nature portfolio

https://doi.org/10.1038/s41559-021-01462-8

Supplementary information

Coexistence holes characterize the assembly and disassembly of multispecies systems

In the format provided by the authors and unedited

Coexistence holes characterize the assembly and disassembly of multispecies systems

Supplementary Notes

Marco Tulio Angulo[1]*, Aaron Kelley[2], Luis Montejano[2], Chuliang Song[3,4,5], and Serguei Saavedra[3]**

[1]CONACyT - Institute of Mathematics, Universidad Nacional Autónoma de México, Juriquilla, México.
[2]Institute of Mathematics, Universidad Nacional Autónoma de México, Juriquilla, México.
[3]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA.
[4]Department of Biology, McGill University, 1205 Dr. Penfield Avenue, Montreal, H3A 1B1 Canada.
[5]Department of Ecology and Evolutionary Biology, University of Toronto, 25 Willcocks Street, Toronto, Ontario M5S 3B2 Canada
correspondence to: *mangulo@im.unam.mx or **sersaa@mit.edu

12 March 2021

Contents

1. Definition of assembly and disassembly hypergraphs

We consider ecological systems where individuals have been organized into S species (or strains, or taxa, or functional groups, or in any other meaningful organization). Let $\mathsf{V} = \{1, 2, \cdots, S\}$ denote the pool of species. Note that the set of all different species collections that can be obtained from this pool is the **power set** 2^{V} —the collection of all subsets of V. Here, we use as convention that sets do not contain repeated elements (e.g., the set $\{2, 1, 2\}$ is not allowed).

Given a species collection $\Sigma \in 2^{\mathsf{V}}$, its coexistence is a dichotomy: the species either coexist together or not. Thus, we formalize species coexistence as follows:

Definition 1. *The* **coexistence function** *of an ecological system is a function* $c : 2^{\mathsf{V}} \to \{0, 1\}$.

For any species collection $\Sigma \in 2^{\mathsf{V}}$, we interpret the condition $c(\Sigma) = 1$ as "species in Σ coexist", and $c(\Sigma) = 0$ as "species in S do not coexist". If Σ contains a single species, then coexistence is interpreted as "surviving in isolation". For mathematical completeness, for the empty set $\emptyset \in 2^{\mathsf{V}}$, we define $c(\emptyset) = 0$.

From Definition 1, the coexistence of species in an ecological system is described by the coexistence relations that one species has with all other members of the species pool[a]. The assembly hypergraph of the system fully captures all those coexistence relations:

Definition 2. *Given a coexistence function* c*, the* **assembly hypergraph** H_c *of the ecological system is the pair* $H_c = (\mathsf{V}, \mathsf{H}_c)$*, where*

$$\mathsf{H}_c := \{\Sigma \in 2^{\mathsf{V}} | c(\Sigma) = 1\}$$

is the set of all species collections that coexist.

Note that $\mathsf{H}_c \subseteq 2^{\mathsf{V}}$. When the coexistence function c is clear from the context, we write H instead of H_c. Hypergraphs are well-studied combinatorial objects in mathematics [2], generalizing the notion of graphs to encode relations between an arbitrary number of components. Recently, hypergraphs are finding important applications across the sciences (see, e.g., [3–7]).

In a hypergraph $H = (\mathsf{V}, \mathsf{H})$, the set $\mathsf{V} \in H$ are its **vertices**, in our case corresponding to isolated species. The collection $\mathsf{H} \in H$ are its **hyperedges**, which for the assembly hypergraph characterize all species collections that coexist. When clear from the context, we also write $h \in H$ to denote a hyperdge h of the hypergraph H. The **dimension** of a hyperedge h is $\dim(h) := |h| - 1$, where $|h|$ is the cardinality of the set h (e.g, for $h = [1, 2, 3]$ its dimension is $\dim h = 2$). The dimension of a hypergraph is defined as $\dim(H) = \max_{h \in H} \dim(h)$.

