local* if s' and t' are contained in some maximal cell of \mathcal{S} . Now, the maximal cells of \mathcal{S} belong to a θ -cluster partition of \mathcal{S}' if and only if

$$e_{h'+h''}(s', t') \begin{cases} < \theta & \text{if } (s', t') \text{ is local dual edge} \\ \geq \theta & \text{otherwise} . \end{cases} \quad (1.9)$$

We observe that setting $h'' \equiv 0$ yields $e_{h'+h''}(s', t') = 0$ for any local dual edge (s', t') since then the $n+2$ vertices of $s' \cup t'$, lifted by $h' = h' + 0$ are contained in some hyperplane. Hence, since the determinant is multilinear (and thus continuous), we can find $\varepsilon_2 > 0$ such that all h'' with $|h''(v)| < \varepsilon_2$ satisfy $e_{h'+h''}(s', t') < \theta$ for all local dual edges (s', t') . An explicit expression for ε_2 can be given in terms of the maximal minors of the matrix formed from all vertices lifted by h ; we leave the details to the reader. In this way, all local dual edges of \mathcal{S}' are contained in a θ -cluster. For a nonlocal dual edge (s', t') of \mathcal{S}' observe that $s' \cup t'$ is a bipyramid in some dual edge of \mathcal{S} and $e_{h'}(s', t') > \theta$ holds by definition (1.8). Again by continuity of $e_{h'+h''}$ in h'' we get an ε_3 -neighbourhood of h' where all nonlocal dual edges of \mathcal{S}' have epistatic weight lying above θ and form singleton θ -clusters. Setting $\varepsilon := \min(\varepsilon_1, \varepsilon_2, \varepsilon_3)$ settles the claim. \square

Example 1.3.4. *We further continue the Example 1.2.1. Suppose $h' \equiv 0$ is the null function and h'' is the height function h given in (1.2). Then $\mathcal{S}(\{0, 1\}^2, h')$ is the trivial subdivision of $[0, 1]^2$ by itself, and $\mathcal{S}(\{0, 1\}^2, h' + h'')$ is the triangulation shown in Figure 1.3. The epistatic weight of the single edge is 9.786 as determined in (1.7). That is, for all $\theta \geq 9.786$ the cluster partition recovers the trivial subdivision of the unit square $[0, 1]^2$.*

Varying θ yields a stepwise coarsening of \mathcal{S} into larger and larger clusters which we call the *cluster filtration* of \mathcal{S} . For sufficiently small values of θ (including, e.g., $\theta = 0$) the θ -cluster partition simply consists of the partition of the node set of Γ into singletons. On the other hand, for sufficiently large values of θ the θ -cluster partition consists of a single cluster which comprises all the nodes. Letting the threshold value vary between these two extremes provides a simple descriptor of the “biologically relevant signal” in the data and allows us to separate epistatic information from noise. Therefore, we also use the name *epistatic filtration* instead of “cluster filtration”. For an illustration of an epistatic filtration see Section 1.3.4.

A threshold value θ is *critical* if the cluster partition $\mathcal{S}(\theta)$ differs from $\mathcal{S}(\theta - \varepsilon)$ for all $\varepsilon > 0$. In this case, θ is necessarily the epistatic value of some dual edge, and $\mathcal{S}(\theta)$ has strictly fewer clusters than $\mathcal{S}(\theta - \varepsilon)$. However, the converse is not true: there may exist dual edges whose epistatic values are not critical. An open interval (θ_0, θ_1) with critical thresholds $\theta_0 < \theta_1$ is *regular* if it does not contain any critical value.

From a more geometric perspective we could also use dihedral angles instead of our epistatic weights. This approach would yield a theoretical result similar to Theorem 1.3.3. Here, we refrain from doing so since in Section 1.4, we will take statistical information into account. There the height of a vertex is actually a mean value, and we do not see a natural way to extend the statistics to dihedral angles.

Remark 1.3.5. *The tight span of an arbitrary n -dimensional polyhedral subdivision \mathcal{S} (of some point configuration) is a CW-complex \mathcal{S}^* whose 0-skeleton is formed by the maximal cells of \mathcal{S} ; cf. [71]. The k -cells of \mathcal{S}^* correspond to those subsets of the maximal cells of \mathcal{S}^* which share a common cell of dimension $n - k$. In this way, the 1-skeleton of \mathcal{S}^* agrees with the dual graph $\Gamma(\mathcal{S})$. If the subdivision is regular, its tight span is a polyhedral complex. Assigning epistatic weights to all cells of \mathcal{S}^* , not only to the edges, would open up a way to study more involved epistatic interactions.*

Remark 1.3.6. *The regular subdivision \mathcal{S} of any (rational) point configuration is dual to a tropical hypersurface [105, §3.1]. In this way our epistatic filtrations, which are defined on the dual graph $\Gamma(\mathcal{S})$, impose an additional structure on the 1-skeleton of any tropical hypersurface.*

Remark 1.3.7. *The graph $\Gamma(\mathcal{S})$ equipped with the epistatic weights induces a finite metric space (on the facets of \mathcal{S}), via taking shortest paths. Considering Vietoris–Rips filtrations then allow for studying the geometry of the lifted points by means of persistent homology [48]. It could be interesting to investigate if there is any connection with the epistatic filtrations.*

Studying the ramifications into higher-dimensional epistatic weights, tropical geometry or persistent homology looks very promising, but all these topics are beyond the scope of this article.

1.3.3 Computing epistatic filtrations

We now explain how to compute the epistatic filtration from the vertex set $V \subset \mathbb{R}^n$ and the height function $h : V \mapsto \mathbb{R}$ as input. In the first step we need to determine the list of maximal cells of the regular subdivision $\mathcal{S} = \mathcal{S}(V, h)$; the standard encoding of each maximal cell is as a subset of V . Determining \mathcal{S} is achieved by computing the convex hull of the lifted points in \mathbb{R}^{n+1} and selecting the facets with upward pointing normals. Computing convex hulls is a standard problem in computational geometry [136] with a somewhat delicate complexity status [136, Open problem 26.3.4]; see [3] for a recent survey from a practical point of view.

The input to the second step is the list of maximal cells of \mathcal{S} as subsets of V ; let k denote their number. If V are the vertices of the n -cube then $k \leq n!$ by Lemma 1.2.4. The k maximal cells form the nodes of the dual graph $\Gamma(\mathcal{S})$. To find the edges one can check the $\binom{k}{2}$ pairs of maximal cells, looking for those pairwise intersections which are maximal with respect to inclusion. This yields the edges of $\Gamma(\mathcal{S})$. In the most relevant special case where \mathcal{S} is a triangulation, the maximal pairwise intersections are precisely those of cardinality n . So the total cost for this step amounts to $\mathcal{O}(k^2 n)$. Let ℓ denote the number of dual edges. If V are the vertices of the n -cube then $\ell \leq k(n+1)/2 - n(2^{n-1} - n + 1) < (n+1)!$ by Lemma 1.2.4. Depending on the method used for the first step, this second step of finding the dual graph may not be necessary, since some convex hull algorithms produce it as a side product [136, 3].

The third step is to find the ℓ epistatic weights, each of which is gotten by computing three determinants of size at most $(n+2) \times (n+2)$. This adds up to a total cost of $\mathcal{O}(\ell n^3)$. Sorting the dual edges ascendingly by epistatic weight takes $\mathcal{O}(\ell \log \ell)$.

In the fourth and final step we create the epistatic filtration as a rooted binary tree. We iterate over the thresholds, which define cluster partitions. On the way we maintain a forest where each tree represents one cluster in the corresponding partition. Initially, each tree in the forest is an isolated node, one for each maximal cell of \mathcal{S} . For each dual edge (s, t) with the next epistatic weight θ we merge clusters of $\mathcal{S}(\theta - \varepsilon)$ into clusters of $\mathcal{S}(\theta)$. Then we remove the trees T and T' containing the leaf nodes corresponding to s and t and add one tree with a new root and T, T' as children. By convention, the left child should always be smaller than the right child with respect to some linear order. In our calculations, we use the lexicographic order on the underlying vertex sets. The process ends with the dual edge of highest epistatic weight, and we obtain a rooted binary tree that is uniquely determined up to the choice of the order. The running time of this step is linear in the number of nodes in the resulting tree. Any binary tree has less than twice as many nodes as leaves; hence the cost adds up to $\mathcal{O}(k)$.

Altogether we arrive at a complexity of

$$\mathcal{O}(k^2 n + \ell n^3 + \ell \log \ell + k) = \mathcal{O}(k^2 n + \ell n^3)$$

for the steps two through four. Note that $\ell \log \ell \leq k^2 n$ in view of $\ell \leq kn$ by Lemma 1.2.4. The first step, which requires a convex hull computation, is the bottleneck.

We implemented our method in `polymake` [56, 3], and this was used to produce all computational results presented in this article.

1.3.4 An extended example

To illustrate the concepts described in Section 1.3, let P be a 3-dimensional cube contained in $[0, 1]^5$ with vertex set $V^{(3)}$ given by

$$\begin{aligned} 0 &= 00000 ; & 1 &= 10000 ; & 5 &= 00001 ; & 9 &= 10001 ; \\ 4 &= 00010 ; & 8 &= 10010 ; & 15 &= 00011 ; & 21 &= 10011 . \end{aligned}$$

The vertex labels with their corresponding bit strings (e.g., 4=00010 means that vertex 4 corresponds to bit string 00010) are the genotypes listed in Table 1.9 and are viewed as points in \mathbb{R}^5 . Consider a height function $\text{ttd}^{(3)}$ assigning the following values to the eight vertices of P :

$$\begin{aligned} 0 &\mapsto 53.25 ; & 1 &\mapsto 46.65 ; & 5 &\mapsto 43.16 ; & 9 &\mapsto 43.48 ; \\ 4 &\mapsto 48.3 ; & 8 &\mapsto 47.79 ; & 15 &\mapsto 43.53 ; & 21 &\mapsto 40.71 . \end{aligned} \tag{1.10}$$

Specifically, this height function $\text{ttd}^{(3)}$, is given by restricting the time to death fitness landscape defined over the whole $[0, 1]^5$ to the 3-cube with vertices as above. Details about this fitness landscape and others are given in Section 1.6.1.

The induced triangulation $\mathcal{S} := \mathcal{S}(V^{(3)}, \text{ttd}^{(3)})$ has six maximal cells:

$$\begin{aligned} A &= \{0 \ 1 \ 8 \ 9\} ; & B &= \{0 \ 5 \ 9 \ 15\} ; & C &= \{0 \ 8 \ 9 \ 15\} ; \\ D &= \{8 \ 9 \ 15 \ 21\} ; & E &= \{0 \ 4 \ 8 \ 15\} . \end{aligned}$$

A basic combinatorial invariant of an n -dimensional polyhedral complex is its *f-vector* $f = (f_0, f_1, \dots, f_n)$, where f_k is the number of k -dimensional cells. Here we have $f(\mathcal{S}) = (8, 18, 16, 5)$. For the tight span, which agrees with the dual graph, we get $f(\mathcal{S}^*) = (5, 4)$.

A direct computation shows that the normalized volume of the cell D is two whereas all the other cells have normalized volume one. This makes \mathcal{S} a non-unimodular triangulation. Its dual graph $\Gamma(\mathcal{S})$ is shown in Figure 1.4. For each edge in $\Gamma(\mathcal{S})$ we can compute the epistatic weight using the determinant expression from (1.6). For instance, the 3-simplices C and D are adjacent in \mathcal{S} , and we have

$$e_{\text{ttd}^{(3)}}(C, D) = |\det E_{\text{ttd}^{(3)}}(C, D)| \cdot \frac{\sqrt{3}}{2} \approx 0.113 .$$

This computation reveals that $\text{ttd}^{(3)}$ almost induces a linear dependence among the lifted points indexed by vertices of C and D , i.e. $\{(v, \text{ttd}^{(3)}(v)) | v \in C \cup D\}$. The computations in Example 1.4.4

also provide no evidence against the vanishing of $e_{\text{ttd}^{(3)}}(C, D)$. Therefore, we can assume that the additive assumption holds, thus knowing the fitness of the genotypes belonging to C allows us to deduce the fitness of the unique (satellite) genotype of D which is not in C .

A similar computation of the epistatic weights of the remaining dual edges yields the epistatic filtration in Table 1.4. The rows are sorted with increasing epistatic weight.

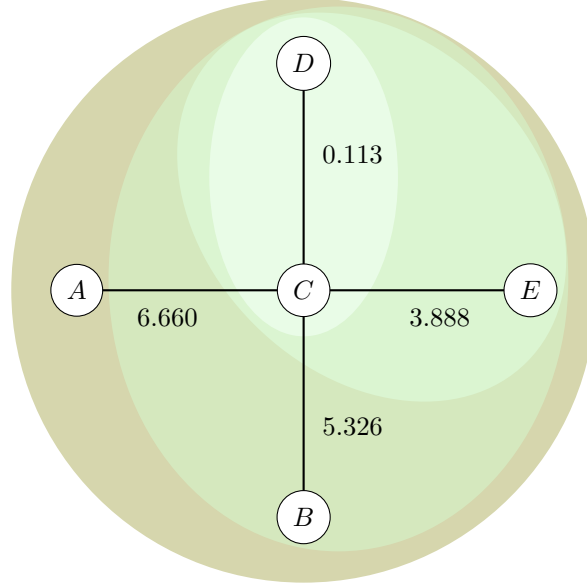






Figure 1.4: Dual graph $\Gamma(\mathcal{S})$ of \mathcal{S} induced by the restricted height function $\text{ttd}^{(3)}$. Each dual edge is labeled with its epistatic value. The shading of the colors indicates the nesting of the cluster partitions.

Table 1.2: Epistatic filtration arising from the restricted height function $\text{ttd}^{(3)}$. The color of the circle to the right of each cluster partition agrees with Figure 1.4.

θ	(s, t)	$\mathcal{S}(\theta)$	
0	—	$A B C D E$	
0.113	(C, D)	$A B CD E$	
3.888	(C, E)	$A B CDE$	
5.326	(B, C)	$A BCDE$	
6.660	(A, C)	$ABCDE$	

The initial cluster partition, for $\theta = 0$, consists of five connected components, one for each maximal cell of \mathcal{S} :

$$\mathcal{S}(0) = A|B|C|D|E ,$$

From Table 1.2 we see that, among the four dual edges of $\Gamma(\mathcal{S})$, the dual edge (C, D) is the one of lowest epistatic weight. This is the second row of Table 1.2 and we have

$$\mathcal{S}(0.113) = A|B|CD|E .$$

After three more steps for $\theta = 3.888, 5.326, 6.660$ we finally arrive at the trivial cluster partition

$$\mathcal{S}(6.660) = ABCDE ,$$

obtained from joining the adjacent simplices A and C . In this triangulation all dual edges are critical. That is, the cluster partitions arising from the epistatic weights of all dual edges are pairwise distinct.

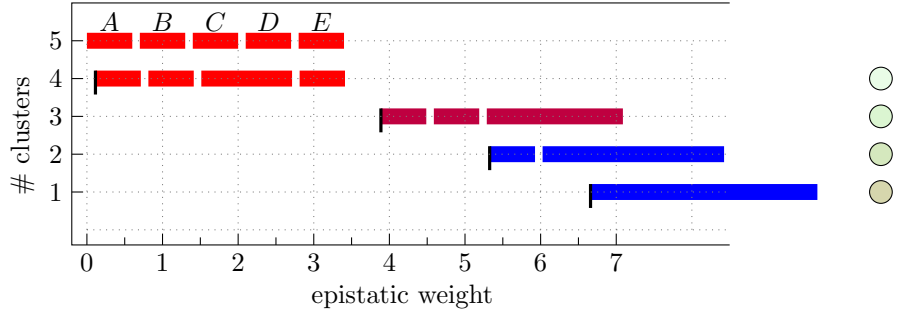


Figure 1.5: Visualizing the epistatic filtration from Table 1.2. The black ticks mark the pairs (θ, ℓ) , where θ is a critical threshold, and ℓ is the corresponding level. The cluster partitions are drawn consistently through all levels. The color of the circle to the right of each cluster partition agrees with Figure 1.4. The colors of the bars represent various levels of statistical significance ($p < 0.05$: blue, $0.05 \leq p < 0.1$: purple, $p \geq 0.1$: red); cf. Section 1.4.

The epistatic filtration is the sequence of nested cluster partitions which arises from increasing the threshold values. The whole process can be visualized as follows. For each critical threshold θ we mark the point (θ, ℓ) in a planar diagram, where ℓ is the *level*, i.e., the number of clusters in the θ -cluster partition. To the right of the marking (θ, ℓ) we draw the ℓ clusters as intervals such that the length of each interval is proportional to the size of the corresponding cluster. Any two subsequent levels are consistently drawn in the following sense, suppose that θ and θ' are two subsequent critical thresholds; i.e., the open interval (θ, θ') is regular. Let ℓ and ℓ' be the θ - and θ' -level, respectively. Then the θ -cluster partition refines the θ' -cluster partition or, conversely, the θ' cluster partition arises from joining clusters. That is, $\ell > \ell'$. To see which clusters get joined one can compare the two sequences of intervals starting from the left (or from the right). The length of each interval on level ℓ' indicates how many clusters of level ℓ get joined. Such a consistent way of drawing an epistatic filtration always exists since the nested cluster partition of all levels form a tree. By labeling the clusters on the top level, such a diagram encodes the entire epistatic filtration.

1.4 Significant cluster partitions

The purpose of our epistatic filtrations is to help separate “biologically interesting” epistatic information from noise; a geometric view is expressed in Theorem 1.3.3. Our next goal is to detect cluster partitions which are significant in a statistical sense. To do so, in Section 1.4.2 we develop a hypothesis test for edges in the dual graph $\Gamma(\mathcal{S})$.

1.4.1 Error analysis and standard deviation

Let P and V be as before. Throughout, for each $v \in V$ let X_v be a positive (absolutely) continuous random variable. Assume the first two moments of X_v exist and are given by $\mathbb{E}(X_v)$ and $\sigma_{X_v} = \sqrt{\mathbb{E}(X_v^2) - (\mathbb{E}(X_v))^2}$. We view $X = (X_v)$, for $v \in V$, as a vector of random variables and assume that each realization of X is a generic height function on V with probability one. Let $s = \text{conv}\{v_1, \dots, v_{n+1}\}$ and $t = \text{conv}\{v_2, \dots, v_{n+2}\}$ be two simplices which are spanned by points in V and which share a common codimension-1-cell. These are candidates for two adjacent maximal cells of $\mathcal{S}(V, X)$; such cells are simplices almost surely, as X is generic with probability one. In addition, let $E_X(s, t)$ be the matrix from (1.5) with h replaced by X . We write $E_i = E_X(s, t)_i$ for the matrix obtained from $E_X(s, t)$ by deleting the i -th row and the last column. Note that the coefficients of E_i are ones or coordinates of vertices in V . In particular, those coefficients do not depend on X or any other entries of the last column of the matrix given in (1.5).

Remark 1.4.1. *The epistatic weight was defined in the situation where the pair (s, t) forms a dual edge of some subdivision. Yet the formula (1.6) makes sense even without that assumption, i.e., for an arbitrary bipyramid.*

To simplify the exposition, in the following claim let $N := \text{nvol}(s \cap t) / (\text{nvol } s \cdot \text{nvol } t)$ and let $i, j \in \{1, n+2\}$. Let $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Y = |X| \sim \mathcal{FN}(\mu, \sigma^2)$ is again defined by μ and σ^2 and is called a *folded normal distribution*, see [145]. The density function of Y is given by:

$$f(y | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \left(e^{-\frac{1}{2\sigma^2}(y-\mu)^2} + e^{-\frac{1}{2\sigma^2}(y+\mu)^2} \right). \quad (1.11)$$

Proposition 1.4.2. *We set*

$$\lambda_i := (-1)^{n+i} N \det(E_i).$$

First, the expectation of the random variable $e_X(s, t)$ satisfies

$$\left| \sum_{i=1}^{n+2} \lambda_i \mathbb{E}(X_{v_i}) \right| \leq \mathbb{E}(e_X(s, t)) \leq \sum_{i=1}^{n+2} |\lambda_i| \mathbb{E}(X_{v_i}),$$

and its variance satisfies

$$\begin{aligned} \sigma_{e_X(s, t)}^2 &\leq \sum_{i=1}^{n+2} (N \det(E_i) \sigma_{X_{v_i}})^2 \\ &\quad + 2N^2 \sum_{1 \leq i < j \leq n+2} |\det(E_i \cdot E_j)| \sigma_{X_{v_i}} \sigma_{X_{v_j}}. \end{aligned} \quad (1.12)$$

Second, if the $n+2$ random variables X_{v_i} are mutually independent then

$$\sigma_{e_X(s, t)}^2 \leq \sum_{i=1}^{n+2} (N \det(E_i) \sigma_{X_{v_i}})^2. \quad (1.13)$$

Third, if in addition to independence the relation $X_{v_i} \sim \mathcal{N}(\mathbb{E}(X_{v_i}), \sigma_{X_{v_i}}^2)$ holds for all i , then $e_X(s, t)$ has folded normal distribution

$$\mathcal{FN} \left(\sum_{i=1}^{n+2} \lambda_i \mathbb{E}(X_{v_i}), \sum_{i=1}^{n+2} (\lambda_i \sigma_{X_{v_i}})^2 \right). \quad (1.14)$$

Proof. For a fixed ordering of the $n+2$ vertices in $s \cup t$, the numbers λ_i are real constants. To deduce the first claim, use the Laplace expansion of $E_X(s, t)$ along the last column and the linearity of the expected value. The left inequality follows by Jensen's inequality $|\mathbb{E}(Y)| \leq \mathbb{E}(|Y|)$, [132, Thm. B.17 in §.B.2.2]. The right inequality follows from the triangle inequality and since X_{v_i} are assumed to be positive. The approximations of the variance of $e_X(s, t)$ in (1.12) follow by the bilinearity of the covariance, Jensen's inequality and the Cauchy-Schwarz inequality [132, Thm. B.19].

If all X_{v_i} and X_{v_j} are mutually independent, then $\text{cov}(X_{v_i}, X_{v_j}) = 0$ in (1.12), and this settles (1.13) in our claim.

The last claim follows, e.g., from the convolution property of the density functions of X_{v_i} , see [132, §. B.1.3] or [38, Prop.7.17]. Passing to the absolute value implies that the distribution of $e_X(s, t)$ is the folded normal distribution, defined by the claimed parameters. \square

1.4.2 Significance test for the epistatic weight

Assume that the distribution mean $\mu = \mathbb{E}(e_X(s, t))$ is unknown. To test if μ is zero or not, we set up a one-sided test of significance. The null hypothesis is $\mu = 0$, and the alternative hypothesis is $\mu > 0$. To define the test statistics, for each $v \in V$, consider a sample $\mathbf{x}(v) = (x_1(v), x_2(v), \dots, x_L(v))$ of size L of independent and equally distributed realizations of X_v . The number L may vary across $v \in V$. Let $\bar{x} : V \rightarrow \mathbb{R}$ be defined by the sample mean \bar{x}_v at each $v \in V$. Let \bar{X}_v be the random variable evaluating to \bar{x} . Since the sample size L is large enough, we assume that $\bar{X}_v \sim \mathcal{N}(\bar{x}_v, s_{\bar{x}_v})$, for each $v \in V$, and where $s_{\bar{x}_v} = \sqrt{\sum_{i=1}^L (x_i(v) - \bar{x}_v)^2 / \sqrt{L}}$. Let s and t be two adjacent simplices of the triangulation induced by \bar{x} . We define the test statistics for the dual edge

(s, t) to be $z = e_{\bar{X}}(s, t)$. Assuming pairwise independence of X_v for $v \in V$, we deduce that the random variable Z evaluating to z satisfies $Z \sim \mathcal{FN}(0, \sigma_{e_{\bar{X}}(s, t)}^2)$ under the null assumption.

The validity of the null hypothesis is then deduced by computing the p -value of the test:

$$P(Z \geq z) = \int_z^\infty \frac{\sqrt{2}}{\sigma_{e_{\bar{X}}(s, t)} \sqrt{\pi}} e^{-\frac{1}{2} \left(\frac{y}{\sigma_{e_{\bar{X}}(s, t)}} \right)^2} dy, \quad (1.15)$$

where the integrand is the density function given in (1.11) and $\sigma_{e_{\bar{X}}(s, t)}$ is estimated as in Prop. 1.4.2. We call the dual edge (s, t) statistically significant if its p -value fulfills $p < 0.05$. In this case the null hypothesis can be rejected. Setting the significance level at 0.05 is a common choice; cf. [50].

Naturally higher epistatic weights are more likely to be significant. However, this does not always have to be the case as also the standard deviation of $e_{\bar{X}}(s, t)$ is taken into consideration in this test.

Remark 1.4.3. *Our assumptions on Z are plausible for the data analyzed in this paper, see Section 1.6.1. At the same time, these assumptions are permissive in terms of significance. Moreover, computing epistatic weights can be of interest, if the lifted point configuration is given by mutually independent random variables, as well as correlated random variables.*

1.4.3 Clusters from significant epistatic weights

We now explain how the notion of significant epistatic weights makes the cluster filtration process discussed in Section 1.3.2 into a biologically meaningful clustering algorithm. For this let $\mathcal{S}(V, \bar{X})$ be as above an induced regular triangulation of an n -polytope P equipped with a height function \bar{X} . Again we assume that \bar{X} assigns to each vertex of P the sample mean over a number of experimental measurements.

We saw that varying the parameter θ partitions \mathcal{S} into clusters ordered according to their epistatic weight. This can be used to discard noisy signal from relevant data. However, that approach does not take into account the dispersion of the experimental measurements. To account for this, we propose to combine the epistatic filtrations with the significance test discussed above. More precisely, we suggest to compute all epistatic weights for the dual graph of \mathcal{S} , contract the edges whose epistatic weight does not reach significance at $p < 0.05$ and unify the labels of the affected vertices. We call the remaining graph the *significant subgraph* $\Gamma_{\text{sig}}(\mathcal{S})$ of $\Gamma(\mathcal{S})$. This induces *significant clusters* and the *significant cluster partition* \mathcal{S}_{sig} . Now the epistatic filtration process from Section 1.3.2 and the algorithm from Section 1.3.3 carry over.

1.4.4 Continuation of the extended example

We now illustrate the above definitions on the example in 1.3.4. Consider again the regular triangulation $\mathcal{S} = \mathcal{S}(V^{(3)}, \text{ttd}^{(3)})$ induced by the height function $\text{ttd}^{(3)}$ from (1.10). Figure 1.4 shows the dual graph of \mathcal{S} ; the five maximal cells are labeled A, B, C, D, E . The value for each vertex is the sample mean over a large number of outcomes of replicated experiments. Therefore we now write \bar{X} instead of $\text{ttd}^{(3)}$. The standard error of the mean for the eight vertices is given by:

$$\begin{array}{cccc} 0 \mapsto 1.450 & 1 \mapsto 1.498 & 5 \mapsto 1.136 & 9 \mapsto 0.988 \\ 4 \mapsto 1.010 & 8 \mapsto 1.098 & 15 \mapsto 1.156 & 21 \mapsto 0.907 \end{array}.$$

Assuming independence, for each dual edge in $\Gamma(\mathcal{S})$ we now compute bounds on the associated standard deviation using (1.13). For instance, for $e_{\bar{X}}(C, D)$ consider





$$E_{\sigma_{\bar{X}}}(C, D) = \begin{pmatrix} 1 & 0 & 0 & 0 & 1.450 \\ 1 & 0 & 1 & 1 & 1.098 \\ 1 & 1 & 0 & 1 & 0.988 \\ 1 & 1 & 1 & 0 & 1.156 \\ 1 & 1 & 1 & 1 & 0.907 \end{pmatrix}.$$

This yields

$$\sigma_{e_{\bar{X}}}(C, D) \leq \left(\sum_{i=1}^5 (N \det(E_{\sigma_{\bar{X}}}(C, D)_i) \sigma_{\bar{X}}(v_i))^2 \right)^{1/2} \approx 2.586$$

with $N := \text{nvol}(C \cap D) / \text{nvol}(C) \cdot \text{nvol}(D)$. Processing the other epistatic weights in a similar fashion yields the values in Table 1.3. The rows are sorted by increasing epistatic weight. The bounds on the p -values in the last column are determined by (1.15).

Table 1.3: Significance of epistatic weights. The dual edge in bold is asserted to reach significance for $p < 0.05$.

(s, t)	$e_{\bar{X}}(s, t)$	$\sigma_{e_{\bar{X}}}(s, t) \leq$	$p\text{-value} \leq$	
(C, D)	0.113	2.586	0.965	
(C, E)	3.888	2.698	0.149	
(B, C)	5.326	2.844	0.061	
(A, C)	6.660	3.309	0.044	

In this example, the significant cluster partition \mathcal{S}_{sig} arising from the restricted height function \bar{X} reads $A|BCDE$.

Remark 1.4.4. *Since (1.13) only provides an upper bound on the p -value, it is useful to also investigate dual edges and epistatic weights whose p -value bounds are near 0.05. In this way, the dual edge (B, C) with epistatic weight 5.326 and p -value bound 0.061 comes into focus. It would be interesting to check if additional experiments involving the five genotypes 0, 5, 8, 9, 15 in $B \cup C$ lead to a higher level of significance or not.*

1.4.5 A synthetic experiment

In this section we describe one synthetic experiment on a quantitative analysis of Theorem 1.3.3 in relationship with the concept of statistical significance from Section 1.4. The purpose is to evaluate the extent to which changes in the height function for a single vertex impact the overall epistatic filtration.

We consider the vertex set $V = \{0, 1\}^5$ of the regular 5-cube and a height function η which takes every vertex to height 5 except for the wild type, which is mapped to $5 + \eta_0$ for some strictly positive real number η_0 . The wild type corresponds to the vertex 0; cf. Table 1.9. The combinatorial type of the regular subdivision $\mathcal{S}(V, \eta)$ does not depend on the precise value $\eta_0 > 0$. In fact, there are precisely two maximal cells, s_0 and t , and $\mathcal{S}(V, \eta)$ is a *vertex split* in the terminology of [69]. The cell s_0 is a simplex spanned by the wild type and its five neighbors (i.e., the standard simplex with the origin and the five unit vectors as its vertices). The other cell, t , is not a simplex; instead this is the convex hull of all 31 vertices different from the wild type. The intersection of s_0 and t is the regular 4-simplex spanned by the five unit vectors. So the dual graph $\Gamma(\mathcal{S}(V, \eta))$ has two nodes connected by the single dual edge (s_0, t) .

Our experiments depend on the choice of η_0 and a second strictly positive real number σ . To each vertex $v \in V$ we assign a normally distributed random variable X_v with zero mean and standard deviation σ . From 100 realizations per vertex we compute the resulting sample means and standard errors. This gives rise to a generic perturbation

$$\eta' = \eta + (\bar{X}_v | v \in V)$$

of the height function η by adding the sample means. For the resulting triangulation $\mathcal{S}(V, \eta')$ we compute the epistatic weights and the p -values (based on the standard errors computed). These are all the ingredients required for the significance test via (1.13). We start with a height function η at level 5 since the mean values \bar{X}_v from the perturbation may be negative; note that the perturbed height function η' needs to be strictly positive in order to qualify for the analysis via the p -values from (1.15).

We only consider perturbed height functions η' such that the simplex s_0 is a maximal cell of $\mathcal{S}(V, \eta')$, just as in $\mathcal{S}(V, \eta)$. For σ sufficiently small compared to η_0 this holds almost always. If s_0 is a maximal cell then s_0 is adjacent to some unique maximal cell of $\mathcal{S}(V, \eta')$. We call the corresponding dual edge the *bridge* of $\Gamma(\mathcal{S}(V, \eta'))$.

Now, for a fixed pair (η_0, σ) we repeat the above random construction 100 times, and we count how often the bridge is significant with respect to $p = 0.05$ and $p = 0.1$. In all the cases that

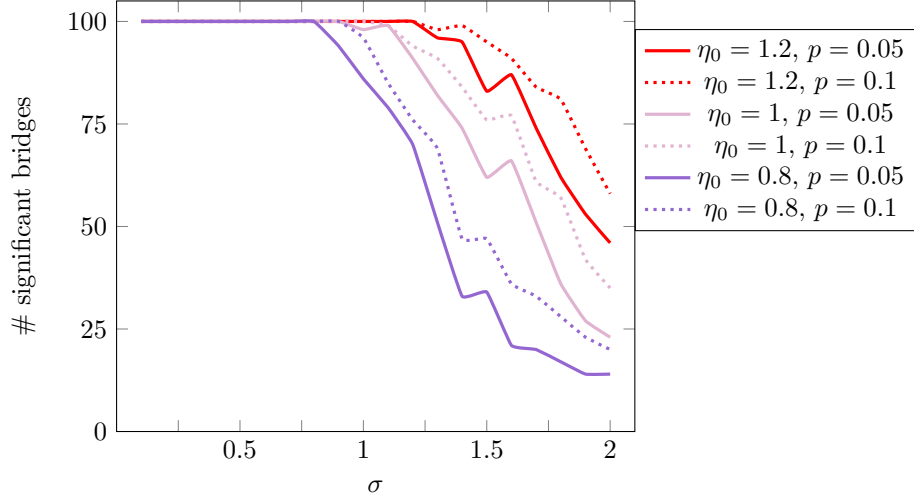


Figure 1.6: Percentage of significant bridges depending on σ , for various choices of η_0 and p .

we saw the simplex s_0 was a maximal cell, and the bridge existed. Further all perturbed height functions η' were nonnegative. Figure 1.6 shows the result for $\eta_0 \in \{0.8, 1.0, 1.2\}$ and $0.1 \leq \sigma \leq 2$.

The experimental results displayed in Figure 1.6 can be summarized as follows: for any fixed choice of η_0 and p the percentage of triangulations with the bridge present approximates a threshold function in the parameter σ . If σ is sufficiently low then the bridge is always significant; this observation can be seen as a variation of Theorem 1.3.3 for this particular setup. The lower the value of η_0 , the steeper the corresponding curve in Figure 1.6. For larger values of σ random fluctuations kick in, and this makes the curve less smooth.

This experiment provides evidence that our concept of significant cluster partitions is suitable to weed out small statistical fluctuations in the height function.

1.5 Epistasis, interaction coordinates and circuit interactions

In this section we compare the cluster partition to previous approaches. To do so, we now recall the biological phenomenon of epistasis as described in the work of Beerenwinkel et al. [10].

1.5.1 Interaction spaces

Let P be any n -dimensional convex polytope with vertex set V . Let \mathbb{R}^V be the real vector space of all height functions on V . Let \mathcal{L}_V be the subspace of \mathbb{R}^V consisting of all height functions on V for which the lifted polytope has dimension n . The *interaction space* is the quotient

$$\mathcal{I}_V := \left(\mathbb{R}^V / \mathcal{L}_V \right)^*.$$

Elements of \mathcal{I}_V are linear forms

$$\begin{aligned} \lambda: \mathbb{R}^V &\longrightarrow \mathbb{R} \\ h &\longmapsto \sum_{v \in V} \alpha_v h(v), \end{aligned}$$

with vanishing restriction $\lambda|_{\mathcal{L}_V}$. In the following, we call elements of \mathcal{I}_V *interactions*. The dimension of the interaction space is $\dim(\mathcal{I}_V) = \dim(\mathbb{R}^V) - \dim(\mathcal{L}_V) = |V| - \dim(P) - 1$.

When V is the vertex set of an n -cube, a basis for \mathcal{I}_V is given by the *interaction coordinates* defined up to multiplication by a scalar as:

$$\begin{aligned} u_{h,w}: \mathbb{R}^V &\longrightarrow \mathbb{R} \\ h &\longmapsto \sum_{v \in V} (-1)^{\langle v,w \rangle} h(v), \end{aligned}$$

where $v, w \in \{0, 1\}^n$ are vertices of P and w is assumed to have at least two coordinates being 1.

When $n = 2$, and a height function is fixed, then $u_{h,11} = \varepsilon(00, 01, 10, 11)$ as defined in Example 1.3.1. A possible generalization of the usual epistasis formula arises by considering *circuit interactions*, as defined in [10]. These are linear forms contained in the interaction space with support given by a minimal affinely dependent set of vertices of P . Notice that if the height function h induces an affine function on $s \cup t$, then the lift of a set of affinely dependent points is also affinely dependent. Some examples of circuit interactions and possible epistatic interpretations are given in [10, Example 3.8, 3.9].

1.5.2 Epistatic weight interactions

In this work, a new distinguished set of interactions inside \mathcal{I}_V are given by the linear forms:

$$\begin{aligned} \mathbb{R}^V &\longrightarrow \mathbb{R} \\ h &\longmapsto e_h(s, t) \end{aligned}$$

where s and t are adjacent n -simplices with vertex set in V . When h is generic, s and t are given as maximal adjacent simplices in the induced regular triangulation $\mathcal{S}(V, h)$.

When $V = \{0, 1\}^2$, there is a unique epistatic weight which agrees with the absolute value of the linear form $u_{h,11}$. This is the only case where the epistatic weight agrees with the notion of interaction coordinates. Now we will summarize the relation to circuit interactions.

Let v_1, v_2, \dots, v_{n+2} be vertices in V such that $s = \text{conv}\{v_1, \dots, v_{n+1}\}$ and $t = \text{conv}\{v_2, \dots, v_{n+2}\}$ are two n -simplices. Then the $n + 2$ vertices in $s \cup t$ are affinely dependent and contain a unique *circuit*, i.e., a minimal affinely dependent set; cf. [39, §2.4.1]. Furthermore, $e_h(s, t) \in \mathcal{I}_V$ is a circuit interaction in the sense of [10].

1.6 Epistasis in *Drosophila melanogaster* fruit fly microbiomes

In this section, we use the above approach and analyze existing *Drosophila* microbiome data. In particular, we demonstrate similarities with and differences from the methods used in [10, 11] and locate interesting epistatic information where the previous approach is less conclusive.

1.6.1 Data

The data we use is published in Tables S1 in [61, p.30, Supplemental Material(SM)]. It consists of experimental measurements of *Drosophila* flies inoculated with all possible combinations of five bacterial species naturally present in the gut of wild flies. This dataset is remarkably complete as all 32 bacterial combinations are considered.

The data set includes measurements of time to death (days), daily fecundity (progeny/day/female) and development time (days). All measurements were repeated many times to give a mean and standard error for each measurement and bacterial combination. More details on the replications of measurements and experimental settings can be found in the Materials and Methods Section of [61, SM]. In this work, we consistently referred to the above fitness landscapes by the labels ttd, fec and dev. When restricting one of these fitness landscapes to smaller sub-genotypes, we add a superscription, as in Example 1.3.4.

1.6.2 Epistatic weight approach

With the approach developed in this paper one asks for general epistatic information in genotype–phenotype mappings. Contrary to previous studies, epistasis here is understood as a general deviation from additivity rather than a specific manifestation of it, quantified for instance by marginal, conditional, 2-, 3-, 4- or 5-way interactions, see [10] for the terminology.

In the specific *Drosophila* data set, the methods developed in this work allowed us to distinguish certain bacterial combinations with vanishing epistatic weights from statistically significant ones. Bacterial combinations with low epistatic weights are not expected to have synergistic or antagonistic interactions between the species, while such effects are expected for bacterial combinations with statistically significant epistatic weight. Filtrations then provide clusters of bacterial combinations, based on adjacent relationships between simplices in a triangulation. Clusters thus

determine interesting regions inside the genotope, which we propose should be the targets for further analysis. An example of such an analysis is given in Section 1.6.3.

Our method intentionally involves a relatively small number of tests, limited by the adjacency relations among maximal simplices in the triangulated genotope and are specific to each phenotype mapping. In Section 1.4 we provided statistical tools for the data analysis. These tools also facilitate comparisons with previous approaches. In [61] a number of specific epistatic formulas were classified as significant. These epistatic formulas include standard tests, contextual tests, interaction coordinates and circuits, as described in [61, §5, Math Supplement, SM, pp.62]. Significance was tested as described in [61, §6, Math Supplement, pp.69 SM]. The results of these tests and their significance was reported in Figure 4 A-E, [61, Main text].

Comparing our work with the results of these tests reveals important differences between the epistatic formulas of [61] and the epistatic filtrations examined here. Examples are given below.

1.6.3 The case of the entire $[0, 1]^5$

Filtrations for the fitness landscapes of $[0, 1]^5$ defined for the time to death (ttd), daily fecundity (fec) and development time (dev) provide insights on the different cluster patterns arising but revealed a unique significant epistatic weights. The significant dual edge in ttd is given by $e_{\text{ttd}}(s, t) \approx 5.435$ with a p -value of approximately 0.0376. This epistatic weight arises over the bipyramid with vertex set $s = \{0, 9, 12, 14, 15, 28\}$ and $t = \{0, 9, 12, 14, 27, 28\}$.

This fact has two biological implications. First, it shows that we have evidence against one 5-dimensional affine relation between the ttd measurements for 15 and 27, given the ttd measurements for $s \cup t \setminus \{15, 27\}$.

Second, we observe that $s \cup t = \{0, 9, 12, 14, 15, 27, 28\}$ is not a minimal affinely dependent set. Yet, the subset $\{14, 15, 27, 28\}$ is a minimal affinely dependent set, which gives rise to the linear form

$$\omega_h = (h(00101) - h(00011)) - (h(11101) - h(11011)) . \quad (1.16)$$

This linear form cannot be written as a 3-cube circuit interaction, and allows for a new biological interpretation. More specifically, the linear form of equation (1.16), compares the effect of two pairs of *Acetobacter* bacteria (*pasteurianus* with *orientalis*, resp. *tropicalis* with *orientalis*) in the joint presence, resp. absence, of both *Lactobacillus* bacteria. Evaluating equation (1.16) at ttd gives $|\omega_{\text{ttd}}| = \frac{1}{2} |\det E_{\text{ttd}}(s, t)| \approx 4.86$ with a p -value of approximately 0.038. Thus, there is evidence to believe that the above form of marginal epistasis is non-additive. This fact refines our first conclusion. By the discussion in Section 1.5.2, there is no other circuit interaction on $s \cup t$.

Significant outcomes for other linear forms on $[0, 1]^5$ studied in the context of epistasis are also possible. For instance, results of recent work imply that for ttd, 124 tests out of 936 resulted to be significantly different than zero, ($p < 0.05$), see [61, §5, Math Supplement, SM]. These 936 tests include: epistatic weights on all 2-faces in $[0, 1]^5$, all 20 circuit interactions, a - ℓ in [10], for all 3-faces in $[0, 1]^5$, and all interaction coordinates for the k -faces of $[0, 1]^5$, for $k \in \{3, 4, 5\}$, defined as in Section 1.5.1.

Remark 1.6.1. In [61, §6, Math Supplement, pp.69 SM] the discovery rate was corrected by Benjamini–Hochberg multiple testing correction method [12]. Correction methods of this type aim at decreasing the number of significant outcomes by varying the p -values of the tests. They are typically used when a large number of tests are made, making false discoveries more likely. Due to the fact that there are very few significant epistatic weights, here we refrained from applying similar correction methods.

1.6.4 The case of parallel facets inside $[0, 1]^5$

Consider the interaction coordinate:

$$\begin{aligned} u_{h,0*101} &= h(0*000) + h(0*010) + h(1*000) + h(0*101) \\ &\quad + h(1*010) + h(0*111) + h(1*101) + h(1*111) \\ &\quad - h(0*001) - h(0*100) - h(0*011) - h(1*001) \\ &\quad - h(1*100) - h(0*110) - h(1*011) - h(1*110) , \end{aligned} \quad (1.17)$$

where $* \in \{0, 1\}$. The two values of $*$ determine so called *four-way interactions*, [10], on parallel facets of $[0, 1]^5$. If $* = 0$, the above interaction is considered in the absence of the *Lactobacillus brevis* bacteria, otherwise the bacteria are present. The computations of [61, §5, Math Supplement, SM, pp.62] determine that $u_{h,0*101}$ is simultaneously significant for h given by fec, dev and ttd, only in the absence of the *Lactobacillus brevis* bacteria.

To further inspect the effect of inoculating *Lactobacillus brevis* bacteria in *Drosophila*, we compute the fitness landscape defined by the vertices in the summands of $u_{h,0*101}$ for both values of $*$ and for h given by fec, dev and ttd. Results are shown in Figure 1.7. For ttd and fec, different cluster patterns appear on the parallel facets. The presence of significant epistatic weights in the absence of the *Lactobacillus brevis* bacteria confirms that these bacteria significantly affect epistatic interactions. Further dissecting the significant epistatic weights, as above, as well as the filtration steps, restricts the set of possible bacterial combinations responsible for this effect. For dev, all epistatic weights are near zero. As a consequence, this phenotype produced no significant dual edges (all bars are red in Figure 5). We conclude that the epistatic filtration effect of *Lactobacillus brevis* on the *Drosophila* microbiome is most pronounced for the phenotypes of fec and ttd.

This result is in contrast to the local significance tests computed in [61, §5, Math Supplement, SM, pp.62]. Out of those tests, 15 interactions can be seen to be simultaneously significant for fec and dev. Thus, the two approaches to discover epistatic interactions reveal different biological insights, and it remains for future empirical investigations to determine which approach best captures the underlying biological phenomena.

1.6.5 The case of three-cubes inside $[0, 1]^5$

Table 1.4: Circuit interactions and interaction coordinates for $\text{ttd}^{(3)}$ on the 3-cube of Example 1.3.4 which reached significance for $p < 0.05$.

Interaction	$ Z $	σ_Z^2	p -value folded	Vertices
a	6.09	2.57	0.02	0 1 4 8
c	6.92	2.58	0.01	0 1 5 9
e	5.32	2.40	0.03	0 4 5 15
m	9.10	3.72	0.01	0 1 4 5 21
ς	7.69	3.83	0.04	0 1 8 9 15
u_{111}	9.23	3.32	0.01	0 1 5 9 4 8 15 21

To make the comparison between the two methods more explicit, consider again Example 1.3.4. As before, let $\text{ttd}^{(3)}$ denote the time to death fitness landscape restricted to the genotypes defining the 3-dimensional cube in $[0, 1]^5$ with vertex set $V^{(3)} = \{0, 1, 4, 5, 8, 9, 15, 21\}$. On the one hand, following [10] we know that there are 20 circuit interactions of interest and four interaction coordinates for such a 3-cube. Out of these 24 tests, six reached statistical significance for $p < 0.05$ and assuming that the test statistic satisfies $|Z| \sim \mathcal{FN}(\mu, \sigma_Z^2)$. These significant tests are reported in Table 1.6.5. Using the terminology of [10], these significant tests capture three forms of conditional epistasis (for a, c, e), an interaction coordinate (u_{111} , the three-way interaction) and the circuit interactions (m, ς , two bipyramids in the 3-cube).

