CLUSTER PARTITIONS AND FITNESS LANDSCAPES OF 
THE DROSOPHILA FLY MICROBIOME

HOLGER EBLE, MICHAEL JOSWIG, LISA LAMBERTI, AND WILL LUDINGTON

Abstract. Beerenwinkel et al. (2007) suggested studying fitness land- 
scapes via regular subdivisions of convex polytopes. Building on their 
approach we propose cluster partitions and cluster filtrations of fitness 
landscapes as a new mathematical tool. In this way, we provide a con- 
cise combinatorial way of processing metric information from epistatic 
interactions. Using existing Drosophila microbiome data, we demon- 
strate similarities with and differences to the previous approach. As one 
outcome we locate interesting epistatic information where the previous 
approach is less conclusive.

Contents

1. Introduction 1
2. Mathematical background and terminology 3
3. Clusters in fitness landscapes 7
4. Significant cluster partitions 15
5. Epistasis, interaction coordinates and circuits 22
6. Epistasis in Drosophila melanogaster fruit fly microbiomes 25
Acknowledgement 30
References 30

1. Introduction

In evolutionary biology, the concept of a fitness landscape plays a promi-
nent role in the study of genetic mutations, evolutionary pathways and be-

yond [7]. These landscapes were introduced by Sewall Wright in [16] and typ-

cally 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 genot
tope. Fitness landscapes then arise by associating the reproductive suc-

cess (or other fitness traits) to genotypes. Shapes of fitness landscapes are 
regular subdivisions of the genotope induced by genotype–phenotype mappings and have been studied by Beerenwinkel et al. [2] [3]. Such subdivisions 
play a key role in determining interaction patterns among altered genes and

Research by M. Joswig is carried out in the framework of Matheon supported by Ein-
stein Foundation Berlin. Further partial support by Deutsche Forschungsgemeinschaft 
(SFB-TRR 109: “Discretization in Geometry and Dynamics” and SFB-TRR 195: “Sym-
bolic Tools in Mathematics and their Application”) is gratefully acknowledged.
pathways on the genotope, and they shed light on the possible orders in which genetic mutations might occur. The study of interaction patterns is a general one that applies also to economies, social networks, and foodwebs in ecology. Recently the gut microbiome has arisen in biology as a major factor shaping the genotype–phenotype mappings in animals. 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. Because the number of species in the microbiome is on the order of hundreds to thousands, the number of interaction patterns is huge in general and there is a big 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. [2,3] and propose such a new method to process metric interaction information, also called 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 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 program is that the filtration process is designed to 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. [2,3]. Cluster partitions and their filtrations are a new mathematical idea, and we hope that this will find applications also beyond *Drosophila* data.

To describe the strength of our approach we consider fitness landscapes for microbiome-modified *Drosophila* flies and 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 such a way that the genotope is given by a 5-dimensional cube. The fitness landscapes we consider are defined by daily fecundity (referred to as fec), time to death (ttt) 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 we obtain in this way among fitness landscapes we use cluster partitions. Our results show that cluster filtrations detect interactions when these are present in the sense of [2,3] but additionally locate statistically relevant and previously undetected epistatic
Comparing our findings with previous studies confirms that cluster filtrations can also be used to strengthen existing analysis and to prompt new possible conclusions in interaction networks.

2. Mathematical background and terminology

Our approach relies on the theoretical framework for revealing epistatic interactions in genetic systems given in [2, 3], and we use the same terminology. The theory of regular subdivisions is developed in the monograph [6].

2.1. Genotopes 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 [2, 3] 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 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 $v$ 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 $v$ 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 subdivision of $P$ into polyhedral cells. This is the regular subdivision $S = S(V, h)$ of $P$ induced by $h$; cf. Definition 2.3.1 and Lemma 2.3.11 in [6]. A polyhedral subdivision for which all cells are simplices is called a triangulation. Triangulations 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 [2 §3]. The genotope subdivision $S(V, h)$ 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. As such we can assume these fittings to be generic, so they induce triangulations of the $n$-cube $[0,1]^n$.

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.** 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, e.g., 01 is shorthand notation for the point with coordinates $(0,1)$. Assume that some measurement gives the height function $h$ which reads

\[(1) \quad h(00) = 53.25 \quad h(10) = 46.65 \quad h(01) = 43.16 \quad h(11) = 43.48.\]

The lifted point 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 shows the upper facets of $P(h)$ and the resulting genotope subdivision $S(V,h)$. The latter is a triangulation with two maximal cells.

**Remark 2.** Most polytopes admit triangulations which are not regular, i.e., not induced by any height function; cf. [6, 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 also apply to the nonregular setting.
2.2. Fittest populations. Again we consider the genotypes \( V = \{0, 1\}^n \) of an \( n \)-biallelic system. A map \( p : V \to \mathbb{R} \) 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 genotope \([0, 1]^n\), which is the allele frequency vector. Its coordinate \( \rho_i \) describes the probability for the population \( p \) to have allele 1 in its \( i \)-th locus. The set of all relative populations, denoted as \( \Delta_V \), is a simplex of dimension \( 2^n - 1 \).

Now we add the height function \( h : V \to \mathbb{R} \) to the picture. For a fixed allele frequency vector \( w \in [0, 1]^n \) this gives rise to the linear program
\[
\text{(LP}(h, w)) \quad \text{maximize} \quad h \cdot p \\
\text{subject to} \quad p \in \Delta_V \text{ and } \rho(p) = w
\]
in the variable \( p \). 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 \( \Delta_V \), and the projection \( \rho(p^*) = w \) gives rise to a representation of \( w = \lambda_1 v_1 + \cdots + \lambda_{n+1} v_{n+1} \), with \( \lambda_i \geq 0 \) and \( \sum \lambda_i = 1 \), as the convex combination of at most \( n + 1 \) genotypes \( v_1, \ldots, v_{n+1} \in V \). These genotypes are precisely the vertices of the unique simplex \( s \) of \( S(V, h) \) which contains \( w \). The genericity of \( h \) implies that \( S(V, h) \) 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
\[
h^* : [0, 1]^n \longrightarrow \mathbb{R} \\
w \longmapsto h \cdot p^*(h, w)
\]
on the genotope. Now the regions of linearity of \( h^* \) coincide with the maximal cells of the regular triangulation \( S(V, h) \).

Remark 3. Finally we can explain our preference of the upper versus the lower facets in the definition of a regular subdivision. This is because we implicitly assume that higher values of \( h \) are better. That is to say, the fittest population is chosen via \( \text{LP}(h, w) \), which is a maximization problem. This convention is consistent with previously adopted ones, see [11].

2.3. The dual graph of a subdivision and its complexity. Let \( S \) be a polyhedral subdivision of some \( n \)-dimensional polytope \( P \). Two maximal cells of \( S \) 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 \( \Gamma(S) \), and we denote it by \( \Gamma(S) \). It is easy to see that \( \Gamma(S) \) is always connected. The edges of \( \Gamma(S) \) are the dual edges of \( S \).
The lemma below gives essential complexity bounds in the case of most interest to us. It is a special case of [6, Theorem 2.6]. Let us denote the minimal number of maximal cells of any triangulation of $[0,1]^n$ by $k_*(n)$.

**Lemma 4.** Let $S$ be a triangulation of the unit cube $[0,1]^n$, and let $k$ be the number of nodes of $\Gamma(S)$. Then

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

The lower bound is attained if and only if $\Gamma(S)$ 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 cube $[0,1]^n$ has $2^n$ vertices and an $n$-dimensional simplex has $n+1$ vertices. Let $s$ be a maximal cell of $S$; i.e., $s$ is an $n$-simplex. Then any adjacent simplex shares $n$ vertices with the previous one, thus these two simplices together occupy $n+2$ of the $2^n$ vertices. Any subsequent $n$-simplex will either use only vertices that were already there or insert a single new one. In particular, we need at least $2^n - n$ simplices other than $s$ to insert all vertices and the lower bound follows.

Further, the lower bound is attained if and only if every simplex after the first one indeed inserts a new vertex, which is easily seen to be equivalent to not having cycles in the dual graph (if there is a cycle, the last simplex of the cycle that we insert has two facets in common with previous simplices, hence it introduces no new vertex).

The Euclidean volume of any $n$-simplex spanned by $n+1$ vertices of $[0,1]$ is at least $1/n!$. So the upper bound follows from the fact that the volume of $[0,1]^n$ equals one.