In the Main Text, we introduce the notion of the **disassembly hypergraph** $D(H)$ associated to an assembly hypergraph H. To formally define this notion, we first need the following concept:

Definition 3. *The* **minimal simplicial complex** $K(H)$ *containing the hypergraph* H *is the simplicial complex with vertex set* V *and simplices* K *defined as*

$$K(H) = \{\sigma \subset \mathsf{V} | \sigma \subset \tau \text{ for some } \tau \in H\}.$$

Recall that a simplicial complex is a collection of subsets of V that is closed under inclusion (i.e., if $\sigma \in K$ and $\tau \subset \sigma$ then $\tau \in K$). In other words, for a simplicial complex K, each element $\tau \in K$ contains all its boundaries. Any simplicial complex is a hypergraph, but not all hypergraphs are simplicial complex.

Next, for a hypergraph H, we define its its **missing boundary** $M(H)$ as

$$M(H) := K(H) \setminus H.$$

[a] Formally, coexistence is a **relation**. To analyze all those relations, it is necessary to apply an *analysis situs* [1] —an analysis of the "situation" describing the coexistence relation that one species with the rest. Such analysis requires a mathematical object whose geometry disregards angles nor distances. Actually, what could be the distance or angle between two species that coexist? Therefore, this geometry, encoded by the assembly/disassembly hypergraph of an ecological system, considers only relations of coexistence between species. Thus the *analysis situs* of species coexistence corresponds to the topological analysis of the assembly hypergraph.

In words, the elements of $M(H)$ are the missing boundaries in the hyperedges of H. The missing boundary of a hypergraph is another hypergraph over the same vertex set. Note that $K(H)$ is uniquely determined given H. This implies that the missing boundary $M(H)$ is well-defined for any hypergraph.

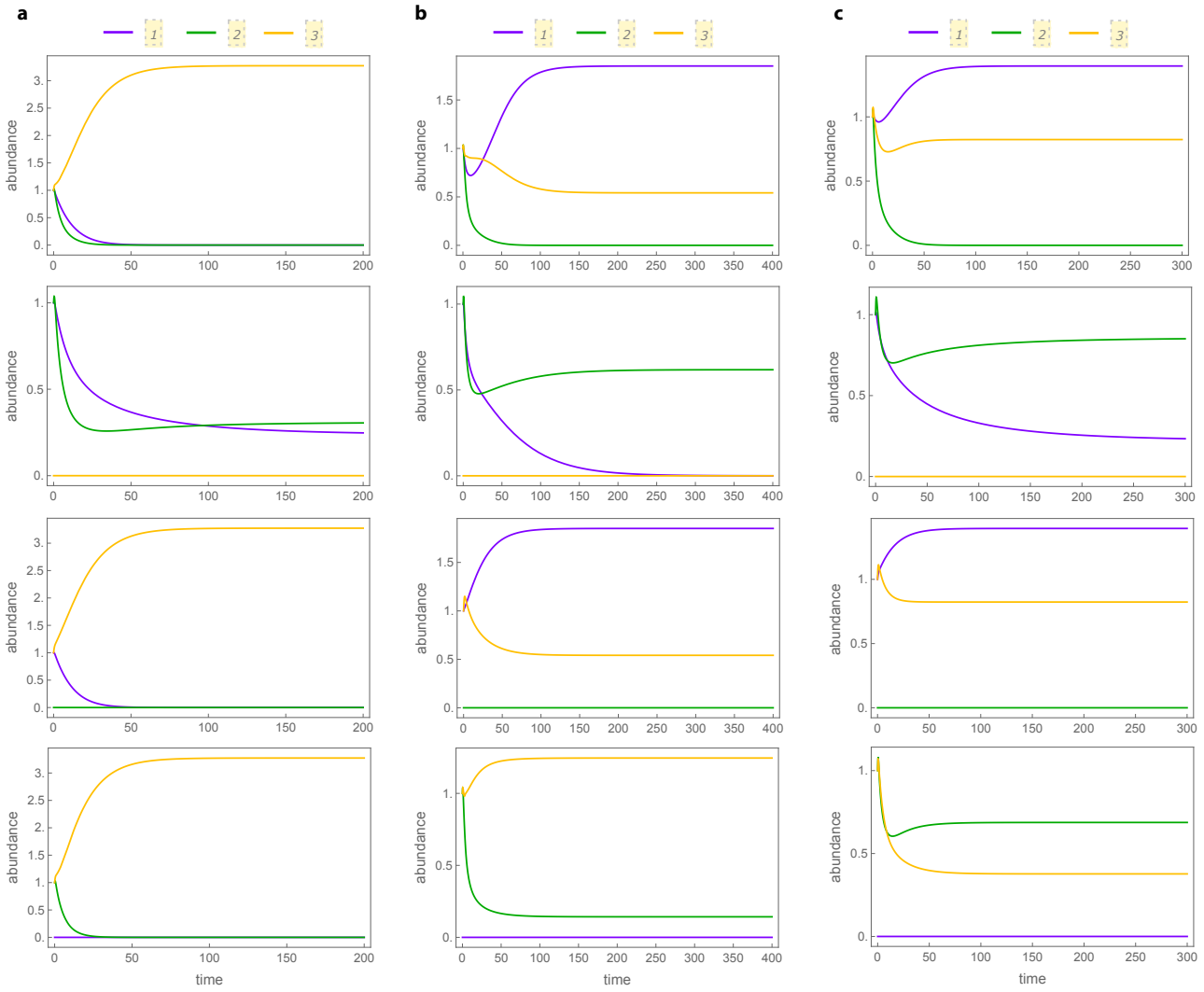
Finally, we define:

Definition 4. *The* **disassembly hypergraph** $D(H)$ *of the hypergraph* $H = (\mathsf{V}, \mathsf{H})$ *is the pair* $(\mathsf{V}, \mathsf{D}(H))$ *where*

$$\mathsf{D}(H) = M(H) \cup \{h \in \mathsf{H} \text{ such that } |h| = 1\}.$$

In words, the disassembly hypergraph $D(H)$ of H contains all missing boundaries of H, together with all species that survive in isolation. Note that, given H, the disassembly hypergraph $D(H)$ is unique. With the exception of zero-dimensional hyperedges, any hyperedge $d \in D(H)$ is a sub-collection that does not coexist dissembled from of a species collection $h \in H$ that coexist. In other words, given a species collection that coexist $h \in H$, a sub-collection $d \subset h$ obtained by "disassembling" h will belong to $D(H)$ if d does not coexist.

2. Coexistence holes in small ecological systems with classic dynamics



Supplementary Figure 1 | Coexistence in Tilman's consumer-resource model. Simulation results for $S = 3$ species and $M = 2$ resources with parameters as in Figure 1 of the Main Text. **a.** Case of two 0-dimensional assembly holes. **b.** Case of one 0-dimensional assembly hole. **c.** Case of one 1-dimensional assembly hole.

3. Identifying holes in hypergraphs using homology theory

3.1 An homology theory for arbitrary hypergraphs

Consider a set of vertices $\mathsf{V} = \{1, 2, \cdots, S\}$. Recall that a hypergraph H is a collection of subset $\{h_i\}$ where $h_i \subseteq \mathsf{V}$. Each subset $h \in H$ is called a **hyperedge**.

Example 1. *For $S = 3$ vertices, an example of a hypergraph with six hyperedges is*

$$H = \big[[1], [2], [3], [1, 2], [2, 3], [1, 2, 3]\big].$$

To study the structure of hypergraphs, and in particular to identify their holes, we need to endow them with the structure of a topological space. We do this by embedding the hypergraph into an S-dimensional space. First, we associate to each vertex in V one unit vector of $\mathbb{R}^S$. That is, the vertex [1] is associated to the vector $e_1 = (1, 0, \cdots, 0) \in \mathbb{R}^S$, the vertex [2] to the vector $e_2 = (0, 1, 0, \cdots, 0) \in \mathbb{R}^S$, and so on. Second, a hyperedge $h \in H$ is associated to the relative interior of the **simplex** spanned by the unit vectors associated to the vertices it contains, a process we denote as relint(h). In this way, the embedding of the hypergraph is formalized as follows:

Definition 5. *The* **geometric realization** $|H| \subseteq \mathbb{R}^S$ *of the hypergraph H is*

$$|H| := \cup_{h \in H} \text{relint}(h).$$

Example 2. *For the hypergraph of Example 1 with $S = 3$ vertices, its geometric realization (in $\mathbb{R}^3$) is shown in Fig. 2a.*