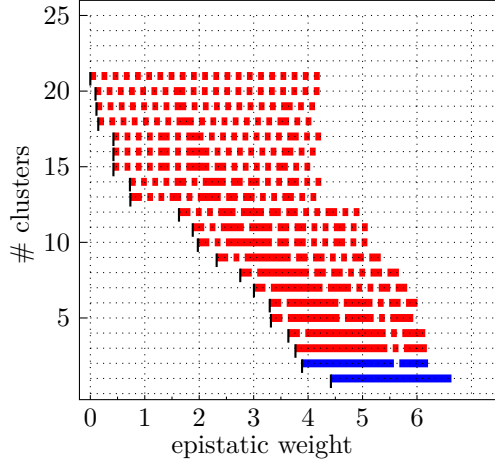
On the other hand, out of the four epistatic weights for $\mathcal{S}(V^{(3)}, \text{ttd}^{(3)})$, computed in Example 1.3.4, only $e_{\text{ttd}^{(3)}}(A, C)$ reached statistical significance; cf. Section 1.4.4. As before, we have

$$\varsigma_{\text{ttd}^{(3)}} = |\det E_{\text{ttd}^{(3)}}(A, C)|.$$

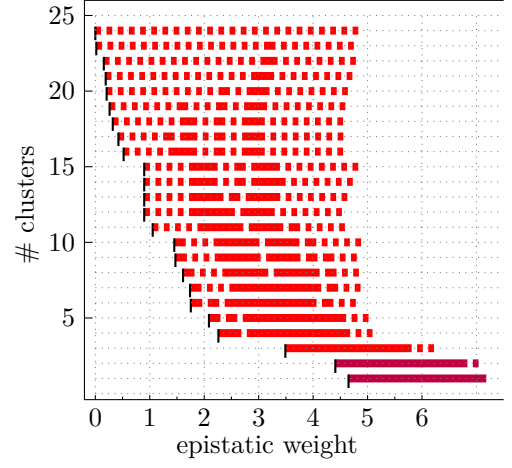
1.6.6 Parallel epistatic weights

In Section 1.6.4 we studied epistatic effects arising from the presence or the absence of one type of bacteria. Here we discuss a different way to address the same.

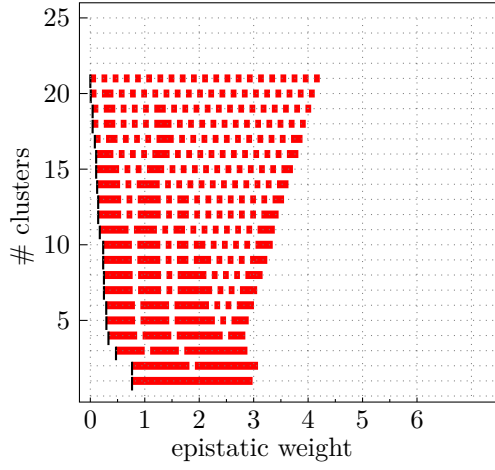
The presence or the absence of the k th type of bacteria defines one pair of parallel facets, which induce a partition on the full vertex set. We denote these facets as $F = * \cdots * 0 * \cdots *$ and $F' = * \cdots * 1 * \cdots *$, respectively; the 0 and 1 are in the k th position. The map, ϕ , which sends a



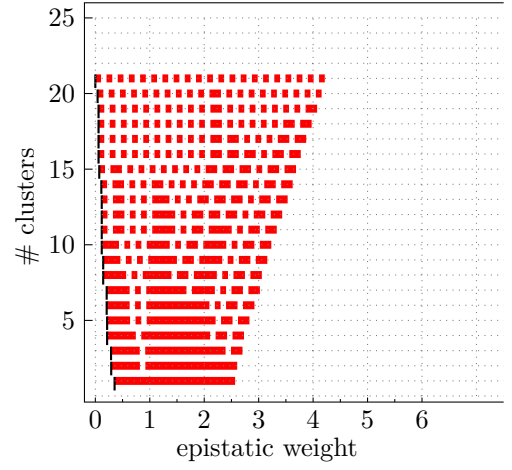
(a) ttd restricted to *0***



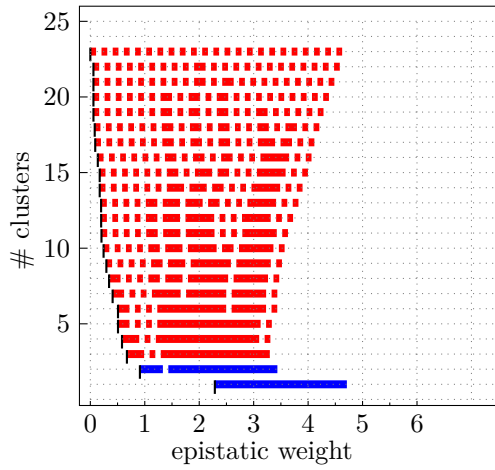
(b) ttd restricted to *1***



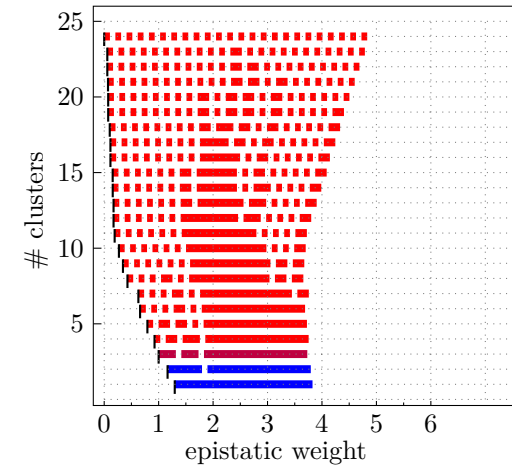
(c) dev restricted to *0***



(d) dev restricted to *1***



(e) fec restricted to *0***



(f) fec restricted to *1***

Figure 1.7: Filtrations of distinguished four-faces. The colors of the bars represent various levels of statistical significance ($p < 0.05$: blue, $0.05 \leq p < 0.1$: purple, $p \geq 0.1$: red); cf. Section 1.4. .

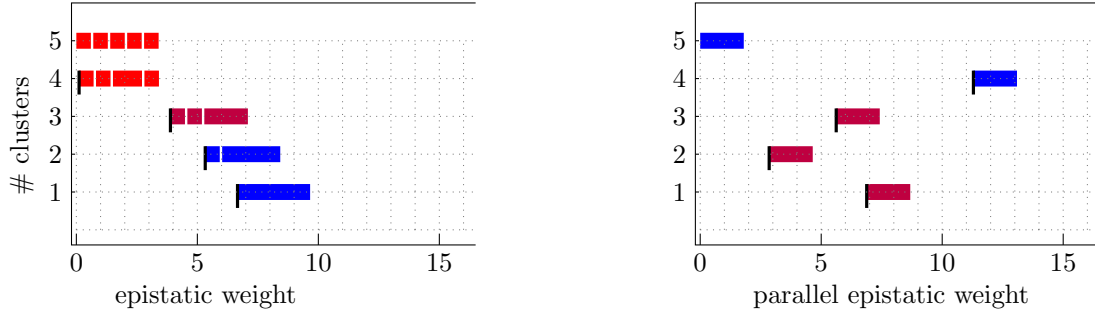


Figure 1.8: Left: Epistatic filtration of the triangulation $\mathcal{S}(V^{(3)}, \text{ttd}^{(3)})$ of the 3-cube $*00**$. Right: Parallel epistatic weights in $*10**$. The colors of the bars represent various levels of statistical significance ($p < 0.05$: blue, $0.05 \leq p < 0.1$: purple, $p \geq 0.1$: red); cf. Section 1.4.

vertex $v = (v_1 v_2 \dots v_n) \in \{0, 1\}^n$ to $v' = (v'_1 v'_2 \dots v'_n)$ where

$$v'_i = \begin{cases} 1 - v_k & \text{if } i = k \\ v_i & \text{otherwise} \end{cases}$$

is a reflection which exchanges F and F' . Restricting any generic height function on the entire cube $[0, 1]^n$ to the two facets induces two triangulations, of F and F' , respectively. Now we can compare the epistatic weights of the bipyramids arising from the triangulation of F with the epistatic weights of their images in F' under the reflection map ϕ . We call these *parallel epistatic weights*. Note that the bipyramids in F' , in general, are not bipyramids of the triangulation induced on F' . Still we can compute their epistatic weights, assess their statistical significance and compare; cf. Remark 1.4.1. Note that this also applies to any face of the n -cube, as faces of cubes are cubes.

Example 1.6.2. In Example 1.3.4, we investigated the epistatic filtration arising from restricting the height function ttd to the 3-face $*00**$. Now we can view $*00**$ as a facet of the 4-cube $**0**$, and $*10**$ is the parallel 3-face. Its vertices are

$$\begin{aligned} 2 &= 01000 ; & 6 &= 11000 ; & 12 &= 01001 ; & 18 &= 11001 ; \\ 11 &= 01010 ; & 17 &= 11010 ; & 24 &= 01011 ; & 28 &= 11011 . \end{aligned}$$

Restricting the height function ttd yields

$$\begin{aligned} 2 &\mapsto 52.175 ; & 6 &\mapsto 50.81 ; & 12 &\mapsto 46.79 ; & 18 &\mapsto 43.1 ; \\ 11 &\mapsto 45.46 ; & 17 &\mapsto 43.37 ; & 24 &\mapsto 44.97 ; & 28 &\mapsto 46.15 . \end{aligned} \tag{1.18}$$

Thus with $k = 2$, the bipyramid (C, D) in the triangulation $\mathcal{S}(V^{(3)}, \text{ttd}^{(3)})$ of $*00**$ gets reflected onto a bipyramid with vertex set $\{2, 17, 18, 24, 28\}$. The epistatic weight of this bipyramid restriction to the 3-cube, i.e., after deleting the second and third coordinate, is the parallel epistatic weight of $e_{\text{ttd}^{(3)}}(C, D)$ for the 3-face $*10**$. It is given by

$$\left| \det \begin{pmatrix} 1 & 0 & 0 & 0 & 52.175 \\ 1 & 0 & 1 & 1 & 43.37 \\ 1 & 1 & 0 & 1 & 43.1 \\ 1 & 1 & 1 & 0 & 44.97 \\ 1 & 1 & 1 & 1 & 46.15 \end{pmatrix} \right| \cdot \frac{\sqrt{3}}{2} \approx 11.289 . \tag{1.19}$$

The epistatic filtration of $\text{ttd}^{(3)}$ and the parallel epistatic weights in $*10**$ are visualized in Figure 1.8, where the bipyramid (C, D) is processed in level 4. Note that the image on the left of that same figure is a horizontally squeezed version of Figure 1.5.

Figure 1.9 shows (the epistatic filtration of) the triangulation of $*0***$ induced by ttd (upper left), the parallel epistatic weights in $*1***$ (upper right), the triangulation of $*1***$ (lower right) and the parallel epistatic weights in $*0***$.

Computing circuit interactions and interaction coordinates, as describe in Section 1.6.5, on *10** yields the significant results reported in Table 1.5. As above, there is a unique circuit interaction with support given by the points $\{2, 17, 18, 24, 28\}$. Its projection to the 3-cube is given by

$$n_h = h(011) + h(101) + h(110) - h(000) - 2h(111) .$$

Evaluating n_h at $\text{ttd}^{(3)}$ agrees with the determinant of Equation (1.19), up to sign.

Table 1.5: Parallel circuit interactions and interaction coordinates for $\text{ttd}^{(3)}$ on the 3-face of *10** which reached significance for $p < 0.05$.

Interaction	$ Z $	σ_Z^2	p -value folded	Vertices
b	4.87	2.23	0.029	12 18 24 28
e	4.90	2.50	0.050	2 11 12 24
f	10.49	2.68	0.000	6 17 18 28
i	9.77	2.66	0.000	2 11 17 28
j	5.62	2.52	0.026	6 12 17 24
k	8.17	2.60	0.002	2 12 17 28
l	7.22	2.58	0.005	6 11 18 24
n	13.04	3.41	0.000	2 17 18 24 28
o	8.89	3.50	0.011	6 11 12 17 28
q	12.09	3.66	0.001	6 11 12 18 28
u_{011}	15.39	3.66	0.000	2 6 11 12 17 18 24 28

1.6.7 The case of parallel squares inside $[0, 1]^5$

Fixing two bacterial species, α and β , defines a 2-dimensional face of $[0, 1]^5$, i.e., a square. This is the set of four points in $\{0, 1\}^5$ where the α - and β -coordinates vary, and all others are set to zero. Now a third bacterial species, γ , defines a *parallel square* in $[0, 1]^5$, where the α - and β -coordinates vary, and the γ -coordinate is set to one. Altogether, α , β and γ define a 3-dimensional face of $[0, 1]^5$. We want to investigate the impact of the presence of γ on the epistatic interaction between α and β . This amounts to comparing the epistatic weights of the two subdivisions induced on the parallel pair of squares.

Empirically, we know that the *Lactobacillus* bacteria compete with one another and *Acetobacters* also compete with one another, while *Acetobacters* form mutualist relationships with the *Lactobacilli*. We therefore focus on the combinations $\alpha \in \{1, 2\}$, $\beta \in \{3, 4, 5\}$ and $\gamma \in \{1, 2, 3, 4, 5\} - \{\alpha, \beta\}$. This means that, for each height function, we obtain $2 \cdot 3 \cdot (5 - 2) = 18$ faces which are spanned by α , β and γ . Each face is a row in each of the three Tables 1.6, 1.8 and 1.7 in the appendix; these tables show the result of our analysis for the three height functions ttd , dev and fec , respectively. The results can also be compared with [61, Figure S.13, p20, SM].

Each table shows the normalized volumes of the 3-dimensional simplices obtained from the four vertices of each square lifted by the respective height function; here ω_0 is the normalized volume of the simplex arising from the square with γ -coordinate zero, and ω_1 corresponds to γ -coordinate equal to one. The epistatic weights of the single dual edges of the two induced subdivisions on the two parallel squares are the absolute values $|\omega_0|$ and $|\omega_1|$. Yet, in order to track the effect of adding γ to α and β , here we take the orientation into account. The sign entry is positive if qualitatively the epistasis is the same, and it is negative if the effect gets reversed. The final column lists the relative increase or decrease (multiplicatively); i.e., a value of 1 would mean that the effect stays the same while larger values mean that the effect gets stronger.

Here are a few observations that we find particularly noteworthy.

1. There are some cases where adding γ seems to almost annihilate the epistasis, meaning that bystander species can disrupt an interacting pair.
2. A distinguished combination appears to be $(\alpha, \beta) = (2, 4)$: for which the time to death and the development time results match, suggesting the interaction affects these two traits in the same manner.

3. Another interesting case is $(\alpha, \beta) = (1, 5)$: adding γ weakens the epistasis for both time to death and development time.

So far, we have not determined biological explanations for these facts, however, the analysis defines the direction of our ongoing experimental investigations. Our results underscore the importance of context in determining microbiome interactions, which is to say that interacting groups of species care a great deal about their neighbors.

1.7 Discussion and outlook

In this paper, we develop a new approach to studying properties of fitness landscapes via regular subdivisions of convex polytopes, building on and extending previous work of Beerenwinkel et al. (2007). Our approach offers a concise combinatorial way of processing and clustering epistatic information in higher dimensions and is based on statistical principles. In Theorem 1.3.3 we present first provable robustness considerations. The main new tools we propose are cluster partitions and cluster filtrations in weighted graphs associated to fitness landscapes. To show that our methods are capable of quantifying epistasis with a higher resolution than previously done, we investigated an existing *Drosophila* microbiome data set. Among other results, we provide new forms of epistasis together with their biological interpretations.

To biologically validate our first promising findings, more biological properties of coexisting bacterial species, completing for instance Table 1.9, have to be determined. Moreover, further research has to be undertaken to connect the approaches discussed in this paper to a variety of other methodologies used within this theory, summarized for instance in [41]. Finally, although our methods are fully scalable, it is a matter of time until potential computational bottlenecks can fully be addressed.

Acknowledgement

The data was collected by Alison Gould and Vivian Zhang in Will Ludington’s Lab at UC Berkeley and is now published in [61]. We are indebted to Christian Haase and Günter Rote for a fruitful discussion concerning epistatic weights, leading to the definition (1.6).

1.8 Additional computations, figures and information for Chapter 1

Table 1.6: Parallel squares in $[0, 1]^5$ for ttd.

$\alpha = 1, \beta = 3$	ω_0	ω_1	sign	quot
$\gamma=2$	3.560	0.485	+	0.136
$\gamma=4$	3.560	-1.590	-	0.447
$\gamma=5$	3.560	-4.500	-	1.264
$\alpha = 1, \beta = 4$	ω_0	ω_1	sign	quot
$\gamma=2$	6.090	-0.725	-	0.119
$\gamma=3$	6.090	0.940	+	0.154
$\gamma=5$	6.090	-3.140	-	0.516
$\alpha = 1, \beta = 5$	ω_0	ω_1	sign	quot
$\gamma=2$	6.920	-2.325	-	0.336
$\gamma=3$	6.920	-1.140	-	0.165
$\gamma=4$	6.920	-2.310	-	0.334
$\alpha = 2, \beta = 3$	ω_0	ω_1	sign	quot
$\gamma=1$	0.715	3.790	+	5.301
$\gamma=4$	0.715	1.620	+	2.266
$\gamma=5$	0.715	8.090	+	11.315
$\alpha = 2, \beta = 4$	ω_0	ω_1	sign	quot
$\gamma=1$	1.765	8.580	+	4.861
$\gamma=3$	1.765	2.670	+	1.513
$\gamma=5$	1.765	2.190	+	1.241
$\alpha = 2, \beta = 5$	ω_0	ω_1	sign	quot
$\gamma=1$	4.705	-4.540	-	0.965
$\gamma=3$	4.705	-2.670	-	0.567
$\gamma=4$	4.705	4.280	+	0.910

Table 1.7: Parallel squares in $[0, 1]^5$ for dev.

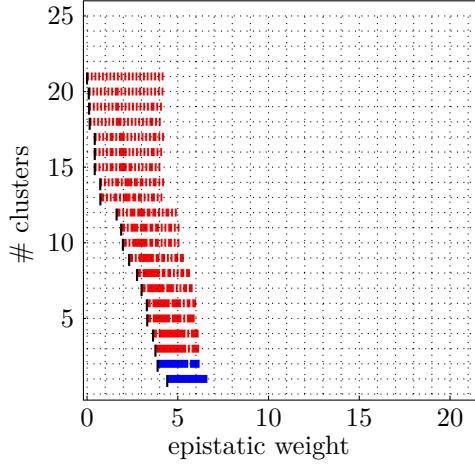
$\alpha = 1, \beta = 3$	ω_0	ω_1	sign	quot
$\gamma=2$	0.712	0.083	+	0.117
$\gamma=4$	0.712	-0.208	-	0.293
$\gamma=5$	0.712	-0.208	-	0.293
$\alpha = 1, \beta = 4$	ω_0	ω_1	sign	quot
$\gamma=2$	0.795	0.167	+	0.210
$\gamma=3$	0.795	-0.125	-	0.157
$\gamma=5$	0.795	-0.083	-	0.105
$\alpha = 1, \beta = 5$	ω_0	ω_1	sign	quot
$\gamma=2$	0.754	-0.208	-	0.276
$\gamma=3$	0.754	-0.167	-	0.221
$\gamma=4$	0.754	-0.125	-	0.166
$\alpha = 2, \beta = 3$	ω_0	ω_1	sign	quot
$\gamma=1$	0.337	-0.292	-	0.865
$\gamma=4$	0.337	0.042	+	0.124
$\gamma=5$	0.337	0.125	+	0.371
$\alpha = 2, \beta = 4$	ω_0	ω_1	sign	quot
$\gamma=1$	0.163	0.792	+	4.860
$\gamma=3$	0.163	0.458	+	2.814
$\gamma=5$	0.163	0.417	+	2.558
$\alpha = 2, \beta = 5$	ω_0	ω_1	sign	quot
$\gamma=1$	0.080	1.042	+	13.095
$\gamma=3$	0.080	0.292	+	3.667
$\gamma=4$	0.080	0.333	+	4.190

Table 1.8: Parallel squares in $[0, 1]^5$ for fec.

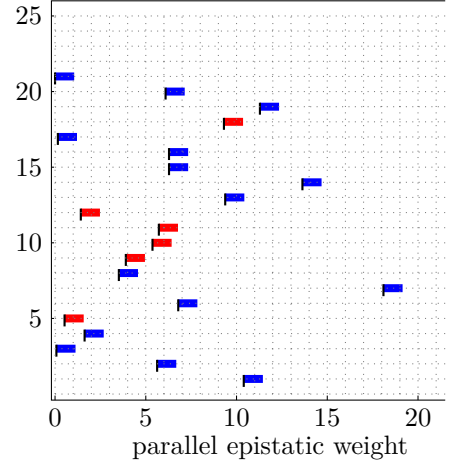
$\alpha = 1, \beta = 3$	ω_0	ω_1	sign	quot
$\gamma=2$	1.152	-0.577	-	0.501
$\gamma=4$	1.152	1.008	+	0.876
$\gamma=5$	1.152	-0.336	-	0.292
$\alpha = 1, \beta = 4$	ω_0	ω_1	sign	quot
$\gamma=2$	0.172	0.823	+	4.784
$\gamma=3$	0.172	0.029	+	0.167
$\gamma=5$	0.172	-0.099	-	0.578
$\alpha = 1, \beta = 5$	ω_0	ω_1	sign	quot
$\gamma=2$	0.508	-1.138	-	2.241
$\gamma=3$	0.508	-0.980	-	1.929
$\gamma=4$	0.508	0.236	+	0.465
$\alpha = 2, \beta = 3$	ω_0	ω_1	sign	quot
$\gamma=1$	0.547	-1.183	-	2.163
$\gamma=4$	0.547	1.701	+	3.111
$\gamma=5$	0.547	0.146	+	0.267
$\alpha = 2, \beta = 4$	ω_0	ω_1	sign	quot
$\gamma=1$	0.577	-0.074	-	0.128
$\gamma=3$	0.577	-0.577	-	1.000
$\gamma=5$	0.577	0.065	+	0.113
$\alpha = 2, \beta = 5$	ω_0	ω_1	sign	quot
$\gamma=1$	0.739	-0.908	-	1.229
$\gamma=3$	0.739	0.338	+	0.458
$\gamma=4$	0.739	1.250	+	1.693

Table 1.9: Numbering of the genotypes.

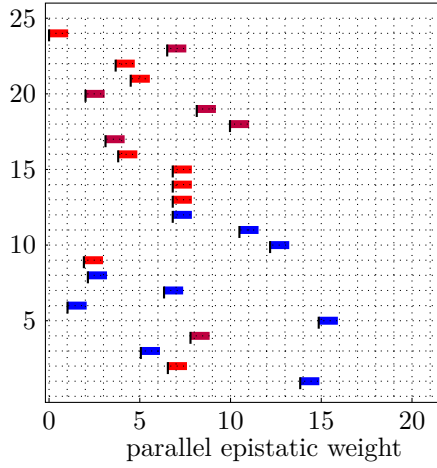
number	genotype	comment
0	0 0 0 0 0	germ-free
1	1 0 0 0 0	<i>Lactobacillus plantarum</i>
2	0 1 0 0 0	<i>Lactobacillus brevis</i>
3	0 0 1 0 0	<i>Acetobacter pasteurianus</i>
4	0 0 0 1 0	<i>Acetobacter tropicalis</i>
5	0 0 0 0 1	<i>Acetobacter orientalis</i>
6	1 1 0 0 0	bacterial growth decreases: competition
7	1 0 1 0 0	bacterial growth increases: synergy
8	1 0 0 1 0	synergy
9	1 0 0 0 1	synergy
10	0 1 1 0 0	synergy
11	0 1 0 1 0	synergy
12	0 1 0 0 1	synergy
13	0 0 1 1 0	competition
14	0 0 1 0 1	competition
15	0 0 0 1 1	competition
16	1 1 1 0 0	
17	1 1 0 1 0	
18	1 1 0 0 1	
19	1 0 1 1 0	
20	1 0 1 0 1	
21	1 0 0 1 1	
22	0 1 1 1 0	
23	0 1 1 0 1	
24	0 1 0 1 1	
25	0 0 1 1 1	
26	1 1 1 1 0	
27	1 1 1 0 1	
28	1 1 0 1 1	
29	1 0 1 1 1	
30	0 1 1 1 1	
31	1 1 1 1 1	



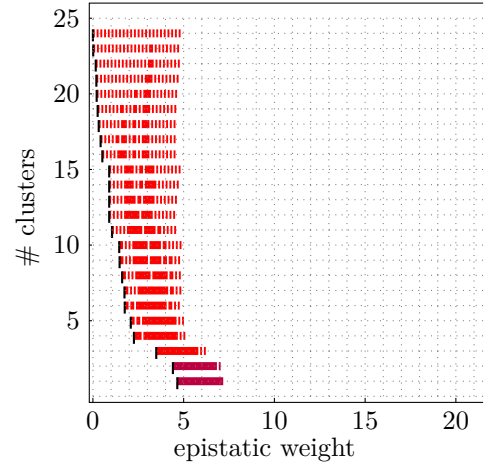
(a) Triangulation of the 4-cube $*0***$ induced by restricting ttd.



(b) Parallel epistatic weights in $*1***$ corresponding to ttd-triangulation of $*0***$.



(c) Parallel epistatic weights in $*0***$ corresponding to ttd-triangulation of $*1***$



(d) Triangulation of the 4-cube $*1***$ induced by restricting ttd.

Figure 1.9: The ttd-triangulations of $*0***$ and $*1***$ with their parallel epistatic weights; see Figures 1.7a and 1.7b. The f -vector for $*0***$ reads $(16, 63, 103, 76, 21)$ and $(21, 29, 9)$ for its tight span. The f -vector for $*1***$ reads $(16, 65, 110, 84, 24)$ and $(24, 36, 14, 1)$ for its tight span. The colors of the bars represent various levels of statistical significance ($p < 0.05$: blue, $0.05 \leq p < 0.1$: purple, $p \geq 0.1$: red); cf. Section 1.4.

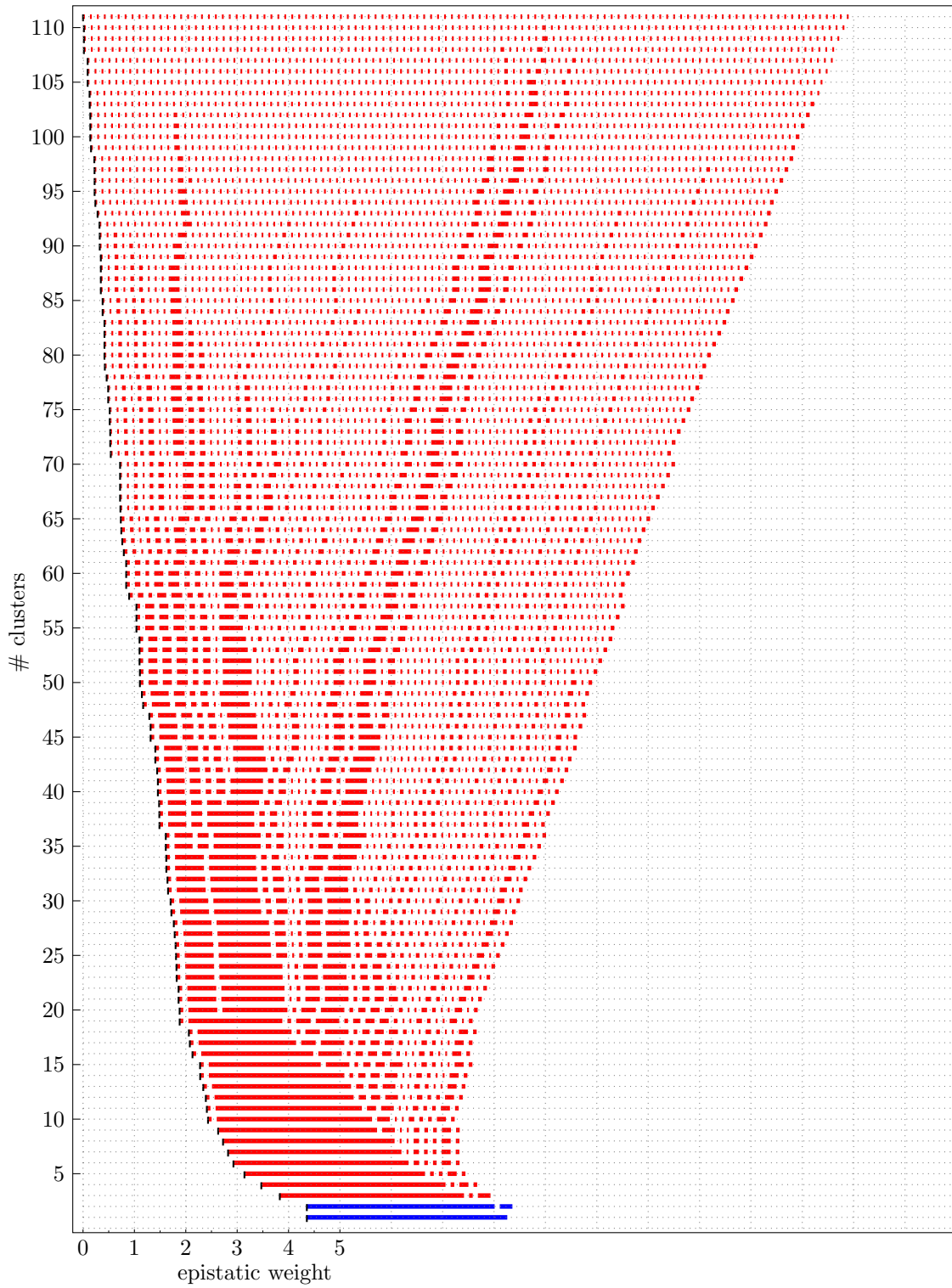


Figure 1.10: Epistatic filtration of the triangulation $\mathcal{S}(\{0, 1\}^5, \text{ttd})$ with 111 maximal cells. The colors of the bars represent various levels of statistical significance ($p < 0.05$: blue, $0.05 \leq p < 0.1$: purple, $p \geq 0.1$: red); cf. Section 1.4. The f -vector for $\mathcal{S}(\{0, 1\}^5, \text{ttd})$ reads (32, 204, 540, 702, 446, 111) and (111, 220, 137, 28, 1) for its tight span.

Master regulators of evolution and the microbiome in higher dimensions

This chapter is based on the preprint “Master regulators of evolution and the microbiome in higher dimensions” [46] by Holger Eble, Michael Joswig, Lisa Lamberti and William Ludington. The preprint can be found on the arXiv with the number 2009.12277.

2.0 Abstract

A longstanding goal of biology is to identify the key genes and species that critically impact evolution, ecology, and health. Network analysis has revealed keystone species that regulate ecosystems [118] and master regulators that regulate cellular genetic networks [134, 147, 32]. Yet these studies have focused on pairwise biological interactions, which can be affected by the context of genetic background [153, 98] and other species present [61, 28, 16, 62] generating higher-order interactions. The important regulators of higher-order interactions are unstudied. To address this, we applied a new high-dimensional geometry approach that quantifies epistasis in a fitness landscape [47] to ask how individual genes and species influence the interactions in the rest of the biological network. We then generated and also reanalyzed 5-dimensional datasets (two genetic, two microbiome). We identified key genes (e.g. the *rbs* locus and *pykF*) and species (e.g. *Lactobacilli*) that control the interactions of many other genes and species. These higher-order master regulators can induce or suppress evolutionary and ecological diversification [5] by controlling the topography of the fitness landscape. Thus, we provide mathematical intuition and justification for exploration of biological networks in higher dimensions.

2.1 Introduction

Master regulators are nodes in a network that control the rest of the network. They are often identified as highly connected nodes. For example, in eukarotic cells, the protein, target of rapamycin (TOR), interacts with many other proteins and pathways to control cellular metabolism [158]. Identifying TOR unified studies in many areas of cell biology, including regulation of transcription, translation, and the cytoskeleton around a central signaling pathway, with druggable targets for therapeutics of cancer, autoimmunity, metabolic disorders, and aging [158]. Ecological master regulators are called keystone species, a classical example being the starfish, *Pisaster*, which regulates the biodiversity of intertidal zone by eating many other species [118]. Identifying these key nodes in biological networks provides control points that can be used for instance in cancer therapy (through TOR) or ecological restoration (through starfish).

Epistasis is a framework to quantify biological networks, specifically gene networks, in terms of which genes (the nodes) interact and are thus connected by an edge. Constructing a gene network using epistasis works by iteratively mutating a set of individual genes and pairs of these genes, and then using the phenotypes of the mutants to construct the network. For instance, if genes A and B both affect a phenotype, C , we make the single mutants a and b and the double mutant ab . By measuring the effects on the output phenotype, e.g. fitness, it can be determined if A and B operate in parallel to affect C ($A \rightarrow C$ and $B \rightarrow C$) or in serial ($A \rightarrow B \rightarrow C$). These two possibilities are differentiated based on the degree of non-additivity: if the phenotypes of a and b add up to the phenotype of ab , the genes do not interact and thus operate in parallel. If they are non-additive, the genes interact and thus operate in serial. More specifically, if $A \rightarrow B \rightarrow C$, then mutants a , b , and ab will each produce the same phenotype, thus, $a + b \neq ab$, indicating

non-additivity or epistasis. The concept has been applied to map pairwise connections for protein structure [157], genetics [27, 153, 98, 32, 30], microbiomes [61], and ecology [28, 16, 62].

Epistatic interactions are important in nature [155], for instance when mutations occur [138, 90, 107] or when sex, recombination, and horizontal gene transfer bring groups of genes together [135, 154, 153, 35, 129, 108], making multiple loci interact. Applying epistasis to genome-wide measurement of pairwise genetic interactions has revealed biochemical pathways composed of discrete sets of genes [32, 30] as well as complex traits, such as human height, that are affected by almost every gene in the genome [103, 18]. New innovations have applied epistasis to broader data types [163, 98] and at different scales, making epistasis a widely valuable tool. For instance, epistasis between bacteria in the microbiome has functional consequences [125, 54, 61, 143, 122, 130] when community assembly combines groups of species in a fecal transplant. In this case, the nodes in the network are bacterial species. The master regulators of biological networks are identified by their position in the network, often as nodes with a higher degree of edges than average [117].

A known challenge of biological networks is that they are high-dimensional, meaning the interactions can change depending on the biological context or the genetic background [150], cf. [95] and references therein. This is important because such networks cannot be fully captured by pairwise interactions. Higher-order epistatic interactions are interactions that require three or more interacting parts, for instance genetic loci. From a network standpoint, loci that affect the interactions of many other loci play a key role in regulation of network structure.

Identifying such regulators requires a high-dimensional formulation of network structure. We recently developed such a formulation based on epistasis of fitness landscapes [47]. Fitness landscapes depict biological fitness as a function of genotype space [156, 138, 90]. Sewall Wright defined the genotype space as a hypercube with each genetic locus represented as an independent dimension [156]. Previous work formalized the fitness landscape of this genotype space and quantified epistasis on the fitness landscape [10, 36, 35, 47]. We developed the **epistatic filtration** technique, which segments the high-dimensional fitness landscape into local subregions and quantifies their epistasis in higher dimensions, allowing a researcher to hone in on important subregions of the landscape.

Here we develop that framework further in order to apply it to identify regulators of high-dimensional interactions. Rather than the traditional approach of assigning significance to a gene or species based on its pairwise interactions [118, 119, 134, 147, 32], we assign significance based on how the presence of that gene or species influences the structure and magnitude of interactions in the rest of the network. In order to compare interaction magnitudes across different dimensions, we develop a dimensionally-normalized definition of epistasis. We also develop a graphical approach to determine whether high-dimensional epistasis has lower-dimensional roots and what they are. We then analyze four data sets for 5-dimensional genotypes. Two are genetic datasets for (i) mutations that arose in *E. coli* evolution [91] and (ii) β -galactosidase antibiotic resistance [144]. Two are microbiome datasets measuring the impact of bacterial interactions on *Drosophila* lifespan, with one previously published [61] and another generated here. Our framework identifies regulators of higher-dimensional network structure in both the genetics and microbiome datasets. We find that specific genes and bacterial species suppress interactions in the rest of the network, meaning they regulate the higher-order network structure.

2.2 Results

2.2.1 Epistatic filtrations describe higher-dimensional biological networks

Our goal is to identify master regulators of biological interactions in higher dimensions. We use epistasis as a measure of interactions, and in higher dimensions, these occur on a fitness landscape. Our approach is to first measure epistasis on the high dimensional fitness landscape and then ask how individual loci, e.g. genes, change the shape of the landscape. We use the epistatic filtration technique to quantify epistasis on the fitness landscape. We use parallel epistatic filtrations to quantify the changes in the landscape due to each locus.

First, we describe epistatic filtrations. Epistatic filtrations are analogous to analyzing the drainage sectors within a watershed (see 1), which is a real physical landscape with altitude as a function of latitude and longitude. The topography sets where water will flow. Boundaries of a watershed are set by ridges, which enclose sectors within the watershed. These sectors feed

tributary creeks, which join with other tributaries to form larger sectors within the watershed. We can think of a fitness landscape as having sectors as well. In a fitness landscape, the topography is set not by altitude but by measurements of organismal fitness as a function of genotype. The longitude and latitude of a watershed correspond to genotypes in the fitness landscape. Because the biological entities are discrete (i.e., a gene is either wildtype or mutant), our framework is discrete too. We represent each gene with a separate dimension as proposed by Wright [156]. The space of all genotypes has many dimensions, one per mutated gene [156, 95]. This high-dimensional space is a *genotype hypercube* [156, 138, 90]. We next quantify the epistasis of the fitness landscape. This requires that we define sets of genotypes to compare. We do so by segmenting the genotype cube into sectors (see Box 1). This approach is different from previous approaches that defined sets of genotypes called circuits that traverse paths across the landscape [10]. An advantage of our approach is that there are orders of magnitude fewer sectors in a landscape than circuits (c.f. Table 2.10 versus Table 2.11), reducing the search space and the associated statistical constraints from multiple testing comparisons. These sectors are sets of adjacent genotypes in the hypercube. Geometrically speaking, these sectors are simplices, meaning each vertex (genotype) is directly connected to every other vertex in the set. For instance in $2D$, each vertex in a triangle is connected to the other two. To perform the segmentation, we use a triangulation. In Box 2, we illustrate how a two dimensional fitness landscape is triangulated using the phenotypes of the genotypes, which form a third dimension that we depict on the vertical axis. We use the topography provided by the phenotype data to uniquely determine the ridges of the landscape. Projecting these ridges back to the $2D$ genotype plane forms a triangulation of the genotypes into sectors (see Box 2). This diagram is similar to previous illustrations of epistasis on a two-dimensional landscape (c.f. [150, 95]), but our approach is unique in that we use the triangulation to sector the fitness landscape. Next, we construct a network representation of the sectored genotype space to depict the pairwise adjacency of neighboring simplices (nodes) [47]. An edge in this network indicates that two simplices are adjacent, meaning they share a face. Next, we locate the epistasis on this network topology. Our definition of epistasis is unique yet consistent with previous ones in lower dimensions (see Box 2). We assess the magnitude of epistasis of each pair of adjacent sectors in the triangulation by calculating the volume spanned by the fitness phenotypes corresponding to the genotypes of the vertices of the adjacent sectors. This definition makes the framework consistent when applying it to higher dimensions. We next rank the magnitudes of the adjacent sectors from smallest to largest. Plotting these merges gives an epistatic filtration (see Box 1 & 3).

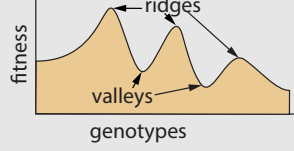
To determine how an individual locus, e.g. gene or species, affects the interactions in the rest of the network, we compare the epistasis for each pair of adjacent sectors with the locus of interest added or removed. This **parallel filtration** quantifies how adding or removing a locus affects the epistasis of the individual sectors of the high-dimensional network (see Box 4). Discovering loci that have outsized effects on their network allows a new approach to identify master regulators that operate in higher dimensions.

2.2.2 A volume-based definition of epistasis is valid across many dimensions

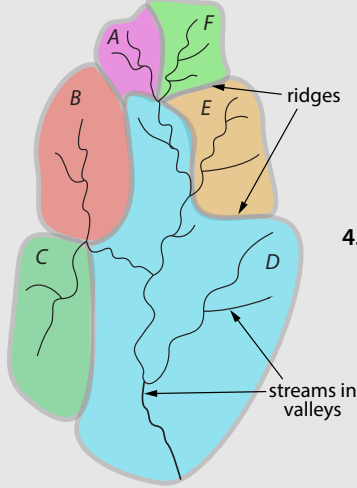
In this section, we explain the definition of epistasis that we employ throughout. We start by explaining the $2D$ genotype case. With two loci and two alleles (0 or 1) at each locus, we plot the genotypes as a unit square in the x-y plane and the measured phenotypes of each genotype on the z-axis (Box 2a). The phenotypes thus *lift* the genotypes into one higher dimension, here going from $2D$ to $3D$. Connecting the four phenotypes gives a simplex, shown as the green polytope in Box 2a. Depending on the relative magnitudes of the phenotypes, the green polytope can be larger or smaller, with the perfectly additive (no epistasis) case giving zero volume (Box 2a inset). We define epistasis as the euclidean volume of the green polytope, which in $2D$ is proportional to the absolute value of the established formula for epistasis, $\varepsilon = h(00) + h(11) - (h(10) + h(01))$ [10]. We call our definition the *epistatic volume* and note that it is of one dimension higher than the genotype space due to the measured phenotype (Box 2). This definition of epistasis based on volume is important because it applies equally well in higher dimensions (Box 2a,b; 2.6.1), as we discuss in the next section.

Box 1. Conceptual introduction to epistatic filtrations.

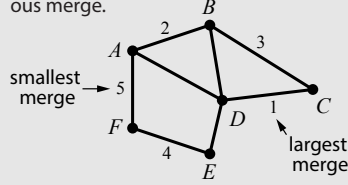
1. **Fitness landscape** plots fitness phenotype as a function of genotype



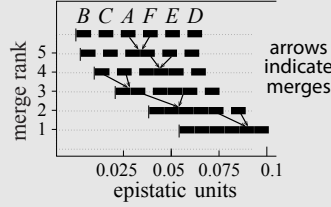
2. Ridges (grey lines) divide a landscape into drainages (colored shapes), which are smaller sectors drained by streams (black lines) that together form a larger **watershed**



3. Sectors are displayed as nodes of a **network** with connecting edges indicating adjacent sectors that share a ridge in the landscape. Edge numbers denote the rank order of the merged sector sizes. An unnumbered edge is superseded by a previous merge.



4. Fitness landscape is displayed as an **epistatic filtration** based on the magnitude of epistasis in each pair of adjacent sectors.



An epistatic filtration depicts the epistasis of a fitness landscape. By analogy with a watershed, producing the filtration can be conceptualized in four steps: (a) the fitness landscape defines topography; (b) the landscape is segmented into sectors based on the topography; (c) epistasis is calculated as the shared area of adjacent sectors and displayed on a graph that depicts the adjacency relationships of sectors; (d) the epistatic filtration depicts the rank order of epistasis magnitude in the adjacent sectors as a set of merges. Formal definitions follow in Box 2, Box 3, and text.

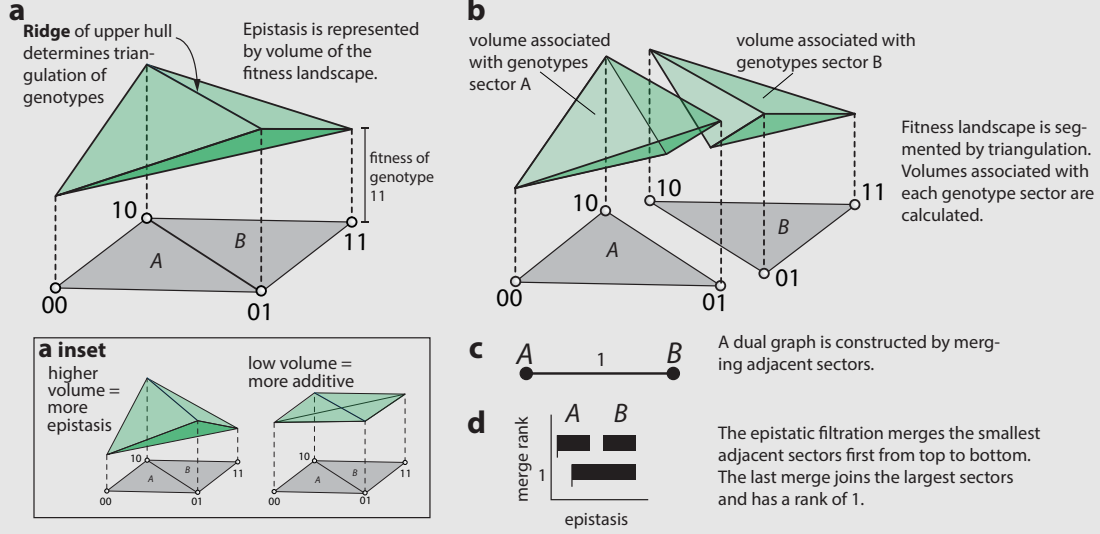
2.2.3 Epistatic filtrations: The n -loci case

In the n -loci case, the genotype set is given by $\{0, 1\}^n$, i.e. every genotype is encoded as a bitstring of length n , and the genotype-phenotype assignment h is a map $h: \{0, 1\}^n \rightarrow \mathbb{R}$, meaning each vertex v in the hypercube of genotype space has an associated phenotype $h(v)$. This is shown in Box 2 and Box 3 which visualize the two smallest cases $n = 2$ and $n = 3$, respectively. As in these lower dimensional cases, the lifted convex body $G^{(n+1)} \subset \mathbb{R}^{n+1}$ is given by the convex hull of the lifted points $(v, h(v))$ for genotypes $v \in \{0, 1\}^n$. The upper hull of $G^{(n+1)}$ consists of many facets and, as before, removing the phenotype coordinate, $h(v)$, from the vertices of the ridges (see Box 2a) yields the regular triangulation $\mathcal{S}(h)$ of the genotype space. Every sector s of $\mathcal{S}(h)$ is an n -dimensional simplex and, as such, it is spanned by $n + 1$ vertices $v^{(1)}, \dots, v^{(n+1)} \in \{0, 1\}^n$, cf. Box 3b). Given another simplex t of $\mathcal{S}(h)$, the pair (s, t) describes a bipyramid if the two are adjacent, which is true when t is spanned by vertices $v^{(2)}, \dots, v^{(n+2)} \in \{0, 1\}^n$. We use the notation

$$\{v^{(1)}\} + \{v^{(2)}, \dots, v^{(n+1)}\} + \{v^{(n+2)}\} \quad (2.21)$$

for the bipyramid (s, t) in order to emphasize its satellite vertices $v^{(1)}$ and $v^{(n+2)}$. As before, the lifted bipyramid $(s, t)^{(n+1)} \subset \mathbb{R}^{n+1}$ is the convex hull of the points $(v^{(i)}, h(v^{(i)}))$ for $1 \leq i \leq n+2$ and the epistatic weight $e_h(s, t)$ of the bipyramid (s, t) , defined in equation (1.6) of Appendix 2.6.1, can be seen as a variant of the euclidean volume of the lifted bipyramid $(s, t)^{(n+1)}$. Since that volume is non-negative, there are only two cases. Either $e_h(s, t) = 0$, which signals perfect additivity. Or we have $e_h(s, t) > 0$, which means that $G^{(n+1)}$ breaks at the ridge $\{v^{(2)}, \dots, v^{(n+1)}\}$. In that case the phenotype of the satellite $v^{(1)}$ lies below the expected value, assuming that h extends additively from the simplex t to the whole bipyramid $(s, t) = \{v^{(1)}\} + t$. A similar statement applies for the

Box 2. Definition of epistatic filtrations for a genotype space with two loci.