Each $n$-simplex is adjacent to at most $n+1$ other simplices as this is the number of its facets. Yet $S$ induces a triangulation on each of the $2n$ facets of $[0,1]$. Therefore, at least $2n \cdot k_*(n-1)$ of the $k(n+1)$ cells of dimension $n-1$ in $S$ lie in the boundary of $[0,1]$. 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 $\Gamma(S)$. Dividing by two gives the upper bound on the number of dual edges. \qed

While the lower bound in (2) is only $2^4 - 4 = 12$ we have $k_*(4) = 16$; cf. [6, Example 6.3.14]. To determine the exact lower bound $k_*(n)$ for $n \geq 8$ is a difficult open problem; cf. [6, §6.3.3] and Table 1 for an overview.

| $n$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|-----|---|---|---|---|---|---|---|---|
| $k_*(n)$ | 1 | 2 | 5 | 16 | 67 | 308 | 1493 | $\leq 11\,944$ |

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

The proof of Lemma 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
Lemma 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
\[
\frac{2^n n!}{(n+1)^{(n+1)/2}} \leq k_*(n)
\]
arises from the Hadamard bound for matrices with ±1 coefficients; cf. [6, §6.3.3]. Starting from $n \geq 7$ the bound [6] is better than the more naive lower bound $2^n - n$ from Lemma [4].

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 genotpe $[0,1]^n$ 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.

3.1. Epistatic weight. As before we will consider an arbitrary $n$-polytope $P$ in $\mathbb{R}^n$, equipped with a generic height function $h$. This induces a regular triangulation $S = S(V,h)$, where $V$ is the vertex set of $P$. Let $s$ and $t$ be two adjacent $n$-simplices in $S$. Then there are altogether $n+2$ vertices $v_1, v_2, \ldots, v_{n+2}$ of $P$ such that
\[
s = \text{conv}\{v_1, v_2, \ldots, v_{n+1}\} \quad \text{and} \quad t = \text{conv}\{v_2, v_3, \ldots, v_{n+2}\}.
\]

For two adjacent $n$-simplices, we consider the $(n+2) \times (n+2)$-matrix
\[
E_h(s,t) := \begin{pmatrix}
1 & v_{1,1} & v_{1,2} & \cdots & v_{1,n} & h(v_1) \\
1 & v_{2,1} & v_{2,2} & \cdots & v_{2,n} & h(v_2) \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
1 & v_{n+1,1} & v_{n+1,2} & \cdots & v_{n+1,n} & h(v_{n+1})
\end{pmatrix},
\]

where $v_{1,1}, v_{1,2}, \ldots, v_{n,n}$ 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
\[
\epsilon_h(s,t) := \left| \frac{\det E_h(s,t)}{\text{nvol } s \cdot \text{nvol } t} \right|,
\]

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 $v_2, v_3, \ldots, v_{n+1}$ and separates the two satellite vertices $v_1$ and $v_{n+2}$. 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.

**Example 6.** We continue the Example 1 where $P = [0, 1]^2$, and the height function $h$ is given by (1). The induced triangulation has precisely two maximal cells, and the dual graph has a single edge which connects these two cells. We compute the epistatic weight

$$e_h([00, 10, 11], [00, 01, 11]) = \det \begin{pmatrix} 1 & 0 & 0 & 53.25 \\ 1 & 1 & 0 & 46.65 \\ 1 & 0 & 1 & 43.16 \\ 1 & 1 & 1 & 43.48 \end{pmatrix} = 6.92.$$  

Notice that the denominator in (5) is one; both triangles are unimodular. For this 2-locus system, the computation (6) thus agrees with the usual epistasis formula:

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

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

**Remark 7.** 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 below solely rests on the epistatic weights from (5); 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 $\Gamma(S)$ 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 $\Gamma(S)$ 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 $\Gamma(S)$ thus encode exactly those $s$-exposed genotypes that can be joined by $n$ genotypes of $s$ to form a fittest population in the above sense.

Finally, the epistatic weight associated with an edge $e = \{s, t\}$ of $\Gamma(S)$ 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. 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 $s$-exposed satellite genotype. When the epistatic weight $e_h(s, t)$ is one there is a perfect linear relationship between the volume of the lifted and underlying point configuration. In the same vein, we observe that paths of lengths two inside $\Gamma(S)$ yield the following epistatic information: knowing the fitness dependency relations of the genotypes occurring in the two fittest populations at both ends of the path does not allow
us to determine fitness dependency relations among all subset of genotypes in their union.

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

Our next goal is to define a filtration process on $S$ by a sequence of nested cluster partitions. For this, consider a threshold value $\theta$, where $\theta \geq 0$. The specification of a threshold value defines a not necessarily connected sub-graph, $\Gamma(\theta)$, of $\Gamma$ by deleting those dual edges whose normalized epistatic weight exceeds $\theta$. The $\Gamma(\theta)$-clusters and the $\Gamma(\theta)$-cluster partition are shortened to $\theta$-clusters and the $\theta$-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 $\epsilon$ such that $|h''(v)| < \epsilon$ for all $v \in V$. Our assumption on $h'$ means that $S(V, h')$ is not a triangulation. We call $h''$ sufficiently generic for $h'$ if $h' + h''$ is generic, i.e., $S(V, h' + h'')$ 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 $h' + (-h') = 0$ is not. Now $\theta$-cluster partitions can undo the perturbation by $h''$ in the following sense.

Theorem 8. For every height function $h'$ there are positive numbers $\epsilon$ and $\theta$ such that the following holds: If $h''$ is sufficiently generic for $h'$ and additionally satisfies $|h''(v)| < \epsilon$ for all $v \in V$, then each maximal cell of $S(V, h')$ corresponds to exactly one $\theta$-cluster of the triangulation $S(V, h' + h'')$.

Proof. First assume that $h'$ is generic, and thus the partition $S = S(V, h')$ induced by $h$ is a triangulation. Pick $\epsilon > 0$ sufficiently small such that for $|h''(v)| < \epsilon$ we have $S(V, h' + h'') = S(V, h')$. That is, $h'$ and $h' + h''$ lie in the same secondary cone. Such an $\epsilon$ 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. [3] §5.2.1]. Then the claim in the generic case becomes trivial, e.g., with $\theta = 0$. The maximal cells of $S(V, h')$ are precisely the $0$-clusters of $S(V, h' + h'')$.

A second easy case arises when $S(V, h')$ is trivial in the sense that there is only one maximal cell, the entire polytope conv$(V)$. Then we can pick $\epsilon > 0$ arbitrary and make $\theta$ sufficiently large such that the $\theta$-cluster partition of $S(V, h' + h'')$ comprises a single cluster.

Now we consider the only interesting case where $h'$ is not generic and $S := S(V, h')$ is a non-trivial regular decomposition of $V$ induced by $h$ which is not necessarily a triangulation. Here we pick $\epsilon_1 > 0$ small enough such that for all sufficiently generic $h''$ with $|h''(v)| < \epsilon_1$ the subdivision $S(V, h' + h'')$ is a triangulation which refines $S$. Such an $\epsilon$ always exists as shown for instance in Lemma 2.3.15 in [3]. This claim can equivalently be
expressed by saying that $h'$ lies in the boundary of the (full-dimensional) secondary cone of $S' := S(V, h + h'')$.

Let $C$ and $D$ be two maximal cells in $S$ which are adjacent, i.e., the intersection $C \cap D$ is a common face of codimension one. A maximal collection of points $v_1, v_2, \ldots, v_{n+2}$ is called a bipyramid in the dual edge $(C, D)$ of $S$ if $v_1, \ldots, v_{n+1}$ are affinely independent vertices of $C$ and $v_2, \ldots, v_{n+2}$ 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 $s \cap t \subseteq C \cap D$. We can form the matrix $E_h(s, t)$ as in (4), and find the corresponding epistatic value $e_h(s, t)$ via (5). Now we let

$$
\theta := \frac{1}{2} \cdot \min \{ e_h(s, t) \mid s \cup t \text{ bipyramid in some dual edge of } \Gamma(S) \},
$$

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 $S'$. We call the dual edge $(s', t')$ local if $s'$ and $t'$ are contained in some maximal cell of $S$. Clearly, this maximal cells of $S$ belong to a $\theta$-cluster of $S'$ if and only if

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

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 $\epsilon_2 > 0$ such that all $h''$ with $|h''(v)| < \epsilon_2$ satisfy $e_{h'+h''}(s', t') < \theta$ for all local dual edges $(s', t')$. An explicit expression for $\epsilon_2$ 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 $S'$ are contained in a $\theta$-cluster. For a nonlocal dual edge $(s', t')$ of $S'$ observe that $s' \cup t'$ is a bipyramid in some dual edge of $S$ and $e_h(s', t') > \theta$ holds by definition (7). Again by continuity of $e_{h'+h''}$ in $h''$ we get an $\epsilon_3$-neighbourhood of $h'$ where all nonlocal dual edges of $S'$ have epistatic weight lying above $\theta$ and form singleton $\theta$-clusters. Setting $\epsilon := \min(\epsilon_1, \epsilon_2, \epsilon_3)$ settles the claim.