In general, it is difficult to visualize the geometric realization of hypergraphs if $S \geq 4$. To manage this difficulty, we will often use a linear equivalent embedding of $|H|$ in some lower n-dimensional space, for some $n \leq S$. We illustrate this process in the following example:

Example 3. *Consider the hypergraph*

$$H = \big[[1], [2], [3], [4], [1, 2], [2, 4], [3, 4], [1, 2, 3], [2, 3, 4]\big].$$

The geometric realization of this hypergraph lives in $\mathbb{R}^4$. However, we can embed this geometric realization into the $n = 2$-dimensional space as shown in Fig. 2b.

The geometric realization of a hypergraph is now a subset of the Euclidean space, with each of its hyperedges forming a simplex of $\mathbb{R}^S$. In particular, the holes of the hypergraph are represented by the elements of the homology of $|H|$. However, from a mathematical viewpoint, such a topological space is difficult to analyze because it is neither open nor closed in $\mathbb{R}^S$. This fact implies that for calculating the holes of $|H|$ it is necessary to use the notion of **singular homology**, which is in general very challenging to calculate. To circumvent this challenge, we built an strategy that allows us to calculate these holes more easily using simplicial homology.

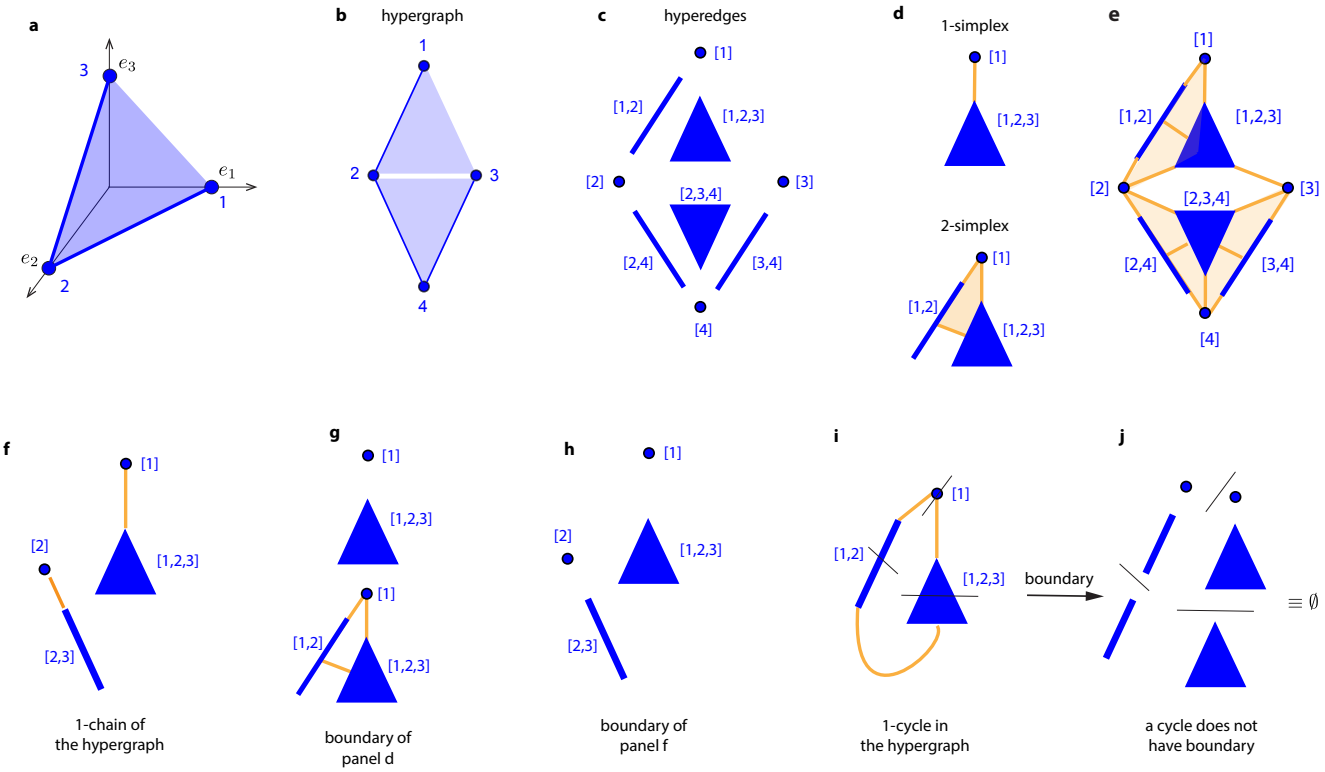
Our strategy is to construct the homology groups (or holes) as follows. We construct our homology by imagining we "explode" the hypergraph into its hyperedges. For example, the hypergraph of Fig. 2b is "exploded" into its hypereges in Fig. 2c. We will now connect back these hyperedges ensuring we keep the original structure of the hypergraph. That is, in this process of connecting back the hyperedges, not all connections will be allowed. To characterize those connections that are allowed, we define the notion of k-simplex of an hypergraph:

Definition 6. *A* **k-simplex** *of an hypergraph H is a collection $\sigma = \{h_1, h_2, \cdots, h_{k+1}\}$ such that:*

(i) $h_i \in H$ *for all i, and*

(ii) $h_i \subseteq h_{i+1}$.

In words, a k-simplex is a collection of k hyperedges of H ordered by inclusion. We illustrate this concept in the following example:



Supplementary Figure 2 | Homology for hypergraphs. Figure shows two examples of hypergraphs (panels a and b), illustrating the construction of their homology theory.

Example 4. *Consider the hypergraph shown in Fig. 2b. A 1-simplex in this hypergraph is*

$$\sigma = \big\{[1],\ [1,2,3]\big\}.$$

This simplex is "valid" because $[1] \subseteq [1,2,3]$. *This 1-simplex can be visualized as a link connecting the hyperedge* $[1]$ *and the hyperedge* $[1,2,3]$, *see top of Fig. 2d. By contrast, note that* $\{[3],[1,2]\}$ *is not a 1-simplex because these two hyperedges cannot be ordered by inclusion.*

Example 5. *A 2-simplex of the hypergraph in Fig. 2b is* $\sigma = \{[1],[1,2],[1,2,3]\}$, *see bottom part of Fig. 2d. There are no bigger simplices for this hypergraph.*

We can put together connections between groups of k hyperedges by "adding together" different k-simplices. This idea is formalized by the notion of a k-chain, which we define as below:

Definition 7. *A* k**-chain** *of the hypergraph* H *is the formal sum*

$$c = \sum_q n_q \sigma_q,$$

where each σ_q *is a* k*-simplex of* H, *and coefficient* $n_q \in \{0,1\}$ *with addition defined modulo 2.*

Recall that modulo-2 addition means that $1+1=0$. In other words, "repeated" hyperedges cancel out each other. We illustrate the above notion in the following example:

Example 6. *For the hypergraph of Fig. 2b, a 1-chain is*

$$c = \big\{[1],\ [1,2,3]\big\} + \big\{[2],\ [2,3]\big\}.$$

Chains in a hypergraph can be visualized as connecting its pieces. For example, the above chain describes the two pieces shown as orange edges in Fig. 2f.

From the definition above, we can construct the k**-th chain group** of the hypergraph H, denoted by $\langle \mathsf{C}_k(H), + \rangle$. This is the group of all k-chains with "+" denoting addition modulo 2.

The final notion that we need is that of boundary, defined as follows:

Definition 8. *The* **boundary** *of a k-simplex $\sigma = \{h_1, h_2, \cdots, h_{k+1}\}$ of H is the $(k-1)$-th chain*

$$\partial_k \sigma = \sum_{i=1}^{k+1} \sigma \setminus h_i = \sum_{i=1}^{k+1} \{h_1, h_2, \cdots, h_{k+1}\} \setminus h_i, \quad k \geq 0.$$

By completeness, we define $\mathsf{C}_{-1}(H)$ as the trivial group $\{0\}$, and the 0-boundary $\partial_0 : \mathsf{C}_0(H) \to \mathsf{C}_{-1}(H)$ as the zero epimorphism.