(a) The biallelic, 2D genotype set has two loci, each of which can be 0 or 1: $\{00, 01, 10, 11\}$. Each genotype gets lifted into 3D space by appending the phenotype $h(v)$ to each genotype coordinate in the set, $v \in \{(00), (01), (10), (11)\} \subset \mathbb{R}^2$. Connecting these lifted phenotype points forms a convex hull, depicted as the green 3D body $G^{(3)}$ above the grey genotype set. The upper surface of the green body is two green triangles, which are divided by the **ridge**. The euclidean volume of the 3D body $G^{(3)}$ yields a measure for epistasis (c.f. [91]). Inset: A higher degree of epistasis produces a larger volume, and lower epistasis produces a lower volume of the green body. (b) The ridge sets a triangulation of the genotype space in grey (a.k.a. genotype [10]). This is done by removing the phenotype dimension from the ridge vertices, which projects it back to the 2D genotype space. The ridge thus splits the space into sectors, which are two adjacent triangles, $\{00, 01, 10\}$ and $\{01, 10, 11\}$, denoted as A and B. We note that the euclidean volume of $G^{(3)}$ equals the absolute value of the established formula $\varepsilon = h(00) + h(11) - (h(10) + h(01))$ for epistasis in the two-dimensional case, scaled by a dimension related constant factor. (c) The dual graph connecting the adjacent triangles A and B is trivial in 2D as is the (d) epistatic filtration. Generalizing to higher dimensions, the triangles become simplices. These are explained further in Box 3 for the 3D case.

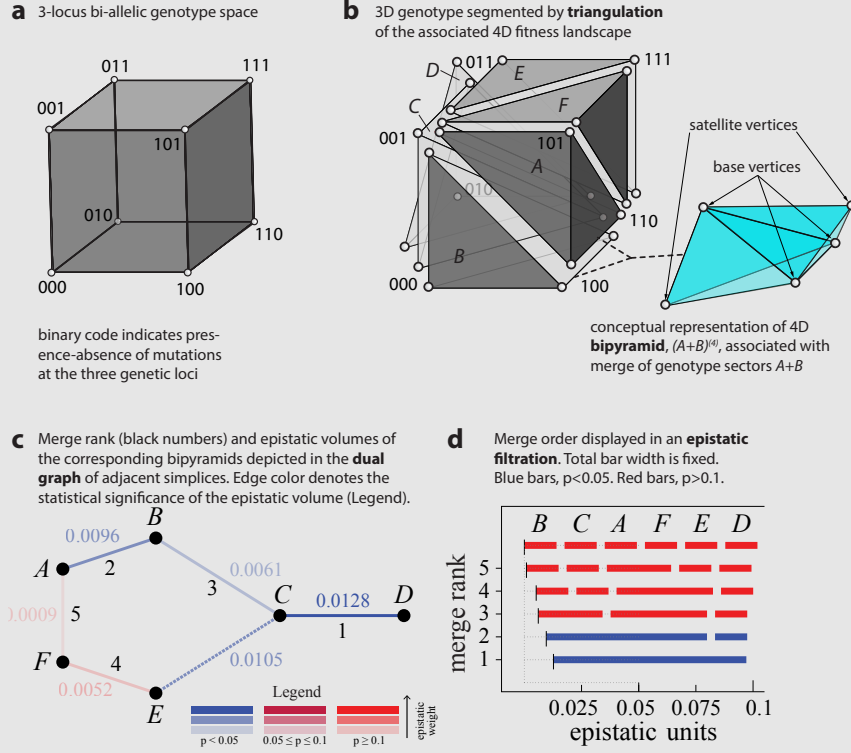
other satellite $v^{(n+2)}$. In this case, the $n + 2$ genotypes of the bipyramid (s, t) form an **epistatic interaction**, and the value $e_h(s, t)$ measures its strength.

Visualizing an n -dimensional polytope can be non-intuitive, but as for the 3-dimensional case, we can visualize the topography of the **epistatic landscape** by forming the **dual graph** of the triangulation $\mathcal{S}(h)$, where the nodes are n -dimensional simplices and the edges are bipyramids formed by adjacent simplices. We then calculate the volume of each bipyramid to determine the epistasis. We rank the bipyramids by their epistasis and depict the order with what we call an epistatic filtration.

As in lower dimensions, this visualization of a fitness landscape, ranked by epistasis, can be thought of intuitively like a watershed. Ridges enclose sectors that are iteratively merged with progressively larger sectors to form the entire landscape. Epistatic filtrations break apart a high-dimensional fitness landscape into sectors using a triangulation to define the ridges. In higher dimensions, the sectors are n -dimensional simplices. The dimensionality of the simplices is the dimensionality of the fitness landscape. Epistasis within these sectors is calculated using the full dimensionality. A statistical test determines significance of each epistatic interaction. The epistatic filtration of the fitness landscape depicts the path from smallest to largest epistasis by merging adjacent simplices to form connected clusters. Therefore, this is not a dimensional reduction but rather an approach that allows a global view of epistasis on a fitness landscape in higher dimensions. This process rests on the mathematical theory of linear optimization, convex polyhedra, and regular subdivisions [39, 47].

It is often useful to restrict the analysis to subsystems which are characterized by assuming

Box 3. Example epistatic filtration for three loci.



(a) The 3D genotype set forms a cube, and, as before, mapping the phenotypes onto the genotypes, $h(v)$, adds an extra dimension. The convex hull of the phenotypes, $h(v)$, forms a convex body $G^{(4)}$ in dimension 4, which yields ridges (see Box 2). (b) The ridges produce a **regular triangulation**, S , which consists of the six tetrahedra, A, B, C, D, E and F. Epistasis is calculated from the union of adjacent tetrahedra, which form a convex body in 4D, cartooned in blue. The blue is called a **bipyramid** because it is comprised of two neighboring tetrahedra that share a face. The vertices of the shared face are called **base vertices**. The unshared vertices of the two tetrahedra are called **satellites**. (c) The adjacency relations of the tetrahedra give rise to a network, which is the **dual graph** of S . In this graph, for instance, the edge (A, F) refers to the **bipyramid** comprised of A and F with vertices $\{010\} + \{011, 110, 001\} + \{111\}$ eqnnum . The set $\{011, 110, 001\}$ is the base where A and F meet, and it separates the two satellites 010 and 111. Analogous to the two-loci case, appending the $h(v)$ phenotypes to the genotypes in (2.20) yields a 4D simplex $(A, F)^{(4)}$. The volume of $(A, F)^{(4)}$ is the **epistatic weight** $e_h(A, F)$ (see Appendix 2.6.1. Color of edges indicates statistical significance (Legend; see Appendix for method; [47]). (d) The **epistatic filtration** of the genotype-phenotype map depicts the iterative process of glueing bipyramids in a non-redundant manner, going from lowest to highest epistatic weight. For example, rank 5 is the merge between A and F and has the lowest epistasis, rank 4 is the merge between E and F, and so forth. The black vertical tick mark at the left end of each row of blocks gives the epistasis added to the filtration at that rank. (e) The epistatic filtration is analogous to merging drainage sectors in a watershed.

the presence or absence of specific genes. These subsystems correspond to **faces** of the fitness cube $[0, 1]^n$, which are cubes of lower dimensions. We denote these faces as a string of zeros, ones and stars. For instance, 0**** in Fig. 2.11 is the 4-loci subsystem where the first gene is wildtype, and only mutations among the remaining four loci are studied. The analysis applies to such subsystems by restricting the genotype-phenotype map, which is important in our approach for identifying master regulators, as discussed later.

2.2.4 Epistatic filtrations reveal higher-order structure in *E. coli* evolution

To illustrate our approach, we examined an existing data set from Lenski's [7] classic experimental evolution of *Escherichia coli*, in a set of strains with each combination of five beneficial mutations [91] (Fig. 2.11a). We first examine $n = 3$ loci, corresponding to biallelic mutations in *topA*, *spoT*, and *pykF*. Epistasis was generally low in magnitude [91, 128], and occurs in two ways: (i) either from merging groups of groups of simplices (c.f. BC + AFE in line #2 of Box 3e, or (ii) from merging a single simplex, c.f. D, with the aggregated rest of the simplices (c.f. line #1 of Box 3e, much like a dominant effect in the NK model [90]. This second way is consistent with a fitness landscape distortion, which occurs when certain mutations influence the interactions of many other genes [78]. Geometrically, such a distortion constitutes a vertex split [69]. We next add a fourth biallelic mutation, in the *glmUS* locus (Fig. 2.11b,c), encoding peptidoglycan availability, which is an essential component of the cell wall.

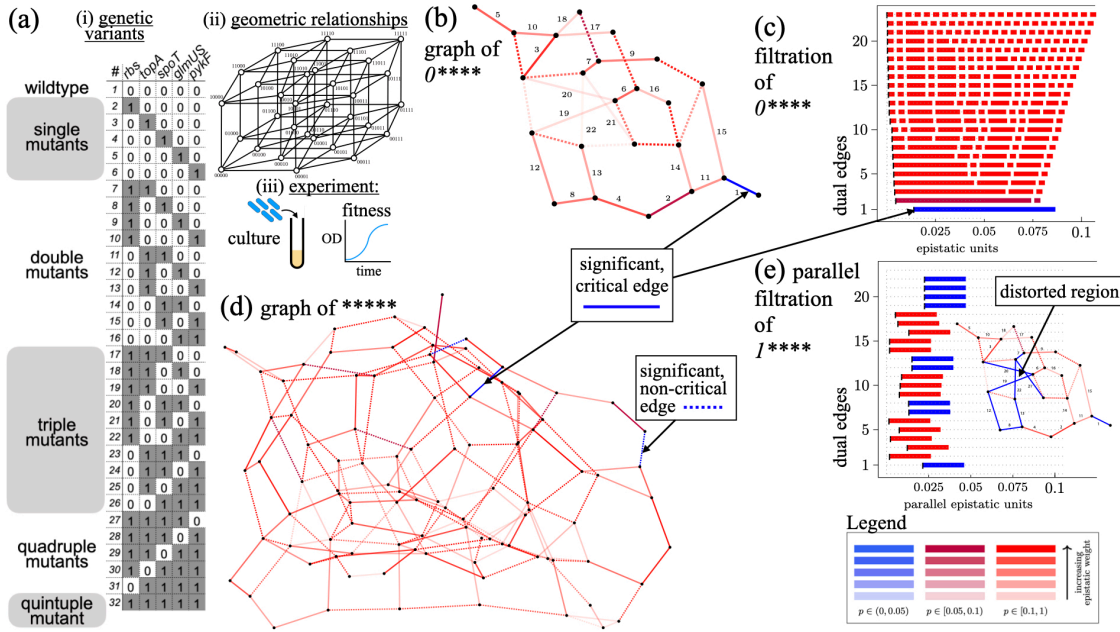


Figure 2.11: *E. coli* evolution is guided by epistatic landscape distortions. (a) (i) *E. coli* mutants examined [91], (ii) their geometric relationships, and (iii) experimental approach to measure fitness. (b) Edge labeled dual graph and (c) epistatic filtration restricted to $n = 4$ mutations in *topA* (locus 2), *spoT* (locus 3), *glmUS* (locus 4) and *pykF* (locus 5). Locus 1, *rbs*, is fixed 0 (*wildtype*). Note that the left edge of the bars in (c) indicates there is very little epistatic weight added to the filtration except for the final merge, where the single genotype 00001 gives weight to the entire filtration. This final interaction corresponds to the vertices $\{00001\} + \{00000, 01001, 00101, 00011\} + \{00010\}$. (d) Dual graph for the complete Khan data set. Black indices in (b) label the critical dual edges of $\mathcal{S}(h)$. (e) In the parallel filtration, for 1****, where the *rbs* mutation is present, the landscape is distorted by a concentrated area of higher epistasis. Inset: graph in (b) recolored with weights from (e). The lengths of the bars in the parallel transport figure (e) have no meaning. Only the horizontal position of the black marks, the vertical position of the bars and its coloring encode information. The horizontal shift represents the value of the epistatic weight, the vertical position of the bar indicates which dual edge is transported and the color expresses if the epistatic weight is significant after parallel transport.

The filtration reveals a smooth, additive landscape with one dominant cell where epistasis arises only in the final merge of the filtration (Fig. 2.11c), meaning the epistatic topography of the entire landscape (Fig. 2.11d) rests upon the single vertex, 00001, *pykF*. While the previous analysis detected a significant, marginal effect of *pykF* [91], filtrations reveal the geometric structure in terms of which specific combinations of loci are responsible for the effect (Fig. 2.11e): we establish an interaction between *glmUS*, {00001}, and *pykF*, {00010}. The interaction depends on the genotypes {00000, 01001, 00101, 00011} in the bipyramid base. Interestingly, the four loci context involves genotypes with the wild type and only up to double mutants. But these double mutants must be present together to yield a higher dimensional interaction. This conclusion is consistent with recent genome-wide work on trans-gene interactions [103], suggesting that complex traits may arise from genome-wide epistasis, where each mutation’s contribution to the trait depends on the presence of other mutations. Additionally, we observe that the interaction of {00001}, {00000, 01001, 00101, 00011}, {00010} in the 4D case (with the first locus wildtype) remains significant in the full 5-locus setting, ***** (see the blue critical edge in the dual graph of Fig. 2.11d), indicating an interaction in lower dimensions that is unaffected when a mutation is introduced in the first locus.

2.2.5 Parallel epistatic filtrations reveal master regulators in *E. coli* evolution

To discern the role of each locus on the 4D network structure, we applied **parallel filtrations** [47, §6.6]. This technique measures context-dependence in the fitness landscape by assessing changes in the epistasis of sectors that occur when a particular locus is mutated versus wildtype. For example, the epistatic filtration can be calculated for 0****, where the first locus is fixed as wildtype and the filtration is performed for the remaining 4 loci. This yields a set of bipyramids for which the epistasis is calculated. In the parallel filtration, we compare the epistasis for 0**** with the epistasis for 1**** using the triangulation set by 0**** as well as the rank order. In this way, two parallel faces of the 5-cube are compared (see Box 4 and Fig. 2.13). Parallel filtrations extend the concepts of conditional, marginal, and sign epistasis [59, 155] into the epistatic filtrations context.

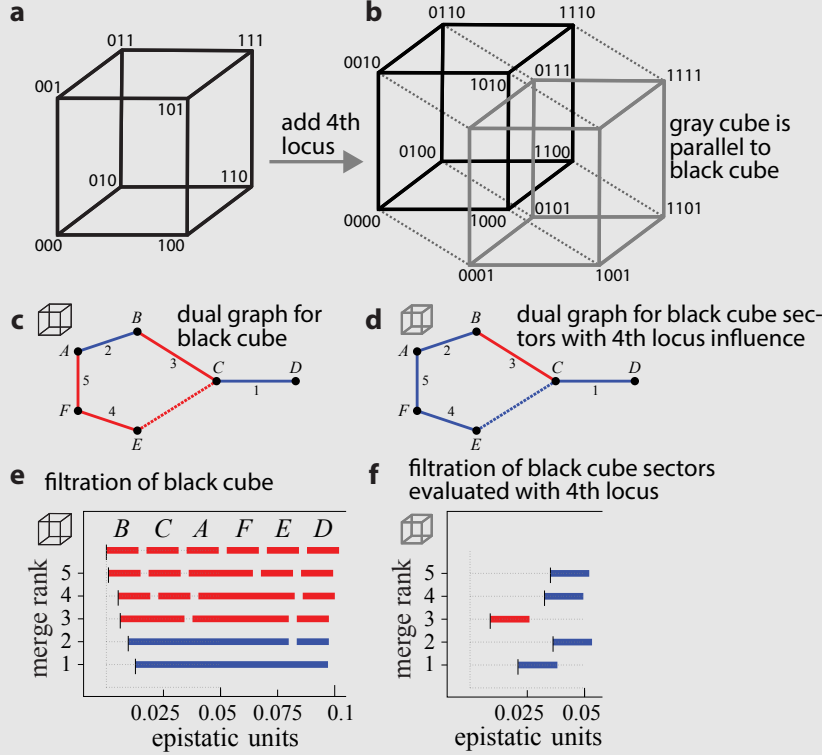
Examining the Khan data with and without the *pykF* mutation [91] (Fig. 2.15) showed increased significance in 8 out of 22 of the dual edges, when *pykF* was mutated. Each bipyramid in Fig. 2.15e) matches a bipyramid in Fig. 2.15c) via the parallel transport operation [47]. In particular, both filtrations have 22 dual edges.

The biological interpretation of the parallel transport operation is simple. It changes the context in which the epistatic weights associated to the dual edges are measured. For Fig. 2.11e) this means that epistatic weights in the genotype system with wildtype *rbs* are different when *rbs* is mutated. Since this locus is fixed in the parallel transport operation, comparing the wildtype and mutant, we call this locus the bystander. Here, changing the bystander state modifies the magnitude and significance status of the epistatic weights (Fig. 2.11c,e), with epistatic weights generally higher when *rbs* is mutated. Thus mutating the *rbs* locus distorts the fitness landscape. We note that the precise locations of the distortions are concentrated as a set of adjacent blue edges in the dual graph (Fig. 2.11e Inset). Examining the restoration of *pykF* to wildtype (Fig. 2.16), only 3 of 22 edges changed significance and just one critical edge lost significance, emphasizing the importance of context in the fitness landscape. Filtrations thus provide a new perspective on how genes regulate biological network structure in higher dimensions.

2.2.6 Lactobacilli produce microbiome distortions

Up to this point, we have focused on genetic epistasis, but our framework is equally valid for interactions of environmental parameters, including bacterial species in the gut microbiome. Like the genome, which is composed of many genes that interact to determine organismal fitness, the microbiome is also composed of many smaller units, i.e. bacterial species, that affect host fitness. Hosts are known to select and maintain a certain core set of microbes [101, 127]; the interactions of these bacteria can affect host fitness [61]; and it is debated to what extent these interactions are of higher order, cf. [54]. See also [95] for a broad overview on papers elaborating on possible meanings and instances of higher-order epistasis. While vertebrates have a gut taxonomic diversity of ≈ 1000 species, precluding study of all possible combinations, the laboratory fruit fly, *Drosophila melanogaster*, has naturally low diversity of ≈ 5 stably associated species [104].

Box 4. Parallel epistatic filtration for three loci when a 4th locus is modified.



(a) The 3D genotype space. (b) Adding a locus produces a 4D genotype space that can be visualized as two parallel 3D genotype spaces, depicted in black and grey, where the grey genotype space has a mutation in the 4th locus and the black is wildtype at the 4th locus. (c) The dual graph of S for the black genotype space. (d) The parallel dual graph for the grey genotype space. Note several edges in **c** (black cube) shift to significant in **d** (grey cube), indicating the context of the 4th locus influences the interactions. (e) The **epistatic filtration** of the black genotype space. (f) The **parallel filtration** calculates epistasis of the black genotype sectors with the phenotypes of the parallel cube (i.e. when the 4th locus is present). This approach measures the influence of the 4th locus on the rest of the epistatic interactions in the network. Specifically, note the shift in the x-values of the black vertical tick marks on the left sides of the left-most colored bars in **e** versus the corresponding tick mark and bar in **f**.

We made gnotobiotic flies inoculated with each combination of a set of $n = 5$ bacteria ($2^5 = 32$ combinations) that were isolated from a single wild-caught *D. melanogaster*, consisting of two members of the *Lactobacillus* genus (*L. plantarum* and *L. brevis*) and three members of the *Acetobacter* genus (Fig. 2.12a). We measured fly lifespan, which we previously identified as a reproducible phenotype that is changed by the microbiome [61]. Overall a reduction of microbial diversity (number of species) led to an increase in fly lifespan as with a taxonomically similar set of bacteria we examined previously, which came from multiple hosts [61].

The dual graph for the 5-loci genotype space revealed a single significant and critical epistatic interaction (Fig. 2.12b). Abundant non-critical edges were distributed throughout the graph (Fig. 2.12c) indicating prevalent interactions that weakly affect the fitness landscape. We note that such interactions were absent from the *E. coli* fitness landscape (compare the number of blue edges in 2.12b versus Fig. 2.11d). Using parallel filtrations to measure the role of individual bacterial species on the overall network, we found that the *Lactobacilli* drive changes in the global structure (Fig 2.12d,e). In 46 out of 128 (36%) interactions, significance changed due to adding or removing a *Lactobacillus* (Fig 2.12c-f, 2.20, 2.21). These changes in significance primarily derive from non-significant interactions when *L. brevis* is present that become significant when it is removed and vice versa, indicating *L. brevis* suppresses epistatic interactions that affect fly lifespan.

Microbiome abundances could drive the effects on host lifespan, however, comparing the epistatic

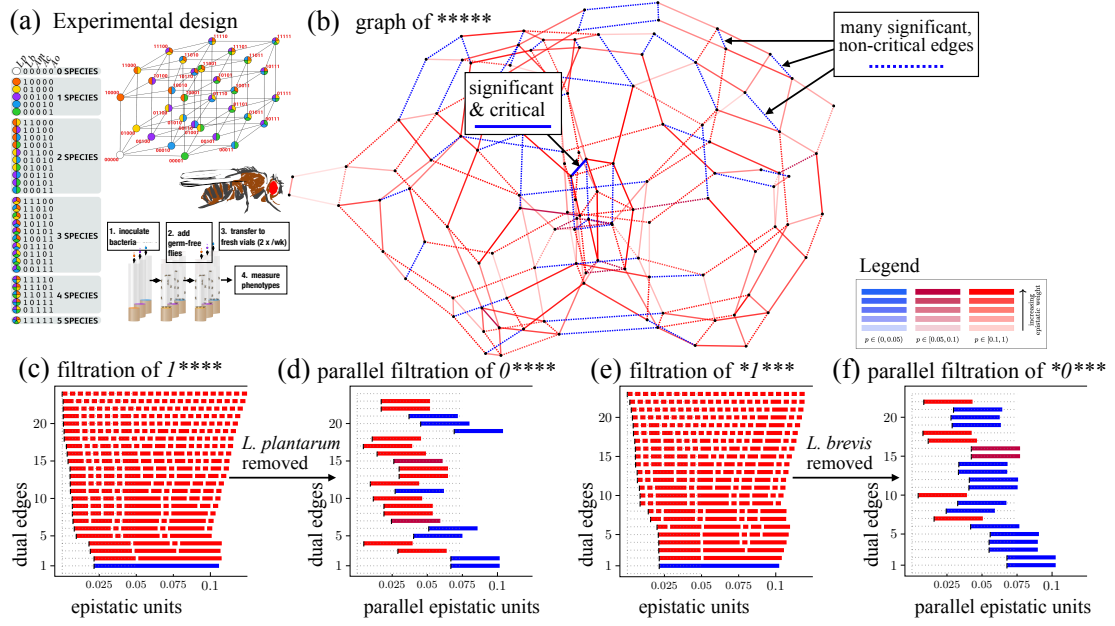


Figure 2.12: **Loss of lactobacilli causes global distortion of the microbiome epistatic landscape.** (a) Experimental design for Eble and Gould [61] microbiome manipulations in flies. (b) Full graph of ***** for the Eble data. (c) Filtration of $S(h)$ for the 4-face, 1****, of Eble data, where *L. plantarum* is present, indicates epistasis where two clusters of maximal cells merge. (d) Parallel filtration with *L. plantarum* removed shows a landscape distortion. (e) Filtration for *1***, where *L. brevis* is present has similar structure to 1****. (f) Parallel filtration with *L. brevis* removed shows a landscape distortion.

landscapes for CFUs and lifespan, we found that only 2 of 99 dual edges were significant for both the bacterial abundance and fly lifespan data sets (Fig. 2.22, 2.23, 2.24, 2.25, Tables 2.12, 2.13, 2.14, 2.15), and there was a lack of correlation between the epistatic weights of the bipyramids (Spearman rank correlations: $p = 0.7$, $p = 0.5$, $p = 0.3$, and $p = 0.3$ respectively). This discord between the epistatic landscapes for microbiome fitness and host fitness could e.g. diminish the rate of co-evolution.

2.2.7 The epistatic landscape within a single enzyme is rugged

As a point of comparison with the Khan data set, we re-analyzed data from a fully factorial 5-mutation data set in the β -lactamase gene, where each mutation is in a separate residue of the same enzyme [144, 151]. We note that the data are discrete (growth/no growth for a given set of antibiotic concentrations), and this type of microbiology experiment does not show variation in general. Thus, we can generally treat the calculated interaction magnitudes as accurate. We therefore discuss the meanings of the magnitudes. Due to a lack of the raw replicate data, our computations are based on the reported mean values, and p -values are not calculated.

The filtration holds a high magnitude of epistasis (Fig. 2.18, 2.19) compared with the Khan data set (Fig. 2.17, 2.15). Note that we can directly compare magnitudes (x -axis) due to the normalization procedure (see section 2.6.3). The epistasis arises in many steps (note slope of filtration adds magnitude in each step; (Fig. 2.18, 2.19)), consistent with the low number of possible evolutionary paths observed by Weinreich [151], and distortions are apparent in the shifted magnitude of epistasis by parallel transport (Fig. 2.18, 2.19). The filtration also reveals a tiered structure to the epistasis, cf. the largest weight merges two clusters of simplices (Fig. 2.18, 2.19) in contrast to the Khan data set, where epistasis came from one individual simplex on the periphery of the dual graph, indicating a more complex epistatic landscape in the β -lactamase.

Comparing the filtrations between the different datasets (Fig. 2.12d), the epistatic weight (i.e. magnitude) for the microbiome data generated $\approx 5\%$ effect, roughly three times the weight in the Khan data and half that in the Tan β -lactamase landscapes [144] (cf. x -axis between Fig. 2.12,

2.17, 2.18), indicating comparable interactions.

2.2.8 Interactions are sparse in higher dimensions

We used epistatic filtrations to systematically evaluate the prevalence of higher-order interactions as a function of the number of dimensions. Critical, significant, higher-order interactions were less frequent than pairwise interactions ($p < 10^{-6}$, Z -test) for each of the Khan, Eble, and Gould data sets, with a decreasing probability as a function of the face dimension (Table 2.10). This occurs for three primary reasons. First, the degrees of freedom increase fast in higher dimensions. Second, the probability of selecting a significant interaction from the set of all possible interactions decreases because the total number of interactions increases with increasing dimensions. Finally, the absolute number of significant interactions decreases in higher dimensions (Table 2.10), meaning they are biologically less prevalent. Overall, $\approx 10\%$ of possible dual edges were significant at higher order, with $\approx 1\%$ significant for $n = 5$ dimensions (Table 2.10), suggesting limits to the dimensions of biological complexity.

Table 2.10: Prevalence of interactions at different levels of complexity in genetics and microbiome data sets. Significant versus all critical dual edges ($p < 0.05$).

Interaction dimension	Dataset: Khan	Dataset: Eble	Dataset: Gould
2:	20/80 (25%)	24/80 (30%)	22/80 (28%)
all higher order:	29/508 (5.7%)	58/540 (10%)	21/520 (4.0%)
3:	21/194 (11%)	35/199 (17%)	14/194 (7.2%)
4:	7/214 (3.2%)	22/226 (10%)	6/216 (2.7%)
5:	1/100 (1.0%)	1/115 (0.8%)	1/110 (0.9%)
total:	49/588 (8.3%)	82/620 (13%)	43/600 (7.1%)

The epistatic filtration of the Eble microbiome data in (Fig. 2.12) has a much richer texture than the epistatic filtration of the Khan data set.

For instance, in the Eble microbiome data there are two top 4-dimensional epistatic weights which greatly impact the topography of the fitness landscape, in the following sense. The two epistatic weights are

$$\begin{aligned} \{01001\} + \{00000, 01000, 01101, 01111\} + \{01100\} & 0.0451 \quad \#2 \\ \{01001\} + \{00000, 01000, 01011, 01111\} + \{01110\} & 0.0485 \quad \#1 \end{aligned}$$

here given with their spanning genotypes, magnitude of the interaction, and edge ID number. The edge ID matches the position of the dual edge in the filtration of the left panel in Fig. 2.20 when counting from down up. The magnitudes of these two interactions combined have a 9% effect on fitness (sum of the magnitudes of the epistatic weights) with the largest accounting for $\simeq 5\%$, indicating a region of the landscape where epistasis is concentrated. Proximal to these genotypes are two additional cells with nearly significant epistatic weight:

$$\begin{aligned} \{01011\} + \{00000, 01001, 00111, 01111\} + \{01101\} & \#8 \\ \{01011\} + \{00000, 01000, 01001, 01111\} + \{01101\} & \#7 \end{aligned}$$

The corresponding dual edges are purple in the left panel in Fig. 2.20.

The genotypes in the interactions form a cluster relating the interactions between *L. brevis* and increasing numbers of *Acetobacters*. Because the interaction is detected based on the phenotype of fly lifespan, it suggests there may be interesting cellular and molecular mechanisms to investigate. For instance, the interactions could derive from metabolic crossfeeding between the *Acetobacters*, which produce many co-factors, and *L. brevis*, which produces lactate, stimulating *Acetobacter* growth [31, 68, 1]. Note that the support sets of all four interactions above contain both the wild type 00000 and 01111, which are the genotypes with maximum and minimum fitness respectively, indicating that all loci contribute to the higher-dimensional epistatic effect, even ones with low fitness.

2.2.9 Higher-order interactions can arise from lower-order interactions

Lower-order interactions can produce interactions in higher dimensions [128]. In examining the higher-order epistasis present in our data sets, we noted that the clusters where significant epistatic weights occur are often preceded by clusters with nearly significant epistatic weights in lower dimensions (Fig. 2.17). These lower dimensional interactions involve fewer genotypes than the higher-order interactions that they set up, meaning that the addition of genotypes pushes nearly significant interactions to significance.

We developed a graphical approach to distinguish these interactions from those that arise *de novo* (Fig. 2.27b,c; Appendix B11). More specifically, these graphics are intended to answer the question of to what extent higher-order epistatic effects are induced by lower dimensional ones or, put in other terms, which lower dimensional epistatic effects maintain significance when embedded into higher dimensions?

In (Fig. 2.27b) we exhibit an example for the Eble data set, with 5 loci, where we take the three 4-dimensional faces 0****, *0*** and **0** into consideration. For each such face, we computed the corresponding filtration of epistatic weights. We then repeat this procedure, and display the filtrations for relevant 3-dimensional subspaces ((Fig. 2.27b) second row), and finally filtrations for 2-dimensional subspaces (Fig. 2.27b) last row). The reasoning behind this is similar to what happens in regression-based epistasis calculations, where one can extract a certain portion of a higher dimensional space into lower dimensional spaces.

Performing the same operations on the Gould data, there are over all fewer significant epistatic weights. In this data set, we also observe examples of lower order interactions inducing higher order ones, as explained above, but for which the statistical significance status changes - here, from not significant (red bars) to significant (blue bars) (Fig. 2.27c). Linking the observed higher-order interactions to their lower-dimensional sources can help design biological experiments into the molecular mechanisms, for instance by designating two interacting bacteria to focus on from a larger community where the higher-order interactions emerge.

We also observe that several higher order interactions in the Eble, Gould and Khan data could not be attributed to lower-order effects (see (Fig. 2.27b,c) as well as Table 2.17). By this we mean that the interactions could not be linked to subsets with four, three, or two loci inside the 5-locus system, regardless of their significance (cf. Fig. 2.27c). Thus, some interactions arise only in the higher dimensional context and cannot be discovered or predicted by studying lower-order interactions.

As we noted, the 4-dimensional interaction in the *E. coli* evolution experiment involved loci with two genes (Fig. 2.11), whereas in the microbiome, interactions involved loci with four species, suggesting there may be different types of underlying geometries for the interactions between genes in evolution versus between species in the microbiome (Table 2.17).

2.3 Discussion and Conclusions

2.3.1 New biological findings

From an evolutionary perspective, the Red Queen hypothesis emphasizes how conflicts with other organisms can drive continuous genetic innovation [146]. In our analysis of the shapes of fitness landscapes, we find that epistasis in higher dimensions reshapes the fitness landscape. Thus, the continuous diversification observed in long term evolution experiments [60] could be generated by the continuously changing fitness landscape as new mutations occur. In particular, we identify master regulators that operate in higher dimensions by significantly enhancing or suppressing interactions in the rest of the biological network. In the microbiome these are lactobacilli, and in *E. coli* evolution we identified *rhs* and *pykF*. While it would require future experiments, it might be expected that such higher-order master regulators may also regulate the onset and progression of cancers.

The prevalence and importance of higher-order interactions is debated, with some studies suggesting pairwise interactions predict the vast majority of interactions in complex communities [54], and others suggesting a large influence of context-dependent effects [61] [142], which would make higher-order interactions unpredictable. Ample evidence that higher-order epistasis has at least some evolutionary impact was established in recent publications, see [95] and its references. Our

analyses suggest limitations on the existence of epistasis in higher dimensions. This could arise due to e.g. limited phenotypic dimensions where interactions can be detected or to a lower dimensional manifold that absorbs the majority of the effects [76] (e.g. lifespan and fecundity are anti-correlated, making fitness robust to changes in one or the other).

In Section 2.2.9, we analyzed how higher-order interactions in three data sets can arise from lower order ones. We found that in the majority of cases, the full biological information can only be obtained by analyzing epistatic weights in the full dimensional genotype space and that lower-order interactions are not sufficient to describe *all* interactions. In a few cases, however, the source of the higher dimensional interaction is rooted in a lower dimensional space and no additional biological information is obtained by increasing the dimension.

Our analysis also shows that significant epistatic interactions are increasingly sparse as the number of dimensions for interaction increase, indicating some limits to biological complexity.

2.3.2 Relation between epistatic filtrations and other measures of epistasis

From a methodological point of view, the present work lays the geometric groundwork for detecting epistasis via interactions of higher-order as well as other geometric properties of large fitness landscapes. Our work relies on polytope theory, following the shape approach of [10, 11], as this is the only framework allowing a mathematical definition of epistasis in a fine grained manner for a general n -locus system. By this we mean, that our interactions involve a minimal number of genotypes in the sense of a minimal set of dependent points [39]. The motivation for this is that these sets generalize the notion of adjacent triangles in a 2-locus system to an n -locus system. Additionally, in this way interactions have a geometric meaning, which makes them comparable across data sets. Although our method has similarities with [10, 11], it also has significant theoretical and computational differences and improvements. For example, our analyses heavily rely on studying the dual graph of the induced triangulation together with colored filtrations. This is a novelty in the theory and provides a number of new biological findings. For example, we localize regions of epistasis in four fitness landscapes, we quantify the sparsity of these regions, we compare portions of fitness landscapes via the parallel transport operation or by changing bystander species. We also further develop [47] by providing a new framework to detect and interpret how higher-order epistasis arises from lower order epistasis via meta-epistatic charts.

More specifically, epistatic weights capture new properties of fitness landscapes even in the 3-locus case. In this case, there are between four and six epistatic weights, as these are the number of adjacent pairs of simplices in the subdivision of the 3D cube, which appear as edges in the dual graph [73, Fig.1]. In contrast, there are 20 circuit interactions [10, Ex.3.9] and many more possible and potentially relevant interactions that must be checked in a randomized, exhaustive search. In addition to reducing the search space, epistatic weights can be localized in the fitness landscape, allowing the occurrence of mutations to be linked to changes in the topography of the epistatic landscape. Furthermore, we can link these changes across dimensions, tracking the source of the interactions.

Our method relates to other measures of epistasis, for example to linear regression approaches, as we explain in Section 2.6.9, see also the recent work [163]. It also relates to methods originating from harmonic analysis, cf. [148, 128, 152]; and to correlations between the effects of pairwise mutations, as we pointed out in [47]. More concretely, in a 2-locus, biallelic system, all these methods can easily be recovered from one another; some of them even agree. This is also true for some ecological approaches, including the generalized Lotka-Volterra equations, which yield a mathematically equivalent form to epistasis for certain situations cf. see equation 9 of [28]. In higher dimensional systems, these methods remain conceptually closely related but they generally yield different insights about the problem, such as which interactions are considered, whether the interactions are significant, what their magnitude is, and what their sign is. Because these previous methods make specific, *a priori* assumptions about the forms of interactions, they are limited by these assumptions. Epistatic filtrations add a global perspective, determining the structure of interactions from the shape of the fitness landscape in a parameter-free approach.

Finally, rank orders play an important role in the recent fitness landscape theory [36, 102]. For an overview and for references to relevant work in the theory, see the review article [95]. It is straightforward to recast the fitness landscapes presented here into a rank-order fitness graph and then count the number of peaks, i.e. the number of sinks in a fitness graph. The technical details

are beyond the scope of the present paper.

2.3.3 Interactions in higher dimensions

We found that biologically-significant epistatic interactions in four and five dimensions are sparse and often rooted in lower order, meaning that a limited number of regions of epistasis and hence of distortion exist in these fitness landscapes. This extends to higher dimensions the trend that 3-way interactions are often predicted from 2-way interactions [98, 54, 61]. However, our finding that key genes and species cause distortions emphasizes the need to identify the significant higher-order interactions from the vast number of possible ones, a task that epistatic filtrations enable.

In a five-loci case, we also found that the fitness landscape in the Eble data set is much more distorted, i.e. non-linear, than the Khan fitness landscape. We also found the precise locations of distortions inside the corresponding fitness landscapes and contextualize them in terms of distortions visible in lower dimensional sub-fitness landscapes. These findings are new and cannot be established with the old methods.

2.3.4 Strength and limitations of epistatic filtrations

A major advance of this work is that we provide a way to discover high dimensional regulators of biological networks. Rather than identifying key nodes as having a high number of low dimensional edges, we developed a method to identify nodes that regulate the higher-dimensional interactions in the rest of the network. This operation is performed by the parallel transport function, and we provide a web-based tool to perform the analysis (see Appendix 2.18). The implications of these findings are that certain genes and species modulate the interactions in the rest of the network, and perturbing these loci can destabilize the network. Destabilizing an unhealthy biological network could be crucial to restoring a degraded ecosystem, a sick microbiome, or curing a cancer, while destabilization of a healthy biological network could have the opposite consequences.

Methodologically, we also improve the framework in which higher-order epistasis can be mathematically formalized and analyzed geometrically. We provide concrete tools to find epistatic interactions in the fitness landscape and to distinguish if the landscape is locally flat, i.e. a hyper-plane of a certain dimension. Our work additionally allows us to localize and contextualize regions inside the fitness landscape which are not flat and hence distorted.

Our approach does not provide a distinction between positive and negative epistasis, but only between presence and absence of epistasis. However, this limitation is shared with other methods including the circuit, linear regression, and Fourier expansion approaches. To give an example, the circuit interactions in [10] can produce positive or negative values, but the sign depends on the choice of a basis for the interaction space, without a real biological motivation. The biallelic case provides an elementary case. In traditional terms, the epistasis in the Example from Box 2 is negative since the lifted genotype 11 lies *below* the plane spanned by the lifted genotypes 00, 10 and 01. Picking that particular plane for choosing the sign rests on the basis where the wild type is 00. If instead we use the genotype 10 as a basis, then the lifts of that genotype and its two neighbors 00 and 11, span a plane such that the lifted fourth genotype 10 lies *above* that plane of reference. However, while circuit interactions use signs to locate epistatic effects, in our approach this is not necessary, as the location information is concisely encoded in the regular triangulation induced by the phenotypes as described (c.f. Box 2). In this sense, the lack of sign is not a limitation of epistatic filtrations but a consequence of the high-dimensional approach.

A second limitation is a computational one which arises when one considers a multi-allelic system. In that setting our method still applies in theory, but the computational bottlenecks are reached rather quickly (at around $n = 10$ alleles without large hardware). However, it should be pointed out that the number of circuits of the cube $[0, 1]^n$ grows even faster with n ; cf. Table 2.11. So methods based on these also suffer from combinatorial explosion.

2.3.5 Outlook

This geometric approach could be extended, e.g. to GWAS [51, 103, 27], ecosystems [28, 16], or neuronal networks [126], to discover non-additive higher-order structures at different scales. It should be noted that the polyhedral geometry methods for analyzing epistasis deserve to be

developed further from the mathematical point of view. We believe that more concepts related to curvature for piecewise linear manifolds will be useful [141].

Taken together, our approach offers a number of new insights on higher-dimensional properties of fitness landscapes and their biological implications, and we think these will be useful as higher throughput experiments enable more combinatorial approaches.

2.4 Acknowledgements

The authors acknowledge L.J. Holt, O. Brandman, and J. Derrick for insightful comments on the manuscript. Research by M.J. is carried out in the framework of Matheon supported by Einstein Foundation Berlin. Further partial support by Deutsche Forschungsgemeinschaft (SFB-TRR 109: “Discretization in Geometry and Dynamics” and SFB-TRR 195: “Symbolic Tools in Mathematics and their Application”. W.B.L. acknowledges NIH grant DP5OD017851, NSF IOS award 2032985, and the Carnegie Institution for Science Endowment.

2.5 Materials and Methods

2.5.1 Fly husbandry

Flies were reared germ-free and inoculated with one combination of bacteria on day 5 after eclosion. $N \geq 100$ flies were assayed for lifespan in $n \geq 5$ independent vials per bacterial combination for a total of 3200 individual flies. Food was 10% autoclaved fresh yeast, 5% filter-sterilized glucose, 1.2% agar, and 0.42% propionic acid, pH 4.5. Complete methods are described in Gould *et al* [61].

2.5.2 Bacterial cultures

Bacteria were cultured on MRS or MYPL, washed in PBS, standardized to a density of 10^7 CFU/mL and 50 μ L was inoculated onto the fly food. Strains are indicated in Table 2.18. See Gould *et al* [61] for complete methods.

2.5.3 Genetics data

Existing genetics data sets were gotten from Sailer and Harms 2017 [128] github repository (<https://github.com/harmslab/epistasis>) or from Tan *et al* [144].

For the Khan data in Fig. 2.11, the fitness function h is defined for (b) by assigning the following normalized values to the 16 genotypes:

00000 \mapsto 0.1524	01000 \mapsto 0.1745	00100 \mapsto 0.1689	00010 \mapsto 0.1569
00001 \mapsto 0.1528	01100 \mapsto 0.1842	01010 \mapsto 0.1756	01001 \mapsto 0.1823
00110 \mapsto 0.1718	00101 \mapsto 0.1810	00011 \mapsto 0.1642	01110 \mapsto 0.1836
01101 \mapsto 0.1956	01011 \mapsto 0.1858	00111 \mapsto 0.1813	01111 \mapsto 0.1987

The Tan data set is different from the other fitness values in that only median and mean values are given, meaning we cannot compute p -values to assess the statistical significance. The fitness values are minimum inhibitory concentrations of antibiotics from a well-standardized assay with little experimental variation. Thus, the measurements and our analysis are believed to be robust. We note that the regular subdivision resulting from the corresponding height function of $[0, 1]^5$ is degenerate in the sense that it is not a triangulation. This degeneracy arises because the data are discrete antibiotic concentrations with 24 possible values. The repetition of exact values in several cases means a triangulation does not occur. We extended our methods to this degenerate case by restricting the analysis to the faces that do have a triangulation, broadening the application of our approach. We focused on the piperacillin with clavulanate data from [144] as it is the better behaved.

2.5.4 Computational analysis

The filtrations code is available as a `polymake` [56] package (c.f. <https://github.com/holgereble/EpistaticFiltration>). We also provide an online client, which processes raw csv data sheets, on https://www3.math.tu-berlin.de/combi/dmg/data/epistatic_filtrations/.

2.6 Terminology

Loci (singular **locus**) refer to individual sites in the genome where a mutation may occur, or in the microbiome sense, a locus is a particular bacterial species. We write $[n] := \{1, \dots, n\}$ for the set of all loci.

Genotypes, $v = (v_1, \dots, v_n)$, are vectors of loci with 0/1-coordinates that form points in some fixed Euclidean space \mathbb{R}^n , where n is the number of genetic loci or bacterial species considered. In this article we focus on **biallelic** n -locus systems, i.e. genotype sets of the form $V = \{0, 1\}^n$ where n is the number of loci and each locus is either 0, absent, or 1, present. For instance, $v = (1, 0, 1)$ denotes a genotype in a 3-locus system \mathbb{R}^3 , where the first and third loci are mutant and the second is wild type. The set of all genotypes will be denoted by V . The convex hull $P := \text{conv}(V)$ of all genotypes is called the **genotope**. In our setting P is the n -dimensional unit cube $[0, 1]^n$ (cf. (Fig. 2.14) for a 2D projection of $[0, 1]^5$).

A **fitness function** (also called **height function**) associates to each genotype $v \in V$ a quantified **phenotype** describing the impact of the genotype on the organism. For example, if the measured phenotype is fitness, h encodes the reproductive output of the genotype.

The **fitness landscape** is the pair (V, h) , which defines the fitness $h(v)$ for each genotype $v \in V$. Let $v = (v_1, \dots, v_n) \in V$ be a genotype. Then its **lift** is given by $(v, h(v)) = (v_1, \dots, v_n, h(v)) \in \mathbb{R}^{n+1}$.

A set of points $W = \{w^{(1)}, \dots, w^{(\ell)}\}$ is **affinely independent** if for every point $x \in \mathbb{R}^n$ which admits real scalars λ_i with $\sum_{i=1}^{\ell} \lambda_i = 1$ and $\sum_{i=1}^{\ell} \lambda_i w^{(i)} = x$ those scalars are uniquely determined. Otherwise W is **affinely dependent**.

An **interaction** with respect to a fitness function h occurs between a collection of $k+2$ affinely dependent genotypes $v^{(1)}, \dots, v^{(k+2)} \in V \subset \mathbb{R}^n$, for $k \leq n$, whose lifts are affinely independent points in \mathbb{R}^{n+1} . This is in line with the standard concept of additive epistasis. The number k is the **dimension** of the interaction; throughout we assume that $k \geq 2$.