**Example 9.** We further continue the Example 1. Suppose $h' \equiv 0$ is the null function and $h''$ is the height function $h$ given in (1). Then $S(\{0, 1\}^2, h')$ is the trivial subdivision of $[0, 1]^2$ by itself, and $S(\{0, 1\}^2, h' + h'')$ is the triangulation shown in Figure 1. The epistatic weight of the single edge is 6.92 as determined in (6). That is, for all $\theta \geq 6.92$ the cluster partition recovers the trivial subdivision of the unit square $[0, 1]^2$.

Varying $\theta$ yields a stepwise coarsening of $S$ into larger and larger clusters which we call the *cluster filtration* of $S$. For sufficiently small values of $\theta$ (including, e.g., $\theta = 0$) the $\theta$-cluster partition simply consists of the partition of the node set of $\Gamma$ into singletons. On the other hand, for sufficiently large values of $\theta$ the $\theta$-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”. 


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

**Remark 10.** The proof of Theorem 8 opens up a path for extending our definition of epistatic weights to subdivisions induced by non-generic height functions. For a dual edge $(C, D)$ where one or both of the maximal cells $C$ and $D$ are non-simplices, we set $e_h(C, D)$ to be the mean over all epistatic weights obtained from bipyramids in $(C, D)$.

From a more geometric perspective we could also use dihedral angles instead of our epistatic weights. And this would yield a theoretical result similar to Theorem 8. The reason for our seemingly more complicated approach here is that, in Section 4 below, 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.

### 3.3. Computing epistatic filtrations.

Now we will explain how to compute the epistatic filtration from the vertex set $V \subset \mathbb{R}^n$ and the height function $h : V \rightarrow \mathbb{R}$ as input. In the first step we need to determine the list of maximal cells of the regular subdivision $S = S(V, h)$; the standard encoding of each maximal cell is as a subset of $V$. Determining $S$ 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 [14] with a somewhat delicate complexity status [14, Open problem 26.3.4]; see [1] for a recent survey from a practical point of view.

The input to the second step is the list of maximal cells of $S$ as subsets of $V$; let $k$ denote their number. If $V$ are the vertices of the $n$-cube then $k \leq n!$ by Lemma 4. The $k$ maximal cells form the nodes of the dual graph $\Gamma(S)$. 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 $\Gamma(S)$. In the most relevant special case where $S$ is a triangulation, the maximal pairwise intersections of precisely those of cardinality $n$. So the total cost for this step amounts to $O(k^2 n!)$. Let $\ell$ 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 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 anyway.

The third step is to find the $\ell$ epistatic weights, each of which is gotten by computing a $(n+2)\times(n+2)$-determinant. This adds up to a total cost of $O(\ell n^3)$. Sorting the dual edges ascendingly by epistatic weight takes $O(\ell \log \ell)$.

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 $S$. For each dual edge $(s, t)$ with the next epistatic weight $\theta$ we merge clusters of $S(\theta - \epsilon)$ into clusters of $S(\theta)$. 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. The process ends with the dual edge of highest epistatic weight, and we obtain a rooted binary tree. 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 $O(k)$.

Altogether we arrive at a complexity of $O(k^2n + \ell n^3 + \ell \log \ell + k) = O(k^2n + \ell n^3)$ for the steps two through four. Note that $\ell \log \ell \leq k^2n$ in view of $\ell \leq k$ by Lemma 4. The first step, which requires a convex hull computation, is the bottleneck.

We implemented our method in polymake [9, 1], and this was used to produce all computational results presented in this article.

3.4. An extended example. To illustrate the concepts described in Section 3 we now let $P$ be a 3-dimensional cube contained in $[0, 1]^3$ whose vertices are given by

$$
0 = (0, 0, 0, 0) \quad 1 = (1, 0, 0, 0) \quad 5 = (0, 0, 0, 1) \quad 9 = (1, 0, 0, 1) \\
4 = (0, 0, 1, 0) \quad 8 = (1, 0, 1, 0) \quad 15 = (0, 0, 1, 1) \quad 21 = (1, 0, 1, 1) .
$$

These numbers correspond to the genotypes listed in Table 8. Consider a height function $ttd(3)$ assigning the following values to the eight vertices of $P$:

$$
(9) \quad 0 \mapsto 53.25 \quad 1 \mapsto 46.65 \quad 5 \mapsto 43.16 \quad 9 \mapsto 43.48 \\
4 \mapsto 48.3 \quad 8 \mapsto 47.79 \quad 15 \mapsto 43.53 \quad 21 \mapsto 40.71 .
$$

This height function induces a triangulation $S := S(V, ttd(3))$ and its six maximal cells read

$$
A = \{0 \ 1 \ 8 \ 9\} \quad B = \{0 \ 5 \ 9 \ 15\} \quad C = \{0 \ 4 \ 8 \ 15\} \\
D = \{0 \ 8 \ 9 \ 15\} \quad E = \{8 \ 9 \ 15 \ 21\} .
$$

A direct computation shows that cell $D$ has normalized volume two whereas all the other cells have normalized volume one. This makes $S$ a non-unimodular triangulation. Its dual graph $\Gamma(S)$ is shown in Figure 2. For each edge in $\Gamma(S)$ we can compute the normalized epistatic weight using the determinant expression from (5). For instance, the 3-simplices $D$ and $E$ are adjacent in $S$, and we have

$$
|\det E_{ttd(3)}(D, E)| = 0.13 .
$$

Since the normalized volume of $D$ is two (and $E$ is unimodular), the normalized epistatic weight $e_{ttd(3)}(D, E)$ amounts to 0.065. A similar computation of the epistatic weights of the remaining dual edges yields the epistatic filtration in Table 2. The rows are sorted with increasing normalized epistatic weight.
Figure 2. Dual graph $\Gamma(S)$ of $S$ induced by the restricted height function $\text{ttd}^{(3)}$. Each dual edge is labeled with its normalized epistatic value. The shading of the colors indicates the nesting of the cluster partitions.

Table 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 2.

| $\theta$ | $(s, t)$ | $S(\theta)$ |
|----------|----------|--------------|
| 0        | $-$      | $A|B|C|D|E$  |
| 0.065    | $(D, E)$ | $A|B|C|DE$   | $\circ$ |
| 2.245    | $(C, D)$ | $A|B|CDE$    | $\circ$ |
| 3.075    | $(B, D)$ | $A|BCDE$     | $\circ$ |
| 3.845    | $(A, D)$ | $ABCDEF$     | $\circ$ |

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

$S(0) = A|B|C|D|E$.

From Table 2 we see that, among the four dual edges of $\Gamma(S)$, the dual edge $(D, E)$ is the one of lowest epistatic weight. This is the second row of Table 2 and we have

$S(0.065) = A|B|C|DE$.

After three more steps for $\theta = 2.245, 3.075, 3.845$ we finally arrive at the trivial cluster partition

$S(3.845) = ABCDE$,

obtained from joining the adjacent simplices $A$ and $D$. In this triangulation all dual edges are critical.
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 $\theta$ we mark the point $(\theta, \ell)$ in a planar diagram, where $\ell$ is the level, i.e., the number of clusters in the $\theta$-cluster partition. To the right of the marking $(\theta, \ell)$ we draw the $\ell$ 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 $\theta$ and $\theta'$ are two subsequent critical thresholds; i.e., the open interval $(\theta, \theta')$ is regular. Let $\ell$ and $\ell'$ be the $\theta$- and $\theta'$-level, respectively. Then the $\theta$-cluster partition refines the $\theta'$-cluster partition or, conversely, the $\theta'$ 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 $\ell'$ indicates how many clusters of level $\ell$ 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.

For a given cluster filtration of $\mathcal{S}$ consider the piecewise linear cluster function $B(\theta)$ in the continuous parameter $\theta$ fitting the number of connected components in the cluster filtration. When $\theta_0$ is an epistatic weight of a dual edge of $\Gamma(\mathcal{S})$, then $B(\theta_0)$ gives the number of clusters at the filtration stage $\theta_0$. In this way $B(\theta)$ is a decreasing function with minimum attained at $\theta \geq e(A, D)$.

Let us now list a few key properties of the cluster function. First, two closed cluster functions come from similar cluster filtrations. This occurs when epistatic weights are similarly distributed and when the dual graphs of $\Gamma(\mathcal{S})$ and $\Gamma(\mathcal{S}')$ have similar topological properties. Notice that this might occur even when the triangulations $\mathcal{S}$ and $\mathcal{S}'$ are different.

Second, long linear pieces of $B(\theta)$ with almost zero slope indicate cluster filtration steps with high epistatic weight variations. These clusters yield
new regions of interest inside $S$ which cannot be perceived by the shape of the triangulation alone nor by the interaction coordinates.