Therefore, the boundary is an operator that decreases the order, i.e., $\partial_k : \mathsf{C}_k(H) \to \mathsf{C}_{k-1}(H)$. By linearity, we extend this definition to chains as follows: if $c = \sum_q n_q \sigma_q$ is a k-chain then its boundary is defined as

$$\partial_k c := \sum_q n_q \, \partial_k \sigma_q.$$

Example 7. *Consider the 1-simplex $\sigma = \big\{[1],\ [1,2,3]\big\}$ of Fig. 2d. Its boundary is*

$$\partial_1 \sigma = \big\{[1],\ [1,2,3]\big\} \setminus [1] + \big\{[1],\ [1,2,3]\big\} \setminus [1,2,3] = [1] + [1,2,3],$$

see Fig. 2g. For the chain $c = \big\{[1],\ [1,2,3]\big\} + \big\{[2],\ [2,3]\big\}$ of Fig. 2f, its boundary is

$$\partial_1 c = [1] + [1,2,3] + [2] + [2,3],$$

see Fig. 2h.

The so-called **fundamental boundary property** states that $\partial_k \partial_{k+1} = 0$. Note that this property applies using to above definition of boundary. In particular, this property implies that the boundary of a **cycle** is zero. We illustrate this latter fact in the following example:

Example 8. *For the hypergraph of Fig. 2b consider the 1-chain*

$$c = \big\{[1], [1,2,3]\big\} + \big\{[1,2,3], [1,2]\big\} + \big\{[1,2], [1]\big\}.$$

This chain starts and ends in the hyperedge $[1]$*, see Fig. 2i. Its boundary is*

$$\partial_1 c = [1] + [1,2,3] + [1,2,3] + [1,2] + [1,2] + [1] = 0,$$

which is empty because addition modulo 2-addition cancels repeated hyperedges, see Fig. 2j.

The idea is now to use cycles to detect holes. Intuitively, a cycle could be the boundary of some chain, in which case it is a "filled" cycle. Cycles could also be the boundary of no chain —and in particular, of no hyperedge— implying they are "empty". Consequently, "empty cycles" characterize the boundary of a hole. On the other hand, note that different cycles may encircle the same hole. Therefore, to count the number of different holes using cycles, it is necessary to construct an equivalence class of all cycles encircling the same hole.

To make the above ideas operative, we use the boundary operator to characterizes two key subgroups of the k-th chain group $\mathsf{C}_k(H)$ of a hypergraph H:

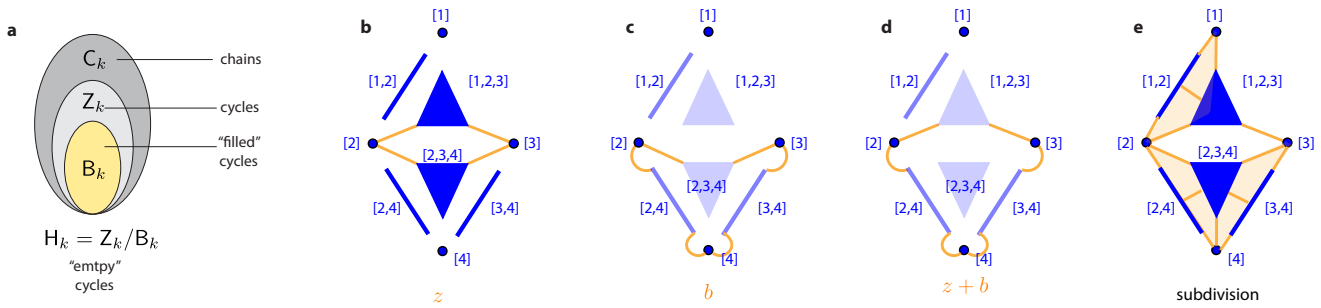
Definition 9.