Let $U = \{v^{(1)}, \dots, v^{(\ell)}\}$ be a set of genotypes. Its **support** is the set

$$\text{supp}(U) := \left\{ k \in [n] \mid \text{there are distinct } 1 \leq i, j \leq \ell \text{ with } v_k^{(i)} \neq v_k^{(j)} \right\}.$$

That is, the support is the set of loci where at least two of the given genotypes differ. For example, if $n = 3$ and $U = \{(0, 0, 0), (1, 0, 1), (1, 0, 0)\}$ then $\text{supp}(U) = \{1, 3\}$.

The number of loci that vary (0 vs 1) in the support is called the **order** of an interaction; this definition agrees with, cf., [152]: “We designate interactions among any subset of k mutations as k th-order epistasis.” We give two examples: First, let $n = 2$ and $U = \{(0, 0), (0, 1), (1, 0), (1, 1)\} = V$ such that U is an interaction with respect to some fitness function. Then U is an interaction of dimension 2 and order 2. Second, let $n = 3$ and $U = \{(0, 0, 0), (0, 1, 1), (1, 0, 0), (1, 1, 1)\}$ such that, again, U is an interaction with respect to some height function. Then the dimension is 2 and the order is 3. In general, the order is at least as large as the dimension, but the two quantities may differ. We say that genes (corresponding to loci) **interact** if they form the support set of an interaction of genotypes.

Remark 2.6.1. *The dimension k of an interaction $v^{(1)}, \dots, v^{(k+2)}$ with respect to some fitness function agrees with the dimension of the affine span of the given points in \mathbb{R}^n . This can be seen as follows. By definition the lifted points $(v^{(1)}, h(v^{(1)})), \dots, (v^{(k+2)}, h(v^{(k+2)}))$ are affinely independent in \mathbb{R}^{n+1} . So their affine span has dimension $k+1$. As $v^{(1)}, \dots, v^{(k+2)}$ are affinely dependent, the dimension of their affine span is at most k . Now the affine dimension can only increase by at most one if one coordinate is appended.*

2.6.1 A primer on epistatic filtrations

We first explain the biallelic case with $n \geq 2$ loci. In the geometric framework [10], two interacting loci give rise to four possible genotypes, which form the vertices of a square and may be written as vectors of zeros and ones, indicating the absence (0, wildtype) or the presence (1, mutant) of each locus respectively (Box 2a) [47, 10]. The measured phenotypes lift the genotype vertices into 3-space, and there is epistasis corresponding to the volume of the simplex enclosed by the lifted points (see green simplex in Box 2a). Geometrically, the four genotypes involved are fully symmetric, meaning that the sign of the epistasis for $n = 2$ is relative to the choice of a coordinate system.

Thus, the sign of epistasis depends on which genotype is considered wildtype. By considering the simplex volume rather than the fold of the upper shell of the simplex, epistatic filtrations do not specify a sign and thus avoid this caveat. However, directionality is considered by parallel transport (see later section). Returning to our explanation, by taking the upper convex hull of all 2^n lifted points and projecting back onto the genotype $[0, 1]^n$ we induce a **subdivision** $\mathcal{S}(h)$; cf. [47, 39, §2.1], into **maximal cells** (Box 2b). Generically, every maximal cell of $\mathcal{S}(h)$ is an n -dimensional simplex, which is the convex hull of $(n + 1)$ affinely independent genotypes. Importantly, these n -dimensional simplices are the most elementary parts into which a fitness landscape can naturally be decomposed.

Our framework generalizes to higher dimensions through a geometric shape called a **bipyramid**, where two satellite vertices, each the apex of one pyramid, are joined to a common set of base vertices. The satellites correspond in the $2D$ example (Box 2) to 00 and 11 and the base to 10 and 01. This is naturally associated with $\mathcal{S}(h)$, set up by the **ridge** (Box 2). For an ordered sequence of $n + 2$ genotypes $(v^{(1)}, v^{(2)}, \dots, v^{(n+2)})$ we let

$$s = \text{conv}\{v^{(1)}, \dots, v^{(n+1)}\} \quad \text{and} \quad t = \text{conv}\{v^{(2)}, \dots, v^{(n+2)}\} .$$

In other words, s and t form convex hulls. We call such a pair (s, t) a bipyramid with vertices $v^{(1)}, v^{(2)}, \dots, v^{(n+2)}$. Then we can find the volume of the lifted bipyramid by forming the $(n + 2) \times (n + 2)$ -matrix

$$E_h(s, t) := \begin{pmatrix} 1 & v_{1,1} & v_{1,2} & \dots & v_{1,n} & h(v^{(1)}) \\ 1 & v_{2,1} & v_{2,2} & \dots & v_{2,n} & h(v^{(2)}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & v_{n+2,1} & v_{n+2,2} & \dots & v_{n+2,n} & h(v^{(n+2)}) \end{pmatrix} , \quad (2.22)$$

where $v_{i,1}, v_{i,2}, \dots, v_{i,n}$ are the coordinates of $v^{(i)} \in \mathbb{R}^n$. The **epistatic weight** of the bipyramid (s, t) is

$$e_h(s, t) := |\det E_h(s, t)| \cdot \frac{\text{nvol}(s \cap t)}{\text{nvol}(s) \cdot \text{nvol}(t)} . \quad (2.23)$$

Here nvol denotes the dimensionally normalized volume. The quantity $\text{nvol}(s \cap t)$ is the relative $(n-1)$ -dimensional normalized volume of the **ridge** of the bipyramid, given by the intersection $s \cap t = \text{conv}(v^{(2)}, \dots, v^{(n+1)})$. We use the notation

$$\{v^{(1)}\} + \{v^{(2)}, \dots, v^{(n+1)}\} + \{v^{(n+2)}\} \quad (2.24)$$

for the bipyramid (s, t) , where the first and last vertices are the satellites and the middle set forms the base. Now the $n + 2$ genotypes of the bipyramid form an interaction of dimension n when $e_h(s, t) > 0$.

In our regular triangulation $\mathcal{S}(h)$, the two n -dimensional simplices, s and t , are **adjacent** because their intersection $s \cap t$ is a common face of dimension $n - 1$.

2.6.2 Constructing a filtration from the epistasis of adjacent simplices

We visualize the topography of the **epistatic landscape** by forming a **dual graph** of $\mathcal{S}(h)$, where the nodes are the maximal simplices and adjacent simplices form the dual edges. A rugged path is one with more blue edges (Box 3d). To each such dual edge we associate an epistatic weight and a label, epistatic weights are in shades of blue and red, while labels are in black). In this way, we construct an epistatic landscape that corresponds to the underlying fitness landscape with the ruggedness specified along the dual graph. The **epistatic filtration** of h (Box 3e) depicts the path from weakest to highest epistasis by merging adjacent simplices. These diagrams summarize the information contained in epistatic weights and dual graphs, and facilitate comparisons across data sets. But there is important new information contained in epistatic filtrations, which is not directly visible from the dual graph and its epistatic weights. Indeed, a step in the epistatic filtration merges adjacent simplices. We build the complete fitness landscape by stepwise merging of maximal cells, starting from the lowest epistatic weight and stepwise merging adjacent simplices to form a connected **cluster** cf. [47]. In this sense, epistatic filtrations encode a global notion of epistasis in higher dimensions by connecting adjacent bipyramids.

To see this, notice that each row of the diagram has a number of bars and a black leftmost line. In the top row the black line marks the epistatic weight of zero (x -coordinate). Each bar is red and corresponds to one maximal simplex of $\mathcal{S}(h)$. In the second row (counting from the top), we see three things: (1) the value of the lowest epistatic weight moves the x -coordinate of the black line slightly to the right. (2) The two maximal simplices of $\mathcal{S}(h)$ corresponding to this epistatic weight are merged into one. These correspond to the two bars in the previous row above the new, longer bar in the row. The lengths of the other bars remain unchanged but are shifted horizontally by the epistatic weight in (1). (3) The statistical significance of the epistatic weight giving rise to the merging step, encoded by the colors of the bars; cf. Section 2.6.4.

The merging procedure is then repeated for each pair of maximal simplices arising in each epistatic weight until one reaches the highest epistatic weight and the last maximal simplex of $\mathcal{S}(h)$ to be merged with the rest. In this way the indentation of the bar charts increases from top to bottom. The total width of the bars stays constant throughout.

Importantly, in the epistatic filtration diagram, not every merging step is displayed; e.g., in Box 3e there are fewer rows than dual edges in Box 3d. This is because some steps do not change the resulting fitness landscape (no actual new portion is merged to the previous one). The reported steps are only the ones increasing the connected components of the fitness landscape obtained from the previous merging steps. The epistatic weights corresponding to these steps are the edges in the dual graph which we call **critical** in [47, §3.2].

2.6.3 Normalized epistatic weights

To gain a perspective on the generality of higher-order interactions, it is desirable to compare epistatic landscapes. Different phenotypes have different metrics, making comparisons difficult for current approaches to epistasis. Filtrations are well-suited in this sense. Scaling the height function h by a positive constant does not change the regular triangulation, and thus it does not change the dual graph. In order to compare different data sets, we scale the height function to Euclidean norm one. The epistatic weights are scaled accordingly. The resulting **normalized epistatic weights** are measured in **epistatic units**, giving a generalized metric for epistasis.

Measuring the effect of context on epistatic interactions is also desirable, e.g. to detect the marginal or conditional effects of a locus [91], and these are a natural feature of filtrations. If we fix some k loci and let the remaining $n - k$ loci vary, we obtain a height function, which is **restricted** to a face of the genotype $[0, 1]^n$. That face has 2^{n-k} vertices, and it is an isomorphic copy of the cube $[0, 1]^{n-k}$. For instance, if $n = 5$ and we fix the first and the fourth locus to 0, we obtain a 3-dimensional face, which we denote $0**0*$. That is, such a face is written as a string of n symbols in the alphabet $\{0, 1, *\}$, where 0 or 1 mark the fixed choices, and $*$ stands for variation. The number of $*$ symbols equals the dimension of the face. Triangulations, their dual graphs, epistatic weights, etc. are well-defined for height functions restricted to faces. This aspect of the theory allows the study of conditional epistatic effects.

2.6.4 Statistics of epistatic weights

We developed a statistical test to quantify the significance of an interaction associated with a fixed bipyrmaid; cf. [47, §4.2]. Here we assume that $h(v)$ is the mean value of the individual phenotype measurements for some number of replicated experiments for the fixed genotype v . To each dual edge we associate a p -value, which is independent of the epistatic weight normalization. If that p -value is below 0.05 we call that dual edge **significant**. It is useful to also consider p -values, which are slightly higher because one can use the shape of the landscape to identify interesting locations for further statistical analysis. To this end we call a dual edge **semi-significant** if $0.05 \leq p < 0.1$.

While it may be possible that this approach misses some biologically relevant interactions (e.g. if they do not correspond to a bipyrmaid selected by our method), those interactions that we identify carry information that is robust and supported by a statistical model. The fact that not all possible interactions can be approached is an inevitable consequence of the higher dimensional nature of fitness landscapes, also reflected by a very high number of possible regular triangulations of $[0, 1]^n$. That number equals 74 for $n = 3$ and 87,959,448 for $n = 4$, whereas the precise numbers for $n \geq 5$ are unknown; cf. [39, §6.3]. Thus, filtrations use the data to greatly condense the number of possible interactions considered.

The bar colorings in the filtrations of epistatic weights, as in (Fig. 2.17), reflect the outcome of multiple simultaneous statistical tests (one for each epistatic weight) [47].

Significant dual edges at $p < 0.05$ are shown in blue, $0.05 \leq p < 0.1$ in purple, and $p \geq 0.1$ in red.

It may happen that a triangulation has a significant dual edge, which is not critical, whence it does not show in the epistatic filtration. In that case the next critical dual edge becomes blue; so a filtration encodes all significant interactions found by our method.

Remark 2.6.2. *By funneling the analysis through the concept of regular triangulations our approach pre-selects interactions, which are most relevant with respect to fitness [47, §2.2]. Via this major deviation from [10] we are able to detect interactions in many data sets, which are biologically plausible; this suggests strongly that our method is particularly good at avoiding false positives. Future work will investigate the relationship to other methods from statistics and signal processing. While most of this is beyond the scope of the present study, in Appendix B12 we offer a first step by comparing with traditional linear regression approaches.*

2.6.5 A synthetic experiment examining how epistatic weights change as a function of the interaction order

Our method calculates significance of detected interactions and normalizes the epistatic weight to the volume of the unit cube of the same dimensionality. We used synthetic data to analyze the method performance. We first examined 468 synthetic filtrations over the 4-dimensional cube, producing 10011 critical dual edges. We found that the epistatic weight is indeed constant as a function of the interaction order, see (Fig. 2.26a). This indicates that the normalization method is effective. Furthermore, the number of significant interactions decreased as the standard deviation of the input data increased, indicating the statistical method is sensitive to noise, see (Fig. 2.26b).

2.6.6 A microbiome example in dimension 4

Here $n = 4$, and the fitness function h is defined by assigning the following values to the 16 genotypes:

$$\begin{aligned} 0000 &\mapsto 0.2484 ; & 1000 &\mapsto 0.2320 ; & 0100 &\mapsto 0.1618 ; & 0010 &\mapsto 0.1698 ; \\ 0001 &\mapsto 0.1943 ; & 1100 &\mapsto 0.1749 ; & 1010 &\mapsto 0.1714 ; & 1001 &\mapsto 0.1929 ; \\ 0110 &\mapsto 0.1668 ; & 0101 &\mapsto 0.1608 ; & 0011 &\mapsto 0.1617 ; & 1110 &\mapsto 0.1643 ; \\ 1101 &\mapsto 0.1677 ; & 1011 &\mapsto 0.1715 ; & 0111 &\mapsto 0.1613 ; & 1111 &\mapsto 0.1594 . \end{aligned}$$

The vertices $U := \{v^{(1)}, \dots, v^{(6)}\} \in V$ given by

$$\begin{aligned} v^{(1)} &= (1, 1, 0, 0) ; & v^{(2)} &= (0, 0, 0, 0) ; & v^{(3)} &= (1, 0, 0, 0) ; \\ v^{(4)} &= (1, 1, 0, 1) ; & v^{(5)} &= (1, 1, 1, 1) ; & v^{(6)} &= (1, 0, 0, 1) \end{aligned}$$

form a bipyramid (s, t) consisting of 4-dimensional simplices s and t as above. The simplices s and t correspond to nodes in the dual graph of $\mathcal{S}(h)$ that share a dual edge recording their adjacency relation as indicated in (Fig. 2.12b).

In this situation, equation (2.23) reads

$$e_h(s, t) = \begin{vmatrix} 1 & 1 & 1 & 0 & 0 & 0.1749 \\ 1 & 0 & 0 & 0 & 0 & 0.2484 \\ 1 & 1 & 0 & 0 & 0 & 0.2320 \\ 1 & 1 & 1 & 0 & 1 & 0.1677 \\ 1 & 1 & 1 & 1 & 1 & 0.1594 \\ 1 & 1 & 0 & 0 & 1 & 0.1929 \end{vmatrix} \cdot \frac{\text{nvol}(s \cap t)}{\text{nvol}(s) \cdot \text{nvol}(t)} = 0.0318 \cdot \frac{\sqrt{2}}{1 \cdot 1} \approx 0.045 .$$

Since $e_h(s, t) > 0$, the genotype set U defines a 4-dimensional interaction with full support $\{1, 2, 3, 4\}$ and of order 4, according to our terminology of Section Terminology. With a p -value of $0.0005 < 0.05$ the significance test established in [47, §.4] rejects the zero hypothesis for $e_h(s, t)$ and therefore proves the effect of the interaction U to be significant. We indicate this fact with the color blue both in the dual graph of $\mathcal{S}(h)$ in (Fig. 2.12b) and in the epistatic filtration of h in (Fig. 2.12c).

This example illustrates the following fact of biological interest. For the bacterial combinations $v^{(1)}, v^{(2)}, \dots, v^{(6)}$ fitness, given by the fitness function h , varies significantly in a non-linear way.

2.6.7 Parallel transport of epistatic weights

The notion of parallel transport in a fitness landscape (V, h) was introduced in [47, §6.6] as a way to geometrically compare biological information between pairs of parallel facets of the convex polytope $\text{conv}V$. In this work, we extended that notion to include the case of two fitness landscapes, (V, h_1) and (V, h_2) , associated to different generic and normalized height functions $h_i : V \rightarrow \mathbb{R}, i \in \{1, 2\}$, defined on the same vertex set $V = \{0, 1\}^n$ for some $n \in \mathbb{N}$. To enable meaningful comparisons, we assume that each h_i is normalized and that there is a larger fitness landscape (W, h) with a generic and normalized height function $h : W \rightarrow \mathbb{R}$ restricting to h_1 and h_2 on the parallel facets V in W , such that the partition of $\text{conv}W$ induced by h is compatible with the one of $\text{conv}V$ induced by h_1 , resp. by h_2 . In this setting, we define **normalized epistatic weights** as with Eq. (2.23) with h the normalized height function and s, t any adjacent simplices forming a bipyramid.

Parallel transports enable us to transport epistatic filtrations along the reflection map

$$\phi: V \rightarrow V; v = (v_1, v_2, \dots, v_n) \mapsto (v'_1, v'_2, \dots, v'_n) ,$$

with $v'_i = 1 - v_k$ if $i = k$ and $v'_i = v_i$ otherwise. More precisely, let $e_{h_1}(s, t)$ be the normalized epistatic weight associated to a bipyramid of $\mathcal{S}(h_1)$ and let $\phi(e_{h_1}(s, t)) := e_{h_2}(\phi(s), \phi(t))$ be the parallel normalized epistatic weight transported by ϕ . Then the filtration of normalized epistatic weights induces a filtration of parallel normalized epistatic weights. Additionally, to $e_{h_1}(s, t)$ and to $\phi(e_{h_1}(s, t))$ a p -value can unambiguously be associated [47, §4.1-4.2]. Notice that by design epistatic filtrations for $\mathcal{S}(h_1)$ only show normalized epistatic weights associated to critical dual edges, defined as in [47]. But normalized epistatic weights and their significance can be defined for all bipyramids including the ones associated to noncritical dual edges. This explains the labelling of the parallel transport tables below. There a row is numbered only if the bipyramid corresponds to a critical dual edge in the dual graph of $\mathcal{S}(h_1)$. Noncritical dual edges whose normalized epistatic weight remains non-significant after the parallel transport are omitted. The normalized epistatic weight before (denoted by $e_o = e_{h_1}(s, t)$) and after (denoted by $e_p = \phi(e_{h_1}(s, t))$) the parallel transport, as well as their p -values (denoted by p_o and p_p) are also reported, as well as ratios of these quantities.

These parallel transport tables are linked to the epistatic filtration diagrams. Indeed, each numbered row in the table corresponds to the row in the epistatic filtration diagram with the black line set at e_o . It also corresponds to the row with black line set at e_p in the parallel transported filtration diagram.

Recall from Section *Statistics of epistatic weights* that there may be dual edges of the triangulations which are significant but not critical. Since only the critical dual edges are labeled (by the row number in the epistatic filtration), in our tables for parallel transport these show up as unlabelled rows.

Examples for the parallel transport of epistatic filtrations are shown in Figures 2.13, 2.15, 2.16, 2.18, and 2.19. The magnitude of the epistasis in the left panels are roughly comparable between data sets due to normalization of the input data. Compare each left panel with its corresponding right panel to observe the relative change in epistasis in the parallel path. Larger changes in epistasis indicate stronger context-dependence of the interaction. For instance, in the first Weinreich comparison (Fig. 2.18), bar 10 in the right panel has a parallel epistasis greater than the original filtration on the left, indicating context-dependence.

2.6.8 Meta-epistatic charts

The **meta-epistatic chart** is a diagram drawn on top of the induced epistatic filtrations for some selection of faces of a fixed cube; higher-order interactions induced by lower order interactions are marked as corresponding.

In (Fig. 2.27b) and (Fig. 2.27c) we exhibit an example for the Eble data set, with 5 loci, where we take the five 4-dimensional faces 0****, *0***, **0**, ***0* and ****0 into consideration. Mathematically, these five 4-faces constitute the face figure of the wild type. Fix one 4-face, say 0****. The induced epistatic filtration on this face shows two blue bars corresponding to dual edges labeled 1 and 2. Each of them refers to the ridge of a bipyramid, which is a 3-dimensional simplex in this case. These two ridges may intersect certain 3-dimensional faces in the right dimension and thus may or may not descend to significant ridges within certain 3-dimensional filtrations. In

case of an incidence with a lower dimensional significant ridge, the significant 4-dimensional effect is induced by a lower dimensional effect and one may picture this fact as a directed assignment pointing from the lower towards the higher dimensional interaction.

2.6.9 Comparison with a simple linear regression approach

In the theory of fitness landscapes many linear regression approaches have been proposed to study higher-order interactions, cf. [13, 149, 163, 128]. In this section, we compare our epistatic weight method to an elementary regression approach using an example from the data.

The regression analysis we have in mind assumes that there is a linear relationship between the predictors X_1, X_2, \dots, X_n (one associated to each locus/dimension of the genotype) and response, or dependent, variables Y (associated to the biological measurements). That is, one assumes that $Y = f(X_1, X_2, \dots, X_n) + \varepsilon$ where $f : \mathbb{R}^n \rightarrow \mathbb{R}$; $(X_1, X_2, \dots, X_n) \mapsto \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$ and where ε is a random error term. The coefficients $\beta_1, \beta_2, \dots, \beta_n$ are unknown but can be estimated by minimizing the sum of squared residuals associated to the observations pairs (x, y) . These observations pairs consisting of a genotype and a measurement associated to it. Notice that more than one measurements are typically associated to a single genotype. With the coefficient estimates one can make predictions for the dependent variable via

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_n x_n . \quad (2.25)$$

The hat symbol $\hat{\cdot}$ indicates a prediction, for instance of Y on the basis of $x_i = X_i$, or an estimate for an unknown coefficient.

Below, we are interested in the differences between the observed measurements y associated to the genotypes of $[0, 1]^n$, expressed in terms of x_1, x_2, \dots, x_n and the predicated values \hat{y} on the regression hyperplane (2.25). Notice that the regression analysis remains unchanged after normalizing the height function to Euclidean norm one. Additionally, computing residues for all replicated measurements (when provided) and then take averages builds on the assumption that measurements associated to different genotypes are statistically independent from each other. This assumption is consistent with the one underlying the computation of statistical significances for epistatic weights, following [47, §. 4.2-4.3].

Remark 2.6.3. *In the regression setting of (2.25) there are hypothesis tests (like the F -statistic, t -statistics and p -value) to answer if at least one regression coefficient $\beta_j, 1 \leq j \leq n$ is nonzero, see for example [79]. Such statistical approaches are different from the one in [47, §. 4.2-4.3], where other hypothesis tests for each epistatic weight were proposed.*

Regression for Eble data

In the following, we perform a regression analysis focusing on the replicated measurements for the lifespan fitness landscape on $[0, 1]^5$ obtained from Eble and subspaces thereof. Numerical measures of model fit (F -statistic: 2357, with p -value essentially zero, and for 3840 observations and 5 predictors) show that the multiple linear regression model can be considered to be appropriated for this data. Since the epistatic weights of the dual edges are close to zero (≤ 0.02) and are mostly not significant, the above regression analysis conclusion is in line with what we see from the filtration of epistatic weights associated to the same fitness landscapes, see (Fig. 2.28).

From this example we see that the regression approach provides some general information on higher-order interactions. However, without further assumptions, only one interaction formula is given in terms of a regression hyperplane (2.25) while the epistatic weight approach gives more fine grained information. This example also illustrate that when the regression model fits the data well (essentially the higher the F -statistics and the more coefficients in the hyperplane equation are significantly non-zero) the epistatic filtration has little horizontal shifts and few significant epistatic weights.

We now proceed repeating the above analysis on some of the bipyramids considered in the parallel analysis for the normalized lifespan Eble data. Regressing over bipyramid 23 in Table 2.16

$$\{0001\} + \{0000, 1001, 1011, 0111\} + \{1111\}$$

in 0**** and 1**** reveals that only two average residues over 0**** are non-zero (associated to the microbiomes 00000 and 00001), and only one is non-zero over 1**** (associated to the

microbiome 10000). This confirms the two non significant epistatic weights over bipyrmaid 23 in Table 2.16.

Remark 2.6.4. *If minimally dependent sets of points in the genotope are fixed, as in the epistatic weight approach, and one regresses above these points, then the corresponding regression hyper-planes equations are learned from data and the equations generally differ from the epistatic weights given as in (2.23), but similar biological and geometric conclusions can be drawn. This idea could then be taken further by considering smoothing splines, instead of linear regression, and their relation to epistatic filtrations. From an application point of view, one would obtain an interesting new extension of the concept of epistasis because intermediate genotypes could be assessed, which would correspond to the case of genetically heterogeneous populations of organisms as occur in nature.*

Other numerical results for the above regressions are summarized in Table 2.19. Over 0**** two coefficients are significantly non-zero (for x_1 and x_4), see top part of Table 2.19. Similarly, over 1**** four coefficients are significantly non-zero (x_1, x_2, x_3, x_4), see bottom part of Table 2.19. The fit of the linear regression models is confirmed by the relatively high values of the F -statistic. Over 0***v the F -statistics is 459.1 for a p -value near zero and 720 observations. Over 1**** the corresponding F -statistics (near zero) is 52.61.

2.6.10 Microbiome data sets

In this work, *Drosophila* microbiome fitness landscapes consist of experimental measurements on germ-free *Drosophila* flies inoculated with different bacterial species. The lifespan of approximately 100 individual flies were measured for each combination of bacterial species, giving roughly 3,200 individual fly lifespans for each of the two data sets presented. The experimental methods are described in [61, 99]. The first data set is the exact data presented in [61, 99]. The second data set is the second set of species with exactly the same methods used in [61, 99]. The bacterial compositions considered consist of all possible combinations of five species. The species considered can all occur naturally in the gut of wild flies: *Lactobacillus plantarum* (LP), *Lactobacillus brevis* (LB), *Acetobacter pasteurianus* (APa), *Acetobacter tropicalis* (AT), *Acetobacter orientalis* (AO), *Acetobacter cerevisiae* (AC), *Acetobacter malorum* (AM). The 5-member communities both stably persist in the fly gut. For the purposes of this work, we define **stable** as maintaining colonization of the gut when ≤ 20 flies are co-housed in a standard fly vial and transferred daily to fresh food containing 10% glucose, 5% live yeast that has subsequently been autoclaved, 1.2% agar, and 0.42% propionic acid, with a pH of 4.5. The total number of species found stably associated with an individual fly is typically between 3 and 8. Consistently, *Lactobacillus plantarum* and *Lactobacillus brevis*, are found with two to three *Acetobacter* species. Less consistently, species of *Enterobacteria* and *Enterococci* occur, and these have been described as pathogens. While more strains may be present, for each of the two data sets in the present work, a set of five non pathogen species was chosen, including the two *Lactobacilli* and three *Acetobacter* species. The combinations of species are shown in Table 2.18. Different strains of the same species were used in the two data sets.

2.7 Additional computations, figures and information for Chapter 2

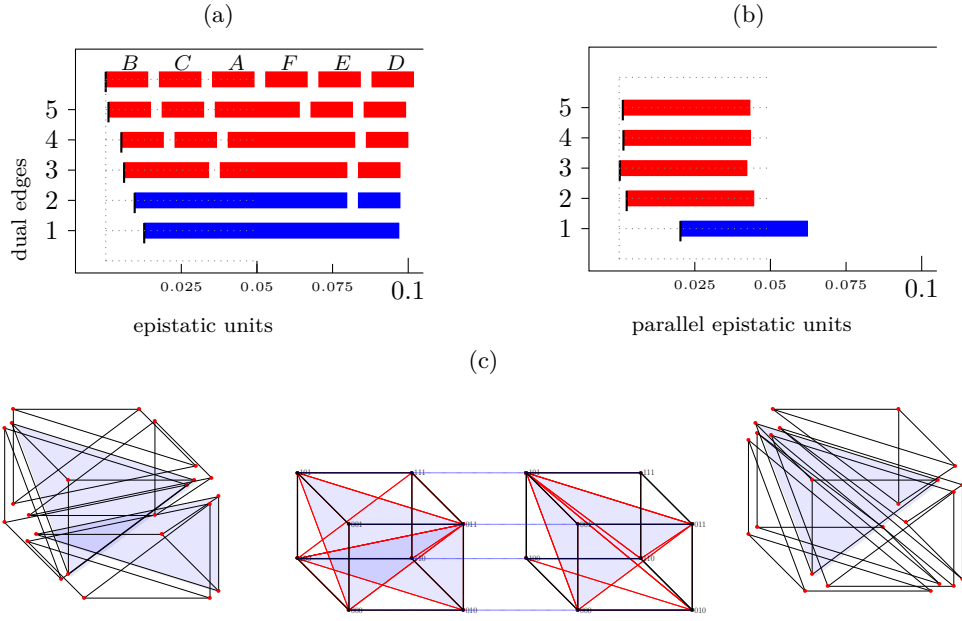


Figure 2.13: **Parallel transport from 0**0* to 1**0* within the Khan dataset.** (a) Filtration based on the triangulation of 0**0*. (b) Parallel epistatic weights computed from 1**0* for the triangulation based on 0**0*. (c) The two parallel triangulations (and exploded copies) are depicted. The partitions in the node set are transferred from the cube on the middle left to the cube on the middle right. Exploded versions of these same triangulation on the far left and far right demonstrate the geometry of the simplices generated by the triangulations.

Table 2.11: Number of circuits of $[0, 1]^n$ and bipyramids among these. This indicates that bipyramids can analyze the majority of all possible interactions, which circuits exhaustively cover. Compare with Table 2.10, which shows the actual number of bipyramids for three datasets, indicating significantly fewer sectors are needed to cover the landscape.

dimensions	circuits	bipyramids	percentage
2	1	1	100.00%
3	20	8	40.00%
4	1348	1088	80.71%
5	353616	309056	87.40%

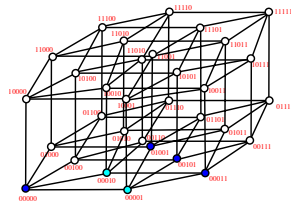


Figure 2.14: **Vertices of the bipyramid $\{00001\} + \{00000, 01001, 00101, 00011\} + \{00010\}$ arising for the Khan data set [91] restricted to $n = 4$ loci.** Dark blue dots correspond to common face $s \cap t$ of the bipyramid and light blue dots correspond to the satellite vertices of s and t .

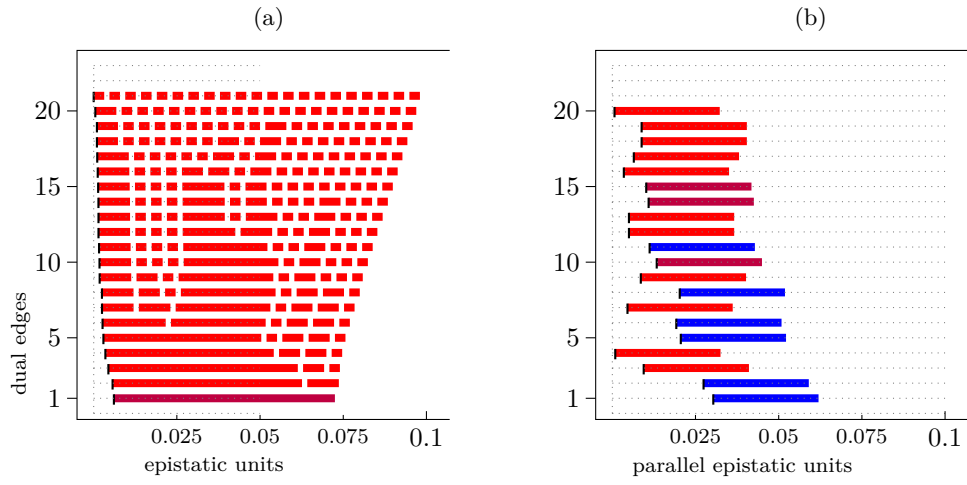


Figure 2.15: Epistatic filtration and parallel epistatic units for transport from ****0 to ****1 within the Khan data.

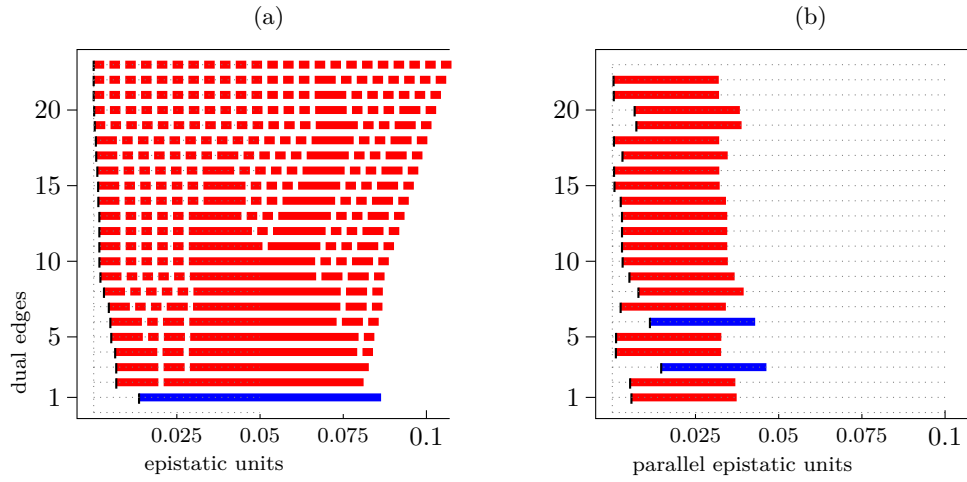


Figure 2.16: Epistatic filtration and parallel epistatic units for transport from ****1 to ****0 within the Khan data.

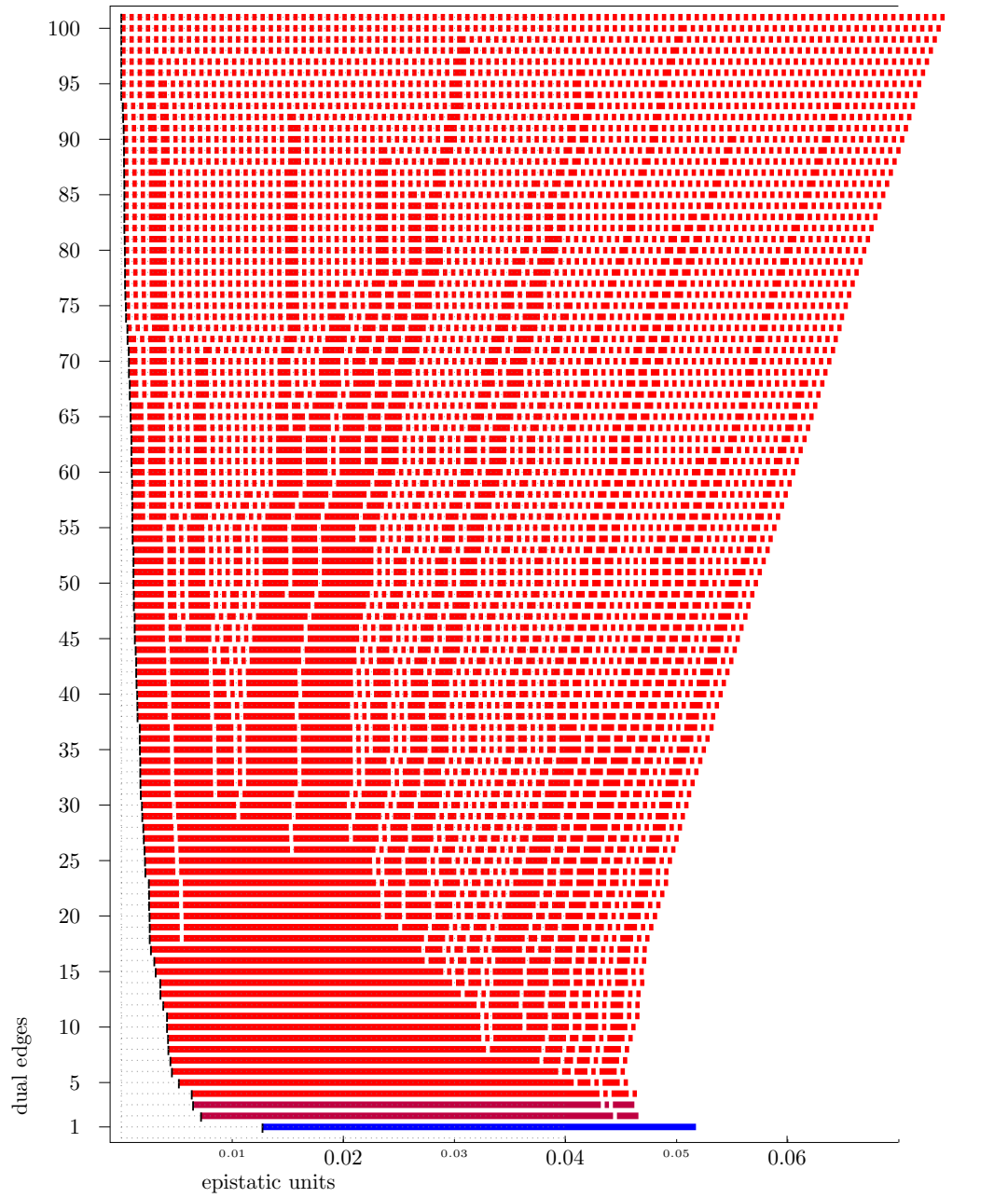


Figure 2.17: **Complete filtration of the Khan data over the whole 5-cube.**

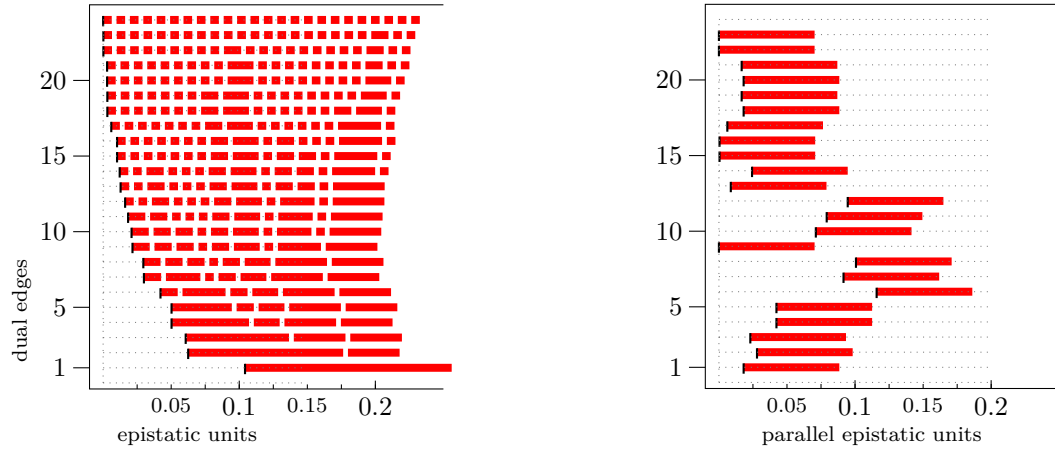


Figure 2.18: Parallel transport from 0**** to 1**** within the Tan data. Analysis based on mean values only; hence there is no color coding for the significance.

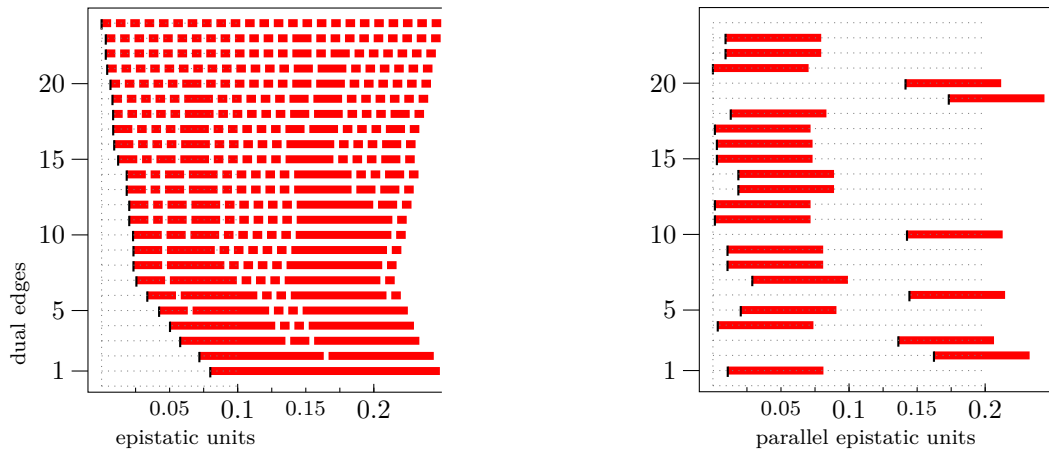


Figure 2.19: Parallel transport from the face **0** to the face **1** within the Tan data. Analysis based on mean values only; hence there is no color coding for the significance.

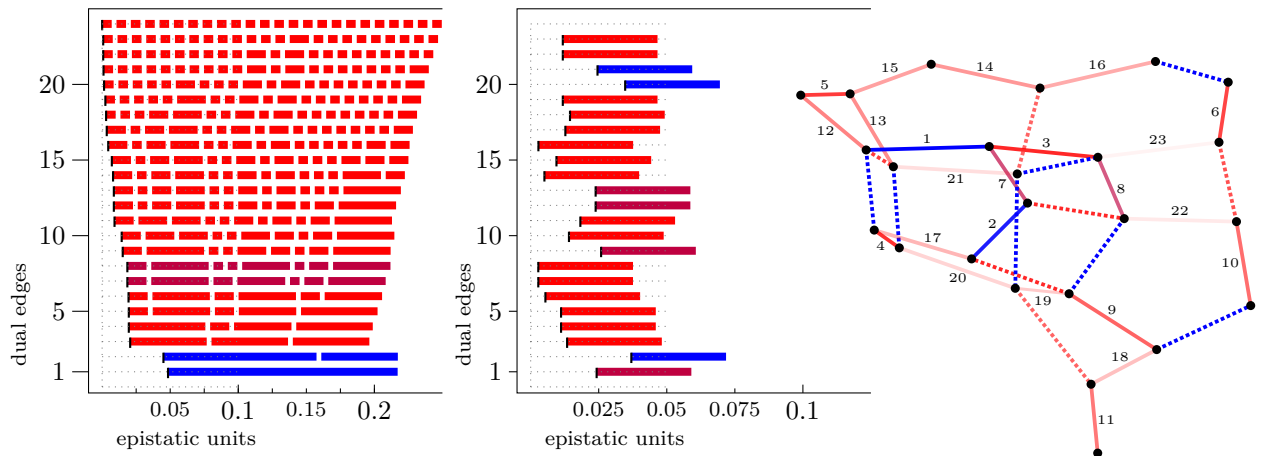


Figure 2.20: Effect of *L. plantarum*. Comparing 0**** to 1**** for Eble data. *Left*. Filtration of 0****. *Middle*. Parallel filtration of 1****. *Right*. Dual graph of 0****.

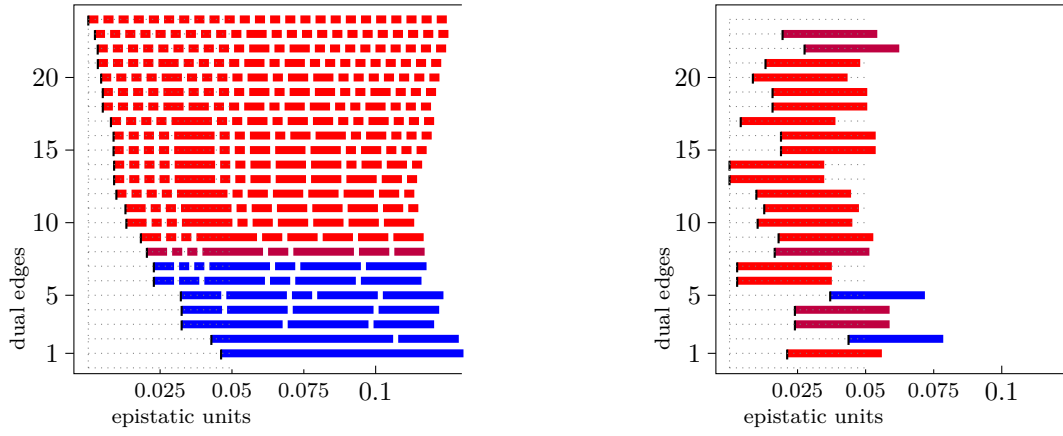


Figure 2.21: Effect of *L. brevis*. Comparing *0*** to *1*** for Eble data. *Left*. Filtration of *0***. *Right*. Parallel filtration of *1***.

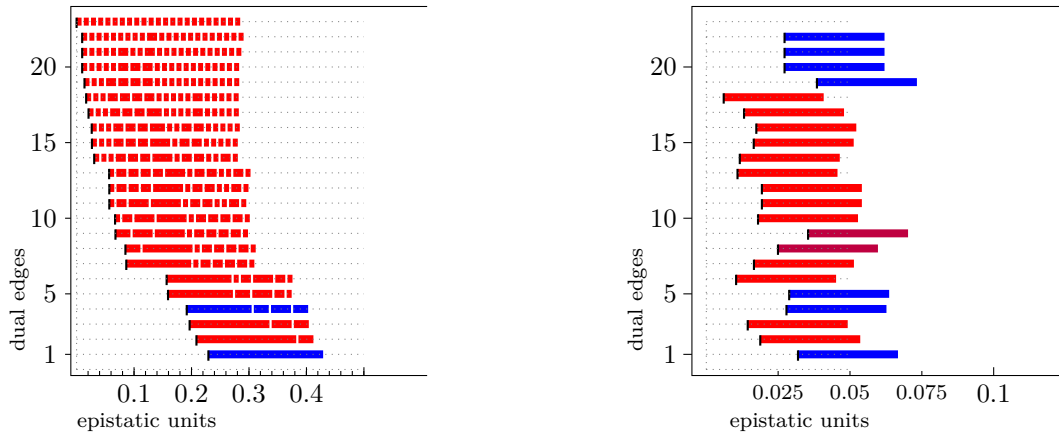


Figure 2.22: Comparing 0****(Gould bacterial CFU counts) to 0****(Gould lifespans). *Left*. Filtration of 0**** CFU counts. *Right*. Parallel filtration of 0**** lifespans.