To see this, let us consider the data listed in Table 2. With the above observations we conclude that the first cluster of interest inside $S$ to be scrutinized further consists of the greatest nontrivial cluster. This cluster is obtained from $S$ by cutting off at the vertices 6 and 8. This give evidence that the mutational pathways from $(0, 0, 0, 0, 0)$ to $(1, 1, 0, 1, 0)$ mainly avoid passing through $(1, 1, 0, 0, 0)$ and $(1, 0, 0, 1, 0)$. Moreover, since the epistatic weights involving the vertices 6 and 8 are low, we deduce that no epistatic information is lost by such a cutting.

Second, we can refine these results to smaller clusters by varying the values of $\theta$ and determine new information on mutational pathways and epistasis. Let us compare the above conclusions with what is already known from [2]. There it is pointed out that all the fittest populations of $P$ are determined by the triangulation $S$, see also Section 2.2 above. With this one deduces that the vertices 6 and 8 don’t coexist in a fittest population, as these vertices belong to two disjoint simplices. In addition, the fittest populations avoid these vertices as all other maximal simplices in $S$ avoid vertex 6 and 8.

Thus our conclusion agrees and extends the conclusion one obtains by looking at the triangulation alone. It is important to point out that it might happen that interesting clusters contain vertices avoided by the fittest populations. That is, cluster and triangulation argumentations can point at different regions of interest inside $S$. For this reason we believe that the biologically most robust results are obtained when the two analyses, cluster filtrations and triangulation reasonings, yield to the same cluster. This occurs when the epistatic weight at the external nodes of $\Gamma(S)$ are minimal.

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 8. Our next goal is to detect cluster partitions which are significant in a statistical sense. To this end we will apply standard tools from statistics to the case when the height function arises for instance from sample means. In particular, Proposition 11 is an adaptation to the case at hand of well known properties satisfied by random variables described for example in [13] and [5, Chapter 6]. We provide the details of the proof for the convenience of the reader.

4.1. Error analysis and standard deviation. As before we consider an $n$-dimensional polytope $P$ with vertex set $V$. Throughout, for each $v \in V$ let $X_v$ be a positive (absolutely) continuous random variable with density function $f_{X_v}$ for which the first two moments exist and are given by $E(X_v)$ and $\sigma_{X_v}^2 = \sqrt{E(X_v^2) - (E(X_v))^2}$. The variance of $X_v$ is given by $\sigma_{X_v}^2$. Here

$$E(X_v) := \int_{-\infty}^{\infty} x f_{X_v}(x)dx$$

and

$$E(X_v X_{vj}) := \int_{R^2} xy f_{X_v, X_{vj}}(x, y)dxdy .$$
We view $X = (X_v)$, for $v \in V$, as a vector of random variables, and we will assume that each realization of $X$ is a generic height function on $V$ with probability one. Let $s = \text{conv}\{v_1, \ldots, v_{n+1}\}$ and $t = \text{conv}\{v_2, \ldots, v_{n+2}\}$ be two simplices which are spanned by points in $V$ and which share a common facet. These are candidates for two adjacent maximal cells of $\mathcal{S}(V,X)$; these are simplices almost surely, as $X$ is generic with probability one. In addition, let $E_X(s,t)$ be the matrix from (4) with $h$ replaced by $X$. We write $E_i = E_X(s,t)$, 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 (4). To simplify the exposition, in the following claim let $N := \text{vol} s \cdot \text{vol} t$ and let $i, j \in \{1, n+1\}$. Let $N(\mu, \sigma^2)$ be the normal distribution defined by the mean $\mu$ and variance $\sigma^2$. The density function of a random variable $X \sim N(\mu, \sigma^2)$ is
\[ g(x | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}. \]
The random variable $Y = |X|$ then follows a folded normal distribution defined again by the parameters $\mu$ and $\sigma^2$ and denoted by $\mathcal{FN}(\mu, \sigma^2)$, see [15] and references therein. The density function of $Y$ is given by:
\[ f(y | \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} (e^{-\frac{1}{2\sigma^2}y-\mu)^2} + e^{-\frac{1}{2\sigma^2}(y+\mu)^2}) \]
The distribution $\mathcal{FN}(0, \sigma^2)$ is also called half-normal distribution, [13, §6.4]. In the above reference, mean and variances of $\mathcal{FN}(\mu, \sigma^2)$ are explicitly given, below we give some useful approximations.

**Proposition 11.** We set
\[ \lambda_i := (-1)^{n+i} \frac{\text{det}(E_i)}{N}. \]
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
\[ \sigma^2_{e_X(s,t)} \leq \sum_{i=1}^{n+2} \left( \frac{\text{det}(E_i)}{N} \sigma_{X_{v_i}} \right)^2 + \frac{2}{N^2} \sum_{1 \leq i < j \leq n+2} |\text{det}(E_i \cdot E_j)| \sigma_{X_{v_i}} \sigma_{X_{v_j}}. \]
Second, if the $n + 2$ random variables $X_{v_i}$ are mutually independent then
\[ \sigma^2_{e_X(s,t)} \leq \sum_{i=1}^{n+2} \left( \frac{\text{det}(E_i)}{N} \sigma_{X_{v_i}} \right)^2. \]
Third, if in addition to independence the relation $X_{v_i} \sim N(\mathbb{E}(X_{v_i}), \sigma^2_{X_{v_i}})$ holds for all $i$, then $e_X(s,t)$ has folded normal distribution

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

**Proof.** In the following it is essential to notice that, for a fixed ordering of the $n + 2$ vertices in $s \cup t$, the numbers $\lambda_i$ are real constants. Using the Laplace expansion of $E_X(s,t)$ along the last column and the linearity of the expected value, we deduce:

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

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

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

The first inequality is Jensen’s inequality $|\mathbb{E}(Y)| \leq \mathbb{E}(|Y|)$ for any random variable $Y$ with first moment and for the convex function $Y \mapsto |Y|$, [13 Thm. B.17 in §B.2.2]. The second one follows from the triangle inequality. The last equality is ensured since $X_{v_i}$ are assumed to be positive. For the variance of $e_X(s,t) = \left| \sum_{i=1}^{n+2} \lambda_i X_{v_i} \right|$ one has:

$$
\sigma^2_{e_X(s,t)} := \text{cov} \left( \sum_{i=1}^{n+2} \lambda_i X_{v_i}, \sum_{i=1}^{n+2} \lambda_i X_{v_i} \right)
$$

$$
\leq \text{cov} \left( \sum_{i=1}^{n+2} \lambda_i X_{v_i}, \sum_{i=1}^{n+2} \lambda_i X_{v_i} \right)
$$

$$
= \sum_{i,j=1}^{n+2} \lambda_i \lambda_j \text{cov} \left( X_{v_i}, X_{v_j} \right)
$$

$$
= \sum_{i=1}^{n+2} (\lambda_i \sigma_{X_{v_i}})^2 + 2 \sum_{1 \leq i < j \leq n+2} \lambda_i \lambda_j \text{cov} \left( X_{v_i}, X_{v_j} \right).
$$

To see the first inequality, observe that for any random variable $Y$ with first moment we get

$$
\text{cov} \left( |Y|, |Y| \right) = \mathbb{E}(|Y|^2) - (\mathbb{E}(|Y|))^2 \leq \mathbb{E}(|Y|^2) - (\mathbb{E}(Y))^2
$$

$$
= (\mathbb{E}(Y^2) - (\mathbb{E}(Y))^2 = \text{cov}(Y,Y),
$$

as $0 \leq |\mathbb{E}(Y)| \leq \mathbb{E}(|Y|)$. The rest follows from the bilinearity of the covariance. If all $X_{v_i}$ and $X_{v_j}$ are mutually independent, then $\text{cov} \left( X_{v_i}, X_{v_j} \right) = 0$, and this settles [12] in our claim. So suppose that $\text{cov} \left( X_{v_i}, X_{v_j} \right) \neq 0$ for some indices $i, j$. Then the Cauchy-Schwarz inequality [13 Thm. B.19] implies that $|\text{cov} \left( X_{v_i}, X_{v_j} \right)| \leq \sigma_{X_{v_i}} \sigma_{X_{v_j}}$, and the inequality in the first claim follows.
Finally, if all random variables $X_{v_i}$ are normally distributed and pairwise independent, it follows for instance from the convolution property of the density functions of $X_{v_i}$ that $\sum_{i=1}^{n+2} \lambda_i X_{v_i}$ is normally distributed with parameters given as:

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

see [13, § B.1.3] or [5, 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. □

We now explain how the above quantities can be deduced from data. For this, let us first recall the following definitions. For each $v \in V$, consider a sample $x(v) = (x_1(v), x_2(v), \ldots, x_L(v))$ of size $L$ of independent and equally distributed realizations of the random variable $X_v$. The number $L$ may depend on $v$. In our setup each sample correspond to measurements obtained from repeating the same experiment $L$ times for the fixed genotype $v$. The sample mean is given by $\bar{X}_v = \frac{1}{L} \sum_{j=1}^{L} x_j(v)$, and the sample standard deviation is $s_{X_v} = \sqrt{\frac{1}{L-1} \sum_{j=1}^{L} (x_j(v) - \bar{X}_v)^2}$. The quantity $s_{\bar{X}_v} = s_{X_v}/\sqrt{L}$ is the standard deviation of the mean (SDOM); this is sometimes also called the “standard error”.

There are several possible scenarios to deduce the standard deviation of $e_{X}(s,t)$ from samples. In the first scenario, we assume that $\bar{X} : V \to \mathbb{R}$ is defined at each $v \in V$ by the corresponding sample mean $\bar{X}_v$. Viewing each $\bar{X}_v$ as a random variable with standard deviation $s_{\bar{X}_v}$ Proposition [11] yields an approximation for the mean and standard deviation of $e_{X}(s,t)$.

A second interesting scenario is the following. Let $X = (X_v)$, for $v \in V$, be a vector of random variables as above. Consider a sample $e(s,t) = ((e_{X}(s,t))_1, \ldots, (e_{X}(s,t))_L)$ of $L$ realizations of the random variable $e_{X}(s,t)$, given by tuples of joint realizations at all $v \in V$. Let $\bar{e}$ be the sample mean of $e(s,t)$. The standard deviation of $\bar{e}$ is approximated by:

$$s_{\bar{e}} = \sqrt{\text{scov}(\bar{e}, \bar{e})} := \sqrt{\frac{1}{L-1} \sum_{k=1}^{L} \left( (e_{X}(s,t))_k - \bar{e} \right)^2}.$$  

Here scov denotes the sample covariance.

The multilinearity of the determinant in the epistatic weight implies that the second scenario agrees with the first, whenever the same number of realizations/experiments occurred. Yet in our examples we always use the first one since the number of Drosophila experiments conducted varies among the bacterial combinations. The significance tests to be explained next are adaptable to both scenarios.

4.2. **Significance test for the epistatic weight.** Assume that the distribution mean $\mu = \mathbb{E}(e_{X}(s,t))$ of the random variable $e_{X}(s,t)$ is unknown and can be estimated in one of the two procedures described above. We
want to assess if $\mu$ is zero or not. If there is no evidence against $\mu = 0$, we consider $e_X(s, t)$ a significant epistatic weight.

To establish this we consider a one-sided test of significance with null hypothesis $\mu = 0$ vs. the alternative hypothesis $\mu > 0$. For the test statistics given by $Z = e_X(s, t)$ we now need to make an assumption on the distribution of $Z$. If the sample sizes are large enough then, for each vertex $v \in V$, the random variable $X_v$ is normally distributed with mean $\bar{X}_v$, and standard deviation $\sqrt{\sigma^{2}}$. Then assuming pairwise independence of $X_v$ for $v \in V$, $Z \sim \mathcal{N}(0, \sigma^2_{e_X(s,t)})$ under the null assumption and with parameter as estimated in Proposition 11.

The validity of the null hypothesis can then be deduced by computing the probability of getting an epistatic weight as extreme or more extreme than $Z$. This probability amounts to

$$ P(X \geq Z) = \int_{Z}^{\infty} \frac{\sqrt{2}}{\sigma_{e_X(s,t)} \sqrt{\pi}} e^{-\frac{1}{2} \left( \frac{x - \bar{X}_v}{\sigma_{e_X(s,t)}} \right)^2} dx ,$$

where the integrand is the density function given in (10). Such a probability is called the $p$-value of the test. The smaller the value of $p$ the stronger the evidence against the null hypothesis provided the data. We call the $p$-value statistically significant if $p < 0.05$, in which case the null hypothesis can be rejected. We then call $e_X(s, t)$ significant. Setting the significance level at 0.05 is a common choice; cf. [8].

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 $e_X(s, t)$ is taken into consideration in this test.

**Remark 12.** Other assumptions on the distribution of $Z$ disregarding normality or independence will require the estimations given in Prop. 11. Our assumptions on $Z$ are plausible in the light of the experiment conducted and at the same time permissive in terms of significance. Let us also point out that epistatic measurement can be of interest if the lifted point configuration is given by mutually independent random variables as well as in the case of random variables with non-vanishing correlations. In the latter case of statistically dependent random variables, we observe that knowing the covariance or correlation matrix does not provide us with a magnitude range for $e_H(s, t)$. Moreover, different sign patterns in these matrices might yield the same epistatic information. Finally, although epistatic information might be deduced from covariances, in general computing the epistatic weight is a more efficient approach.

### 4.3. Clusters from significant epistatic weights.

We now explain how the notion of significant epistatic weights can improve the cluster filtration process discussed in Section 3.2. For this let $S(V, X)$ be as above an induced regular triangulation of an $n$-polytope $P$ equipped with a height function $X$. Again we assume that $X$ assigns to each vertex of $P$ the sample mean over a number of experimental measurements.

We saw that varying the parameter $\theta$ partitions $S$ 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 of the dual edge of $S$ and delete the edges whose epistatic weight does not reach significance at $p < 0.05$. We call the remaining graph the *significant subgraph* $\Gamma_{\text{sig}}(S)$ of $\Gamma(S)$. This induces significant clusters and the significant cluster partition $S_{\text{sig}}$. The epistatic filtration process from Section 3.2 and the algorithm from Section 3.3 carry over, and we obtain nested cluster partitions which coarsen the significant cluster partition.

4.4. Continuation of the extended example. We now illustrate the above definitions on the example in 3.4. Consider again the regular triangulation $S = S(V, \text{ttd}^{(3)})$ induced by the height function $\text{ttd}^{(3)}$ from (9). Figure 2 shows the dual graph of $S$; 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 $\text{ttd}^{(3)}$ in order to highlight the connection to the previous section.

Assuming independence, for each dual edge in $\Gamma(S)$ we now compute bounds on the associated standard deviation using (12). For instance, for $e \bar{X}(D, E)$ consider

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

This yields

$$ \sigma_{e \bar{X}}(D, E) \leq \left( \sum_{i=1}^{5} \left( \frac{\det(E_{\sigma \bar{X}}(D, E)) \sigma \bar{X}(v_i)}{nvol(D) \cdot nvol(E)} \right)^2 \right)^{1/2} $$

$$ = \frac{1}{2} \cdot \left( \left( \det \begin{bmatrix} 1 & 0 & 1 & 1 & 1.450 \end{bmatrix} \right)^2 + \ldots + \left( \det \begin{bmatrix} 1 & 0 & 1 & 1 \end{bmatrix} \cdot 0.907 \right)^2 \right)^{1/2} $$

$$ = 1.493. $$

Processing the other epistatic weights in a similar fashion yields the values in Table 3. The rows are sorted by increasing epistatic weight. The $p$-values in the last column are determined by (14).

Based on Table 3 the significant cluster partition $S_{\text{sig}}$ arising from the restricted height function $\text{ttd}^{(3)}$ reads $A|BCDE$.

Remark 13. Since (12) 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 $(B, D)$ with epistatic weight 3.075 and $p$-value bound 0.061 comes into focus. It would be interesting to
check if additional experiments involving the five genotypes 0, 2, 8, 9, 15 in B ∪ D lead to a higher level of significance or not.

4.5. A synthetic experiment. In this section we describe one synthetic experiment on a quantitative analysis of Theorem 8 in relationship with the concept of statistical significance from Section 4.

We consider the vertex set \( V = \{0, 1\}^5 \) of the regular 5-cube and a height function \( \eta \) 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 \( \eta_0 \). The wild type corresponds to the vertex 0; cf. Table 8. The combinatorial type of the regular subdivision \( S(V, \eta) \) does not depend on the precise value \( \eta_0 > 0 \). In fact, there are precisely two maximal cells, \( s_0 \) and \( t \), and \( S(V, \eta) \) is a vertex split in the terminology of [12]. 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 \( \Gamma(S(V, \eta)) \) has two nodes connected by the single dual edge \((s_0, t)\).

Our experiments depend on the choice of \( \eta_0 \) and a second strictly positive real number \( \sigma \). To each vertex \( v \in V \) we assign a normally distributed random variable \( X_v \) with zero mean and standard deviation \( \sigma \). From 100 realizations per vertex we compute the resulting sample means and standard errors. This gives rise to a generic perturbation

\[
\eta' = \eta + (\bar{X}_v \mid v \in V)
\]

of the height function \( \eta \) by adding the sample means. For the resulting triangulation \( S(V, \eta') \) 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 (12). We start with a height function \( \eta \) at level 5 since the mean values \( \bar{X}_v \) from the perturbation may be negative; note that the perturbed height function \( \eta' \) needs to be strictly positive in order to qualify for the analysis via the \( p \)-values from (14).