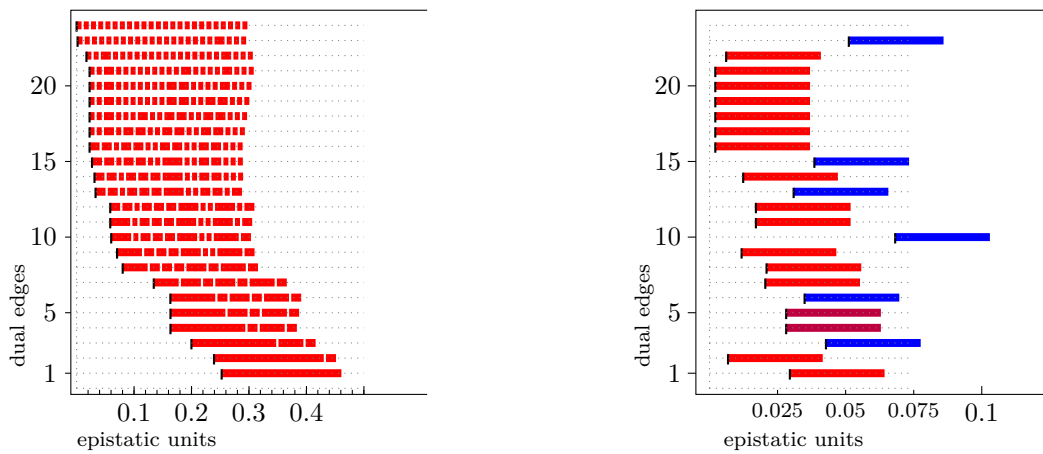


Figure 2.23: Comparing 1****(Gould bacterial CFU counts) to 1****(Gould lifespans). *Left*. Filtration of 1**** CFU counts. *Right*. Parallel filtration of 1**** lifespans.

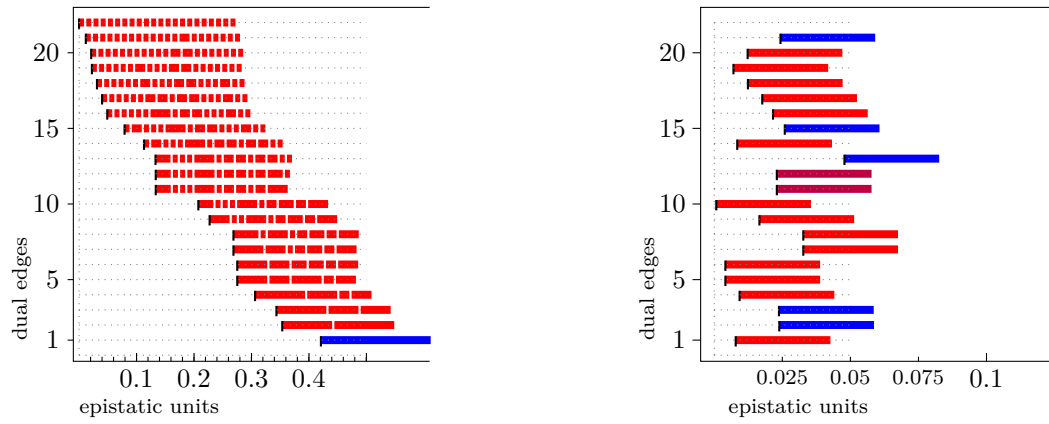


Figure 2.24: Comparing *0*** (Gould bacterial CFU counts) to *0*** (Gould lifespans). *Left.* Filtration of *0*** CFU counts. *Right.* Parallel filtration of *0*** lifespans.

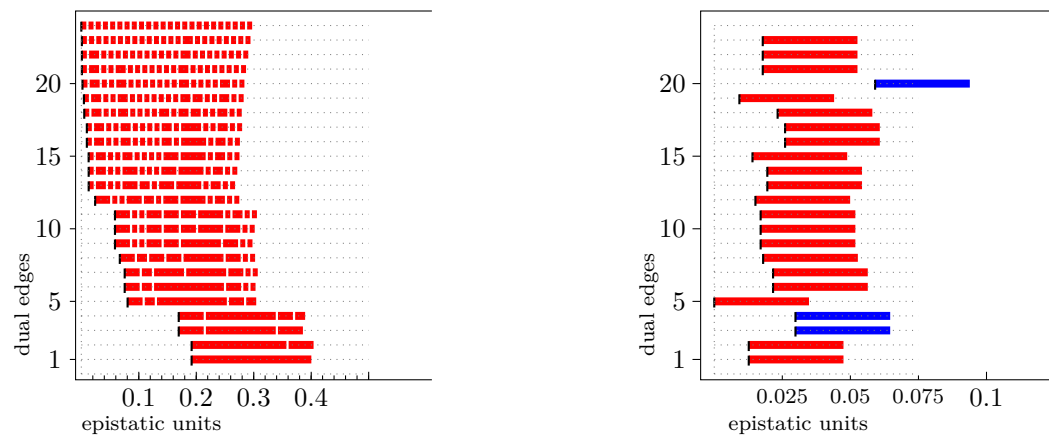


Figure 2.25: Comparing *1*** (Gould bacterial CFU counts) to *1*** (Gould lifespans). *Left.* Filtration of *1*** CFU counts. *Right.* Parallel filtration of *1*** lifespans.

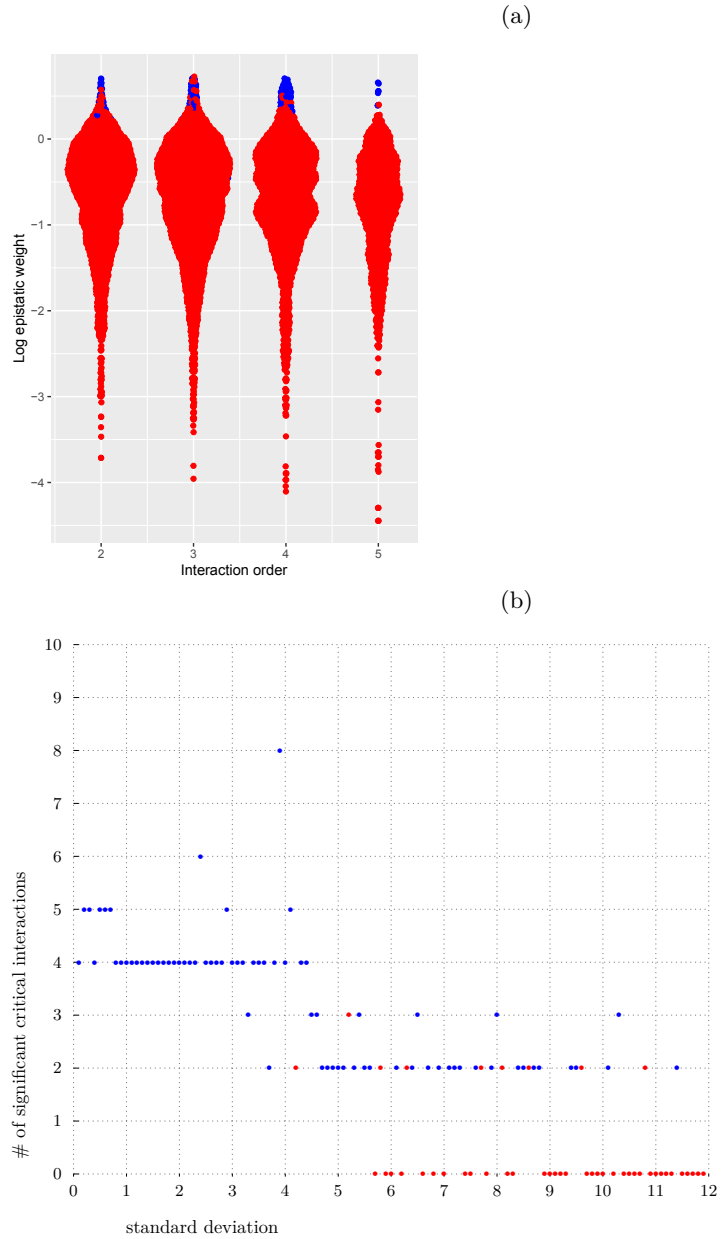


Figure 2.26: Synthetic data demonstrate method performance. Synthetic height functions over the 4-dimensional cube are generated with 100 replicates each and standard deviation as indicated. The heights of the wild type 0000 and 0001 are sampled with mean 53, all the other vertices with mean 50. (a) The distribution of \log_{10} -transformed epistatic weights is roughly constant as a function of interaction order, indicating the dimensional normalization is effective. (b) The number of significant interactions decreases as the standard deviation of the input data for each genotype increases. A blue dot is drawn if the interaction is significant and a red dot is drawn otherwise.

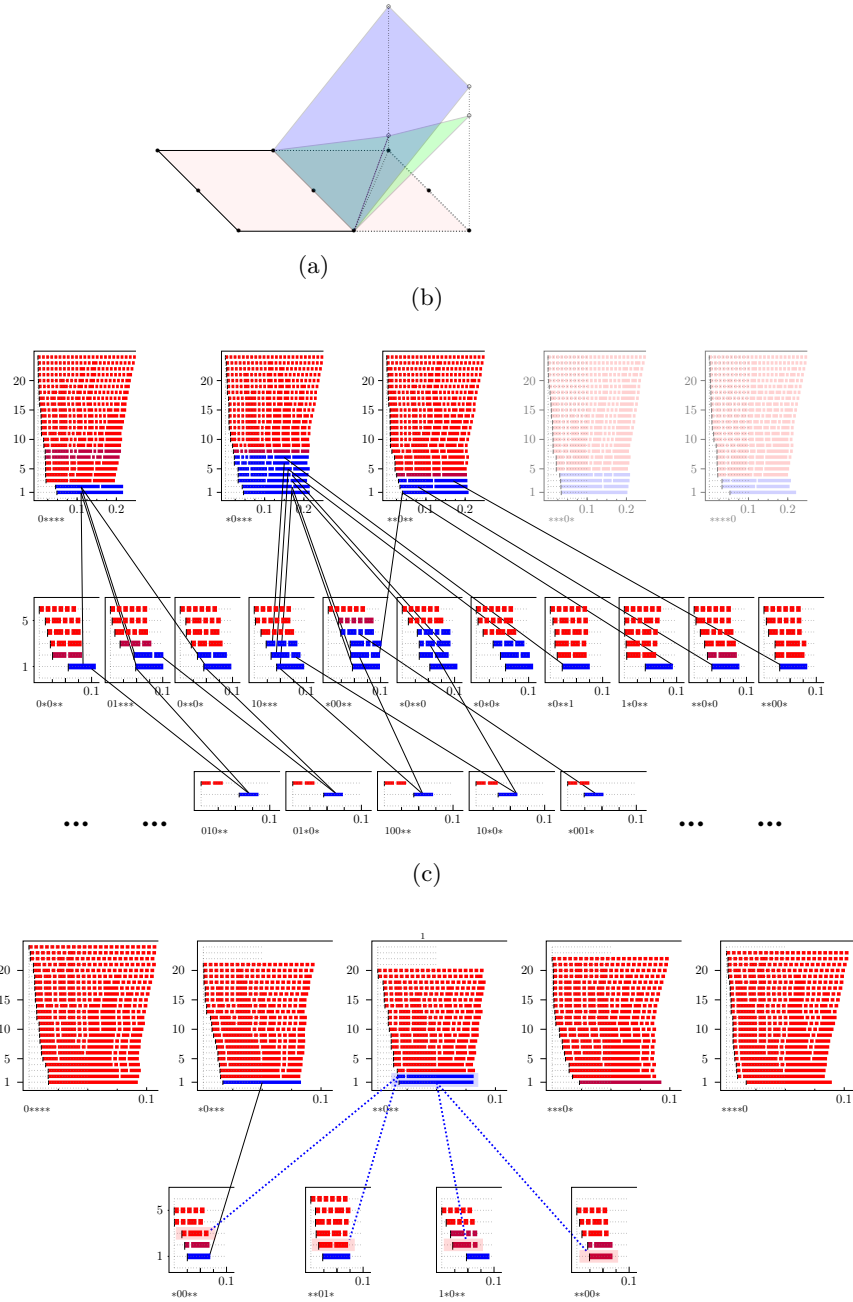


Figure 2.27: **Meta-epistatic charts illustrate whether or not higher-order interactions arise from lower-order interactions.** (a) Cartoon of the principle underlying meta-epistatic charts. The important loci in the interaction are depicted as black dots in a hyperplane through the genotypes, where the true dimensions of the genotypes are flattened onto the cartoon plane (pink). Higher-order interactions that derive from lower-order interactions occur in a new hyperplane (blue), which magnifies the weights of a subset of the landscape. In contrast, novel higher-order interactions that only arise in higher dimensions do not lie in a single additional hyperplane but instead require at least two additional hyperplanes (green). In (b) and (c) two meta-epistatic charts are represented. In each chart we identify the source of a higher-order interaction for the Eble and Gould data respectively. The results are compiled in Table 2.17.

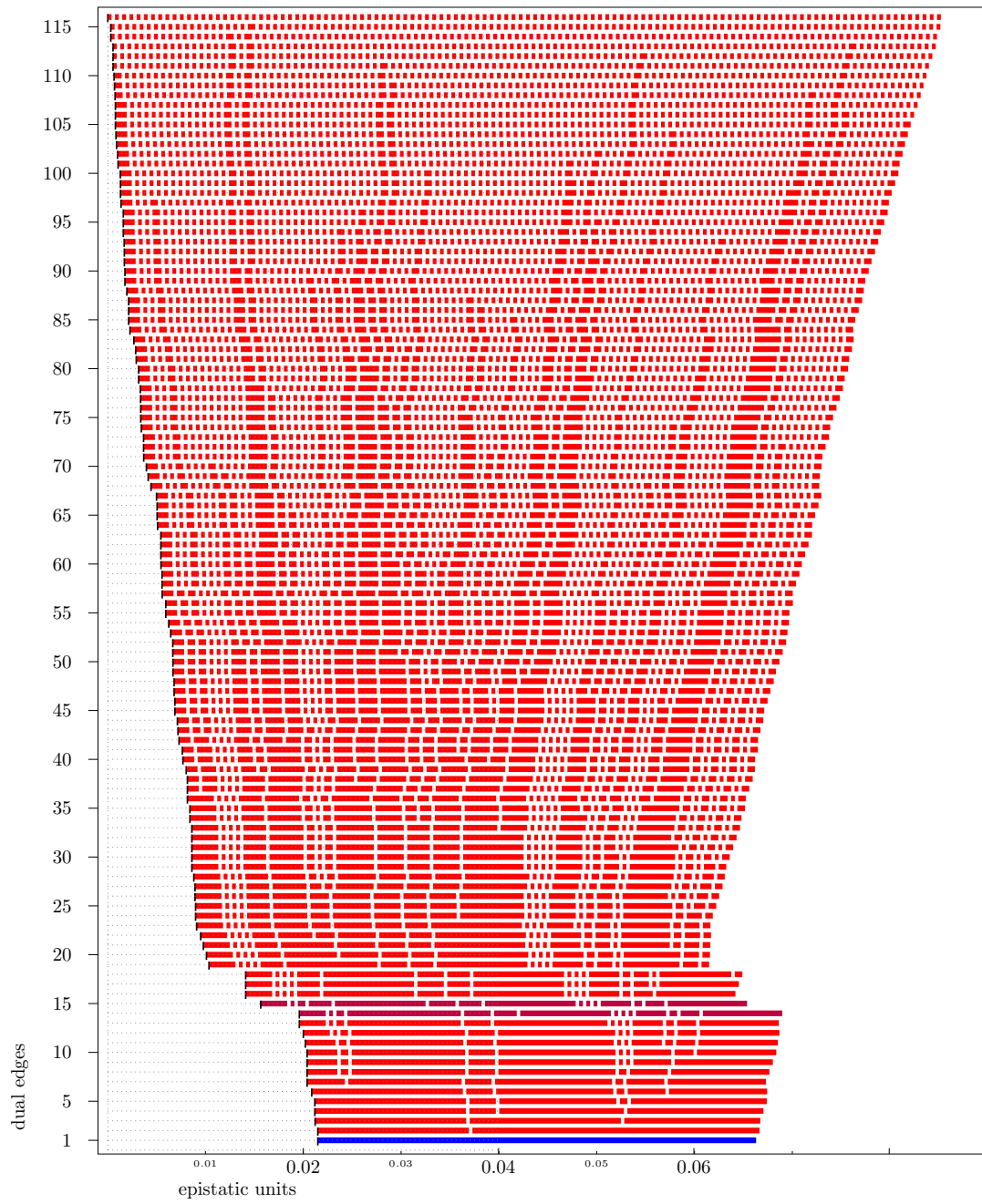


Figure 2.28: Complete filtration of the Eble fitness landscape over the whole 5-cube.

Table 2.12: Parallel analysis GouldCFU 0**** \rightarrow Gould 0****, non-critical red/red-case omitted.

No.	bipyramid	type	e_o	e_p	e_o/e_p	p_o	p_p	p_o/p_p
22	{01001}+{01000,01100,01010,00111}+{00110}	red/blue	0.010	0.027	0.357	0.978	0.038	25.873
21	{01001}+{01000,00100,01100,00111}+{00110}	red/blue	0.010	0.027	0.357	0.978	0.038	25.873
20	{01001}+{01000,00010,01010,00111}+{00110}	red/blue	0.010	0.027	0.357	0.978	0.038	25.873
19	{01001}+{01000,00100,00010,00111}+{00110}	red/blue	0.014	0.039	0.357	0.978	0.038	25.873
18	{01100}+{01001,01110,01101,00111}+{01111}	red/red	0.017	0.006	2.747	0.815	0.677	1.204
17	{01000}+{01100,01010,00110,00111}+{01110}	red/red	0.021	0.013	1.584	0.783	0.433	1.808
16	{00100}+{01100,01001,00101,00111}+{01101}	red/red	0.026	0.017	1.514	0.807	0.302	2.672
15	{01001}+{01100,01010,01110,00111}+{00110}	red/red	0.027	0.017	1.619	0.941	0.231	4.074
14	{00001}+{00010,01001,00011,00111}+{01011}	red/red	0.031	0.012	2.630	0.905	0.312	2.901
13	{01000}+{00100,00010,00001,01001}+{00111}	red/red	0.057	0.011	5.217	0.869	0.479	1.814
12	{00010}+{01000,01010,00110,00111}+{01100}	red/red	0.057	0.019	2.943	0.531	0.148	3.588
11	{00010}+{01000,00100,00110,00111}+{01100}	red/red	0.057	0.019	2.943	0.531	0.148	3.588
	{00010}+{01000,01010,01001,00111}+{01100}	red/blue	0.067	0.047	1.431	0.853	0.032	27.079
	{00010}+{01000,00100,01001,00111}+{01100}	red/blue	0.067	0.047	1.431	0.853	0.032	27.079
10	{01010}+{01001,01110,01011,00111}+{01111}	red/red	0.067	0.018	3.722	0.323	0.186	1.737
9	{00000}+{01000,00100,00010,00001}+{01001}	red/red	0.068	0.035	1.911	0.851	0.086	9.872
8	{00100}+{01000,01100,00110,00111}+{01010}	red/red	0.085	0.025	3.408	0.317	0.083	3.819
7	{01000}+{00100,00010,01001,00111}+{00001}	red/red	0.087	0.017	5.217	0.869	0.479	1.814
	{00100}+{01000,01100,01001,00111}+{01010}	red/blue	0.095	0.052	1.816	0.791	0.019	40.984
	{00100}+{01000,00010,01001,00111}+{01010}	red/blue	0.095	0.052	1.816	0.791	0.019	40.984
6	{01101}+{01001,01110,00111,01111}+{01011}	red/red	0.157	0.010	15.097	0.533	0.362	1.472
5	{00001}+{00100,01001,00101,00111}+{01100}	red/blue	0.159	0.029	5.516	0.541	0.028	19.049
4	{01010}+{00010,01001,01011,00111}+{00011}	blue/blue	0.192	0.028	6.871	0.032	0.042	0.758
3	{00010}+{00100,00001,01001,00111}+{00101}	red/red	0.197	0.014	13.654	0.262	0.211	1.242
2	{01100}+{01010,01001,01110,00111}+{01011}	red/red	0.209	0.019	11.109	0.502	0.175	2.869
1	{01000}+{00010,01010,01001,00111}+{01011}	blue/blue	0.229	0.032	7.188	0.458	0.049	9.271
	{00100}+{00010,00001,01001,00111}+{00011}	blue/red	0.365	0.007	53.243	0.026	0.526	0.049

Table 2.13: Parallel analysis GouldCFU 1**** \rightarrow Gould 1****, non-critical red/red-case omitted.

No.	bipyramid	type	e_o	e_p	e_o/e_p	p_o	p_p	p_o/p_p
23	{11001}+{11000,10101,11101,11011}+{11111}	red/blue	0.002	0.051	0.033	0.962	0.001	1286.096
22	{11100}+{11000,10101,11110,11101}+{11111}	red/red	0.017	0.006	2.799	0.773	0.689	1.122
21	{10000}+{11010,10101,10011,10111}+{11111}	red/red	0.023	0.002	10.615	0.967	0.875	1.105
20	{10000}+{11010,10101,10011,11011}+{11111}	red/red	0.023	0.002	10.615	0.967	0.875	1.105
19	{10000}+{11000,11010,10101,11011}+{11111}	red/red	0.023	0.002	10.615	0.967	0.875	1.105
18	{10000}+{11010,10110,10101,10111}+{11111}	red/red	0.023	0.002	10.615	0.967	0.875	1.105
17	{10000}+{11010,10110,10101,11110}+{11111}	red/red	0.023	0.002	10.615	0.967	0.875	1.105
16	{10000}+{11000,11010,10101,11110}+{11111}	red/red	0.023	0.002	10.615	0.967	0.875	1.105
15	{11011}+{11010,10101,10011,11111}+{10111}	red/blue	0.027	0.039	0.695	0.580	0.012	47.154
14	{10110}+{10000,10010,11010,10111}+{10011}	red/red	0.031	0.012	2.513	0.693	0.277	2.502
13	{11001}+{10000,10001,10101,11011}+{10011}	red/blue	0.033	0.031	1.066	0.388	0.007	54.190
12	{11010}+{11000,10101,11011,11111}+{11101}	red/red	0.059	0.017	3.428	0.905	0.318	2.846
11	{11010}+{11000,10101,11110,11111}+{11101}	red/red	0.059	0.017	3.428	0.905	0.318	2.846
10	{11010}+{10000,11000,10101,11011}+{11001}	red/blue	0.060	0.068	0.881	0.902	0.000	∞
9	{10000}+{11000,11100,10101,11110}+{11101}	red/red	0.070	0.012	5.959	0.897	0.426	2.106
8	{10100}+{10000,11100,10110,10101}+{11110}	red/red	0.080	0.021	3.820	0.430	0.274	1.569
7	{11110}+{11010,10110,10101,11111}+{10111}	red/red	0.134	0.021	6.534	0.130	0.227	0.573
6	{11000}+{10000,11010,10101,11110}+{10110}	red/blue	0.163	0.035	4.659	0.737	0.019	38.586
5	{10010}+{10000,11010,10110,10111}+{10101}	red/red	0.163	0.028	5.788	0.776	0.075	10.402
4	{10010}+{10000,11010,10011,10111}+{10101}	red/red	0.163	0.028	5.788	0.776	0.075	10.402
	{11000}+{11010,10101,11110,11111}+{10110}	red/blue	0.186	0.037	5.000	0.695	0.026	26.834
3	{11000}+{10000,11100,10101,11110}+{10110}	red/blue	0.200	0.043	4.659	0.737	0.019	38.586
2	{11000}+{10000,11010,10101,11011}+{10011}	red/red	0.239	0.007	35.102	0.621	0.628	0.989
1	{11000}+{10000,11001,10101,11011}+{10001}	red/red	0.253	0.030	8.530	0.671	0.104	6.452
	{11110}+{11000,11010,10101,11111}+{11011}	red/blue	0.301	0.026	11.785	0.288	0.035	8.348
	{10001}+{10000,10101,10011,11011}+{11010}	red/blue	0.313	0.039	8.062	0.598	0.014	43.650

Table 2.14: Parallel analysis GouldCFU *0*** \rightarrow Gould *0***, non-critical red/red-case omitted.

No.	bipyramid	type	ϵ_o	ϵ_p	ϵ_o/ϵ_p	P_o	P_p	P_o/P_p
21	{10001}+{10000,00001,10101,10011}+{00111}	red/blue	0.012	0.024	0.481	0.963	0.026	36.756
20	{00010}+{10000,10010,00011,00111}+{10011}	red/red	0.021	0.012	1.714	0.797	0.270	2.952
19	{10100}+{10000,00100,10110,10101}+{00110}	red/red	0.022	0.007	3.155	0.869	0.717	1.212
18	{10110}+{10000,10010,00111,10111}+{10011}	red/red	0.031	0.012	2.513	0.693	0.277	2.502
17	{00100}+{10000,00110,10101,00111}+{10110}	red/red	0.040	0.018	2.266	0.915	0.290	3.155
16	{00100}+{10000,00110,10110,10101}+{00111}	red/red	0.049	0.022	2.266	0.915	0.290	3.155
15	{00001}+{10000,00100,00010,00111}+{00110}	red/blue	0.079	0.026	3.047	0.698	0.023	30.749
14	{00010}+{10000,10010,00110,00111}+{10110}	red/red	0.113	0.008	13.352	0.295	0.461	0.640
13	{00000}+{10000,00100,00010,00001}+{00111}	red/blue	0.133	0.048	2.775	0.476	0.001	707.281
12	{10010}+{10000,10011,00111,10111}+{10101}	red/red	0.133	0.023	5.788	0.776	0.075	10.402
11	{10010}+{10000,10110,00111,10111}+{10101}	red/red	0.133	0.023	5.788	0.776	0.075	10.402
10	{00011}+{10000,00001,10011,00111}+{10101}	red/red	0.208	0.001	275.689	0.413	0.949	0.435
9	{00010}+{10000,00100,00001,00111}+{00101}	red/red	0.227	0.017	13.654	0.262	0.211	1.242
8	{00100}+{10000,00001,00101,00111}+{10101}	red/red	0.269	0.033	8.193	0.579	0.101	5.733
7	{00001}+{10000,00100,00101,00111}+{10101}	red/red	0.269	0.033	8.193	0.579	0.101	5.733
6	{00001}+{10000,00010,00011,00111}+{10010}	red/red	0.275	0.004	67.167	0.493	0.755	0.653
5	{00001}+{10000,00011,10011,00111}+{10010}	red/red	0.275	0.004	67.167	0.493	0.755	0.653
4	{00110}+{10000,00100,10101,00111}+{00101}	red/red	0.306	0.009	32.813	0.413	0.610	0.677
3	{00110}+{10000,10010,10110,00111}+{10111}	red/blue	0.344	0.024	14.403	0.186	0.035	5.345
2	{00110}+{10000,00010,10010,00111}+{00011}	red/blue	0.354	0.024	14.760	0.175	0.030	5.853
	{00001}+{10000,10101,10011,00111}+{10111}	red/blue	0.408	0.028	14.815	0.108	0.013	8.308
1	{00100}+{10000,00010,00001,00111}+{00011}	blue/red	0.421	0.008	53.243	0.026	0.526	0.049
	{00110}+{10000,10110,10101,00111}+{10111}	red/blue	0.486	0.034	14.403	0.186	0.035	5.345

Table 2.15: Parallel analysis GouldCFU *1*** \rightarrow Gould *1***, non-critical red/red-case omitted.

No.	bipyramid	type	ϵ_o	ϵ_p	ϵ_o/ϵ_p	P_o	P_p	P_o/P_p
23	{01100}+{11000,01110,01101,11110}+{11010}	red/red	0.001	0.018	0.054	0.998	0.292	3.418
22	{01100}+{11000,01010,01001,01110}+{11010}	red/red	0.001	0.018	0.054	0.998	0.292	3.418
21	{01100}+{11000,01001,01110,01101}+{11010}	red/red	0.001	0.018	0.054	0.998	0.292	3.418
20	{11001}+{11000,01001,11101,11011}+{11111}	red/blue	0.002	0.059	0.033	0.962	0.001	1286.096
19	{11110}+{11010,01110,01101,11111}+{01111}	red/red	0.005	0.009	0.488	0.945	0.576	1.641
18	{11100}+{11000,01100,11110,11101}+{01101}	red/red	0.005	0.023	0.218	0.952	0.193	4.933
17	{11000}+{11010,01110,01101,11110}+{11111}	red/red	0.010	0.026	0.369	0.981	0.106	9.255
16	{01110}+{11000,11010,01101,11110}+{11111}	red/red	0.010	0.026	0.369	0.981	0.106	9.255
15	{01100}+{11000,01101,11110,11101}+{11111}	red/red	0.013	0.014	0.905	0.866	0.346	2.503
14	{11000}+{01001,11010,01110,01101}+{01111}	red/red	0.013	0.020	0.656	0.974	0.160	6.087
13	{11000}+{01001,11010,01101,11111}+{01111}	red/red	0.013	0.020	0.656	0.974	0.160	6.087
12	{01000}+{11000,01100,01010,01001}+{01110}	red/red	0.024	0.015	1.584	0.783	0.433	1.808
11	{11010}+{11000,01001,01101,11111}+{11101}	red/red	0.059	0.017	3.428	0.905	0.318	2.846
10	{11010}+{11000,01001,11011,11111}+{11101}	red/red	0.059	0.017	3.428	0.905	0.318	2.846
9	{11010}+{11000,01101,11110,11111}+{11101}	red/red	0.059	0.017	3.428	0.905	0.318	2.846
8	{01010}+{01001,11010,01110,01011}+{01111}	red/red	0.067	0.018	3.722	0.323	0.186	1.737
7	{01110}+{01001,11010,01101,01111}+{11111}	red/red	0.075	0.022	3.483	0.841	0.136	6.184
6	{01001}+{11010,01110,01101,01111}+{11111}	red/red	0.075	0.022	3.483	0.841	0.136	6.184
5	{11011}+{01001,11010,01011,11111}+{01111}	red/red	0.081	0.000	1235.241	0.126	0.996	0.127
4	{11000}+{01010,01001,11010,01110}+{01011}	red/blue	0.170	0.030	5.666	0.718	0.026	27.722
3	{11000}+{01001,11010,11011,11111}+{01011}	red/blue	0.170	0.030	5.666	0.718	0.026	27.722
2	{01101}+{01001,11010,01110,01111}+{01011}	red/red	0.192	0.013	15.097	0.533	0.362	1.472
1	{01101}+{01001,11010,01111,11111}+{01011}	red/red	0.192	0.013	15.097	0.533	0.362	1.472

Table 2.16: Parallel analysis Eble 0**** \rightarrow 1****, non-critical red/red-case omitted.

No.	bipyramid _s	type	e_o	e_p	e_o/e_p	p_o	p_p	p_o/p_p
23	{00001}+{00000,01001,01011,00111}+{01111}	red/red	0.001	0.012	0.066	0.953	0.390	2.444
22	{00001}+{00000,01001,01101,00111}+{01111}	red/red	0.001	0.012	0.066	0.953	0.390	2.444
21	{01110}+{00000,00110,01011,01111}+{00111}	red/blue	0.001	0.025	0.041	0.923	0.038	24.226
20	{01110}+{00000,01100,00110,01111}+{00111}	red/blue	0.001	0.035	0.041	0.923	0.038	24.226
19	{00110}+{00000,01100,00111,01111}+{01101}	red/red	0.002	0.012	0.201	0.827	0.303	2.729
18	{00110}+{00000,01100,00101,00111}+{01101}	red/red	0.003	0.014	0.201	0.827	0.303	2.729
17	{01110}+{00000,01000,01100,01111}+{01101}	red/red	0.003	0.013	0.264	0.742	0.251	2.956
16	{00110}+{00000,00010,01011,00111}+{00011}	red/red	0.004	0.003	1.568	0.755	0.843	0.896
15	{00010}+{00000,01010,00110,01011}+{01110}	red/red	0.007	0.010	0.748	0.606	0.488	1.242
14	{01010}+{00000,00010,00110,01011}+{00111}	red/red	0.008	0.005	1.583	0.443	0.639	0.693
13	{01010}+{00000,00110,01110,01011}+{01111}	red/red	0.009	0.024	0.359	0.475	0.062	7.686
12	{01010}+{00000,01000,01110,01011}+{01111}	red/red	0.009	0.024	0.359	0.475	0.062	7.686
11	{00100}+{00000,01100,00110,00101}+{00111}	red/red	0.009	0.018	0.498	0.533	0.269	1.981
10	{01001}+{00000,00001,01101,00111}+{00101}	red/red	0.014	0.014	1.018	0.288	0.313	0.920
9	{00101}+{00000,01100,01101,00111}+{01111}	red/red	0.015	0.026	0.584	0.228	0.062	3.695
	{00101}+{00000,01100,00110,00111}+{01111}	red/blue	0.018	0.040	0.446	0.321	0.035	9.119
8	{01101}+{00000,01001,00111,01111}+{01011}	red/red	0.019	0.003	6.623	0.068	0.800	0.085
7	{01101}+{00000,01000,01001,01111}+{01011}	red/red	0.019	0.003	6.623	0.068	0.800	0.085
6	{01001}+{00000,00001,01011,00111}+{00011}	red/red	0.019	0.005	3.571	0.153	0.689	0.222
5	{01000}+{00000,01010,01110,01011}+{00110}	red/red	0.020	0.011	1.750	0.169	0.443	0.381
4	{01000}+{00000,01100,01110,01111}+{00110}	red/red	0.020	0.011	1.750	0.169	0.443	0.381
3	{01000}+{00000,01001,01011,01111}+{00111}	red/red	0.021	0.013	1.535	0.140	0.339	0.413
2	{01100}+{00000,01000,01101,01111}+{01001}	blue/blue	0.045	0.037	1.215	0.000	0.003	0.176
1	{01001}+{00000,01000,01011,01111}+{01110}	blue/red	0.048	0.024	1.993	0.000	0.056	0.002
	{01100}+{00000,01000,01110,01111}+{01011}	blue/blue	0.064	0.034	1.855	0.000	0.005	0.000
	{00010}+{00000,00011,01011,00111}+{00001}	blue/blue	0.065	0.043	1.518	0.000	0.001	0.001
	{01100}+{00000,01101,00111,01111}+{01001}	blue/red	0.066	0.024	2.775	0.000	0.105	0.000
	{00001}+{00000,00101,01101,00111}+{01100}	blue/blue	0.066	0.033	1.989	0.000	0.009	0.000
	{01001}+{00000,01011,00111,01111}+{00110}	blue/blue	0.068	0.036	1.917	0.000	0.007	0.000
	{01100}+{00000,00110,01110,01111}+{01011}	blue/blue	0.083	0.045	1.829	0.000	0.002	0.000
	{01100}+{00000,00110,00111,01111}+{01011}	blue/red	0.084	0.021	4.035	0.000	0.210	0.000

Table 2.17: Significant 4-dimensional interactions, which cannot be seen in lower dimensions, cf. (Fig. 2.27). The value $p \uparrow$ refers to the p -value of the 4-dimensional bipyramid in question whereas $p \downarrow$ is the p -value of its ridge intersected with the \cap -face, cf. (Fig. 2.27c) for the Gould data.

Data	significant bipyramid	\cap -face	$p \uparrow$	$p \downarrow$
Eble	-	-	-	-
Gould				
0	{00010} + {00000, 10010, 00011, 11011} + {10001}	**01*	0.041	0.270
		*00**	0.041	0.149
	{10010} + {00000, 11000, 10001, 11011} + {01001}	1*0**	0.041	0.076
		**00*	0.041	0.063
Khan				
0****	{00010} + {00000, 01001, 00101, 00011} + {00001}	0***1	0.009	0.052

Table 2.18: Bacterial species considered in the two microbiome data sets.

	Gould data set	Eble data set
Species 1	<i>L. plantarum</i>	<i>L. plantarum</i>
Species 2	<i>L. brevis</i>	<i>L. brevis</i>
Species 3	<i>A. pasteurianus</i>	<i>A. cerevisiae</i>
Species 4	<i>A. tropicalis</i>	<i>A. malorum</i>
Species 5	<i>A. orientalis</i>	<i>A. orientalis</i>

Table 2.19: Regressions over $\{0001\} + \{0000, 1001, 1011, 0111\} + \{1111\}$ for normalized lifespan data for Eble 0**** and Eble 1****.

	Coefficient	Std. error	<i>t</i> -statistic	<i>p</i> -value
β_0	0	0	nan	nan
x_1	-0.0270	0.009	-2.987	0.003
x_2	-0.0149	0.012	-1.246	0.213
x_3	-0.0156	0.012	-1.306	0.192
x_4	0.2039	0.008	26.022	0.000
β_0	0.2320	0.005	44.642	0.000
x_1	0.0310	0.005	5.957	0.000
x_2	0.0610	0.007	8.874	0.000
x_3	-0.0185	0.007	-2.692	0.007
x_4	-0.0861	0.007	-12.518	0.000

The MST-fan of a regular subdivision

This chapter is based on the preprint “The MST-fan of a regular subdivision” [45] by Holger Eble. The preprint can be found on the arXiv with the number 2205.10424.

3.0 Abstract

The dual graph $\Gamma(h)$ of a regular triangulation $\Sigma(h)$ carries a natural metric structure. The minimum spanning trees of $\Gamma(h)$ recently proved to be conclusive for detecting significant data signal in the context of population genetics. In this paper we prove that the parameter space of such minimum spanning trees is organized as a polyhedral fan, called the MST-fan of $\Sigma(h)$, which subdivides the secondary cone of $\Sigma(h)$ into parameter cones. We partially describe its local face structure and examine the connection to tropical geometry in virtue of matroids and Bergman fans.

3.1 Introduction

Triangulations, or more generally polyhedral complexes, play a crucial role both in theoretical and applied scientific fields related to mathematics, such as genetics [9] and machine learning [162]. See [40, Chapter 1] for an overview of triangulations appearing within mathematics. Given a finite point configuration $\mathbf{A} \subset \mathbb{Z}^n$, in practice one often considers a specific class of subdivisions of \mathbf{A} , known as *regular subdivisions*. Their maximal cells can be described as shadows of n -dimensional polytopes in \mathbb{R}^{n+1} , namely facets of the upper convex envelope associated to height functions $h: \mathbf{A} \rightarrow \mathbb{R}$ lifting \mathbf{A} into \mathbb{R}^{n+1} . The cellular incidences of a fixed regular triangulation $\Sigma(h)$ of \mathbf{A} are recorded in its dual polyhedral complex, the tropical hypersurface $V(H)$ given by the tropical polynomial

$$H(x) := \bigoplus_{a \in \mathbf{A}} h(a) \odot x^a := \max_{a \in \mathbf{A}} \{h(a) + \langle a, x \rangle\} ,$$

which is defined to be the set of points $x \in \mathbb{R}^n$ where at least two of the linear forms $h(a) + \langle a, \cdot \rangle$ of H attain the maximum $H(x)$, cf. [86, 106]. The bounded cells of $V(H)$ form the *tight span* of $\Sigma(h)$ [72], whose 1-skeleton is called the *dual graph* $\Gamma(h)$ of $\Sigma(h)$. The nodes of $\Gamma(h)$, i.e. the 0-cells of $V(H)$, correspond to maximal cells of $\Sigma(h)$ and their adjacency relations are recorded by the edges $E(\Gamma(h))$, which in addition are naturally weighted: If $s = \text{conv}(v_1, \dots, v_{n+1})$ and $t = \text{conv}(v_2, \dots, v_{n+2})$ are adjacent maximal simplices of $\Sigma(h)$, the *lattice length* [22] or *tropical edge length* [112] of the edge $(s, t) \in E(\Gamma(h))$ is given by

$$L_h(s, t) := |\tilde{L}_h(s, t)| := \left| \det \begin{pmatrix} 1 & v_{1,1} & v_{1,2} & \dots & v_{1,n} & h(v_1) \\ 1 & v_{2,1} & v_{2,2} & \dots & v_{2,n} & h(v_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & v_{n+2,1} & v_{n+2,2} & \dots & v_{n+2,n} & h(v_{n+2}) \end{pmatrix} \right| . \quad (3.26)$$

Now, for a fixed height function $h: \mathbf{A} \rightarrow \mathbb{R}$ the greedy algorithm chooses some minimum spanning tree $T_{\min}(h) \subset \Gamma(h)$, cf. [133, Chapter 50], and in the sequel of this article we will study the fibers of the operation $T_{\min}(\cdot)$. It is well-known [40, Section 5.2] that the parameter space $\mathbb{R}^{\mathbf{A}}$ for regular subdivisions $\Sigma(h)$ of \mathbf{A} is organized in a nicely behaved union of secondary cones $\text{sc}(h) \subset \mathbb{R}^{\mathbf{A}}$, called the secondary fan of \mathbf{A} . The following Main Theorem is Corollary 3.4.3 and it identifies the parameter space $\mathcal{K}(h) = \cup \mathcal{K}(T)$ of *realizable minimum spanning trees* $T \subset \Gamma(h)$, i.e. those trees

T with $T = T_{\min}(g) \subset \Gamma(h)$ for some $g \in \text{sc}(h)$, as a conic subdivision $\mathcal{K}(h)$ of $\text{sc}(h)$. We refer to $\mathcal{K}(h)$ as the *MST-fan* structure on $\text{sc}(h)$.

Main Theorem. *Let $\Sigma(h)$ be a regular triangulation of the point configuration \mathbf{A} . The MST-fan $\mathcal{K}(h)$ of $\Sigma(h)$ is a pure fan in $\text{sc}(h) \subset \mathbb{R}^{\mathbf{A}}$ with full support. Its maximal cones are in bijection with realizable minimum spanning trees of $\Gamma(h)$.*

Let again $h: \mathbf{A} \rightarrow \mathbb{R}$ be a fixed height function and let $G = G(V, E) = \Gamma(h)$. A collection $\mathcal{C} = \{\mathcal{K}(T^{(\mathcal{C}_i)})\}$ of maximal cones in $\mathcal{K}(h)$ is called *G-saturating* if the spanning trees $T^{(\mathcal{C}_i)}$ cover G , i.e. if $E = \cup_i E(T^{(\mathcal{C}_i)})$ holds. Let $C \subset \mathbb{R}^{\mathbf{A}}$ be the polyhedral complex with cells $\bigcap \mathcal{K}(T^{(\mathcal{C}_i)})$ where \mathcal{C} runs through the *G-saturating* collections.

Corollary. *The complex C naturally embeds into the tropical linear space $\text{trop}(\mathcal{M})$, where $\mathcal{M} := \mathcal{M}(G)$ is the cycle matroid of G , and equals the Bergman fan of \mathcal{M} restricted to valuations given by tropical edge lengths as in equation (3.26).*

Motivation from population genetics

Regular subdivisions admit a reformulation in terms of linear optimization: A subdivision Σ of a point configuration \mathbf{A} with *support* $|\Sigma| := \bigcup_{\sigma \in \Sigma} \sigma$ is regular if and only if there is a convex support function $f_{\Sigma}: |\Sigma| \rightarrow \mathbb{R}$, whose domains of linearity coincide with the maximal cells of Σ , cf. [23, Section 1.F].

This interplay of optimization and polyhedral combinatorics was the grounding for [46] and [47], where the theory of regular subdivisions was applied to statistical population genetics in order to detect and study non-additive mutational gene effects, called *epistasis*. Note that the tropical edge length (3.26) measures the degree to which the points $(x, h(x))$, where x runs through the vertex set of $s \cup t$, fail to lie on a hyperplane in \mathbb{R}^{n+1} and thus provides an adequate measure for epistasis.

Put in a biological wording, the point set \mathbf{A} represents an n -loci biallelic (or multi-allelic) system of genotypes, i.e. $\mathbf{A} = \{0, 1\}^n$ or more generally, \mathbf{A} is the vertex set of a product of simplices, and the function $h: \mathbf{A} \rightarrow \mathbb{R}$ is a genotype-phenotype map which showcases some experimentally obtained physical quantity. Since in this case the support function f_{Σ} can be chosen to record *fittest populations* on \mathbf{A} with respect to h , cf. Section 3.7 and [9], so do the maximal cells of $\Sigma(h)$. With a slight modification of the tropical edge length (3.26) named *epistatic weight*, the minimum spanning tree $T_{\min}(h)$ was introduced as *epistatic filtration* in [47]. The edges of $T_{\min}(h)$, especially the *significant* ones satisfying a statistical p -test, depict a selective choice of *epistatic interactions* and proved to be able to reveal biological relevant information.

Outline of the paper

In Section 3.2 we review facts on regular subdivisions and tropical geometry. In Section 3.3 we introduce the MST-fan of some given regular subdivision. Section 3.4 provides an algorithm for computing the single MST-cones from trees and in Section 3.5 we discuss the effect of changing the edge order of the spanning trees. In Section 3.6 we reinterpret the MST-fan in terms of matroid theory and tropical geometry. Section 3.7 shortly explains some use-case for applying tropical geometry to biology from where the motivation to study MST-fan structures initially rebounded. In Section 3.8 we provide some computational results.

I am indebted to Michael Joswig and Marta Panizzut for many helpful discussions.

3.2 Regular subdivisions and tropical hypersurfaces

Let $\mathbf{A} \subset \mathbb{R}^n$ be a point configuration with point labels J . Following [40, Section 2.3], a *polyhedral subdivision* of \mathbf{A} is a collection \mathcal{C} of subsets of J , such that the convex hulls of \mathcal{C} cover \mathbf{A} and intersect nicely. A *regular subdivision* of \mathbf{A} is a polyhedral subdivision of \mathbf{A} , which is induced by a height function $h: \mathbf{A} \rightarrow \mathbb{R}$ taking the upper facets of the lifted configuration $\text{conv}(\mathbf{A}^h) := \text{conv}\{(v, h(v)): v \in \mathbf{A}\}$ with the induced label set. For instance, the 3-cube has 74 triangulations (6 triangulations up to symmetry) and all of them are regular. The 4-cube has 92,487,256 triangulations (247,451 triangulations up to symmetry) and 87,959,448 of them are regular, cf. [74, 124] for the original results and [84] for a recomputation using parallelized reverse search.