We only consider perturbed height functions \( \eta' \) such that the simplex \( s_0 \) is a maximal cell of \( S(V, \eta') \), just as in \( S(V, \eta) \). For \( \sigma \) sufficiently small compared to \( \eta_0 \) this holds almost always. If \( s_0 \) is a maximal cell then \( s_0 \) is adjacent to some unique maximal cell of \( S(V, \eta') \). We call the corresponding dual edge the bridge of \( \Gamma(S(V, \eta')) \).
Now, for a fixed pair \((\eta_0, \sigma)\) 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 we saw the simplex \(s_0\) was a maximal cell, and the bridge existed. Further all perturbed height functions \(\eta'\) were nonnegative. Figure 4 shows the result for \(\eta_0 \in \{0.8, 1.0, 1.2\}\) and \(0.1 \leq \sigma \leq 2\).

![Figure 4](image-url)

**Figure 4.** Percentage of significant bridges depending on \(\sigma\), for various choices of \(\eta_0\) and \(p\).

The experimental results displayed in Figure 4 can be summarized as follows. For any fixed choice of \(\eta_0\) and \(p\) the percentage of triangulations with the bridge present approximates a threshold function in the parameter \(\sigma\). If \(\sigma\) is sufficiently low then the bridge is always significant; this observation can be seen as a variation of Theorem 8 for this particular setup. The lower the value of \(\eta_0\), the steeper the curve is. For larger values of \(\sigma\) random fluctuations kick in, and this makes the curve less smooth.

### 5. Epistasis, interaction coordinates and circuits

In this section we compare the cluster partition to previous approaches. To do so, we now recall how the biological phenomenon of epistasis as described in the work of Beerenwinkel et al. [2].

#### 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 \(n\). Define the interaction space as

\[
\mathcal{I}_V := \left( \mathbb{R}^V / \mathcal{L}_V \right)^*.
\]
Elements of $I_V$ are linear forms

$$\lambda: \mathbb{R}^V \rightarrow \mathbb{R}$$

$$h \mapsto \sum_{v \in V} \alpha_v h(v),$$

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

When $V$ is the vertex set of an $n$-cube, a basis for $I_V$ is given by the interaction coordinates defined up to scalar multiplication as:

$$u_{h,w}: \mathbb{R}^V \rightarrow \mathbb{R}$$

$$h \mapsto \sum_{v \in V} (-1)^{(v,w)} h(v),$$

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 $u_{h,11} = \epsilon(00, 01, 10, 11)$. A second related generalization of the usual epistasis formula arises by considering circuits interactions, see [2]. These are linear forms whose support is given by a minimal (non-empty) affinely dependent set of vertices of $P$. Since these are elements of the interaction space, these can be expressed as linear combinations of interaction coordinates. Circuit interactions then capture more intricate epistatic patterns such as conditional epistasis or relations among the previous forms of epistasis.

5.2. Normalized epistatic weight interactions. In this work, a new distinguished set of interactions inside $I_V$ are given by the linear forms:

$$\mathbb{R}^V \rightarrow \mathbb{R}$$

$$h \mapsto e_h(s, t),$$

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

When $V = \{0, 1\}^2$, there is a unique normalized epistatic weight which agrees with the absolute value of the linear form $u_{11}$. This is the only case where the epistatic weight agrees with the notion of interaction coordinates. The relation to the other forms of interactions can be summarized as follows:

- bipyramid circuit interactions inside a 3-cube are normalized epistatic weights over the same set of vertices, up to renumbering.
- If all minors $E_h(s, t)$ of $E_h(s, t)$ are non-zero, then $e_h(s, t)$ is a circuit in the sense of [2]. However, there are normalized epistatic weights which are not circuit interactions and circuit interactions which cannot be described as normalized epistatic weights. For the first claim, examine the epistatic weights in a triangulated 3-cube, see [6]. For the second, consider circuits $a$-$\ell$ in [2].
- In Lemma [2] we provided bounds for epistatic weight computations. Computing just the 20 circuits $a$-$\ell$ inside each 3-cube in an $n$-cube amounts to $\binom{n}{3}2^{n-2} + 14\binom{n}{3}2^{n-3}$ computations exceeding the upper
Moreover, the opposite inequality holds for large $x$.

**Proof.** Consider the expressions

$$A e^{-\frac{x^2}{2}} := f(x \mid 0, \sigma^2_U)$$

and

$$C e^{-\frac{x^2}{2}} := f(x \mid 0, \sigma^2_E),$$

for $A, B, \ldots, D$ as one deduces from (1) together with Proposition 11 and Lemma 14. Since $\sigma_U$ is a sum over all vertices of $V$, we immediately have that $\sigma_U > \sigma_E > 0$; cf. Lemma 14 and Proposition 11. Thus $C > A > 0$ and the first claim follows. In addition, as $-\frac{1}{B}x^2 > -\frac{1}{D}x^2$ also $e^{-\frac{1}{B}x^2} > e^{-\frac{1}{D}x^2}$ for $x > 0$. Then, for $x$ large enough, we have $\log(A e^{-\frac{x^2}{2}}) = \log(A) - \frac{1}{B}x^2 > \log(C e^{-\frac{x^2}{2}})$, and the second claim follows.

To better illustrate Lemma 15 let us consider the following...
Example 16. Let $\bar{X}$ be a generic height function on $[0,1]^5$ such that $e_{\bar{X}}(s,t)$ is large enough and agrees with the absolute value of an interaction coordinate $u_{\bar{X},w}$ on $[0,1]^5$, for some $s,t$ and $w$. Then our standard deviations approximations imply that the same test statistic $Z = e_{\bar{X}}(s,t) = u_{h,w}$ has higher significance for $u_{\bar{X},w}$ than for $e_{\bar{X}}(s,t)$, since for $u_{\bar{X},w}$ the $p$-value is smaller.

A measure of dissimilarity between two continuous distributions $p(x)$ and $q(x)$ over $(0,\infty)$ is given by the Kullback-Leibler information - sometimes called Kullback-Leibler divergence:

$$-\text{KL}(p\|q) := \int_0^\infty p(x) \cdot \ln \left( \frac{q(x)}{p(x)} \right) \, dx.$$ 

Notice that in general the Kullback-Leibler information is not a metric as $\text{KL}(p\|q) \neq \text{KL}(q\|p)$. Moreover, $\text{KL}(p\|q) \geq 0$ with equality if and only if $p(x) = q(x)$, see [13, Prop. 2.92]. The smaller the Kullback-Leibler information, the more similar two distributions are. That is, the less likely it is that an observation coming from one distribution did not come from the other distribution.

**Proposition 17.** If $U = E = 0$ then

$$0 \leq \text{KL}(f(x \mid 0, \sigma^2_U) \parallel f(x \mid 0, \sigma^2_E)) \leq -\ln \left( \frac{\sigma_E}{\sigma_U} \right).$$

Similarly,

$$0 \leq \text{KL}(f(x \mid 0, \sigma^2_E) \parallel f(x \mid 0, \sigma^2_U)) \leq \left( \frac{1}{\sigma_U^2} - \frac{1}{\sigma_E^2} \right) \frac{\sigma_E \sigma_U}{\sqrt{2\pi}}.$$

**Proof.** We compute

$$\text{KL}(f(x \mid 0, \sigma^2_U) \parallel f(x \mid 0, \sigma^2_E)) = -\ln \left( \frac{\sigma_E}{\sigma_U} \right) + \left( \frac{1}{\sigma_U^2} - \frac{1}{\sigma_E^2} \right) \frac{\sigma_E \sigma_U}{\sqrt{2\pi}},$$

using that the first moment of the half-normal distribution $f(x \mid 0, \sigma^2_U)$ is given by $\frac{\sigma_U}{\sqrt{2\pi}}$, see [15]. To deduce the upper bound, recall that $\sigma_E \leq \sigma_U$. Thus the first summand is always positive and the second negative. The second claim follow by exchanging the rule of $E$ and $U$ in $\text{KL}$. Then the second summand becomes positive while the first $-\ln \left( \frac{\sigma_E}{\sigma_U} \right)$ is negative. $\square$

More results on the Kullback-Leibler information for folded normal and related distributions can be found in [15]. Estimates for the Kolmogorov–Smirnov distance between $F_N(E, \sigma^2_E)$ and $F_N(U, \sigma^2_U)$ are obtained in a similar fashion. In this section, we have focused on the $n$-dimensional interaction coordinates, arguing in a similar way, one compares the distribution of epistatic weights with lower dimensional interaction coordinates and circuits. The details are left to the reader.

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 previous approach and locate interesting epistatic information where the previous approach is less conclusive.
6.1. **Data.** The data we use consists of experimental measurements of *Drosophila* flies inoculated with all possible combinations of five bacterial species naturally present in the gut of wild flies. The data set includes measurements of time to death (days), daily fecundity (progeny/day/female) and development time (days). All measurements were repeated several times so that the mean and standard error associated to each measurement and bacterial combination could be taken. This dataset is remarkably complete as all 32 bacterial combinations are considered.

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 global deviation from additivity rather than a specific manifestation of it, quantified for instance by marginal, conditional, 2, 3, 4 and 5-way interactions.

In the specific *Drosophila* data set, the methods developed in this work allowed us to distinguish various 7-tuples of bacterial combinations with vanishing epistatic weights from non-vanishing ones. Bacterial combinations in the first set are not expected to have accentuated synergistic and antagonistic effects, while such effects are expected for bacterial combinations with non-zero epistatic weight. Filtrations then provide clusters of bacterial combinations of the second type respecting the complicated neighboring relations satisfied by these bacteria encoded in adjacent relations of simplices in a triangulation. Clusters also provide previously unnoticed regions inside the genotope which we propose to be further dissected using empirical methods.

Our method intentionally involves a relatively small number of tests, limited by the adjacency relations among maximal simplices in triangulated genotope and are specific to each phenotype mapping. To enable first statistical considerations within this theory we provided statistical tools facilitating comparisons with previous approaches. Specifically, in [10] significance tests were performed on a number of specific epistatic formulas. Comparing our work with the results of these tests reveals that a fitness landscape carrying significant epistatic information in the latter sense, does not necessarily imply the existence of significant epistatic weights and vice-versa. Examples are given below.

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, daily fecundity and development time provide insights on the different cluster patterns arising but revealed no significant epistatic weights. We conclude that by the present metric, there is globally not a global, 5-dimensional, epistatic effect for this data set.

Although filtrations established that globally the system is additive, specific significant outcomes of local epistatic formulas do occur. For instance, our recent work shows that 124 tests out of 936 resulted to be significant at a level of at least 0.05 for time to death, after Benjamini-Hochberg multiple testing correction and in a testing settings comparable to the one used in this work. Similarly, 107 tests resulted significant for daily fecundity and
36 for development time. Out of these, 28 were significant for at least two phenotypes, see [10] for the details.

On large scales these two approaches will remain complementary, and to fully understand the biological implications of both, more data has to be analyzed. Below, we present first interesting observations in this direction.

6.4. **The case of opposite facets inside** \([0,1]^5\). Consider the four-way interaction

\[
\begin{align*}
&u_{0^*101} = h(v_{0^*000}) + h(v_{0^*010}) + h(v_{1^*000}) + h(v_{0^*101}) \\
&\quad + h(v_{1^*010}) + h(v_{0^*111}) + h(v_{1^*101}) + h(v_{1^*111}) \\
&\quad - h(v_{0^*001}) - h(v_{0^*100}) - h(v_{0^*011}) - h(v_{1^*001}) \\
&\quad - h(v_{1^*100}) - h(v_{0^*110}) - h(v_{1^*011}) - h(v_{1^*110}) ,
\end{align*}
\]

(15)

where \(^* \in \{0,1\}\). If \(^* = 0\), the above interaction is considered in the absence of the *Lactobacillus brevis* bacteria, otherwise the bacteria are present. The analyses of [10] determine that \(u_{0^*101}\) is simultaneously significant for daily fecundity, development time and time to death only in the absence of the *Lactobacillus brevis* bacteria (after Benjamini-Hochberg multiple testing correction).

To further inspect the rule of the *Lactobacillus brevis* bacteria, we compute the fitness landscape defined by the vertices in the summands of \(u_{0^*101}\) and the three phenotype mappings for both values of \(^*\). For time to death different cluster pattern appear on the opposite facets. Moreover, the presence of significant epistatic weights in the absence of the *Lactobacillus brevis* bacteria confirms that these bacteria significantly affects epistatic interactions. Further dissecting the significant epistatic weights, as well as the filtration steps restricts the set of possible bacterial combinations responsible for this effect.

On the other hand, for development time and fecundity all epistatic weights are near zero. As a consequence there are no significant dual edges (all bars are red in Figure 5). We conclude that the epistatic effect of the *Lactobacillus brevis* bacteria is not equally strong on all phenotypes. However, among the local significant tests described above, 15 remain significant for both fecundity and development time, see [10].

---

**Table 4.** Interactions for ttd on the 3-cube of Example 3.4 which reached significance for \(p < 0.05\) are in bold.

| Interaction | Result | SE  | \(p\)-value folded | Vertices |
|-------------|--------|-----|---------------------|----------|
| \(a_{000}\beta\gamma\) | 6.09   | 2.57| 0.02                | 0 1 4 8  |
| \(c_{000}\beta\gamma\) | 6.92   | 2.58| 0.01                | 0 1 5 9  |
| \(e_{000}\beta\gamma\) | 5.32   | 2.40| 0.03                | 0 4 5 15 |
| \(m_{000}\beta\gamma\) | -9.10  | 3.72| 0.01                | 0 1 4 5 21|
| \(s_{000}\beta\gamma\) | 7.69   | 3.83| 0.04                | 0 1 8 9 15|
| \(u_{000}\beta\gamma\) | 9.23   | 3.32| 0.01                | 0 1 5 9 4 8 15 21|
| \(\ell_{000}\beta\gamma\) | 1.60   | 2.37| 0.50                | 1 4 9 19 |
Figure 5. Filtrations of distinguished four-faces. The colors of the bars represent various levels of statistical significance; cf. Section 4.
6.5. **The case of three-cubes inside** $[0, 1]^5$. To make the comparison between methods more explicit, focus again on the example discussed in 3.4. Following [2], there are 20 circuit interactions and four interaction coordinates. Out of these four reached statistical significance assuming a normal distribution and six taking the absolute value of the result and assuming a folded normal distribution. There are highlighted in bold in Table 4. After Benjamini-Hochberg multiple testing correction, none of the above remain significant.

These formulas are independent of the phenotype mapping and capture three forms of conditional epistasis (for $a, c, e$) the three-way interaction ($u$) and bipyramidal circuits $m, s$. On the other hand, out of the four epistatic weights of $S(V, \text{ttd}^{(3)})$ only the epistatic weight $e_{\bar{X}}(A, D)$ involving the vertices 0, 1, 8, 9, 15 reached significance. This example shows that a significant epistatic weight does not necessarily imply that circuit interactions defined on the vertex set of the two adjacent simplices is significant. In fact, out of the 24 cube interactions just one is a circuit interaction for the bipyramid $A \cup D$, namely:

$$
\ell_{a00\beta\gamma} : h(v_{00010}) - h(v_{00011}) - h(v_{10000}) + h(v_{10001})
$$

The sign pattern of this interaction yields a particularly small result if compared to the statistically significant ones; cf. Table 4.

6.6. **The case of parallel squares inside** $[0, 1]^5$. Fixing two bacterial species, $\alpha$ and $\beta$, 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 $\alpha$- and $\beta$-coordinates vary, and all others are set to zero. Now a third bacterial species, $\gamma$, defines a parallel square in $[0, 1]^5$, where the $\alpha$- and $\beta$-coordinates vary, and the $\gamma$-coordinate is set to one. Altogether, $\alpha$, $\beta$, and $\gamma$ define a 3-dimensional face of $[0, 1]^5$. We want to investigate the impact of the presence of $\gamma$ on the 2-way interactions between $\alpha$ and $\beta$; cf. [10] and [2]. This amounts to comparing the epistatic weights of the two subdivisions induced on the parallel pair of squares.

Knowing that the *Lactobacillus* bacteria compete with one another and *Acetobacters* also compete with one another, while *Acetobacters* form mutualist relationships with the *Lactobacilli*, we 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 $\alpha$, $\beta$ and $\gamma$. Each face is a row in each of the three Tables 5, 7 and 6 in the appendix; these tables show the result of our analysis for the three height functions ttd, dev and fec, respectively.

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 $\omega_0$ is the normalized volume of the simplex arising from the square with $\gamma$-coordinate zero, and $\omega_1$ corresponds to $\gamma$-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 $\gamma$ to $\alpha$ and $\beta$, 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, 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 $\gamma$ seems to almost annihilate the epistasis.

2. A distinguished combination appears to be $(\alpha, \beta) = (2, 4)$: for which the time to death, development time and fecundity results match, suggesting the interaction affects all three traits in the same manner.

3. Another interesting case is $(\alpha, \beta) = (1, 5)$: for which adding $\gamma$ weakens the epistasis. Yet the epistasis for ttd is negative, while it is positive for dev; it is not fully conclusive for fec.

So far, we have not determined biological explanations for these facts, and more experimental examinations are needed. However, 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.

Acknowledgement

The data was collected by Alison Gould and Vivian Zhang in Will Ludington’s Lab at UC Berkeley.

References

[1] 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(1):1–38, 2017.

[2] Niko Beerenwinkel, Lior Pachter, and Bernd Sturmfels. Epistasis and shapes of fitness landscapes. *Statist. Sinica*, 17(4):1317–1342, 2007.