Remark 3.2.1. *The regularity of a polyhedral subdivision can be characterized in terms of convex functions as follows, cf. [23, Section 1.F]. For a given convex set $S \subset \mathbb{R}^n$, a function $f: S \rightarrow \mathbb{R}$ is called convex if for any two points $x, y \in S$ the graph of the restriction of f to the line segment $[x, y] \subset S$ lies below the line segment $[(x, f(x)), (y, f(y))]$ in \mathbb{R}^{n+1} . Now, a polyhedral subdivision Σ of \mathbf{A} is regular precisely if it has a convex support function f_Σ , i.e. the convex hulls of the maximal cells in Σ are the regions of linearity of f_Σ .*

3.2.1 The parameter space for regular subdivision

Let $\Sigma(h)$ be a regular triangulation of \mathbf{A} . The parameter space of all $g \in \mathbb{R}^{\mathbf{A}}$ inducing the same triangulation $\Sigma(g) = \Sigma(h)$ is a full-dimensional cone in $\mathbb{R}^{\mathbf{A}}$, called *secondary cone* $\text{sc}(h)$ of $\Sigma(h)$, cf. [40, Section 5.2]. It can be understood by looking at local folding constraints on h , which push the lifts $(s, t)^h$ of bipyramids (s, t) in \mathbf{A} to the upper hull of \mathbf{A}^h . For a given simplex $s = (v_1, \dots, v_{n+1})$ in \mathbf{A} , the conditions on a height function $h \in \mathbb{R}^{\mathbf{A}}$ to show s in its induced subdivision $\Sigma(h)$ are installed by the linear *folding constraints* $\Psi_{s,j} \geq 0$ for $v_j \in \mathbf{A} \setminus s$, where

$$\Psi_{s,j}(h) := \text{sign}(\det(\tilde{L}_h(s))) \cdot \det \begin{pmatrix} 1 & v_{1,1} & v_{1,2} & \dots & v_{1,n} & h(v_1) \\ 1 & v_{2,1} & v_{2,2} & \dots & v_{2,n} & h(v_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & v_{n+1,1} & v_{n+1,2} & \dots & v_{n+1,n} & h(v_{n+1}) \\ 1 & v_{j,1} & v_{j,2} & \dots & v_{j,n} & h(v_j) \end{pmatrix} \quad (3.27)$$

and where $\tilde{L}_h(s)$ denotes the right matrix of (3.27) with the last row and column omitted. Hence, the full-dimensional cone $H_s := \bigcap_{v_j \in \mathbf{A} \setminus s} \{\Psi_{s,j} \geq 0\} \subset \mathbb{R}^{\mathbf{A}}$ is the parameter space for s to appear in the induced triangulation. Geometrically, this system requires $(\mathbf{A} - s)^h$ to lie below the hyperplane spanned by s^h .

3.2.2 Tropical hypersurfaces

Tropical Geometry is a discrete variant of Algebraic Geometry and is established over the tropical semifield $(\mathbb{R} \cup \{-\infty\}, \oplus, \odot)$, where $a \oplus b := \max(a, b)$ and $a \odot b := a + b$. Similar to ordinary Algebraic Geometry, tropical zero sets of tropical polynomials are the central geometric objects of study. Given a set of exponents $S \subset \mathbb{Z}^d$ and coefficients $h(u) \in \mathbb{R}$ for $u \in S$, the associated d -variate *tropical polynomial* reads

$$F_h := \bigoplus_{u \in S} h(u) \odot x_1^{u_1} \odot x_2^{u_2} \odot \dots \odot x_d^{u_d} = \max_{u \in S} \{h(u) + \langle u, x \rangle\}$$

and can be understood as a function $F_h(x): \mathbb{R}^d \rightarrow \mathbb{R}$ in x . The *tropical hypersurface* $V(F_h) \subset \mathbb{R}^d$ is defined to be the locus where F_h tropically vanishes, i.e. $V(F_h)$ is the set of all points $x \in \mathbb{R}^d$ where at least two tropical summands of F_h evaluate at x to $F_h(x)$.

Remark 3.2.2. *Let (s, t) be a bounded edge of $V(F_h)$. Then the tropical edge length $\ell_h(s, t)$ equals the euclidean distance between the two 0-cells s and t of $V(F_h)$ upto a constant factor not depending on h .*

An important result of tropical geometry states that the k -dimensional cells of the hypersurface $V(F_h)$ are in bijection with the $(n - k)$ -dimensional cells of the regular subdivision $\Sigma(h)$ for $0 \leq k < n$, see [86, Theorem 1.13] for instance. This bijection is inclusion reversing and hence establishes a duality between the combinatorics of $V(F_h)$ and the combinatorics of $\Sigma(h)$. The bounded subcomplex of $V(F_h)$ is called *tight span* of $\Sigma(h)$, cf. [72]. The tight span showcases the adjacency relations of interior cells of $\Sigma(h)$ as will be indicated in Figures 3.29 and 3.30. The 1-cells, i.e. the edges, of the tight span yield the dual graph $\Gamma(h)$ of $\Sigma(h)$ with natural edge weights given by the tropical edge lengths as in (3.26).

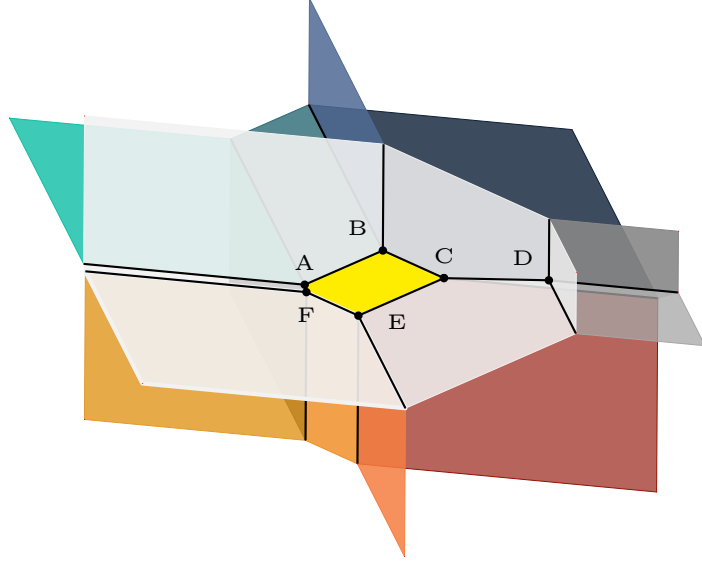


Figure 3.29: The tropical hypersurface associated to a height function on $\mathbf{A} = S = \{0, 1\}^3$, which was obtained in a population genetics experiment on *Escherichia coli* mutations in the genes *topA*, *spoT*, and *pykF* running Lenski's long-term evolution experiment, cf. [91] and Section 3.7. The bounded subcomplex of the hypersurface, i.e. the tight span of the regular subdivision $\Sigma(h)$ of \mathbf{A} introduced in [72] generalizing the tight spans of finite metric spaces from [44] and [77], has six vertices enumerated A to F , six bounded edges $(A, B), (B, C), (C, D), (C, E), (E, F)$ and (A, F) and one bounded 2-dimensional face colored yellow.

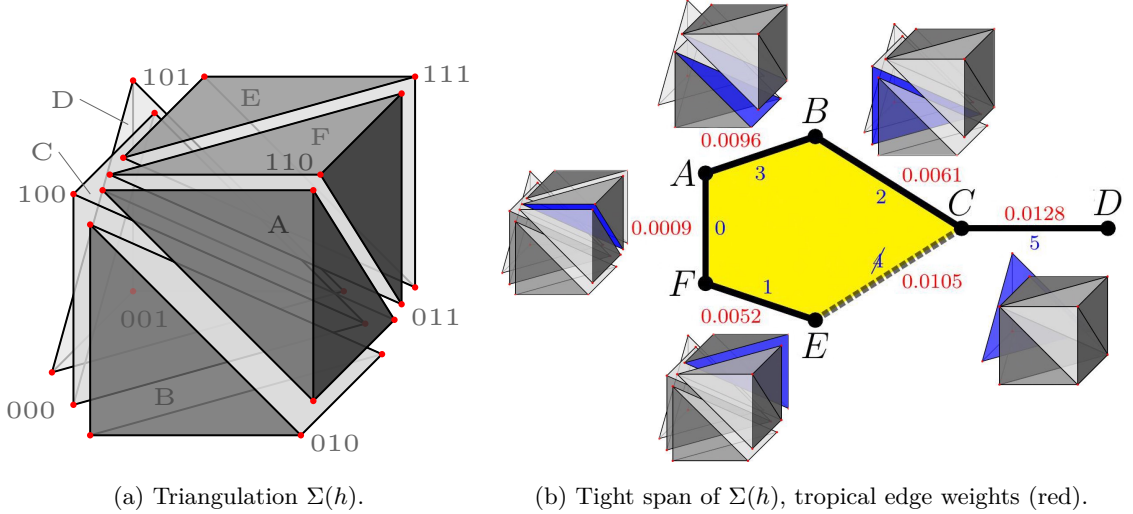


Figure 3.30: Running example. (A) The regular triangulation $\Sigma(h)$ dual to Figure 3.29. (B) The tight span of $\Sigma(h)$. Its 1-skeleton, i.e. with the yellow 2-cell removed, is the dual graph $\Gamma(h)$ of $\Sigma(h)$ and records the adjacency relations among the maximal simplices of $\Sigma(h)$. Tropical edge weights of the dual edges are in red, their order with respect to how the greedy algorithm chooses the edges are in blue. Note that the edge (C, E) with label 4 is omitted. Geometrically, the glueing information of this edge is already recorded in the partition $ABCEF|D$ of the cube in (A), cf. Section 3.7.

3.3 The MST-Fan $\mathcal{K}(h)$

Let again \mathbf{A} be a fixed finite point configuration in \mathbb{R}^n . In order to study the class of realizable minimum spanning trees of $\Gamma(h)$ for varying height functions $h: \mathbb{R}^{\mathbf{A}} \rightarrow \mathbb{R}$, we first consider the tropical edge length as a linear functional modulo sign.

Definition 3.3.1. *The edge length linear form $\ell_{\bullet}(r): \mathbb{R}^{\mathbf{A}} \rightarrow \mathbb{R}$ of an \mathbf{A} -ridge $r = (s, t)$ is given by the determinant $\det(\tilde{L}_{\bullet}(r))$ of the matrix $\tilde{L}_{\bullet}(r)$ from equation (3.26) in Section 3.1. Thus, evaluating $\ell_{\bullet}(r)$ at the height function $h \in \mathbb{R}^{\mathbf{A}}$ yields the tropical edge length $L_h(r)$ up to sign.*

3.3.1 Linear forms and circuits

A matroid $\mathcal{M} = (E, \mathcal{B})$ on a finite ground set E is given by a non-empty collection $\mathcal{B} \subset 2^{\mathcal{M}}$ of subsets of E , called the *set of bases* of \mathcal{M} , which fulfills the following *base exchange axiom* (BE), cf. [116, Section 1.2]:

(BE) For any two members B_1 and B_2 of \mathcal{B} and any element $x \in B_1 \setminus B_2$ there is an element $y \in B_2 \setminus B_1$ such that $(B_1 \setminus \{x\}) \cup \{y\}$ is a member of \mathcal{B} .

A collection \mathcal{B} of bases gives rise to a system $\mathcal{I} \subset 2^{\mathcal{M}}$ of *independent sets* of \mathcal{M} collecting all subsets of $I \subset E$ lying in some base $B_I \in \mathcal{B}$ and, vice versa, bases are independent sets of maximal cardinality. For our purposes, we consider two matroids. First, given an undirected graph $G = (V, E)$, the *cycle matroid* $\mathcal{M}(G)$ on the ground set E has the spanning trees of G as its bases. Second, the *vector matroid* $\mathcal{M}(\mathbf{A})$ on the ground set \mathbf{A} depicts affine independences over \mathbf{A} , i.e. the bases of $\mathcal{M}(\mathbf{A})$ are full-dimensional simplices spanned by points of \mathbf{A} .

For a given \mathbf{A} -ridge r , the coefficients $(x_i)_{i \in \mathbf{A}}$ of $\ell_{\bullet}(r)$ regarded as a row vector in $\mathbb{R}^{\mathbf{A}}$ are obtained by Laplacian expansion of $\tilde{L}_{\bullet}(r)$ along its i -th row and last column. The support $\text{supp}(\tilde{L}_{\bullet}(r)) := \{i \in \mathbf{A} : x_i \neq 0\}$ is seen by the matroid $\mathcal{M}(\mathbf{A})$ as we will show now.

Lemma 3.3.2. [116, Corollary 1.2.6] *Every \mathbf{A} -ridge $r = (s, t)$ contains a unique circuit Z^r .*

The circuit Z^r of Lemma 3.3.2 is called the *fundamental circuit* of r . Now, fix an \mathbf{A} -ridge $r = (s, t)$ and denote the coefficients of the linear form $\ell_{\bullet}(r)$ by $(d_v)_{v \in \mathbf{A}}$.

Proposition 3.3.3. *The circuit Z^r equals the support $\{v \in \mathbf{A} : d_v \neq 0\}$ of $\ell_{\bullet}(r)$.*

Proof. If $d_v = 0$ holds, then $\ker(\ell_{\bullet}(r))$ yields relations in $Z^r \subset (s \cup t) \setminus \{v\}$. If $d_v \neq 0$ holds, then $Z^r \not\subset ((s \cup t) \setminus \{w\}) \in \mathcal{I}$, and thus we get $v \in Z^r$. \square

Remark 3.3.4. *Consequently, one may have $\ell_h(r) = \ell_h(\tilde{r})$ for distinct ridges r and \tilde{r} and generic h . In this case the fundamental circuits of r and \tilde{r} need to coincide.*

Proposition 3.3.5. *Every circuit Z of $\mathcal{M}(\mathbf{A})$ equals Z^r for some \mathbf{A} -ridge (s, t) .*

Proof. Let $Z = \{z_1, \dots, z_r\} \subset \mathbf{A}$. Then $Z \setminus \{z_1\} \subset t$ for some $t \in \mathcal{B}$. Now, choose $z_t \in Z \setminus \{z_1\}$ and set $s := t \setminus \{z_t\} \cup \{z_1\}$. Assume $s \notin \mathcal{I}$. Then there is some circuit $Z^* \subset s \cup t$ with $z_t \notin Z^*$ contradicting Lemma 3.3.2. \square

In fact, the fundamental circuit Z^r of the \mathbf{A} -ridge r in Proposition 3.3.3 is subdivided into a positive and a negative part, $Z^r = (Z_+^r, Z_-^r)$, according to the sign of the coefficients d_v . This turns $\mathcal{M}(\mathbf{A})$ into an *oriented matroid*, cf. [17]. Moreover, the regular triangulations of \mathbf{A} are encoded in the *flip graph*, which is the 1-skeleton of the *secondary polytope* $\Sigma\text{-poly}(\mathbf{A})$ of \mathbf{A} , cf. [40, Section 5.3]. An edge of the flip graph connecting vertices of $\Sigma\text{-poly}(\mathbf{A})$, i.e. triangulations whose secondary cones $\Sigma(h)$ and $\Sigma(g)$ share a common facet, corresponds to an oriented fundamental circuit $Z^r = (Z_+^r, Z_-^r)$ for some ridge $r = (s, t) \in \Gamma(h)$ and represents a split of the bipyramid (s, t) of \mathbf{A} in two different ways.

3.3.2 Reproducing spanning trees of $\Gamma(h)$

In order to reproduce an edge-ordered spanning tree T of $\Gamma(h)$ as output $T_{\min}(g) = T$ of the greedy algorithm on the weighted dual graph $\Gamma(h)$, one needs to impose conditions on the height function g such that the algorithm is forced to choose the right edges in the correct order.

Lemma 3.3.6. *Let g and h be distinct linear forms on \mathbb{R}^m . Then*

$$F^{(g,h)} := \{x \in \mathbb{R}^m : |g(x)| \leq |h(x)|\}$$

is a polyhedral fan consisting of four maximal cones.

Proof. The choices for $\sigma_g, \sigma_h \in \{+, -\}$ yield a subdivision of $F^{(g,h)}$ into the four cones

$$F_{\sigma_g \sigma_h}^{(g,h)} := \{\sigma_g g \geq 0\} \cap \{\sigma_h h \geq 0\} \cap \{(\sigma_h h - \sigma_g g) \geq 0\} .$$

□

Definition 3.3.7. *An \mathbf{A} -ridge graph R is a finite set of \mathbf{A} -ridges, such that there exists a regular subdivision $\Sigma(h)$ of \mathbf{A} which shows R among its adjacencies, i.e. $R \subset \Gamma(h)$ holds. An \mathbf{A} -ridge forest is a cycle-free \mathbf{A} -ridge graph and an \mathbf{A} -ridge tree is a connected \mathbf{A} -forest.*

Lemma 3.3.8. *For an \mathbf{A} -ridge graph R , the set of height functions $h \in \mathbb{R}^{\mathbf{A}}$ such that R occurs in $\Gamma(h)$ is given by the interior of a full-dimensional cone.*

Proof. Following 3.2.1, we see $R \subset \Gamma(h)$ precisely if h lies in the interior of $\bigcap_{s \in V(R)} H_s$. □

The parameter cone of Lemma 3.3.8 is the union of those secondary cones whose triangulations realize the adjacencies of R and we define it itself as the *secondary cone* $\text{sc}(R) := \bigcap_{s \in V(R)} H_s$ of R . For instance, if R is a spanning tree of $\Gamma(h)$, then $\text{sc}(R) = \text{sc}(h)$ holds since all maximal cells of $\Sigma(h)$ are required to appear. Yet, the graph R can be chosen significantly smaller in order to fulfill $\text{sc}(R) = \text{sc}(h)$, for instance in Figure 3.30 the ridge graphs $\{2, 3\}$, $\{1, 5\}$, $\{1, 3\}$ and $\{2, 4, 5\}$ are inclusion minimal with this property.

Proposition 3.3.9. *Given an \mathbf{A} -ridge $r \in \Gamma(h)$, the map $\text{sign}(\ell_{\bullet}(r))$ is constant and non-zero on the interior $\text{int}(\text{sc}(R))$ of $\text{sc}(R)$ for any \mathbf{A} -ridge graph R containing r .*

Proof. If there are $h, h' \in \mathbb{R}^{\mathbf{A}}$ with $\text{sign}(\ell_h(r)) < 0 < \text{sign}(\ell_{h'}(r))$, then there is some $g \in \text{conv}\{h, h'\} \subset \text{sc}(R)$ with $\ell_g(r) = 0$ contradicting Lemma 3.3.8. □

For an \mathbf{A} -ridge $r \in R$, set $\sigma_r \in \{-1, 1\}$ to be $\text{sign}(\ell_{\bullet}(r))$ on $\text{int}(\text{sc}(R))$.

Proposition 3.3.10. *For \mathbf{A} -ridges $r_1, r_2 \in R$ of an \mathbf{A} -ridge graph R , we have*

$$\# \left\{ c \in F^{(l(r_1), l(r_2))} : \dim(c \cap \text{sc}(R)) = \#\mathbf{A} \right\} \leq 1 .$$

If this set has a unique full-dimensional cone, we denote it by $F^{r_1, r_2}(R)$.

Proof. By Proposition 3.3.9, the only candidate for c is $F_{\sigma_{r_1} \sigma_{r_2}}^{(l(r_1), l(r_2))}$. It appears in full dimension precisely if $\{\sigma_{r_2} l(r_2) - \sigma_{r_1} l(r_1) \geq 0\} \cap \text{sc}(R)$ is full-dimensional. □

Remark 3.3.11. *The proof of Proposition 3.3.10 shows that the hyperplane $\sigma_{r_2} l(r_2) = \sigma_{r_1} l(r_1)$, provided that it touches the interior of $\text{sc}(R)$, subdivides $\text{sc}(R)$ into the two subcones $\ell_{\bullet}(r_1) \leq \ell_{\bullet}(r_2)$ and $\ell_{\bullet}(r_1) \geq \ell_{\bullet}(r_2)$.*

With regard to Proposition 3.3.10, we define an *ordered \mathbf{A} -ridge graph* $R_{<}$ to be an \mathbf{A} -ridge graph R equipped with an enumeration of its edges by an ascending sequence of subsequent integers. Further, we call an ordered spanning tree $T_{<}$ of the dual graph $\Gamma(h)$ *realizable* if there exists a height function $g \in \text{sc}(h)$, which satisfies $T_{<} = T_{\min}(g)$. Note that for a fixed spanning tree, one order may be realizable while the other is not as we will see in Figure 3.31. A ridge $r = (s, t)$ in the dual graph $\Gamma(h)$ is called *stable* if no other ridge of $\Gamma(h)$ generically, i.e. for a generically chosen $h \in \mathbb{R}^{\mathbf{A}}$, has the same tropical edge length as r . Otherwise, the ridge r is called *unstable at some circuit Z* , or *Z -unstable*, and its fundamental circuit Z needs to occur within some other ridge in this case, cf. Lemma 3.3.2 and Remark 3.3.4. For instance, if Z -unstable ridges do not receive subsequent integers in the edge order, the tree is not realizable. Hence any set of Z -unstable ridges has an interval of integers as *instability range*.

Definition 3.3.12. For a generic height function $h \in \mathbb{R}^{\mathbf{A}}$ and a realizable ordered spanning tree $T_{<}$ of $\Gamma(h)$, let the MST-cone $\mathcal{K}(T_{<})$ of $T_{<}$ be given by

$$\mathcal{K}(T_{<}) := \{g \in \text{sc}(h) : g \text{ is generic and } T_{\min}(g) =_{\diamond} T_{<}\} \subset \mathbb{R}^{\mathbf{A}}$$

as the set of all height functions g on \mathbf{A} such that the greedy algorithm on $\Gamma(g)$ possibly chooses $T_{<}$ as minimum spanning tree. This is reflected in the notation $T_{\min}(g) =_{\diamond} T_{<}$ as adaption from modal logic.

The \mathcal{K} in the notation of the MST-cone goes back to Kruskal, who described in [96] the greedy algorithm in the graph context, today known as Kruskal's algorithm, with regard to the travelling salesperson problem (TSP). In fact, minimum spanning trees have been used ever since in heuristics and approximations for the TSP, see e.g. [66, 67]. In the current paper, we only consider edge weights which are dictated by geometric relations in tropical hypersurfaces. As such, the framework as described in Definition 3.3.1 is linear.

In the sequel we will show that $\mathcal{K}(T_{<})$ is a full-dimensional polyhedral subcone of the secondary cone $\text{sc}(h)$ of $\Sigma(h)$. By Proposition 3.3.10, every edge of $T_{<}$ contributes at most one linear hyperplane to $\mathcal{K}(T_{<})$ and the MST-cones give rise to a fan structure.

3.4 An algorithm for computing $\mathcal{K}(T_{<})$

Let us fix a realizable ordered spanning tree $T_{<} := (r_1, \dots, r_m) \subset \Gamma(h)$ for some generic height function $h \in \mathbb{R}^{\mathbf{A}}$. For $v = 0, \dots, m$ we define $T_v := (r_1, \dots, r_v)$ to be the tree with the same node set as T but only with edges r_1, \dots, r_v inserted. Further, we set

$$c(T_v) := \{e \in E(\Gamma(h)) \setminus E(T_v) : T_v \cup e \text{ contains a circuit}\}$$

and we define the r_v -relevant edges $H(r_v)$ to be

$$H(r_v) := E(\Gamma(h)) - E(T_v) - c(T_v) .$$

Thus, the edges in $H(r_v)$ are precisely the edges in $\Gamma(h)$ which do not close a circuit in T_v . We say that an edge $r \in \Gamma(h)$ is *relevant* for r_v if r lies in $H(r_v)$. Further, for every edge $r \in E(\Gamma(h)) \setminus T$ there is a unique minimal *relevance index* $i(r) < m$ such that r is relevant for $T_{i(r)-1}$ but irrelevant for $T_{i(r)}$. In this case we call $\underline{r} = r_{i(r)}$ the *cut edge* of r .

Example (Example of Figure 3.30 continued.). In Figure 3.30, edge 4 is relevant for the edges 0, 1 and 2. Once edge 3 is added, edge 4 closes a circuit in the tree $\{0, 1, 2, 3\}$. Thus $i(4) = 4$ and edge 3 is the cut edge for edge 4.

In order to cut the MST-cone $\mathcal{K}(T_{<})$ out of $\text{sc}(h)$ along Proposition 3.3.10, we need to consider two kinds of conditions. First, the $(m-1)$ *in-tree conditions* are given by

$$\ell_{\bullet}(r_1) \leq \ell_{\bullet}(r_2) \leq \ell_{\bullet}(r_3) \leq \dots \leq \ell_{\bullet}(r_{m-1}) \leq \ell_{\bullet}(r_m) .$$

Second, for any $r \in E(\Gamma(h)) \setminus T$, we make sure that the greedy algorithm never choses r by requiring the *cut-edge condition* $\ell_{\bullet}(\underline{r}) \leq \ell_{\bullet}(r)$. As a consequence, the number of facets of $\mathcal{K}(T)$ is always lower than $\#E(\Gamma)$ and in practice, the actual number of facets is almost always a lot lower since for instance by Remark 3.3.4 some of the in-tree conditions might be redundant due to multiple

occurrences of instability ranges.

Algorithm 1: Compute the MST-cone $\mathcal{K}(T_{<})$

Input:

- (1) a generic height function $h \in \mathbb{R}^{\mathbf{A}}$
- (2) the secondary cone $\text{sc}(h)$ of h
- (3) a realizable ordered spanning tree $T_{<} = (r_1, \dots, r_m) \subset \Gamma(h)$

Output: the MST-cone $\mathcal{K}(T_{<}) = \{g \in \mathbb{R}^{\mathbf{A}} : T_{\min}(g) =_{\diamond} T_{<}\}$

MST-cone

```

1   $C \leftarrow \mathbb{R}^{\mathbf{A}}$ 
2  for  $k = 1, \dots, m-1$  do
3     $C \leftarrow C \cap F^{r_k, r_{k+1}}(T)$                                 // Add  $(m-1)$  in-tree conditions
4  for  $r \in E(\Gamma(h)) \setminus E(T)$  do
5     $C \leftarrow C \cap F^{\perp, r}(T)$                                     // Add cut-edge conditions
6  return  $C \cap \text{sc}(h)$ 

```

Proposition 3.4.1. *Algorithm 1 outputs $\mathcal{K}(T_{<})$, which is indeed a full-dimensional cone.*

Proof. The set of in-tree and cut-edge conditions is necessary and sufficient for T to possibly appear as output of the greedy algorithm, and defines a cone by Proposition 3.3.10. \square

Let m_* be the number $\#E(\Gamma(h))$ of edges of $\Gamma(h)$ and let v_* be the number $\#V(\Gamma(h))$ of its vertices. In the discussion from Section 3.2.1 and in the proof of Lemma 3.3.6, we see that the linear forms $l_{\bullet}(r_1), \dots, l_{\bullet}(r_m)$ are byproducts of the secondary cone computation. Concerning the relevance indices needed in the second for-loop of Algorithm 1, every edge which is not in the spanning tree needs at most m cycle checks on subgraphs of $\Gamma(h)$, which takes $\mathcal{O}(m_* + v_*)$ each using depth-first search. Thus, the complexity of Algorithm 1 adds up to $\mathcal{O}(m_*^2 + m_* v_*)$. Concerning the quantities m_* and v_* , for $\mathbf{A} = \{0, 1\}^n$ the n -dimensional unit cube the estimate $2^n - n \leq v_* \leq n!$ follows from [40, Theorem 2.6.1], where $2^n - n = v_*$ holds precisely if $\Gamma(h)$ is a tree, and we have the bound $m_* \leq \frac{1}{2}(n+1)! - n(2^{n-1} - n + 1)$. In a variant of Algorithm 1 where any ordered input tree $T_{<} \subset \Gamma(h)$ is allowed, in order to verify its realizability one additionally needs to ensure that the cone $C \cap \text{sc}(h)$ in line 6 is full-dimensional.

We remark that if $T_{<}$ is realizable and has only stable edges as in Figure 3.30, then the interior $\mathcal{K}(T_{<})$ describes the locus where the greedy algorithm has only one possibility for choosing a minimum spanning tree of $\Gamma(h)$, i.e. $\text{int}(\mathcal{K}(T_{<})) = \{g \in \text{sc}(h) : T_{\min}(g) = T_{<}\}$. If $T_{<}$ has instable edges, the greedy algorithm on $g \in \mathcal{K}(T_{<})$ may possibly choose minimum spanning trees whose edge sets differ from $E(T_{<})$. However, the tree $T_{<}$ is among the candidates for the optimum and the other outcomes can be controlled.

Proposition 3.4.2. *Let $T_{<} =_{\diamond} T_{\min}(h)$ and $T'_{<} =_{\diamond} T_{\min}(h)$ be two valid outputs of the greedy algorithm for the same height function h . Then each stable edge of $T_{<}$ also appears in $T'_{<}$.*

Proof. Assume that the stable edge $r = \{s, t\} \in E(T_{<})$ does not appear in $T'_{<}$, i.e. it closes a cycle when considered to be added to the forest $\tilde{T}'_{<} \subset T'_{<}$. Then every s - t path in $\tilde{T}'_{<}$ uses instable edges r_I of $T'_{<}$ which are not in $T_{<}$. These edges themselves close cycles in some subtree of $\tilde{T}'_{<}$ of $T_{<}$. Replacing each of the r_I by a path in $\tilde{T}'_{<}$ implies that the nodes s and t belong to the same component of $\tilde{T}'_{<}$ before r is added. Contradiction. \square

We now define the *MST-fan* $\mathcal{K}(h)$ of h to be the collection of all *MST-cones* $\mathcal{K}(T_{<})$ where $T_{<}$ runs through the realizable ordered spanning trees of $\Gamma(h)$.

Theorem 3.4.3. *Let $\Sigma(h)$ be a regular triangulation of the point configuration \mathbf{A} . The MST-fan $\mathcal{K}(h)$ is a pure fan in $\mathbb{R}^{\mathbf{A}}$ with full support $\text{supp}(\mathcal{K}(h)) := \bigcup_{c \in \mathcal{K}(h)} c = \text{sc}(h)$. Two height functions $f, g \in \text{sc}(h)$ are in the same cone of $\mathcal{K}(h)$ if and only if the greedy algorithm possibly produces the same spanning tree on $\Gamma(f)$ and $\Gamma(g)$, i.e. precisely if there is an ordered spanning tree $T_{<}$ of $\Gamma(h)$ with $T_{\min}(f) =_{\diamond} T_{<} =_{\diamond} T_{\min}(g)$.*

Proof. If $\mathcal{K}(T_<)$ and $\mathcal{K}(T'_<)$ meet, then by Remark 3.3.11 their intersection is induced by some commonly active edge inequalities and therefore it is a face of both $\mathcal{K}(T)$ and $\mathcal{K}(T')$. The bijectivity is a consequence of Proposition 3.3.10. \square

For further study, we might also consider subfans of the form $\mathcal{K}(T) = \cup \mathcal{K}(T_<)$ collecting all realizable orders on a fixed tree T .

Example (Example of Figure 3.30 continued). *The spanning tree we previously considered reads $T_< = T_{\min}(h) = (0, 1, 2, 3, 5)$. Its MST-cone $\mathcal{K}(T_<)$ has four facets given by the four in-tree conditions $0 < 1 < 2 < 3 < 5$. The facets $(0 < 1)$, $(3 < 5)$ and $(2 < 3)$ meet the interior of $sc(h)$, whereas the facet $(1 < 2)$ lies in the boundary of $sc(h)$, i.e. no tree with the reversed condition $2 < 1$ is realizable. In Figure 3.31 the dotted lines describe the intersection of $\mathcal{K}(\{0, 1, 2, 3, 5\})$ and $\mathcal{K}(\{0, 1, 3, 4, 5\})$, which lies in the hyperplane $\ell_\bullet(2) = \ell_\bullet(4)$. The supports of the fans $\mathcal{K}(\{0, 1, 2, 3, 5\})$ and $\mathcal{K}(\{0, 1, 3, 4, 5\})$ are two adjacent full-dimensional cones.*

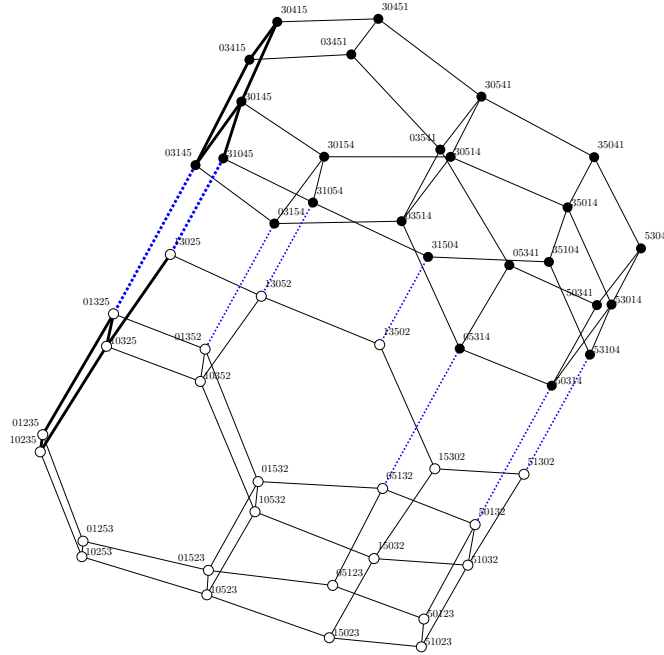


Figure 3.31: The intersection of $\mathcal{K}(\{0, 1, 2, 3, 5\})$ and $\mathcal{K}(\{0, 1, 3, 4, 5\})$, cf. Figure 3.30. The dual graph $\Gamma(h)$ has five spanning trees, each necessarily featuring edge 5 and leaving out one of the other five edges. As a computation shows, the only realizable ordered spanning trees arise as orders on the trees $\{0, 1, 2, 3, 5\}$ (white nodes) and $\{0, 1, 3, 4, 5\}$ (black nodes). Hence, from $5 \cdot 5! = 600$ possible ordered spanning trees of $\Gamma(h)$, only $\frac{1}{12} \cdot 600 = 50$ are realizable. The thick subgraph on the upper left corresponds to the subfan of $\mathcal{K}(h)$ leaving edge 5 fixed and the hyperplane $\ell_\bullet(2) = \ell_\bullet(4)$ is indicated by the dotted blue line segments. All computations were done using `polymake` [56].

3.5 Permutation groups acting on tree orders

Fix a realizable ordered spanning tree $T_< := (r_1, \dots, r_m) := T_{\min}(h) \subset \Gamma(h)$. As explained in Section 3.3, every unstable edge r_k of $T_<$ has an interval $I_k := [i, i+j]$ with $i \leq k \leq i+j$, positive length j and integer endpoints as its instability range, i.e. where the linear forms $\ell_\bullet(r_i) = \dots = \ell_\bullet(r_{i+j})$ coincide, cf. Remark 3.3.4. We may define the instability ranges of stable edges to be the singleton sets depicting their index in the edge order. This convention yields a partition of $\{1, \dots, m\}$ into an ordered collection $I(T_<)$ of $w \leq m$ instability ranges. Note that in practice, occurrences of instability ranges of positive length are frequently encountered in dimension $n \geq 4$, even several of them in the same tree.

Now, the permutation group $G := \mathcal{P}(w)$ naturally acts on $I(T_{<})$. For a fixed permutation $\sigma \in G$, let $\mathbf{m}(\sigma)$ be the minimum number of adjacent transpositions that the permutation σ can be decomposed into.

Proposition 3.5.1. *Let $\sigma \in G$ and let the partitions $I(T_{<})$ and $\sigma I(T_{<})$ of $\{1, \dots, m\}$ both be realizable, the latter by the tree $\sigma T_{<}$. If the MST-cones $\mathcal{K}(T_{<})$ and $\mathcal{K}(\sigma T_{<})$ intersect in codimension k , then $\mathbf{m}(\sigma) < k + 1$ holds. In particular, if the MST-cones $\mathcal{K}(T_{<})$ and $\mathcal{K}(\sigma T_{<})$ share a common facet, then σ is an adjacent transposition.*

Proof. Each adjacent transposition reverts one non-empty in-tree condition. \square

Remark 3.5.2. *Let τ be an adjacent transposition. Then $\dim(\mathcal{K}(T_{<}) \cap \mathcal{K}(\tau T_{<}))$ may be of codimension less than one, since there might be cut-edge conditions in the first place which change their direction when reverting an in-tree condition.*

3.6 Matroids and Bergman fans

3.6.1 Bergman fans

We follow [2] in our presentation. Let $\mathcal{M} = (E, \mathcal{B})$ be a matroid. Any $w \in \mathbb{R}^E$ can be seen as a *valuation* of bases by setting $w(B) = \sum_{b \in B} w(b)$ for any base $B \in \mathcal{B}$. The *initial matroid* \mathcal{M}_w is the matroid on the same ground set E with bases given precisely by those $B \in \mathcal{B}$, whose evaluation $w(B)$ is minimal among all evaluations of bases in \mathcal{B} . The *Bergman fan* $\mathfrak{B}(\mathcal{M})$ is the set of all $w \in \mathbb{R}^E$ such that the initial matroid \mathcal{M}_w has no loops. For instance, if we fix some undirected graph G , then $\mathfrak{B}(\mathcal{M}(G))$ is the set of all edge weights $w \in \mathbb{R}^{E(G)}$ such that any edge of G lies in some w -minimum spanning tree of G . On $\mathfrak{B}(\mathcal{M})$ there are two established fan structures. First, the *coarse subdivision* considers $v, w \in \mathbb{R}^E$ to be equivalent if their initial matroids \mathcal{M}_v and \mathcal{M}_w coincide. It is the coarsest possible fan structure on $\mathfrak{B}(\mathcal{M})$. Second, any valuation $w \in \mathbb{R}^E$ gives rise to a *flag* $\mathcal{F}(w) := \{\emptyset =: F_0 \subset F_1 \subset \dots \subset F_{k+1} := E\}$ of subset of E , such that w is constant on each $F_i \setminus F_{i-1}$ and is strictly increasing with i , i.e. $w(x) < w(y)$ for any $x \in F_i$ and $y \in F_{i+1} \setminus F_i$. The *fine subdivision* of $\mathfrak{B}(\mathcal{M})$ considers valuations $v, w \in \mathbb{R}^E$ to be equivalent if their flags $\mathcal{F}(v)$ and $\mathcal{F}(w)$ coincide. The mere set $\mathfrak{B}(\mathcal{M}(G)) \subset \mathbb{R}^E$ is also called the *tropical linear space* $\text{trop}(\mathcal{M})$ of the matroid \mathcal{M} .

An important attribute of matroid theory is that there are several ways, called *cryptomorphisms*, to define a matroid, all being equivalent but not in an obvious way. For instance, one may use *rank functions* to do so. Given a matroid $\mathcal{M} = (E, \mathfrak{I})$ with independence system \mathfrak{I} , cf. Section 3.3.1, the rank $\rho(A)$ for any subset $A \subset E$ is defined to be the maximal cardinality of an independent set lying inside A . A *flat* of \mathcal{M} is a subset of $F \subset E$ such that $\rho(F \cup \{x\}) > \rho(F)$ holds for any $x \notin F$. The set of flats of \mathcal{M} forms a geometric lattice which fully describes the combinatorics of \mathcal{M} , i.e. giving the lattice of flats is another way to cryptomorphically define matroids.

Proposition 3.6.1. [2, Theorem 1] *The fine subdivision realizes the lattice of flags of \mathcal{M} .*

In other words, a flag \mathcal{F} of subsets of E can be written as $\mathcal{F}(w)$ for some $w \in \mathbb{R}^E$ precisely if every member of \mathcal{F} is a flat of \mathcal{M} . Hence, the faces of $\mathfrak{B}(\mathcal{M})$ are given by $\text{pos}(e_{F_0}, \dots, e_{F_{k+1}})$ where $F_0 \subset \dots \subset F_{k+1}$ is a chain of flats of \mathcal{M} .

3.6.2 The MST-fan as part of a Bergman fan

Let again $\mathbf{A} \subset \mathbb{R}^n$ be a finite point configuration and let $h: \mathbf{A} \rightarrow \mathbb{R}$ be a height function on \mathbf{A} . We consider the dual graph $G := \Gamma(h) \subset \Sigma(h)$ and the cycle matroid $\mathcal{M}(G)$. Consider the map

$$\begin{aligned} l: \mathbb{R}^{\mathbf{A}} &\longrightarrow \mathbb{R}^{E(G)} \\ h &\longmapsto (\ell_h(r))_{r \in E(G)} . \end{aligned}$$

Since the graph G is naturally weighted by the tropical edge lengths, one may consider the restriction $\hat{\mathfrak{B}}(\mathcal{M}(G)) := \{w \in l(\mathbb{R}^{\mathbf{A}}): \mathcal{M}(G)_w \text{ has no loops}\}$, where we only allow tropical edge lengths as valuations.

A collection $\mathcal{C} = \{\mathcal{K}(T^{(C_i)})\}$ of maximal cones in $\mathcal{K}(h)$ is called *G-saturating* if the ordered spanning trees $T^{(C_i)}$ cover G , i.e. if $E = \cup_i E(T^{(C_i)})$ holds. Let $C \subset \mathbb{R}^{\mathbf{A}}$ be the polyhedral complex with cells $\bigcap \mathcal{K}(T^{(C_i)})$ where \mathcal{C} runs through the *G-saturating* collections.

Corollary 3.6.2. *The map l embeds the complex C into $\mathfrak{B}(\mathcal{M})$ with image $\hat{\mathfrak{B}}(\mathcal{M}(G))$. The induced fine subdivision on $\hat{\mathfrak{B}}(\mathcal{M}(G)) \subset \mathfrak{B}(\mathcal{M}(G))$ comes from the cones $\mathcal{K}(T_{<})$, whereas the coarse subdivision is perceived by dropping the order and considering $\mathcal{K}(T) = \cup \mathcal{K}(T_{<})$.*

Example. *In Figure 3.31, the coarse subdivision induced on $\hat{\mathfrak{B}}(\mathcal{M}(G))$ corresponds to the dotted blue part given by the hyperplane $\ell_{\bullet}(2) = \ell_{\bullet}(4)$ and is therefore trivial. There are eight distinct dotted blue line segments and they constitute the fine subdivision.*

3.7 Application in population genetics: epistasis and spanning trees

3.7.1 Regular subdivisions and epistatic weights

Epistatic filtrations are introduced in [46] and [47] to track significant epistatic interactions in a fitness landscape (\mathbf{A}, h) . There, the function $h: \mathbf{A} \rightarrow \mathbb{R}$ is an experimentally obtained genotype-phenotype map on a biallelic (or multiallelic) n -loci *genotype set* $\mathbf{A} = \{0, 1\}^n$, resp. $\mathbf{A} = \text{vert}(\prod_i \Delta_i)$ in the multiallelic case. As in the biology literature can be found [53], an affinely independent genotype subsystem $r = \{v_i: i \in I\} \subset \mathbf{A}$ is called an *epistatic interaction* if the lifted genotype set $r^h = \{(v_i, h(v_i)): i \in I\} \subset \mathbf{A}^h \subset \mathbb{R}^{n+1}$ is affinely independent as well.

A special kind of epistatic interactions can be indicated by \mathbf{A} -bipyramids, or \mathbf{A} -ridges, i.e. pairs (s, t) of adjacent n -simplices $s = \text{conv}\{v_1, \dots, v_{n+1}\}$ and $t = \text{conv}\{v_2, \dots, v_{n+2}\}$ with vertices $v_i \in \mathbf{A}$, sharing a common facet. The *epistatic weight* $e_h(s, t)$ of the bipyramid $r = (s, t)$ measures the degree of non-additivity of h on r and is defined in [47] as

$$e_h(r) := e_h(s, t) := L_h(s, t) \cdot \lambda(s, t)$$

with $\lambda(s, t) := \text{nvol}(A \cap B) / (\text{nvol}(A) \cdot \text{nvol}(B))$, which doesn't depend on h . Thus, the bipyramid r is an epistatic interaction precisely if $e_h(r) > 0$.

Bipyramids occur in a regular triangulation $\Sigma(h)$ as adjacent maximal cells. The transfer from the regular subdivision $\Sigma(h)$ to an according biological statement is now made via the following linear optimization problem $\text{LP}(h, w)$, introduced in [9]:

$$\begin{aligned} & \text{maximize} && h \cdot p \\ & \text{subject to} && p \in \Delta_{2^n} \text{ and } \rho(p) := \sum_{v \in \mathbf{A}} p(v)v = w, \end{aligned} \quad (\text{LP}(h, w))$$

for some given *allele frequency* $w \in [0, 1]^n$ and with decision variable p ranging over all *relative populations* $p \in \Delta_{2^n} = \{p \in \mathbb{R}^{\mathbf{A}}: \sum p_i = 1\}$ on $\mathbf{A} = \{0, 1\}^n$, presented here in the biallelic case only. Since h is experimentally obtained and hence generic, there is a unique optimal solution $p^*(h, w) \in \rho^{-1}(w) \subset \Delta_{2^n}$ of $(\text{LP}(h, w))$. The map

$$\begin{aligned} h^*: [0, 1]^n &\rightarrow \mathbb{R} \\ w &\mapsto h \cdot p^*(h, w) \end{aligned}$$

equals the convex support function $f_{\Sigma(h)}$ from Remark 3.2.1 and therefore its linear regions coincide with the maximal simplices of $\Sigma(h)$. Further, given s and t as above as cells in $\Sigma(h)$, the quantity $e_h(s, t)$ measures the degree to which the exposed vertex v_1 resp. v_{n+2} has a lower than expected phenotype assuming that h extends affinely from t resp. s to the bipyramid (s, t) . In this sense $e_h(s, t)$ measures *negative epistasis*, cf. [121]. The polyhedral subdivision $\{\rho^{-1}(A): A \text{ maximal simplex of } \Sigma(h)\}$ of Δ_{2^n} now stratifies Δ_{2^n} into *fittest populations* with respect to varying allele frequencies.

3.7.2 Epistatic filtrations

The *epistatic filtration* of h serves as a tool to separate data noise from significant epistatic signal. It operates on the ascendingly ordered e_h -weighted dual graph $\Gamma(h)$. Here, we explicitly interpret the nodes of $\Gamma(h)$ as simplices in \mathbb{R}^n .

Geometrically, the set \sum_0 of maximal cells of $\Sigma(h)$ yields a decomposition of $[0, 1]^n$ into $\#V(\Gamma(h))$ simplicial cells and marks the zero signal starting point of the filtration process. Assume \sum_{k-1} is given as a cellular decomposition of the genotype $[0, 1]^n$ and the k -th edge reads $r_k = (s_k, t_k)$. The decomposition \sum_k is now constructed by unifying the cells c_{ks} and c_{kt} of \sum_{k-1}

which contain s_k resp. t_k . If one has $c_{ks} \neq c_{kt}$, the edge r_k is called *critical* since the epistatic information can be considered as irredundant in this case. The *epistatic filtration* of h is the collection of all critical edges. In the biological context, late cell agglutinations reflect strong epistatic effects and, since dual graphs are connected, the critical edge with the highest epistatic weight always glues a two-cell split, cf. Figure 3.30(B).

Lemma 3.7.1. *Epistatic filtrations are precisely the minimum spanning trees of $\Gamma(h)$.*

Proof. The geometric process depicted in 3.7.2 describes the greedy algorithm on the underlying weighted dual graph $\Gamma(h)$. \square

As described in [47, Section 3.3], each epistatic filtration of h gives rise to a unique rooted binary tree $\mathcal{T}(h)$ with leaves labeled by the maximal cells of the regular triangulation $\Sigma(h)$. In the biological context, one refers to such trees with a fixed number of labeled leaves as *phylogenetic trees*. Moreover, there is a natural distance function on the leaf set of $\mathcal{T}(h)$ measuring genetic distances, which is given by the unique path lengths in the corresponding minimum spanning tree inside $\Gamma(h)$.