[3] 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(1):60, Apr 2007.

[4] Joel Brenner and Larry Cummings. The Hadamard maximum determinant problem. *Amer. Math. Monthly*, 79:626–630, 1972.

[5] J. Dauxois and C. Hassenforder. *Toutes les probabilités et les statistiques: cours et exercices corrigés*. Ellipses, 2004.

[6] Jesús A. De Loera, Jörg Rambau, and Francisco Santos. *Triangulations*, volume 25 of *Algorithms and Computation in Mathematics*. Springer-Verlag, Berlin, 2010. Structures for algorithms and applications.

[7] de Visser and Joachim Krug. Empirical fitness landscapes and the predictability of evolution. *Nat Rev Genet*, 15(7):480–490, July 2014.

[8] B. S. Everitt and A. Skrondal. *The Cambridge dictionary of statistics*. Cambridge University Press, Cambridge, fourth edition, 2010.

[9] Ewgenij Gawrilow and Michael Joswig. polymake: a framework for analyzing convex polytopes. In *Polytopes—combinatorics and computation* (Oberwolfach, 1997), volume 29 of *DMV Sem.*, pages 43–73. Birkhäuser, Basel, 2000.

[10] Alison L Gould, Vivian Zhang, Lisa Lamberti, Eric W Jones, Benjamin Obadia, Alex Gavryushkin, Jean M Carlson, Niko Beerenwinkel, and William B Ludington. High-dimensional microbiome interactions shape host fitness. *bioRxiv*, 2017.

[11] Ingeleif B. Hallgrímsdóttir and Debbie S. Yuster. A complete classification of epistatic two-locus models. *BMC Genetics*, 9(1):17, 2008.

[12] Sven Herrmann and Michael Joswig. Splitting polytopes. *Münster J. Math.*, 1:109–141, 2008.
Figure 6. Epistatic filtration of the triangulation $S(\{0,1\}^5, \text{ttd})$ with 111 maximal cells. The colors of the bars represent various levels of statistical significance; cf. Section 4.
Table 5. Parallel squares in $[0,1]^5$ for ttd.

| $\alpha = 1$, $\beta = 3$ | $\omega_0$ | $\omega_1$ | 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           |

| $\alpha = 1$, $\beta = 4$ | $\omega_0$ | $\omega_1$ | 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           |

| $\alpha = 1$, $\beta = 5$ | $\omega_0$ | $\omega_1$ | 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           |

| $\alpha = 2$, $\beta = 3$ | $\omega_0$ | $\omega_1$ | 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          |

| $\alpha = 2$, $\beta = 4$ | $\omega_0$ | $\omega_1$ | 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           |

| $\alpha = 2$, $\beta = 5$ | $\omega_0$ | $\omega_1$ | 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 6. Parallel squares in $[0, 1]^5$ for dev.

| $\alpha = 1, \beta = 3$ | $\omega_0$ | $\omega_1$ | sign | $|$quot$|
|-------------------------|------------|------------|------|---------|
| $\gamma=2$             | 0.034      | -0.033     | -    | 0.971   |
| $\gamma=4$             | 0.034      | 0.015      | +    | 0.441   |
| $\gamma=5$             | 0.034      | 0.013      | +    | 0.382   |

| $\alpha = 1, \beta = 4$ | $\omega_0$ | $\omega_1$ | sign | $|$quot$|
|-------------------------|------------|------------|------|---------|
| $\gamma=2$             | 0.006      | -0.035     | -    | 5.833   |
| $\gamma=3$             | 0.006      | -0.013     | -    | 2.167   |
| $\gamma=5$             | 0.006      | -0.023     | -    | 3.833   |

| $\alpha = 1, \beta = 5$ | $\omega_0$ | $\omega_1$ | sign | $|$quot$|
|-------------------------|------------|------------|------|---------|
| $\gamma=2$             | 0.061      | 0.013      | +    | 0.213   |
| $\gamma=3$             | 0.061      | 0.040      | +    | 0.656   |
| $\gamma=4$             | 0.061      | 0.032      | +    | 0.525   |

| $\alpha = 2, \beta = 3$ | $\omega_0$ | $\omega_1$ | sign | $|$quot$|
|-------------------------|------------|------------|------|---------|
| $\gamma=1$             | 0.009      | -0.058     | -    | 6.444   |
| $\gamma=4$             | 0.009      | 0.015      | +    | 1.667   |
| $\gamma=5$             | 0.009      | 0.039      | +    | 4.333   |

| $\alpha = 2, \beta = 4$ | $\omega_0$ | $\omega_1$ | sign | $|$quot$|
|-------------------------|------------|------------|------|---------|
| $\gamma=1$             | 0.056      | 0.097      | +    | 1.732   |
| $\gamma=3$             | 0.056      | 0.050      | +    | 0.893   |
| $\gamma=5$             | 0.056      | 0.000      |     |         |

| $\alpha = 2, \beta = 5$ | $\omega_0$ | $\omega_1$ | sign | $|$quot$|
|-------------------------|------------|------------|------|---------|
| $\gamma=1$             | 0.026      | 0.074      | +    | 2.846   |
| $\gamma=3$             | 0.026      | -0.004     | -    | 0.154   |
| $\gamma=4$             | 0.026      | -0.030     | -    | 1.154   |
Table 7. Parallel squares in $[0, 1]^5$ for fec.

| $\alpha = 1$, $\beta = 3$ | $\omega_0$ | $\omega_1$ | sign | $|\text{quot}|$ |
|----------------|--------|--------|------|---------|
| $\gamma = 2$ | 0.191  | -0.037 | -    | 0.194   |
| $\gamma = 4$ | 0.191  | 0.051  | +    | 0.267   |
| $\gamma = 5$ | 0.191  | 0.045  | +    | 0.236   |

| $\alpha = 1$, $\beta = 4$ | $\omega_0$ | $\omega_1$ | sign | $|\text{quot}|$ |
|----------------|--------|--------|------|---------|
| $\gamma = 2$ | 0.057  | 0.002  | +    | 0.035   |
| $\gamma = 3$ | 0.057  | -0.083 | -    | 1.456   |
| $\gamma = 5$ | 0.057  | 0.029  | +    | 0.509   |

| $\alpha = 1$, $\beta = 5$ | $\omega_0$ | $\omega_1$ | sign | $|\text{quot}|$ |
|----------------|--------|--------|------|---------|
| $\gamma = 2$ | 0.091  | 0.016  | +    | 0.176   |
| $\gamma = 3$ | 0.091  | -0.055 | -    | 0.604   |
| $\gamma = 4$ | 0.091  | 0.063  | +    | 0.692   |

| $\alpha = 2$, $\beta = 3$ | $\omega_0$ | $\omega_1$ | sign | $|\text{quot}|$ |
|----------------|--------|--------|------|---------|
| $\gamma = 1$ | 0.100  | -0.128 | -    | 1.280   |
| $\gamma = 4$ | 0.100  | 0.092  | +    | 0.920   |
| $\gamma = 5$ | 0.100  | 0.068  | +    | 0.680   |

| $\alpha = 2$, $\beta = 4$ | $\omega_0$ | $\omega_1$ | sign | $|\text{quot}|$ |
|----------------|--------|--------|------|---------|
| $\gamma = 1$ | 0.058  | 0.003  | +    | 0.052   |
| $\gamma = 3$ | 0.058  | 0.050  | +    | 0.862   |
| $\gamma = 5$ | 0.058  | 0.014  | +    | 0.241   |

| $\alpha = 2$, $\beta = 5$ | $\omega_0$ | $\omega_1$ | sign | $|\text{quot}|$ |
|----------------|--------|--------|------|---------|
| $\gamma = 1$ | 0.032  | -0.043 | -    | 1.344   |
| $\gamma = 3$ | 0.032  | 0.000  | +    | 0.375   |
| $\gamma = 4$ | 0.032  | -0.012 | -    | 0.375   |
Table 8. Numbering of the genotypes.

| number | genotype | comment                        |
|--------|----------|--------------------------------|
| 0      | 0 0 0 0 0| germ-free                      |
| 1      | 1 0 0 0 0| *Lactobacillus plantarum*      |
| 2      | 0 1 0 0 0| *Lactobacillus brevis*         |
| 3      | 0 0 1 0 0| *Acetobacter pasteurianus*     |
| 4      | 0 0 0 1 0| *Acetobacter tropicalis*       |
| 5      | 0 0 0 0 1| *Acetobacter orientalis*       |
| 6      | 1 1 0 0 0| competition                    |
| 7      | 1 0 1 0 0| visible 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| antagonistic                   |
| 14     | 0 0 1 0 1| antagonistic                   |
| 15     | 0 0 0 1 1| antagonistic                   |
| 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|                              |