3.8 Computations

An extensive study about which trees are realizable as minimum spanning trees inside some dual graph $\Gamma(h)$ demands for an encoding of the regular triangulations of the underlying polytope. Thus, we need to look at classes of polytopes with known secondary fan. These are the *totally splittable polytopes*, introduced and studied in [70], whose regular subdivisions arise exclusively as refinements of splits, i.e. subdivisions with exactly two cells. Totally splittable polytopes comprise simplices, polygons, regular cross polytopes, and prisms and joins over simplices. The code used for the following computations can be found on the author's webpage <https://holgereble.github.io/>.

3.8.1 Case study I: polygons,

Proposition 3.8.1. *Let P_n be an n -gon. The MST-fan structure, neglecting tree orders, is the same as the secondary fan structure.*

Proof. As the discussion in [40, Section 1.1] shows, the dual graph of a triangulation of P_n is itself a tree. \square

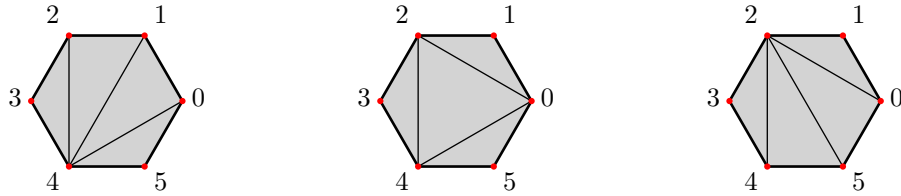


Figure 3.32: Three out of fourteen triangulations of the hexagon. The number of triangulations of the n -gon P_n is given by the Catalan number C_{n-2} , cf. [40]. The underlying dual graphs are trees with $(n-3)$ edges. Each of the C_4 -many maximal cones of the secondary fan of P_6 , i.e. the normal fan of a 3-dimensional associahedron, is split by the MST-fan structure into $3!$ MST-cones according to the permutations of the edge order.

3.8.2 Case study II: totally splittable polytopes

Let ΠP_n denote the prism $P_n \times I$ over the n -gon P_n . Moreover, for any polytope P let $\Delta^{\text{reg}}(P)$ be the set of all regular triangulations of P and let $\mathcal{T}(P)$ be the set of all ordered spanning trees that occur inside the dual graph of some regular subdivision of P . Similarly, let $\mathcal{T}^{\text{real}}(P)$ be the set of those ordered spanning trees, which occur inside some dual graph of P and which are realizable as

minimum spanning trees. The following computations were done in `polymake` [56] on an ordinary office machine (`Intel Core i7-8700`).

Polytope	$\#\Delta^{\text{reg}}(P)$	$\#\mathcal{T}(P)$	$\#\mathcal{T}^{\text{real}}(P)$	timings
P_5	5	$5 \cdot 2! = 10$	10	1s
P_6	14	$14 \cdot 3! = 84$	84	16s
P_7	42	$42 \cdot 4! = 1008$	1008	289s
P_8	132	$132 \cdot 5! = 15840$	15840	1.5h
ΠP_3	6	12	12	2s
ΠP_4	74	37632	4944	3.3h
octahedron	3	72	24	15s

Table 3.20: Calculating realizable minimum spanning trees. Polytopes with a more complicated secondary fan structure as the 3-dimensional cube ΠP_4 , which is not totally splittable, feature many non-realizable spanning trees, cf. Figure 3.31.

Computing characteristic polynomials of hyperplane arrangements with symmetries

This chapter is based on the preprint “Computing characteristic polynomials of hyperplane arrangements with symmetries” [24] by Taylor Brysiewicz, Holger Eble and Lukas Kühne. The preprint can be found on the arXiv with the number 2105.14542.

4.0 Abstract

We introduce a new algorithm computing the characteristic polynomials of hyperplane arrangements which exploits their underlying symmetry groups. Our algorithm counts the chambers of an arrangement as a byproduct of computing its characteristic polynomial. We showcase our `julia` implementation, based on `OSCAR`, on examples coming from hyperplane arrangements with applications to physics and computer science.

4.1 Introduction

The problem of enumerating chambers of hyperplane arrangements is a ubiquitous challenge in computational discrete geometry [65, 89, 109, 137]. A well-known approach to this problem is through the computation of characteristic polynomials [4, 75, 92, 114, 139, 161]. We develop a novel for computing characteristic polynomials which takes advantage of the combinatorial symmetries of an arrangement. While most arrangements admit few combinatorial symmetries [120], most arrangements of interest do [63, 123, 164].

We implemented our algorithm in `julia` [14] and published it as the package `CountingChambers.jl`¹. Our implementation relies heavily on the cornerstones of the new computer algebra system `OSCAR` [115] for group theory computations (`GAP` [55]) and the ability to work over number fields (`Hecke` and `Nemo` [52]). While other algorithms and pieces of software exist for studying hyperplane arrangements (see, for instance, [29, 43, 89, 100, 140]), either their chamber-enumeration computations appear as byproducts of more difficult calculations, the code does not use symmetry, or it only pertains to very specific types of arrangements. For example, [89] computes the associated zonotope, whose vertices are in bijection with the chambers of the arrangement, containing much more information than the characteristic polynomial. A similar approach is suggested in [43] involving a search algorithm relying upon linear programming. To the best of our knowledge, our implementation is the first publicly available software for counting chambers which uses symmetry.

We showcase our algorithm and its implementation on a number of well-known examples, such as the resonance and discriminantal arrangements. Additionally, we study sequences of hyperplane arrangements which come from the problem of linearly separating vertices of regular polytopes. In particular, we investigate one corresponding to the hypercube $[0, 1]^d$ whose chambers are in bijection with linearly separable Boolean functions.

In the presence of symmetry, our implementation outperforms the existing software by several orders of magnitude (cf. Table 4.21). Moreover, its output is guaranteed to be correct since we compute symbolically over the integers or exact number fields and avoid overflow errors thanks to the package `SaferIntegers.jl` [131].

The ninth resonance arrangement (511 hyperplanes in \mathbb{R}^9) approaches the limit of what is possible with our implementation: the computation of its characteristic polynomial took 10 days on

¹available at <https://mathrepo.mis.mpg.de/CountingChambers>

42 processors. Our computation confirms that its chamber-count is 1955230985997140 as independently and concurrently computed by Chroman and Singhar with different methods [29].

We first give background on hyperplane arrangements in Section 4.2. The ideas outlined in Section 4.3, regarding deletion and restriction algorithms, form the basic structure of our algorithm. We explain the relevant results regarding symmetries of arrangements in Section 4.4. The algorithm and its implementation details reside in Section 4.5. In Section 4.6 we construct and discuss examples of arrangements exhibiting large symmetry groups. We conclude in Section 4.7 with timings and comparisons to other software.

Acknowledgements

We are very grateful to Tommy Hofmann, Christopher Jefferson, and Marek Kaluba for their support regarding the implementation and to Michael Cuntz for initial verifications of our computations. We would also like to thank Michael Joswig for his helpful comments throughout the project and Bernd Sturmfels for suggesting the discriminantal arrangement. Lastly, we thank the referees for their careful reading and helpful comments.

4.2 Hyperplane arrangements

We begin by discussing background on the theory of hyperplane arrangements related to the problem of enumerating chambers: the main goal of this article and the associated software. Our notation will mostly follow the textbook by Orlik and Terao [114].

For any field \mathbb{K} , a hyperplane in \mathbb{K}^d is an affine linear space of codimension one. Throughout this article, we denote by $\mathbf{A} = \{H_1, \dots, H_n\}$ a (hyperplane) arrangement where H_i is a hyperplane in \mathbb{K}^d .

Definition 4.2.1. Suppose \mathbf{A} is an arrangement in \mathbb{R}^d . The connected components of the complement $\mathbb{R}^d \setminus \bigcup_{H \in \mathbf{A}} H$ are called chambers of \mathbf{A} and the set of chambers of \mathbf{A} is denoted by $\text{ch}(\mathbf{A})$.

Example 4.2.2. We use the arrangement

$$\underbrace{\{y - x = 1\}}_{H_1}, \underbrace{\{x = 0\}}_{H_2}, \underbrace{\{x + y = 1\}}_{H_3}, \underbrace{\{y = 0\}}_{H_4}$$

in \mathbb{R}^2 as a running example. This arrangement is depicted in Figure 4.33. It has 10 chambers: 2 bounded and 8 unbounded.

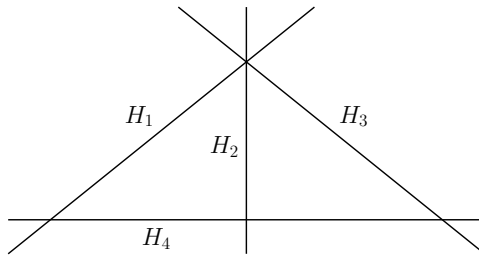


Figure 4.33: The arrangement introduced in Example 4.2.2.

Given a subset $I \subseteq [n] := \{1, \dots, n\}$, we write the set $\{H_i\}_{i \in I}$ as H_I and its intersection as $L_I = \bigcap_{i \in I} H_i$. The collection of these intersections form the set $L(\mathbf{A}) = \{L_I \mid I \subseteq [n], L_I \neq \emptyset\}$, a combinatorial shadow of \mathbf{A} known as its intersection poset. This poset is ordered by reverse inclusion and graded by the rank function, $r : L(\mathbf{A}) \rightarrow \mathbb{Z}_{\geq 0}$, where $r(L_I) = \text{codim}(L_I)$. As a notational convention, we set $r(I) = r(L_I)$ for $I \subseteq [n]$ whenever $L_I \neq \emptyset$.

4.2.1 The characteristic polynomial

Our algorithm counts chambers of an arrangement by computing a more refined count, namely the characteristic polynomial. The coefficients of this polynomial are known as the unsigned Whitney numbers of the first kind of the intersection poset $L(\mathcal{A})$, which we simply refer to as the Whitney numbers of the arrangement.

Definition 4.2.3. *The characteristic polynomial of an arrangement \mathcal{A} in \mathbb{K}^d is the polynomial*

$$\chi_{\mathcal{A}}(t) = \sum_{I \subseteq [n]: L_I \neq \emptyset} (-1)^{|I|} t^{d-r(I)} = \sum_{i=0}^d (-1)^i b_i(\mathcal{A}) t^{d-i}. \quad (4.28)$$

The integers $b_i(\mathcal{A})$, defined via (4.28), are non-negative and are called the Whitney numbers of \mathcal{A} . We denote the vector of Whitney numbers by $b(\mathbf{A})$.

The characteristic polynomial and Whitney numbers of an arrangement \mathbf{A} depend only on the intersection poset $L(\mathbf{A})$ and have various interpretations depending on the field \mathbb{K} as detailed below.

Real For an arrangement \mathbf{A} in \mathbb{R}^d , Zaslavsky [161] proved that

$$|\text{ch}(\mathbf{A})| = (-1)^d \chi_{\mathbf{A}}(-1) = \sum_{i=0}^d b_i(\mathbf{A}).$$

Thus, the Whitney numbers are a refined count of the chambers of \mathbf{A} . They have the following geometric interpretation. Given a generic flag $\mathcal{F}_{\bullet} : F_0 \subset F_1 \subset \dots \subset F_d = \mathbb{R}^d$ of affine linear subspaces F_i (where $\dim(F_i) = i$) the number of chambers of \mathcal{A} which meet F_i but do not meet F_{i-1} is equal to $b_i(\mathbf{A})$ [159, Proposition 2.3.2].

Complex If \mathbf{A} is an arrangement in \mathbb{C}^d where all hyperplanes contain the origin, then $b_i(\mathbf{A})$ is the i -th topological Betti number of the complement $\mathbb{C}^d \setminus \bigcup_{H \in \mathbf{A}} H$ with rational coefficients [113]. Because of this, some papers refer to the Whitney numbers $b_i(\mathcal{A})$ as the Betti numbers of the arrangement \mathcal{A} [160].

Finite When \mathbf{A} is an arrangement over a finite field \mathbb{F}_q , Crapo and Rota proved that $\chi_{\mathbf{A}}(q) = |\mathbb{F}_q^d \setminus \bigcup_{H \in \mathbf{A}} H|$ [34]. Moreover, if \mathbf{A} is a hyperplane arrangement in \mathbb{Q}^d one may consider its reduction modulo q : $\mathbf{A} \otimes \mathbb{F}_q = \{H_1 \otimes \mathbb{F}_q, \dots, H_n \otimes \mathbb{F}_q\}$. When q is sufficiently large, we have that $L(\mathbf{A}) = L(\mathbf{A} \otimes \mathbb{F}_q)$ and thus computing $\chi_{\mathbf{A}}(t)$ for rational arrangements also yields the number of points in the complement after reducing modulo large primes.

Example 4.2.4. *Let \mathbf{A} be the arrangement introduced in Example 4.2.2. Its characteristic polynomial is $\chi_{\mathbf{A}}(t) = t^2 - 4t + 5$. Figure 4.34 shows a generic flag \mathcal{F}_{\bullet} intersecting this arrangement verifying that $b(\mathcal{A}) = (1, 4, 5)$.*

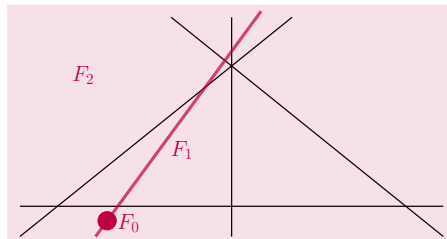


Figure 4.34: The intersections of a generic flag (purple) in \mathbb{R}^2 with the chambers of \mathcal{A} . The point F_0 intersects one chamber, F_1 intersects four others, and F_2 intersects the remaining 5, and so $b(\mathcal{A}) = (1, 4, 5)$.

4.3 A deletion-restriction algorithm

To compute the Whitney numbers of an arrangement \mathcal{A} in \mathbb{K}^d , we take advantage of the behavior of $\chi_{\mathcal{A}}(t)$ under the operations of deletion and restriction. These operations reduce computations about \mathcal{A} to computations about two smaller arrangements. Thus at its core, our main algorithm is a divide-and-conquer algorithm.

Given a hyperplane $H \in \mathcal{A}$, the deletion of H in \mathcal{A} is the arrangement $\mathcal{A} \setminus \{H\}$. The restriction of H in \mathcal{A} is the arrangement in $H \cong \mathbb{K}^{d-1}$ defined by $\mathcal{A}^H = \{K \cap H \mid K \in \mathcal{A} \setminus \{H\}\}$. The following lemma provides the basic foundation of our algorithm.

Lemma 4.3.1 ([114, Corollary 2.57]). *Given a hyperplane $H \in \mathcal{A}$, we have that $\chi_{\mathcal{A}}(t) = \chi_{\mathcal{A} \setminus \{H\}}(t) - \chi_{\mathcal{A}^H}(t)$. In particular, $b(\mathcal{A}) = b(\mathcal{A} \setminus \{H\}) + 0|b(\mathcal{A}^H)$ where $0|b$ means prepending the vector b with a zero.*

4.3.1 A simple deletion-restriction algorithm

Lemma 4.3.1 along with the fact that the empty arrangement in \mathbb{K}^d has the vector of Whitney numbers $(1, 0, \dots, 0) \in \mathbb{N}^{d+1}$ suggests the following well-known recursive algorithm for computing $b(\mathcal{A})$.

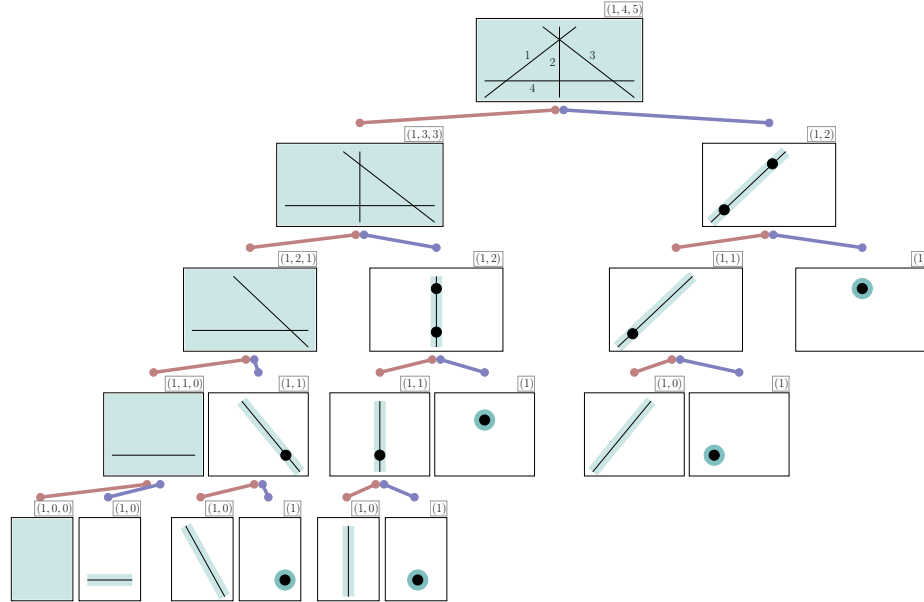


Figure 4.35: The tree structure of Algorithm 2 on the hyperplane arrangement from Example 4.2.2. Hyperplanes are chosen (Line 2) according to the ordering $\{1, 2, 3, 4\}$. In each box, the ambient space of the arrangement is shaded green. Deletions are marked with red edges (left children) and restrictions with blue edges (right children). Each arrangement box has the Whitney numbers above its upper right corner.

Algorithm 2: Whitney numbers via simple deletion and restriction

Input: A hyperplane arrangement \mathcal{A} in \mathbb{K}^d

Output: The vector of Whitney numbers $b(\mathcal{A})$

WhitneyNumbers (\mathcal{A})

```

1  if  $\emptyset \neq \mathcal{A}$  then
2    |   choose  $H \in \mathcal{A}$ 
3    |   return WhitneyNumbers( $\mathcal{A} \setminus \{H\}$ ) +  $0|$ WhitneyNumbers ( $\mathcal{A}^H$ )
4  else
5    |   return  $(1, 0, \dots, 0)$ 
```

Structurally, Algorithm 2 is a depth-first binary tree algorithm on arrangements, rooted at the initial input: one child represents a deletion and the other a restriction, as shown in Figure 4.35.

The implementation of Algorithm 2 is already nontrivial as it is often the case that some hyperplanes become the same after a restriction. Thus, its proper implementation requires care in representing an arrangement on a computer.

4.3.2 Computationally representing deletions and restrictions

An arrangement \mathcal{B} coming from \mathcal{A} via deletions and restrictions may be represented by an encoding of the restricted hyperplanes. To be precise, the pair

$$B = (\{H_{i_1}, \dots, H_{i_k}\}, \{H_{j_1}, \dots, H_{j_\ell}\}) =: (H_I, H_J)$$

represents the hyperplane arrangement \mathcal{B} in $L_I \cong \mathbb{K}^{d-r(L_I)}$ given by the *hyperplanes* in $\{H_j \cap L_I\}_{j \in J}$. Note that $H_j \cap L_I$ may be empty for some $j \in J$, in which case this intersection does not correspond to any hyperplane. We extend notation regarding \mathcal{B} to its representation B (i.e. $\chi_B(t) := \chi_{\mathcal{B}}(t)$ and $b(B) := b(\mathcal{B})$).

If $H_{j_1} \cap L_I$ is a hyperplane which occurs uniquely with respect to the tuple $(H_{j_1} \cap L_I, \dots, H_{j_\ell} \cap L_I)$, then $\mathcal{B}^{H_{j_1} \cap L_I}$ and $\mathcal{B} \setminus \{H_{j_1} \cap L_I\}$ are represented by

$$\begin{aligned} B^{H_{j_1}} &:= (\{H_{i_1}, \dots, H_{i_k}, H_{j_1}\}, \{H_{j_2}, \dots, H_{j_\ell}\}) \\ B \setminus \{H_{j_1}\} &:= (\{H_{i_1}, \dots, H_{i_k}\}, \{H_{j_2}, \dots, H_{j_\ell}\}), \end{aligned}$$

respectively. Whereas if $H_{j_1} \cap L_I$ is either empty or does not occur uniquely, then $B \setminus \{H_{j_1}\}$ trivially represents the same arrangement as B , namely \mathcal{B} .

The following computational analogue of Lemma 4.3.1 establishes how such representations behave under deletion and restriction.

Lemma 4.3.2. *Let $B = (H_I, H_J)$ represent an arrangement \mathcal{B} and fix $H \in H_J$. If $H \cap L_I$ is a hyperplane which occurs uniquely in the tuple $(H_j \cap L_I)_{j \in J}$ then $\chi_B(t) = \chi_{B \setminus \{H\}}(t) - \chi_{B^H}(t)$ and $b(B) = b(B \setminus \{H\}) + 0|b(B^H)$. Otherwise, we have $\chi_B(t) = \chi_{B \setminus \{H\}}(t)$ and $b(B) = b(B \setminus \{H\})$.*

Proof. The first case follows from Lemma 4.3.1. In the second case, B and $B \setminus \{H\}$ represent the same hyperplane arrangement and the result is trivial. \square

The following algorithm is equivalent to Algorithm 2.

Algorithm 3: Whitney numbers via extended deletion and restriction

Input: A representation $B = (H_I, H_J)$ of an arrangement in \mathbb{K}^d

Output: The vector of Whitney numbers $b(B)$

WhitneyNumbers $B = (H_I, H_J)$

```

1  if  $\emptyset \neq H_J$  then
2      choose  $H \in H_J$ 
3      if  $H \cap L_I \neq \emptyset$  occurs uniquely in  $(H_j \cap L_I)_{j \in J}$  then
4          return WhitneyNumbers( $B \setminus \{H\}$ ) +  $0|$ WhitneyNumbers ( $B^H$ )
5      else
6          return WhitneyNumbers( $B \setminus \{H\}$ )
7  else
8      return  $(1, 0, \dots, 0)$ 
```

Given a hyperplane arrangement $\mathcal{A} = \{H_1, \dots, H_n\}$ in \mathbb{K}^d , Algorithm 3 computes the Whitney numbers $b_i(\mathcal{A})$ when given $A = (\emptyset, \{H_1, \dots, H_n\})$ as input. This algorithm traverses a binary tree which is essentially the same as the one from Algorithm 2. The only difference is that some edges are extended with nodes that have only one child and so we say it computes the Whitney numbers via *extended* deletion and restriction.

Algorithm 3 has the advantage that the representations of the original hyperplanes in \mathcal{A} need not be updated upon restriction, and that representations of hyperplanes in \mathcal{A}^H need not be unique. As a consequence, structural aspects of \mathcal{A} such as its symmetries extend trivially to the

4.4 Automorphisms of hyperplane arrangements

Let \mathfrak{S}_n be the permutation group on $[n]$. Elements of a subgroup $G \leq \mathfrak{S}_n$ act on subsets of $[n]$. Given $g \in G$ and $I \subseteq [n]$, we fix the notation

- Definition 4.4.1.** *The automorphism group of $\mathcal{A} = \{H_1, \dots, H_n\}$ is*

Given a representation $B = (H_I, H_J)$ of an arrangement coming from \mathbf{A} , the automorphism group $\text{Aut}(\mathbf{A})$ acts as $gB = (H_{gI}, H_{gJ})$.

94

Lemma 4.4.3. *Let $\mathcal{A} = \{H_1, \dots, H_n\}$ be an arrangement in \mathbb{K}^d and let B_1 and B_2 represent arrangements coming from deletions and restrictions. If B_1 and B_2 are in the same orbit under $\text{Aut}(\mathbf{A})$ then $b(B_1) = b(B_2)$.*

Proof. The conclusion of the lemma is equivalent to showing that the characteristic polynomials of B_1 and B_2 are the same. This follows directly from the fact that the characteristic polynomial depends only on the intersection poset (graded by rank) and that B_1 and B_2 are in the same orbit under $\text{Aut}(\mathbf{A})$ if and only if they are related by a rank-preserving permutation. \square

Our algorithm relies upon the following corollary of Lemma 4.4.3.

Corollary 4.4.4. *Let $B = (H_I, H_J)$ represent a hyperplane arrangement coming from $\mathcal{A} = \{H_1, \dots, H_n\}$. For $g \in J^{\text{Aut}(\mathbf{A})}$ we have that $gB = (H_{gI}, H_J)$ and B have the same Whitney numbers.*

4.5 Enumeration algorithm with symmetry

Our main algorithm augments Algorithm 3, making particular use of Corollary 4.4.4. It is essentially a breadth-first tree algorithm except that at each level, nodes may be identified up to symmetry and so the algorithmic structure is no longer that of a tree. The output is the vector of Whitney numbers $b(\mathcal{A})$ of an arrangement \mathcal{A} , refining its chamber count. We remark that despite the fact that our algorithm takes advantage of symmetry and counts the number of chambers, it does not reveal any information about the sizes of orbits of chambers under this symmetry group.

Given an arrangement $\mathbf{A} = \{H_1, \dots, H_n\}$ in \mathbb{K}^d , we represent the nodes of the algorithm at depth k by a dictionary T_k . The keys of T_k are orbits $G_k \cdot I$ for $I \subseteq [k]$ where G_k is a subgroup of the stabilizer of $\{k+1, \dots, n\}$ in $\text{Aut}(\mathbf{A})$. The value of $G_k \cdot I$ in this dictionary is a pair $(B_I, \omega(B_I))$ where B_I represents the hyperplane arrangement $(H_I, H_{\{k+1, \dots, n\}})$ and $\omega(B_I)$ is some multiplicity, tracking how many arrangements indexed by elements of the orbit $G_k \cdot I$ have appeared. We refer to T_k as a k -th orbit-node dictionary.

Algorithm 4 presents the breadth-first structure of the algorithm.

Algorithm 4: Whitney numbers using symmetry

Input: A hyperplane arrangement $\mathcal{A} = \{H_1, \dots, H_n\}$ in \mathbb{K}^d

A subgroup $G \leq \text{Aut}(\mathbf{A})$

Output: The vector of Whitney numbers $b(\mathcal{A})$

WhitneyNumbers (\mathcal{A})

```

1  // compute the stabilizers of G
   compute  $\{G_i\}_{i=0}^n$  where  $G_i = \{i+1, \dots, n\}^G$  and  $G_n = G$ 
   // initialize orbit-node dictionaries
2  initialize  $\{T_i\}_{i=0}^n$  and set  $T_0 = \{G_0 \cdot \emptyset \Rightarrow ((\emptyset, \mathbf{A}), 1)\}$ 
3  for  $k = 1, \dots, n$  do
4    | set  $T_k = \text{NextGeneration}(\mathcal{A}, G_k, T_{k-1})$ 
5  initialize  $b = (0, 0, \dots, 0)$ 
6  for  $(B_I, \omega(B_I)) \in T_n$  do
7    | increment the entry  $b_{|I|}$  by  $\omega(B_I)$ 
8  return  $b$ 

```

Moving from depth $k - 1$ to k is performed by Algorithm 5.

Algorithm 5: NextGeneration

Input: A hyperplane arrangement $\mathcal{A} = \{H_1, \dots, H_n\}$ in \mathbb{K}^d
A subgroup $G_k \leq \{k + 1, \dots, n\}^{\text{Aut}(\mathbf{A})}$
An orbit-node dictionary T_{k-1}
Output: An orbit-node dictionary T_k
NextGeneration ($\mathcal{A}, G_k, T_{k-1}$)

```

1  set  $J = \{k + 1, \dots, n\}$ 
2  for  $(B_I, \omega(B_I)) \in \text{values}(T_{k-1})$  do
3      if  $H_k \cap L_I$  is a unique hyperplane amongst  $(H_j \cap L_I)_{j=k}^n$  then
4          // produce the restriction as the right child
5          set  $I' = I \cup \{k\}$ 
6          compute the orbit  $\mathcal{O} = G_k \cdot I'$ 
7          if  $\mathcal{O} \in \text{keys}(T_k)$  then
8              | increment the multiplicity of  $T_k[\mathcal{O}]$  by  $\omega(B_I)$ 
9          else
10             |  $T_k[\mathcal{O}] = ((H_{I'}, H_J), \omega(B_I))$ 
11         // produce the deletion as the left child
12         compute the orbit  $\mathcal{O} = G_k \cdot I$ 
13         if  $\mathcal{O} \in \text{keys}(T_k)$  then
14             | increment the multiplicity of  $T_k(\mathcal{O})$  by  $\omega(B_I)$ 
15         else
16             |  $T_k[\mathcal{O}] = ((H_I, H_J), \omega(B_I))$ 
17     return  $T_k$ 

```

Example 4.5.1. The structure underlying Algorithm 4 applied to the arrangement in Example 4.2.2 is shown in Figure 4.37. It is no longer a tree but may be obtained from the tree in Figure 4.36 by identifying nodes under the stabilizers of $\text{Aut}(\mathbf{A})$. Each identification accumulates multiplicity in the node and that multiplicity is passed down to its children.

4.5.1 Representing orbits

The computations of orbits in Line 5 and Line 10 require elaboration; specifically in regards to representing an orbit $G \cdot I$ on a computer. One option is to use a canonical element of $G \cdot I$, which can be computed using the `MinimalImage` or `CanonicalImage` functions from GAP [82, 81]. An alternative approach is to provide any function $\varphi : 2^{[n]} \rightarrow S$ taking values in an arbitrary set S such that $\varphi(I) = \varphi(J)$ only if $G \cdot I = G \cdot J$. Equivalently, φ is any factor of the projection $\pi : 2^{[n]} \rightarrow 2^{[n]}/G$ as a map of sets where $2^{[n]}/G$ is the set of orbits. In this case, the value of $\varphi(I)$ may be used to represent the orbit $G \cdot I$ as a key in the orbit-node dictionaries. While this approach may fail to identify all nodes in the same orbit, nodes in distinct orbits are never identified and so the algorithm remains correct. The benefit is that it may be significantly more efficient to evaluate φ than it is to compute minimal or canonical images.

Our default option for identifying orbits is called `pseudo_minimal_image`. Given a subset $I \subseteq [n]$ and a collection of elements $g_1, \dots, g_m \in G \leq \mathfrak{S}_n$, this function sequentially computes $g_i I$ and recursively calls itself on $g_i I$ whenever $g_i I < I$ lexicographically. If no such g_i produces a smaller subset, I itself is returned. Options are implemented for choosing m to be a proportion of $|G|$ subject to maximum and minimum values. For our computations, we take $m = n$ random elements of G . Although this greedy procedure does not make all possible identifications in the algorithm, we have found that it is quicker than `MinimalImage` to evaluate and produces a comparably small algorithmic structure.

Example 4.5.2. We compare the effect of three choices of identifications in Algorithm 4 (either `pseudo_minimal_image`, the `MinimalImage` function in GAP, or no identifications at all) on the resonance arrangement \mathcal{R}_7 (see Definition 4.6.3) consisting of 127 hyperplanes in \mathbb{R}^7 . We compare

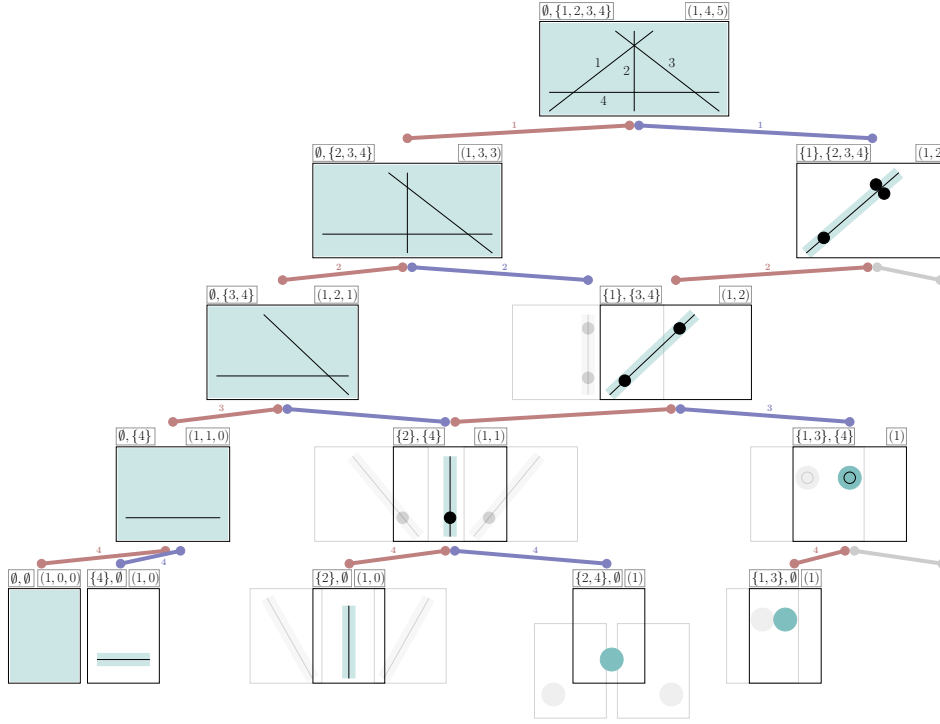


Figure 4.37: The algorithmic structure underlying Algorithm 4. Starting at the top node, each call of Algorithm 5 produces the next depth of this graph.

the number of leaves of the algorithm at some depth, as well as the time per depth of the algorithm and display the results in Figure 4.38.

As depicted, the cost (in number of leaves) of using *pseudo_minimal_image* compared to *MinimalImage* is negligible, while the benefits in terms of speed are significant. Similarly, while the timing of our algorithm with *MinimalImage* is comparable to the timing without any identifications (Algorithm 3), the memory usage is significantly reduced as conveyed by the number of leaves (a reasonable proxy for memory usage). This difference becomes even more dramatic for larger arrangements.

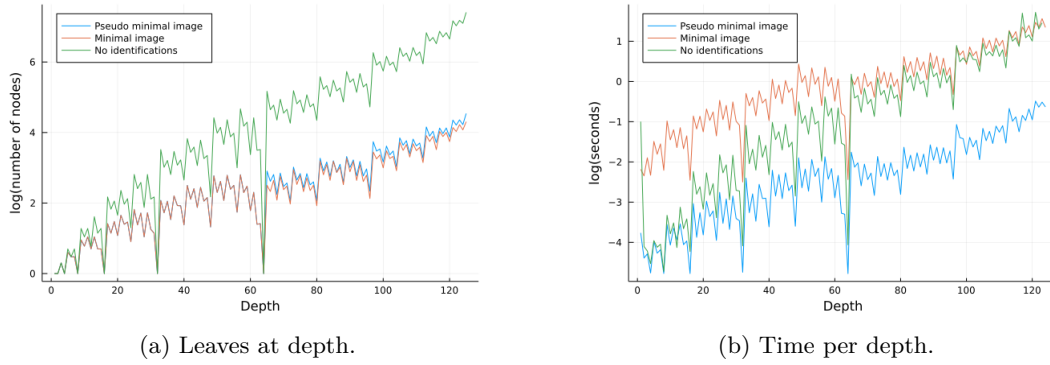


Figure 4.38: The leaves per depth and time per depth of Algorithm 4 on the arrangement \mathcal{R}_7 using *pseudo_minimal_image*, *MinimalImage*, and no identifications.

4.5.2 Accumulating the Whitney numbers and skipping levels

Much of the computational burden occurs in Line 3 of Algorithm 5 and involves projecting the normal vectors of the hyperplanes in \mathbf{A} along those hyperplanes which have been restricted. When implementing Algorithm 5, one may choose whether to save such computations at the cost of

memory, or to perform redundant computations throughout the algorithm. We found that, for our benchmark examples, recomputation held the most benefit.

Nonetheless, from the linear algebra involved in the evaluation of Line 3, one can read off j_{\min} , the smallest $j \in J$ for which this uniqueness condition is true. Hence, one may immediately place the left child of the corresponding node in level j_{\min} rather than k to avoid redundancy in Line 3 later on. This comes at the cost of missing some identifications between the layers k and j_{\min} .

Another implementation choice we made was to keep a running count of the Whitney numbers of the arrangement throughout the algorithm. Whenever $j_{\min} = n$ while computing the children of $(B_I, \omega(B_I))$, we increment $b_{|I|}$ by $\omega(B_I)$ and delete the node altogether since no other deletions or restrictions are possible. Similarly, if \mathbf{A} is a hyperplane arrangement where each hyperplane contains the origin, $b_{|I|}$ and $b_{|I|+1}$ are incremented by $\omega(B_I)$ whenever $j_{\min} = n - 1$ by a similar reasoning. In this way, we can free memory occupied by nodes throughout the algorithm.

4.5.3 Relation to OSCAR

The new computer algebra system OSCAR in `julia` combines the existing systems GAP [55], Singular [42], Polymake [57, 87], and Antic (Hecke, Nemo) [115]. Our software is written in `julia` and builds heavily on these cornerstones. Specifically, we use the number theory components Nemo [52] and Hecke to work with arrangements defined over algebraic field extensions of \mathbb{Q} . For example the separability arrangement of the vertices of the 600-cell is defined over $\mathbb{Q}(\sqrt{5})$.

Secondly, we use GAP [55] for group theoretic computations in Algorithm 5. Concretely, we compute stabilizers and minimal images using the GAP packages `ferret` [80] and `images` [81], respectively.

4.5.4 Functionality of CountingChambers.jl

The `julia` package titled `CountingChambers.jl` contains our implementation and is available at

<https://mathrepo.mis.mpg.de/CountingChambers>

The following code snippet shows some standard functions of our package applied to the arrangement introduced in Example 4.2.2. A collection of hyperplanes defined by the equations $\ell_i(x_1, \dots, x_d) = c_i$ for $1 \leq i \leq n$ is encoded by a $d \times n$ matrix A having the coefficients of ℓ_i as columns and a vector c .

```
julia> A = [-1 1 1 0; 1 0 1 1];
julia> c = [1, 0, 1, 0];
julia> whitney_numbers(A; ConstantTerms=c)
3-element Vector{Int64}:
 1 4 5
julia> characteristic_polynomial(A; ConstantTerms=c)
t^2 - 4*t + 5
julia> number_of_chambers(A; ConstantTerms=c)
10
```

Note that the automorphism group of this arrangement is $\mathfrak{S}_3 \hookrightarrow \mathfrak{S}_4$ consisting of permutations of the first three hyperplanes. This group can be passed to our algorithm via a list of generators in one-line notation:

```
julia> G = [[2,3,1,4],[2,1,3,4]];
julia> whitney_numbers(A; ConstantTerms=c, SymmetryGroup=G)
3-element Vector{Int64}:
 1 4 5
```

As it is easy to run `julia` on multiple threads, we also implemented our algorithm to take advantage of this. By starting `julia` via the command `julia -threads NUM_THREADS` and passing the optional parameter `multi_threaded=true` to our methods, the `for` loop in Algorithm 5 is executed in parallel. Table 4.23 shows how the multithreading scales.

4.6 Examples and integer sequences

We apply our algorithm to a number of examples. Many of these arise from the following construction of separability arrangements.

4.6.1 Separability arrangements

Fix a finite set $\mathbf{V} \subset \mathbb{R}^d$. We associate to every $v \in \mathbf{V}$ the hyperplane $H_v \subset (\mathbb{R}^{d+1})^*$ comprised of linear forms which vanish on $(1, v)$. Equivalently, H_v represents the affine hyperplanes in \mathbb{R}^d which contain v . We call the arrangement $\mathcal{H}_{\mathbf{V}} := \{H_v \mid v \in \mathbf{V}\}$ the separability arrangement of \mathbf{V} . We point out that by increasing the dimension d by one, this construction is distinct from the one which defines *real reflection arrangements* from root systems. In particular, translating \mathbf{V} does not change the combinatorics of $\mathcal{H}_{\mathbf{V}}$.

A hyperplane H_v partitions the points in $(\mathbb{R}^{d+1})^* \setminus H_v$ into the sets H_v^+ of linear forms which are positive on v and H_v^- which are negative on v . Consequently, all affine hyperplanes corresponding to points in a chamber of $\mathcal{H}_{\mathbf{V}}$ are positive on some subset $V_1 \subset \mathbf{V}$ and negative on its complement $V_2 = \mathbf{V} \setminus V_1$. Such a partition $V_1 \sqcup V_2 = \mathbf{V}$ is called linearly separable. Hence, chambers of $\mathcal{H}_{\mathbf{V}}$ are in bijection with linearly separable partitions of \mathbf{V} , motivating the terminology for $\mathcal{H}_{\mathbf{V}}$. This point of view, which connects linear separability and hyperplane arrangements, appears in [6, Section 2].

One purpose for introducing separability arrangements is that it immediately provides us with a zoo of arrangements admitting considerable symmetry; for example, those \mathbf{V} which are the vertices of regular polytopes.

4.6.2 The threshold arrangement

The following arrangement appears in the study of neural networks [110, 111, 165] and algebraic statistics [37].

Definition 4.6.1. *The threshold arrangement², \mathcal{T}_d is the separability arrangement associated to the vertices of the hypercube $[0, 1]^d$. That is,*

$$\mathcal{T}_d := \{\{x_0 + c_1x_1 + \cdots + c_dx_d = 0\} \text{ with } c_i \in \{0, 1\} \text{ for all } c_i\}.$$

As a consequence of the definition of \mathcal{T}_d , the linear automorphisms of the hypercube $[0, 1]^d$, namely the hyperoctahedral group of order $d!2^d$, is a subgroup of $\text{Aut}(\mathcal{T}_d)$. The true size of $\text{Aut}(\mathcal{T}_d)$ is $(d+1)!2^d$.

We computed the Whitney numbers of \mathcal{T}_d for $1 \leq d \leq 8$, and thus their number of chambers. The results are collected in Table 4.24 and the timings appear in Table 4.22. The values of $|\text{ch}(\mathcal{T}_d)|$ for $1 \leq d \leq 9$ are listed in entry A000609 of the Online-Encyclopedia of Integer Sequences (OEIS), whereas the Whitney numbers of \mathcal{T}_d , to the best of our knowledge, have not been published before. Zuev showed that asymptotically $|\text{ch}(\mathcal{T}_d)| \sim 2^{d^2}$ [164].

Remark 4.6.2. *Using similar proof techniques as in [97] one can show that the values of $b_i(\mathcal{T}_d)$ for $1 \leq d \leq 2^i$ determine a formula for $b_i(\mathcal{T}_d)$ for all d . Applying this to the case of $b_2(\mathcal{T}_d)$ and $b_3(\mathcal{T}_d)$ and using the results in Table 4.24 we obtain $b_2(\mathcal{T}_d) = \frac{1}{2}(4^d - 2^d)$ and $b_3(\mathcal{T}_d) = \frac{1}{24}(4 \cdot 8^d - 3 \cdot 6^d - 6 \cdot 4^d + 5 \cdot 2^d)$. For $i \geq 4$ this technique requires knowledge of $b_i(\mathcal{T}_d)$ for at least $1 \leq d \leq 16$.*

4.6.3 The resonance arrangement

The next arrangement we consider appears as a restriction of the threshold arrangement.

Definition 4.6.3. *The resonance arrangement is the restriction of \mathcal{T}_d to the hyperplane $H_{(0, \dots, 0)}$. Equivalently, for $d \geq 1$ the resonance arrangement is*

$$\mathcal{R}_d := \{\{c_1x_1 + c_2x_2 + \cdots + c_dx_d = 0\} \text{ with } c_i \in \{0, 1\} \text{ and not all } c_i \text{ are zero}\}.$$

The chambers of the resonance arrangements are in bijection with generalized retarded functions in quantum field theory [49]. An overview of the applications of the resonance arrangement is given in [97, Section 1]. A formula for their number of chambers remains elusive, let alone one for their Whitney numbers. Nonetheless, partial formulas and bounds exist [15, 63, 97, 164].

The numbers of chambers of the resonance arrangements are listed in the sequence A034997 in the OEIS up to $d = 9$. The Whitney numbers are published in [88] up to $d = 7$. Our software was able to determine the Whitney numbers of \mathcal{R}_8 and \mathcal{R}_9 confirming the concurrent computations in [29]. The computation for \mathcal{R}_9 took ten days, running multithreaded on 42 Intel Xeon E7-8867 v3 CPUs. All Whitney numbers of \mathcal{R}_d up to $d = 9$ are given in Table 4.25 and the timings are listed in Table 4.22.

²The arrangement $\{x_i + x_j\}_{1 \leq i < j \leq d}$ in \mathbb{R}^d is also referred to as a threshold arrangement in the literature. We discuss the arrangement \mathcal{T}_d only as in Definition 4.6.1.

4.6.4 Separability arrangements of the cross-polytopes

The cross-polytope of dimension d is the polytope with the $2d$ vertices $\{\pm e_i\}_{i=1}^d$. Its symmetry group is the hyperoctahedral group of order $d!2^d$. We define the arrangement \mathcal{C}_d in \mathbb{R}^{d+1} to be the separability arrangement of its vertices.

Our computations show that $|\text{ch}(\mathcal{C}_d)| = 2 \cdot 3^d - 2^d$ for $d \leq 20$, suggesting that $|\text{ch}(\mathcal{C}_d)|$ agrees with this sequence (A027649 in the OEIS). This can indeed be proven by applying Athanasiadis' finite field method [4] and seems to be a new result obtained through experiments with our algorithm.

4.6.5 Separability arrangements of permutohedra

The permutohedron of dimension $d-1$ is the convex hull of the $d!$ points $\sigma(1, \dots, d)$ for all $\sigma \in \mathfrak{S}_d$. The separability arrangements \mathcal{P}_d of these points in \mathbb{R}^{d+1} consist of $d!$ hyperplanes. We record their Whitney numbers in Table 4.27 for $1 \leq d \leq 6$.

4.6.6 Separability arrangements of demicubes

The d -demicube is the convex hull of those vertices of the hypercube $[0, 1]^d$ which have an odd number of 1's. For instance, the 3-demicube is a regular tetrahedron. We denote by \mathcal{D}_d the corresponding separability arrangement consisting of 2^{d-1} hyperplanes in \mathbb{R}^{d+1} . Table 4.26 contains the Whitney numbers of \mathcal{D}_d up to $d = 9$.

4.6.7 Separability arrangements of some regular polytopes

In Table 4.28, we provide the Whitney numbers for the separability arrangements corresponding to the remaining two Platonic solids: the icosahedron and the dodecahedron. This table also contains the Whitney numbers of the separability arrangements of the vertices of the regular 24-cell, 600-cell, and 120-cell. Except for the 24-cell, each of these computations uses irrational realizations.

4.6.8 Discriminantal arrangements

Given n points in \mathbb{R}^d in general position, the discriminantal arrangement $\text{Disc}_{d,n}$ is the hyperplane arrangement in \mathbb{R}^d consisting of the $\binom{n}{d}$ hyperplanes spanned by d -subsets of such points. This arrangement, originally called the “geometry of circuits” was introduced by Crapo [33]. We verify the Whitney numbers of $\text{Disc}_{4,n}$ for $5 \leq n \leq 16$ given in [93, Section 4.4]. From this data, we recover their formula for the characteristic polynomial of $\text{Disc}_{4,n}$ for all n . A deformation of this arrangement appears in physics [25, 26] and we were able to confirm the chamber counts given in these papers.

4.7 Timings

While other pieces of software for counting chambers of arrangements exist, they do not take advantage of symmetry and some compute significantly more data than our algorithm does. Consequently, our software outperforms them with respect to the calculation of Whitney numbers as shown below.

Software	$d = 3$	4	5	6	7	8
CountingChambers.jl w/ <i>symm.</i>	0.0038s	0.011s	0.035s	0.12s	2.89s	19.8m
CountingChambers.jl w/o <i>symm.</i>	0.0002s	0.0004s	0.004s	0.53s	6.2m	*
polymake	0.3s	4.31s	3.9m	*		
sage	0.05s	0.21s	5.45s	9.2m	*	
GAP	0,006s	0.035s	1.09s	1.9m	12.87h	*

Table 4.21: Timings for computing the number of chambers $|\text{ch}(\mathcal{R}_d)|$ of the resonance arrangement \mathcal{R}_d for $3 \leq d \leq 8$ on a single thread (Intel Core i7-8700). An asterisk * indicates that the computation was terminated after a day.

The implementation [89] in **polymake** computes much more information than the Whitney numbers, namely a chamber decomposition of the arrangement. The **sage** implementation, on the other hand, uses basic deletion and restriction as in Algorithm 3. Similarly, the **GAP** package **alcove** [100] computes the Tutte polynomial by simple deletion and restriction and then specializes this to the characteristic polynomial.

To illustrate the performance of our software on the arrangements from Section 4.6, we collect our timings in Table 4.22. This table also shows the growth in complexity for computing the number of chambers of these arrangements. Based on our profiling, the main bottleneck in our implementation is the identifications of orbits. Thus, improving **pseudo_minimal_image** would be the most direct method for making our code faster.

A	$ \text{Aut}(\mathbf{A}) $	$d = 3$	4	5	6	7	8	9
\mathcal{T}_d	$(d+1)!2^d$	0.005s	0.013s	0.041s	0.28s	33.17s	8.16h	
\mathcal{R}_d	$(d+1)!$	0.004s	0.011s	0.035s	0.12s	2.89s	19.8m	$\sim 10d^+$
\mathcal{C}_{2d}	$(2d)!2^{2d}$	0.015s	0.039s	0.085s	0.183s	0.42s	1.158s	4.50s
\mathcal{P}_d	$d!$	0.003s	0.013s	6.398s	$\sim 8d^+$			
\mathcal{D}_d	$(d)!2^{d-1}$	0.002s	0.005s	0.018s	0.049s	0.54s	1.9m	$\sim 8d^+$
$\text{Disc}_{4,n}$	$n!$	—	0.0003s	0.0047s	0.055s	0.71s	7.62s	41.14s

Table 4.22: Our timings on examples from Section 4.6. Computations ran on a single thread (Intel Core i7-8700) except for \mathcal{R}_9 which ran on 42 threads (Intel Xeon E7-8867).

\mathcal{A}	#threads = 1	2	4	8	12
\mathcal{R}_8	19.8m	10.5m	6.3m	5.9m	5.1m
\mathcal{T}_8	8.16h	3.9h	2.4h	1.8h	1.6h

Table 4.23: Comparison of the effect of number of threads on run times (Intel Core i7-8700).

4.8 Tables of Whitney numbers

d	1	2	3	4	5	6	7	8
$b_0(\mathcal{T}_d)$	1	1	1	1	1	1	1	1
$b_1(\mathcal{T}_d) = \mathcal{T}_d $	2	4	8	16	32	64	128	256
$b_2(\mathcal{T}_d)$	1	6	28	120	496	2016	8128	32640
$b_3(\mathcal{T}_d)$		3	44	460	4240	36848	310464	2569920
$b_4(\mathcal{T}_d)$			23	820	19660	400400	7493808	133492800
$b_5(\mathcal{T}_d)$				465	43014	2453248	112965776	4626016752
$b_6(\mathcal{T}_d)$					27129	7111650	987779688	103818315888
$b_7(\mathcal{T}_d)$						5023907	4075759064	1382897843304
$b_8(\mathcal{T}_d)$							3193753807	8676817935144
$b_9(\mathcal{T}_d)$								7393243346241
$ \text{ch}(\mathcal{T}_d) $	2	14	104	1882	94572	15028134	8378070864	17561539552946

Table 4.24: The values of $b_i(\mathcal{T}_d)$ and $|\text{ch}(\mathcal{T}_d)|$ of the threshold arrangement for $1 \leq d \leq 9$ and $0 \leq i \leq d$.

d	1	2	3	4	5	6	7	8	9
$b_0(\mathcal{R}_d)$	1	1	1	1	1	1	1	1	1
$b_1(\mathcal{R}_d) = \mathcal{R}_d $	1	3	7	15	31	63	127	255	511
$b_2(\mathcal{R}_d)$		2	15	80	375	1652	7035	29360	120975
$b_3(\mathcal{R}_d)$			9	170	2130	22435	215439	1957200	17153460
$b_4(\mathcal{R}_d)$				104	5270	159460	3831835	81029004	1582492380
$b_5(\mathcal{R}_d)$					3485	510524	37769977	2076831708	96834110730
$b_6(\mathcal{R}_d)$						371909	169824305	30623870732	3829831100340
$b_7(\mathcal{R}_d)$							135677633	207507589302	89702833260450
$b_8(\mathcal{R}_d)$								178881449368	973784079284874
$b_9(\mathcal{R}_d)$									887815808473419
$ \text{ch}(\mathcal{R}_d) $	2	6	32	370	11292	1066044	347326352	419172756930	1955230985997140

Table 4.25: The values of $b_i(\mathcal{R}_d)$ and $|\text{ch}(\mathcal{R}_d)|$ of the resonance arrangement for $1 \leq d \leq 9$ and $0 \leq i \leq d$. We submitted these Whitney numbers to the OEIS as the sequence A344494.

d	2	3	4	5	6	7	8	9
$b_0(\mathcal{D}_d)$	1	1	1	1	1	1	1	1
$b_1(\mathcal{D}_d) = \mathcal{D}_d $	2	4	8	16	32	64	128	256
$b_2(\mathcal{D}_d)$	1	6	28	120	496	2016	8128	32640
$b_3(\mathcal{D}_d)$	0	4	50	500	4480	38304	319200	2622400
$b_4(\mathcal{D}_d)$		1	44	1160	24340	461496	8283744	143504320
$b_5(\mathcal{D}_d)$			15	1362	76364	3486448	143595816	5483536464
$b_6(\mathcal{D}_d)$				597	120942	15440376	1615624080	145378334304
$b_7(\mathcal{D}_d)$					64903	33803416	10878083096	2574289938400
$b_8(\mathcal{D}_d)$						21424343	35828091880	27816202212040
$b_9(\mathcal{D}_d)$							26430009593	146101801794362
$b_{10}(\mathcal{D}_d)$								120719853808577
$ \text{ch}(\mathcal{D}_d) $	4	16	146	3756	291558	74656464	74904015666	297363155783764

Table 4.26: The values of $b_i(\mathcal{D}_d)$ and $|\text{ch}(\mathcal{D}_d)|$ of the demicube arrangement for $2 \leq d \leq 9$ and $0 \leq i \leq d+1$.

d	1	2	3	4	5	6
$b_0(\mathcal{P}_d)$	1	1	1	1	1	1
$b_1(\mathcal{P}_d) = \mathcal{P}_d $	1	2	6	24	120	720
$b_2(\mathcal{P}_d)$		1	15	276	7140	258840
$b_3(\mathcal{P}_d)$			10	1423	246605	59577390
$b_4(\mathcal{P}_d)$				1170	4290610	9271534305
$b_5(\mathcal{P}_d)$					4051026	834595018036
$b_6(\mathcal{P}_d)$						825382803000
$ \text{ch}(\mathcal{P}_d) $	2	4	32	2894	8595502	1669309192292

Table 4.27: The values of $b_i(\mathcal{P}_d)$ and $|\text{ch}(\mathcal{P}_d)|$ of the permutohedron arrangement for $1 \leq d \leq 6$ and $0 \leq i \leq d$.

polytope	Icosahedron	Dodecahedron	24-cell	600-cell	120-cell
$b_0(\mathbf{A})$	1	1	1	1	1
$b_1(\mathbf{A}) = \mathbf{A} $	12	20	24	120	600
$b_2(\mathbf{A})$	66	166	276	7140	179700
$b_3(\mathbf{A})$	157	577	1630	225782	31972550
$b_4(\mathbf{A})$	102	430	4308	3118740	2979870540
$b_5(\mathbf{A})$	—	—	2931	2899979	2948077091
$ \text{ch}(\mathbf{A}) $	338	1194	9170	6251762	5960100482

Table 4.28: The values of $b_i(\mathbf{A})$ and $|\text{ch}(\mathbf{A})|$ of the icosahedral and dodecahedral arrangements as well as arrangements stemming from regular 4-polytopes for $0 \leq i \leq 5$.

Bibliography

- [1] Andrés Aranda-Díaz, Benjamin Obadia, Ren Dodge, Tani Thomsen, Zachary F Hallberg, Zehra Tüzün Güvener, William B Ludington, and Kerwyn Casey Huang, *Bacterial inter-species interactions modulate ph-mediated antibiotic tolerance*, eLife **9** (2020), e51493.
- [2] Federico Ardila and Caroline J. Klivans, *The Bergman complex of a matroid and phylogenetic trees*, J. Comb. Theory, Ser. B **96** (2006), no. 1, 38–49.
- [3] Benjamin Assarf, Ewgenij Gawrilow, Katrin Herr, Michael Joswig, Benjamin Lorenz, Andreas Paffenholz, and Thomas Rehn, *Computing convex hulls and counting integer points with polymake*, Math. Program. Comput. **9** (2017), no. 1, 1–38. MR 3613012
- [4] Christos A. Athanasiadis, *Characteristic polynomials of subspace arrangements and finite fields*, Adv. Math. **122** (1996), no. 2, 193–233. MR 1409420
- [5] Djordje Bajic, Jean C C Vila, Zachary D Blount, and Alvaro Sanchez, *On the deformability of an empirical fitness landscape by microbial evolution.*, Proceedings of the National Academy of Sciences **115** (2018), no. 44, 11286–11291.
- [6] P. Baldi and R. Vershynin, *Polynomial threshold functions, hyperplane arrangements, and random tensors*, SIAM J. Math. Data Sci. **1** (2019), no. 4, 699–729.
- [7] Jeffrey E. Barrick and Richard E. Lenski, *Genome dynamics during experimental evolution*, Nature Reviews Genetics **14** (2013), no. 12, 827–839.
- [8] William Bateson and Gregor Mendel, *Mendel’s principles of heredity*, Cambridge :University Press, 1909, <https://www.biodiversitylibrary.org/bibliography/44575> — "Part II: 1. Biographical notice of Mendel. 2. Translation of the paper on hybridisation. 3. Translation of the paper on Hieracium": p. [307]-368.
- [9] Niko Beerenwinkel, Lior Pachter, and Bernd Sturmfels, *Epistasis and shapes of fitness landscapes*, Statist. Sinica **17** (2007), no. 4, 1317–1342. MR 2398598
- [10] ———, *Epistasis and shapes of fitness landscapes*, Statist. Sinica **17** (2007), no. 4, 1317–1342. MR 2398598
- [11] Niko Beerenwinkel, Lior Pachter, Bernd Sturmfels, Santiago F. Elena, and Richard E. Lenski, *Analysis of epistatic interactions and fitness landscapes using a new geometric approach*, BMC Evolutionary Biology **7** (2007), no. 1, 60.
- [12] Yoav Benjamini and Daniel Yekutieli, *The control of the false discovery rate in multiple testing under dependency*, The Annals of Statistics **29** (2001), no. 4, 1165–1188.
- [13] Amy Berrington de González and D. R. Cox, *Interpretation of interaction: A review*, Ann. Appl. Stat. **1** (2007), no. 2, 371–385.
- [14] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B. Shah, *Julia: a fresh approach to numerical computing*, SIAM Rev. **59** (2017), no. 1, 65–98. MR 3605826
- [15] L. J. Billera, J. Tatch Moore, C. Dufort Moraites, Y. Wang, and K. Williams, *Maximal unbalanced families*, arXiv:1209.2309 (2012).
- [16] Ian Billick and Ted Case, *Higher Order Interactions in Ecological Communities : What Are They and How Can They be Detected?*, Ecology **75** (1994), no. 6, 1529–1543.

- [17] Anders Björner, Michel Las Vergnas, Bernd Sturmfels, Neil White, and Gunter M. Ziegler, *Oriented matroids*, 2 ed., Encyclopedia of Mathematics and its Applications, Cambridge University Press, 1999.
- [18] Evan A Boyle, Yang I Li, and Jonathan K Pritchard, *An Expanded View of Complex Traits: From Polygenic to Omnigenic*, *Cell* **169** (2017), no. 7, 1177–1186.
- [19] David Bremner, Mathieu Dutour Sikirić, Dmitrii V. Pasechnik, Thomas Rehn, and Achill Schürmann, *Computing symmetry groups of polyhedra*, *LMS J. Comput. Math.* **17** (2014), no. 1, 565–581. MR 3356046
- [20] David Bremner, Mathieu Dutour Sikirić, and Achill Schürmann, *Polyhedral representation conversion up to symmetries*, *Polyhedral computation*, CRM Proc. Lecture Notes, vol. 48, Amer. Math. Soc., Providence, RI, 2009, pp. 45–71. MR 2503772
- [21] Joel Brenner and Larry Cummings, *The Hadamard maximum determinant problem*, *Amer. Math. Monthly* **79** (1972), 626–630. MR 0301030
- [22] Sarah Brodsky, Michael Joswig, Ralph Morrison, and Bernd Sturmfels, *Moduli of tropical plane curves*, *Res. Math. Sci.* **2** (2015), Art. 4, 31. MR 3333700
- [23] Winfried Bruns and Joseph Gubeladze, *Polytopes, rings, and K-theory*, Springer Monographs in Mathematics, Springer, Dordrecht, 2009. MR 2508056
- [24] Taylor Brysiewicz, Holger Eble, and Lukas Kühne, *Computing characteristic polynomials of hyperplane arrangements with symmetries*, arXiv:2105.14542 (2021).
- [25] Freddy Cachazo, Nick Early, Alfredo Guevara, and Sebastian Mizera, *Scattering equations: from projective spaces to tropical grassmannians*, *J. High Energy Phys.* **2019** (2019), no. 6, 39.
- [26] Freddy Cachazo, Bruno Umbert, and Yong Zhang, *Singular solutions in soft limits*, *Journal of High Energy Physics* **2020** (2020), no. 5, 148.
- [27] Örjan Carlborg and Chris S Haley, *Epistasis: too often neglected in complex trait studies?*, *Nature reviews Genetics* **5** (2004), no. 8, 618–625.
- [28] Ted J. Case and Edward A. Bender, *Testing for Higher Order Interactions*, *The American Naturalist* **118** (1981), no. 6, 920–929.
- [29] Zachary Chroman and Mihir Singhal, *Computations associated with the resonance arrangement*, arXiv:2106.09940 (2021).
- [30] Sean R Collins, Kyle M Miller, Nancy L Maas, Assen Roguev, Jeffrey Fillingham, Clement S Chu, Maya Schuldiner, Marinella Gebbia, Judith Recht, Michael Shales, Huiming Ding, Hong Xu, Junhong Han, Kristin Ingvarsdottir, Benjamin Cheng, Brenda Andrews, Charles Boone, Shelley L Berger, Phil Hieter, Zhiguo Zhang, Grant W Brown, C James Ingles, Andrew Emili, C David Allis, David P Toczyski, Jonathan S Weissman, Jack F Greenblatt, and Nevan J Krogan, *Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map*, *Nature* **446** (2007), no. 7137, 806–810.
- [31] Jessika Consuegra, Théodore Grenier, Houssam Akherraz, Isabelle Rahioui, Hugo Gervais, Pedro da Silva, and François Leulier, *Metabolic cooperation among commensal bacteria supports *Drosophila* juvenile growth under nutritional stress*, *ISCIENCE* (2020), 101232.
- [32] Michael Costanzo, Anastasia Baryshnikova, Jeremy Bellay, Yungil Kim, Eric D Spear, Carolyn S Sevier, Huiming Ding, Judice L Y Koh, Kiana Toufighi, Sara Mostafavi, Jeany Prinz, Robert P St Onge, Benjamin VanderSluis, Taras Makhnevych, Franco J Vizeacoumar, Solmaz Alizadeh, Sonda Bahr, Renee L Brost, Yiqun Chen, Murat Cokol, Raamesh Deshpande, Zhijian Li, Zhen-Yuan Lin, Wendy Liang, Michaela Marback, Jadine Paw, Bryan-Joseph San Luis, Ermira Shuteriqi, Amy Hin Yan Tong, Nydia van Dyk, Iain M Wallace, Joseph A Whitney, Matthew T Weirauch, Guoqing Zhong, Hongwei Zhu, Walid A Houry, Michael Brudno,

- Sasan Ragibizadeh, Balázs Papp, Csaba Pál, Frederick P Roth, Guri Giaever, Corey Nislow, Olga G Troyanskaya, Howard Bussey, Gary D Bader, Anne-Claude Gingras, Quaid D Morris, Philip M Kim, Chris A Kaiser, Chad L Myers, Brenda J Andrews, and Charles Boone, *The genetic landscape of a cell*, Science (New York, NY) **327** (2010), no. 5964, 425–431.
- [33] Henry Crapo, *The combinatorial theory of structures*, Matroid theory (Szeged, 1982), Colloq. Math. Soc. János Bolyai, vol. 40, North-Holland, Amsterdam, 1985, pp. 107–213. MR 843374
 - [34] Henry H. Crapo and Gian-Carlo Rota, *On the foundations of combinatorial theory: Combinatorial geometries*, preliminary ed., The M.I.T. Press, Cambridge, Mass.-London, 1970. MR 0290980
 - [35] Kristina Crona, *Rank orders and signed interactions in evolutionary biology*, eLife **9** (2020), 1–12.
 - [36] Kristina Crona, Alex Gavryushkin, Devin Greene, and Niko Beerenwinkel, *Inferring genetic interactions from comparative fitness data*, eLife **6** (2017), e28629.
 - [37] María Angélica Cueto, Jason Morton, and Bernd Sturmfels, *Geometry of the restricted Boltzmann machine*, Algebraic methods in statistics and probability II, Contemp. Math., vol. 516, Amer. Math. Soc., Providence, RI, 2010, pp. 135–153. MR 2730746
 - [38] J. Dauxois and C. Hassenforder, *Toutes les probabilités et les statistiques: cours et exercices corrigés*, Ellipses, 2004.
 - [39] Jesús A. De Loera, Jörg Rambau, and Francisco Santos, *Triangulations*, Algorithms and Computation in Mathematics, vol. 25, Springer-Verlag, Berlin, 2010, Structures for algorithms and applications. MR 2743368
 - [40] Jesús A. De Loera, Jörg Rambau, and Francisco Santos, *Triangulations. Structures for algorithms and applications*, vol. 25, Berlin: Springer, 2010 (English).
 - [41] de Visser and Joachim Krug, *Empirical fitness landscapes and the predictability of evolution*, Nat Rev Genet **15** (2014), no. 7, 480–490.
 - [42] Wolfram Decker, Gert-Martin Greuel, Gerhard Pfister, and Hans Schönemann, SINGULAR 4-2-0 — A computer algebra system for polynomial computations, <http://www.singular.uni-kl.de>, 2020.
 - [43] Antoine Deza and Lionel Pournin, *A linear optimization oracle for zonotope computation*, Comput. Geom. Theory Appl. **100** (2022), no. C.
 - [44] Andreas W. M. Dress, *Trees, tight extensions of metric spaces, and the cohomological dimension of certain groups: a note on combinatorial properties of metric spaces*, Adv. in Math. **53** (1984), no. 3, 321–402. MR 753872
 - [45] Holger Eble, *The MST-fan of a regular subdivision*, arXiv:2205.10424 (2022).
 - [46] Holger Eble, Michael Joswig, Lisa Lamberti, and Will Ludington, *Master regulators of evolution and the microbiome in higher dimensions*, arXiv:2009.12277 (2020).
 - [47] Holger Eble, Michael Joswig, Lisa Lamberti, and William B. Ludington, *Cluster partitions and fitness landscapes of the Drosophila fly microbiome*, J. Math. Biol. **79** (2019), no. 3, 861–899. MR 3987182
 - [48] Herbert Edelsbrunner and John L. Harer, *Computational topology*, American Mathematical Society, Providence, RI, 2010, An introduction. MR 2572029
 - [49] Tim Evans, *What is being calculated with thermal field theory?*, pp. 343–352, World Scientific, 1995.
 - [50] B. S. Everitt and A. Skrondal, *The Cambridge dictionary of statistics*, fourth ed., Cambridge University Press, Cambridge, 2010. MR 2723410

- [51] Gang Fang, Wen Wang, Vanja Paunic, Hamed Heydari, Michael Costanzo, Xiaoye Liu, Xiaotong Liu, Benjamin Vandersluis, Benjamin Oatley, Michael Steinbach, Brian Van Ness, Eric E Schadt, Nathan D Pankratz, Charles Boone, Vipin Kumar, and Chad L Myers, *Discovering genetic interactions bridging pathways in genome-wide association studies*, Nature Communications **10** (2019), no. 4274, 1–18.
- [52] Claus Fieker, William Hart, Tommy Hofmann, and Fredrik Johansson, *Nemo/Hecke: Computer algebra and number theory packages for the julia programming language*, Proceedings of the 2017 ACM on International Symposium on Symbolic and Algebraic Computation (New York, NY, USA), ISSAC '17, ACM, 2017, pp. 157–164.
- [53] Ronald A. Fisher, *The correlation between relatives on the supposition of Mendelian inheritance*, Trans. Roy. Soc. Edinb. **52** (1918), 399–433.
- [54] Jonathan Friedman, Logan M Higgins, and Jeff Gore, *Community structure follows simple assembly rules in microbial microcosms*, Nature Publishing Group **1** (2017), 1–7.
- [55] *GAP – Groups, Algorithms, and Programming, Version 4.10.2*, <https://www.gap-system.org>, Jun 2019.
- [56] Ewgenij Gawrilow and Michael Joswig, *polymake: a framework for analyzing convex polytopes*, Polytopes—combinatorics and computation (Oberwolfach, 1997), DMV Sem., vol. 29, Birkhäuser, Basel, 2000, pp. 43–73. MR MR1785292 (2001f:52033)
- [57] ———, *polymake: a framework for analyzing convex polytopes*, Polytopes—combinatorics and computation (Oberwolfach, 1997), DMV Sem., vol. 29, Birkhäuser, Basel, 2000, pp. 43–73. MR 1785292
- [58] James E. Gentle, *Elements of computational statistics*, Statistics and Computing, Springer-Verlag, New York, 2002. MR 1948588
- [59] J. L. Gill, *Effects of Finite Size on Selection Advance in Simulated Genetic*, Australian journal of biological sciences **18** (1965), no. 1508, 599–617.
- [60] Benjamin H. Good, Michael J. McDonald, Jeffrey E. Barrick, Richard E. Lenski, and Michael M. Desai, *The dynamics of molecular evolution over 60,000 generations*, Nature **551** (2017), no. 7678, 45–50.
- [61] Alison L. Gould, Vivian Zhang, Lisa Lamberti, Eric W. Jones, Benjamin Obadia, Nikolaos Korasidis, Alex Gavryushkin, Jean M. Carlson, Niko Beerenwinkel, and William B. Ludington, *Microbiome interactions shape host fitness*, Proceedings of the National Academy of Sciences **115** (2018), no. 51, E11951–E11960.
- [62] Jacopo Grilli, György Barabás, Matthew J Michalska-Smith, and Stefano Allesina, *Higher-order interactions stabilize dynamics in competitive network models*, Nature (2017), 1–5.
- [63] Samuel C. Gutekunst, Karola Mészáros, and T. Kyle Petersen, *Root cones and the resonance arrangement*, Electron. J. Combin. **28** (2021), no. 1, Paper No. 1.12, 39. MR 4245245
- [64] Ingileif B. Hallgrímsdóttir and Debbie S. Yuster, *A complete classification of epistatic two-locus models*, BMC Genetics **9** (2008), no. 1, 17.
- [65] Dan Halperin and Micha Sharir, *Arrangements*, Handbook of discrete and computational geometry, 3rd edition, CRC Press, 2017, pp. 49–119.
- [66] Michael Held and Richard M. Karp, *The traveling-salesman problem and minimum spanning trees*, Operations Res. **18** (1970), 1138–1162. MR 278710
- [67] ———, *The traveling-salesman problem and minimum spanning trees. II*, Math. Programming **1** (1971), no. 1, 6–25. MR 289119

- [68] Sílvia F Henriques, Darshan B Dhakan, Lúcia Serra, Ana Patrícia Francisco, Zita Carvalho-Santos, Célia Baltazar, Ana Paula Elias, Margarida Anjos, Tong Zhang, Oliver D K Maddocks, and Carlos Ribeiro, *Metabolic cross-feeding in imbalanced diets allows gut microbes to improve reproduction and alter host behaviour*, Nature Communications **11** (2020), no. 1, 4236.
- [69] Sven Herrmann and Michael Joswig, *Splitting polytopes*, Münster J. Math. **1** (2008), 109–141.
- [70] ———, *Totally splittable polytopes*, Discrete Comput. Geom. **44** (2010), no. 1, 149–166. MR 2639822
- [71] Sven Herrmann, Michael Joswig, and David Speyer, *Dressians, tropical Grassmannians, and their rays*, Forum Math. **26** (2014), no. 6, 1853–1882.
- [72] Sven Herrmann, Michael Joswig, and David E. Speyer, *Dressians, tropical Grassmannians, and their rays*, Forum Math. **26** (2014), no. 6, 1853–1881. MR 3334049
- [73] Peter Huggins, Bernd Sturmfels, Josephine Yu, and Debbie S. Yuster, *The hyperdeterminant and triangulations of the 4-cube*, Mathematics of Computation **77** (2008), no. 263, 1653–1679.
- [74] Peter Huggins, Bernd Sturmfels, Josephine Yu, and Debbie S. Yuster, *The hyperdeterminant and triangulations of the 4-cube*, Math. Comp. **77** (2008), no. 263, 1653–1679. MR 2398786
- [75] June Huh and Eric Katz, *Log-concavity of characteristic polynomials and the Bergman fan of matroids*, Math. Ann. **354** (2012), no. 3, 1103–1116. MR 2983081
- [76] Kabir Husain and Arvind Murugan, *Physical Constraints on Epistasis*, Molecular Biology and Evolution **37** (2020), no. 10, 2865–2874.
- [77] John R. Isbell, *Six theorems about injective metric spaces*, Comment. Math. Helv. **39** (1964), 65–76. MR 182949
- [78] Hiroshi C. Ito and Akira Sasaki, *Evolutionary branching in distorted trait spaces*, Journal of Theoretical Biology **489** (2020), 110152.
- [79] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, *An introduction to statistical learning: With applications in r*, Springer Publishing Company, Incorporated, 2014.
- [80] C. Jefferson, *ferret, backtrack search in permutation groups, Version 1.0.2*, <https://gap-packages.github.io/ferret/>, Jan 2019, GAP package.
- [81] C. Jefferson, M. Pfeiffer, R. Waldecker, and E. Jonauskyte, *images, minimal and canonical images, Version 1.3.0*, <https://gap-packages.github.io/images/>, Mar 2019, GAP package.
- [82] Christopher Jefferson, Eliza Jonauskyte, Markus Pfeiffer, and Rebecca Waldecker, *Minimal and canonical images*, J. Algebra **521** (2019), 481–506. MR 3906181
- [83] Anders Nedergaard Jensen, *Traversing symmetric polyhedral fans*, Mathematical Software – ICMS 2010 (Berlin, Heidelberg) (Komei Fukuda, Joris van der Hoeven, Michael Joswig, and Nobuki Takayama, eds.), Springer Berlin Heidelberg, 2010, pp. 282–294.
- [84] Charles Jordan, Michael Joswig, and Lars Kastner, *Parallel enumeration of triangulations*, Electron. J. Combin. **25** (2018), no. 3, Paper No. 3.6, 27. MR 3829292
- [85] ———, *Parallel enumeration of triangulations*, Electron. J. Combin. **25** (2018), no. 3, Paper No. 3.6, 27. MR 3829292
- [86] Michael Joswig, *Essentials of tropical combinatorics*, Graduate Studies in Mathematics, vol. 219, American Mathematical Society, Providence, RI, 2021.

- [87] Marek Kaluba, Benjamin Lorenz, and Sascha Timme, *Polymake.jl: A new interface to polymake*, Mathematical Software – ICMS 2020 (Cham) (Anna Maria Bigatti, Jacques Carette, James H. Davenport, Michael Joswig, and Timo de Wolff, eds.), Springer International Publishing, 2020, pp. 377–385.
- [88] Hidehiko Kamiya, Akimichi Takemura, and Hiroaki Terao, *Ranking patterns of unfolding models of codimension one*, Adv. in Appl. Math. **47** (2011), no. 2, 379–400. MR 2803809
- [89] Lars Kastner and Marta Panizzut, *Hyperplane arrangements in polymake*, Mathematical software – ICMS 2020 (Anna Maria Bigatti, Jacques Carette, James H. Davenport, Michael Joswig, and Timo de Wolff, eds.), Lecture Notes in Computer Science, vol. 12097, Springer, 2020, pp. 232–240.
- [90] S Kauffman and S Levin, *Towards a general theory of adaptive walks on rugged landscapes.*, Journal of Theoretical Biology **128** (1987), no. 1, 11–45.
- [91] Aisha I. Khan, Duy M. Dinh, Dominique Schneider, Richard E. Lenski, and Tim F. Cooper, *Negative epistasis between beneficial mutations in an evolving bacterial population*, Science **332** (2011), no. 6034, 1193–1196.
- [92] Caroline J. Klivans and Ed Swartz, *Projection volumes of hyperplane arrangements*, Discrete Comput. Geom. **46** (2011), no. 3, 417–426. MR 2826960
- [93] Hiroshi Koizumi, Yasuhide Numata, and Akimichi Takemura, *On intersection lattices of hyperplane arrangements generated by generic points*, Ann. Comb. **16** (2012), no. 4, 789–813. MR 3000446
- [94] Julia H Kreznar, Mark P Keller, Lindsay L Traeger, Mary E Rabaglia, Kathryn L Schueler, Donald S Stapleton, Wen Zhao, Eugenio I Vivas, Brian S Yandell, Aimee Teo Broman, Bruno Hagenbuch, Alan D Attie, and Federico E Rey, *Host Genotype and Gut Microbiome Modulate Insulin Secretion and Diet-Induced Metabolic Phenotypes*, CellReports **18** (2017), no. 7, 1739–1750.
- [95] Joachim Krug, *Epistasis and evolution*, 2021.
- [96] Joseph B. Kruskal, Jr., *On the shortest spanning subtree of a graph and the traveling salesman problem*, Proc. Amer. Math. Soc. **7** (1956), 48–50. MR 78686
- [97] Lukas Kühne, *The universality of the resonance arrangement and its Betti numbers*, arXiv:2008.10553 (2020).
- [98] Elena Kuzmin, Benjamin VanderSluis, Wen Wang, Guihong Tan, Raamesh Deshpande, Yiqun Chen, Matej Usaj, Attila Balint, Mojca Mattiazzi Usaj, Jolanda van Leeuwen, Elizabeth N. Koch, Carles Pons, Andrius J. Dagilis, Michael Pryszlak, Jason Zi Yang Wang, Julia Hanchard, Margot Riggi, Kaicong Xu, Hamed Heydari, Bryan-Joseph San Luis, Ermira Shuteriqi, Hongwei Zhu, Nydia Van Dyk, Sara Sharifpoor, Michael Costanzo, Robbie Loewith, Amy Caudy, Daniel Bolnick, Grant W. Brown, Brenda J. Andrews, Charles Boone, and Chad L. Myers, *Systematic analysis of complex genetic interactions*, Science **360** (2018), no. 6386, eaao1729.
- [99] Hye-Yeon Lee, Shin-Hae Lee, Ji-Hyeon Lee, Won-Jae Lee, and Kyung-Jin Min, *The role of commensal microbes in the lifespan of Drosophila melanogaster*, Aging **11** (2019), no. 13, 4611–4640.
- [100] Martin Leuner, *alcove*, <https://github.com/martin-leuner/alcove>, 2019.
- [101] R E Ley, M Hamady, C Lozupone, P J Turnbaugh, R R Ramey, J S Bircher, M L Schlegel, T A Tucker, M D Schrenzel, R Knight, and J I Gordon, *Evolution of Mammals and Their Gut Microbes*, Science (New York, NY) **320** (2008), no. 5883, 1647–1651.
- [102] Caitlin Lienkaemper, Lisa Lamberti, James Drain, Niko Beerenwinkel, and Alex Gavryushkin, *The geometry of partial fitness orders and an efficient method for detecting genetic interactions*, Journal of Mathematical Biology **77** (2018), no. 4, 951–970.

- [103] Xuanyao Liu, Yang I Li, and Jonathan K Pritchard, *Trans Effects on Gene Expression Can Drive Omnigenic Inheritance.*, Cell **177** (2019), no. 4, 1022–1034.e6.
- [104] William B. Ludington and William W. Ja, *Drosophila as a model for the gut microbiome*, PLOS Pathogens **16** (2020), no. 4, 1–6.
- [105] Diane Maclagan and Bernd Sturmfels, *Introduction to tropical geometry*, Graduate Studies in Mathematics, vol. 161, American Mathematical Society, Providence, RI, 2015. MR 3287221
- [106] ———, *Introduction to tropical geometry*, Graduate Studies in Mathematics, vol. 161, American Mathematical Society, Providence, RI, 2015. MR 3287221
- [107] David M. McCandlish, *Long-term evolution on complex fitness landscapes when mutation is weak*, Heredity **121** (2018), no. 5, 449–465.
- [108] Michael J. McDonald, Daniel P. Rice, and Michael M. Desai, *Sex speeds adaptation by altering the dynamics of molecular evolution*, Nature **531** (2016), no. 7593, 233–236.
- [109] Tilman Möller and Gerhard Röhrle, *Counting chambers in restricted Coxeter arrangements*, Arch. Math. (Basel) **112** (2019), no. 4, 347–359. MR 3928360
- [110] Guido Montúfar, Nihat Ay, and Keyan Ghazi-Zahedi, *Geometry and expressive power of conditional restricted boltzmann machines*, J. Mach. Learn. Res. **16** (2015), no. 73, 2405–2436.
- [111] Guido F. Montúfar and Jason Morton, *When does a mixture of products contain a product of mixtures?*, SIAM J. Discrete Math. **29** (2015), no. 1, 321–347. MR 3310972
- [112] Navid Nabijou and Dhruv Ranganathan, *Gromov-Witten theory with maximal contacts*, 2021.
- [113] Peter Orlik and Louis Solomon, *Combinatorics and topology of complements of hyperplanes*, Invent. Math. **56** (1980), no. 2, 167–189. MR 558866
- [114] Peter Orlik and Hiroaki Terao, *Arrangements of hyperplanes*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 300, Springer-Verlag, Berlin, 1992. MR 1217488
- [115] *OSCAR Computer Algebra System, Version 0.5.2*, <https://oscar.computeralgebra.de>, April 2021.
- [116] James Oxley, *Matroid theory*, second ed., Oxford Graduate Texts in Mathematics, vol. 21, Oxford University Press, Oxford, 2011. MR 2849819
- [117] Megha Padi and John Quackenbush, *Integrating transcriptional and protein interaction networks to prioritize condition-specific master regulators*, BMC Systems Biology **9** (2015), no. 1, 1–17.
- [118] R. T. Paine, *A note on trophic complexity and community stability*, The American Naturalist **103** (1969), no. 929, 91–93.
- [119] R T Paine, *Food-web analysis through field measurement of per capita interaction strength*, Nature (1992), 1–3.
- [120] Rudi Pendavingh and Jorn van der Pol, *Asymptotics of symmetry in matroids*, J. Combin. Theory Ser. B **135** (2019), 349–365. MR 3926274
- [121] Patrick C. Phillips, *Epistasis - the essential role of gene interactions in the structure and evolution of genetic systems*, Nat Rev Genet **9** (2008), no. 11, 855–867.
- [122] Philippe Piccardi, Björn Vessman, and Sara Mitri, *Toxicity drives facilitation between 4 bacterial species*, Proceedings of the National Academy of Sciences **116** (2019), no. 32, 15979–15984.

- [123] Alexander Postnikov and Richard P. Stanley, *Deformations of Coxeter hyperplane arrangements*, J. Combin. Theory Ser. A **91** (2000), no. 1-2, 544–597. MR 1780038
- [124] Lionel Pournin, *The flip-graph of the 4-dimensional cube is connected*, Discrete Comput. Geom. **49** (2013), no. 3, 511–530. MR 3038527
- [125] Christoph Ratzke, Julien Barrere, and Jeff Gore, *Strength of species interactions determines biodiversity and stability in microbial communities*, Nature Ecology & Evolution (2020), 1–21.
- [126] Michael Reimann, Max Nolte, Martina Scolamiero, Katharine Turner, Rodrigo Perin, Giuseppe Chindemi, Pawel Dlotko, Ran Levi, Kathryn Hess, and Henry Markram, *Cliques of Neurons Bound into Cavities Provide a Missing Link between Structure and Function*, Frontiers in Computational Neuroscience **11** (2017), 1–16.
- [127] Alice Risely, *Applying the core microbiome to understand host–microbe systems*, Journal of Animal Ecology **89** (2020), no. 7, 1549–1558.
- [128] Zachary R Sailer and Michael J Harms, *Detecting High-Order Epistasis in Nonlinear Genotype-Phenotype Maps.*, Genetics **205** (2017), no. 3, 1079–1088.
- [129] Zachary R. Sailer and Michael J. Harms, *High-order epistasis shapes evolutionary trajectories*, PLOS Computational Biology **13** (2017), no. 5, 1–16.
- [130] Alicia Sanchez-Gorostiaga, Djordje Bajić, Melisa L. Osborne, Juan F. Poyatos, and Alvaro Sanchez, *High-order interactions distort the functional landscape of microbial consortia*, PLoS Biology **17** (2019), no. 12, 1–34.
- [131] Jeffery Sarnoff, *SaferInteger, julia package, version 2.5.3*, <https://github.com/JeffreySarnoff/SaferIntegers.jl>, 2021.
- [132] M.J. Schervish, *Theory of statistics*, Springer Series in Statistics, Springer New York, 1996.
- [133] Alexander Schrijver, *Combinatorial optimization. Polyhedra and efficiency. Vol. B, Algorithms and Combinatorics*, vol. 24, Springer-Verlag, Berlin, 2003, Matroids, trees, stable sets, Chapters 39–69. MR 1956925
- [134] Maya Schuldiner, Sean R Collins, Natalie J Thompson, Vladimir Denic, Arunashree Bhamidipati, Thanuja Punna, Jan Ihmels, Brenda Andrews, Charles Boone, Jack F Greenblatt, Jonathan S Weissman, and Nevan J Krogan, *Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile*, Cell **123** (2005), no. 3, 507–519.
- [135] Molly Schumer, Chenling Xu, Daniel L Powell, Arun Durvasula, Laurits Skov, Chris Holland, John C Blazier, Sriram Sankararaman, Peter Andolfatto, Gil G Rosenthal, and Molly Przeworski, *Natural selection interacts with recombination to shape the evolution of hybrid genomes*, Science **360** (2018), no. May 11, 656–660.
- [136] Raimund Seidel, *Convex hull computations*, Handbook of Discrete and Computational Geometry (Csaba D. Tóth, Jacob E. Goodman, and Joseph O’Rourke, eds.), CRC Press, 2018, 3rd edition.
- [137] Nora Sleumer, *Output-sensitive cell enumeration in hyperplane arrangements*, Algorithm theory—SWAT’98 (Stockholm), Lecture Notes in Comput. Sci., vol. 1432, Springer, Berlin, 1998, pp. 300–309. MR 1678416
- [138] John Maynard Smith, *Natural selection and the Concept of a Protein Space*, Nature **225** (1970), no. February 7, 563–564.
- [139] L. Solomon and H. Terao, *A formula for the characteristic polynomial of an arrangement*, Adv. in Math. **64** (1987), no. 3, 305–325. MR 888631
- [140] W. A. Stein et al., *Sage Mathematics Software (Version x.y.z)*, The Sage Development Team, 2021, <http://www.sagemath.org>.

- [141] John M. Sullivan, *Curvatures of smooth and discrete surfaces*, Discrete differential geometry, Oberwolfach Semin., vol. 38, Birkhäuser, Basel, 2008, pp. 175–188. MR 2405666
- [142] Deepika Sundarraman, Edouard A. Hay, Dylan M. Martins, Drew S. Shields, Noah L. Pettinari, and Raghuvveer Parthasarathy, *Higher-order interactions dampen pairwise competition in the zebrafish gut microbiome*, mBio **11** (2020), no. 5, 1–15.
- [143] Deepika Sundarraman, Edouard A Hay, Dylan M Martins, Drew S Shields, Noah L Pettinari, and Raghuvveer Parthasarathy, *Quantifying multi-species microbial interactions in the larval zebrafish gut*, bioRxiv (2020), 1–23.
- [144] Longzhi Tan, Stephen Serene, Hui Xiao Chao, and Jeff Gore, *Hidden randomness between fitness landscapes limits reverse evolution*, Phys. Rev. Lett. **106** (2011), 198102.
- [145] Michail Tsagris, Christina Beneki, and Hossein Hassani, *On the folded normal distribution*, Mathematics **2** (2014), no. 1, 12–28.
- [146] Leigh Van Valen, *Molecular evolution as predicted by natural selection*, Journal of Molecular Evolution **3** (1974), no. 2, 89–101.
- [147] Kavitha Venkatesan, Jean-François Rual, Alexei Vazquez, Ulrich Stelzl, Irma Lemmens, Tomoko Hirozane-Kishikawa, Tong Hao, Martina Zenkner, Xiaofeng Xin, Kwang-Il Goh, Muhammed A Yildirim, Nicolas Simonis, Kathrin Heinzmann, Fana Gebreab, Julie M Sahalie, Sebiha Cevik, Christophe Simon, Anne-Sophie de Smet, Elizabeth Dann, Alex Smolyar, Arunachalam Vinayagam, Haiyuan Yu, David Szeto, Heather Borick, Amélie Dricot, Niels Klitgord, Ryan R Murray, Chenwei Lin, Maciej Lalowski, Jan Timm, Kirstin Rau, Charles Boone, Pascal Braun, Michael E Cusick, Frederick P Roth, David E Hill, Jan Tavernier, Erich E Wanker, Albert-László Barabási, and Marc Vidal, *An empirical framework for binary interactome mapping*, Nature methods **6** (2009), no. 1, 83–90.
- [148] E. D. Weinberger, *Fourier and Taylor series on fitness landscapes*, Biological Cybernetics **65** (1991), no. 5, 321–330.
- [149] ———, *Fourier and Taylor series on fitness landscapes*, Biological Cybernetics **65** (1991), no. 5, 321–330.
- [150] Daniel Weinreich, Richard Watson, and Lin Chao, *Perspective: Sign epistasis and genetic constraint on evolutionary trajectories*, Evolution **59** (2007), 1165 – 1174.
- [151] Daniel M Weinreich, Nigel F Delaney, Mark A Depristo, and Daniel L Hartl, *Darwinian evolution can follow only very few mutational paths to fitter proteins.*, Science (New York, NY) **312** (2006), no. 5770, 111–114.
- [152] Daniel M Weinreich, Yinghong Lan, Jacob Jaffe, and Robert B Heckendorn, *The influence of higher-order epistasis on biological fitness landscape topography*, Journal of statistical physics **172** (2018), no. 1, 208–225.
- [153] ———, *The Influence of Higher-Order Epistasis on Biological Fitness Landscape Topography*, Journal of Statistical Physics **172** (2018), no. 1, 208–225.
- [154] Daniel M. Weinreich, Yinghong Lan, C Scott Wylie, and Robert B. Heckendorn, *Should evolutionary geneticists worry about higher-order epistasis?*, Current Opinion in Genetics & Development **23** (2013), no. 6, 700 – 707, Genetics of system biology.
- [155] Daniel M. Weinreich, Richard A. Watson, and Lin Chao, *Perspective: Sign epistasis and genetic constraint on evolutionary trajectories*, Evolution **59** (2005), no. 6, 1165–1174.
- [156] Sewall Wright, *The roles of mutation, inbreeding, crossbreeding and selection in evolution*, Proceedings of the Sixth International Congress of Genetics **1** (1932), 356–366.
- [157] Nicholas C Wu, Lei Dai, C Anders Olson, James O Lloyd-Smith, and Ren Sun, *Adaptation in protein fitness landscapes is facilitated by indirect paths*, eLife **5** (2016), e16965.

- [158] Stephan Wulschleger, Robbie Loewith, and Michael N. Hall, *TOR signaling in growth and metabolism*, Cell **124** (2006), no. 3, 471–484.
- [159] Masahiko Yoshinaga, *Hyperplane arrangements and Lefschetz’s hyperplane section theorem*, Kodai Math. J. **30** (2007), no. 2, 157–194. MR 2343416
- [160] ———, *Freeness of hyperplane arrangements and related topics*, Ann. Fac. Sci. Toulouse Math. (6) **23** (2014), no. 2, 483–512. MR 3205600
- [161] Thomas Zaslavsky, *Facing up to arrangements: face-count formulas for partitions of space by hyperplanes*, Mem. Amer. Math. Soc. **1** (1975), no. issue 1, 154, vii+102. MR 0357135
- [162] Liwen Zhang, Gregory Naitzat, and Lek-Heng Lim, *Tropical geometry of Deep Neural Networks*, Proceedings of the 35th International Conference on Machine Learning (Jennifer Dy and Andreas Krause, eds.), Proceedings of Machine Learning Research, vol. 80, PMLR, 10–15 Jul 2018, pp. 5824–5832.
- [163] Juannan Zhou and David M. McCandlish, *Minimum epistasis interpolation for sequence-function relationships*, Nature Communications **11** (2020), no. 1, 1782.
- [164] Yu. A. Zuev, *Methods of geometry and probabilistic combinatorics in threshold logic*, Discrete Math. Appl. **2** (1992), no. 4, 427 – 438.
- [165] J. Zunic, *On encoding and enumerating threshold functions*, IEEE Transactions on Neural Networks **15** (2004), no. 2, 261–267.