(i) The k-th **cycle group** Z_k *is*

$$\mathsf{Z}_k = \ker \partial_k := \{c \in \mathsf{C}_k | \partial_k c = \emptyset\}.$$

(ii) The k-th **boundary group** B_k *is*

$$\mathsf{B}_k = \operatorname{im} \partial_{k+1} := \{c \in \mathsf{C}_k | \exists d \in \mathsf{C}_{k+1} \text{ such that } c = \partial_{k+1} d\}.$$



Supplementary Figure 3 | Defining the homology of hypergraphs.

A chain $c \in \mathsf{B}_k$ is the boundary of some higher-dimensional chain $d \in \mathsf{C}_{k+1}$. Therefore, such a k-chain c is a k**-boundary** or a **bounding cycle**. Cycles that are not in B_k are **non-bounding** cycles. Therefore, bounding cycles bound higher-dimensional chains so they are "filled cycles". Non-bounding cycles are "empty cycles". Note also that, since $\partial_k \partial_{k+1} = 0$, it follows that $\mathsf{B}_k \subseteq \mathsf{Z}_k \subseteq \mathsf{C}_k$, as illustrated in Fig. 3a.

It turns out that both subgroups Z_k and B_k are normal (because they are abelian), allowing the construction of quotient spaces. To illustrate the implications of this fact, consider again the hypergraph of Fig. 2b. Let $b \in \mathsf{B}_1$ and $z \in \mathsf{Z}_1$ be bounding- and non-bounding cycles of Fig. 3b and Fig. 3c, respectively. Glueing both cycles results in the cycle $z + b$ of Fig. 3d. Crucially, note that $z + b$ is **homologous** to z. Namely, we can **retract** $z + b$ along the solid triangle $\{[2,4],[2,3,4],[4],[3,4]\}$ to obtain exactly z. In this sense, there exists only **one** cycle of the form $z + \mathsf{B}_1$.

The above idea is formalized by the notion of quotient space $\mathsf{Z}_1/\mathsf{B}_1$. Extending this idea from $k = 1$ to an arbitrary $k \geq 1$ leads to the definition of homology groups:

Definition 10. *The k-th* **homology group** H_k *is*

$$\mathsf{H}_k = \mathsf{Z}_k/\mathsf{B}_k = \ker \partial_k / \operatorname{im} \partial_{k+1}.$$

Thus, if $z_1 = z_2 + \mathsf{B}_k$ for some $z_1, z_2 \in \mathsf{Z}_k$, we say that the empty cycles z_1 and z_2 are **homologous**.

The number of "different" empty cycles provide a characterization of the "holes" of the space. This is formalized by the Betti numbers. Specifically, we have:

Definition 11. *The k-th* **Betti number** β_k *of the hypergraph H is* $\beta_k = \operatorname{rank} \mathsf{H}_k$.

Therefore, the k-th Betti number corresponds to the number of k-dimensional holes of H. In particular, β_0 provides the number of connected components of H.

3.2 The constructed homology captures exactly all holes

To finish, we prove that the above constructed homology theory will capture exactly all holes of the hypergraph. With this aim, let us first define:

Definition 12. *For a hypergraph H, its associated* **subdivision** $S(H)$ *is the hypergraph such that:*

(i) $S(H)$ has one vertex for each hyperedge of H.

(ii) $S(H)$ contains as hyperedges all the k-simplices of H for $k \geq 0$.

Note that $S(H)$ is by construction a simplicial complex since any subset of a k-simplex is also a k-simplex. Moreover, the geometric realization of $S(H)$ satisfies $|S(H)| \subseteq |H|$ by embedding every vertex of $h \in S(H)$ as the barycenter of the relint(h) in $|H|$. We illustrate this latter fact in the following example:

Example 9. *For the hypergraph of Fig. 4a, its subdivision is shown in Fig. 4b. Note the subdivision $S(H)$ is contained in H.*

Note that our notion of subdivision is analogous to the well-known first barycentric subdivision[8] of a simplicial complex.

Our main result proves that the constructed homology captures exactly all holes of the geometric realization of an hypergraph. That is, any hole in the geometric realization of a hypergraph corresponds to a "empty cycle" family, and every "empty cycle" family corresponds to a hole in the geometric realization of the hypergraph. Mathematically, such property means that the homology of a hypergraph and its subdivision are identical:

Theorem 1. *The homology of H and $S(H)$ are identical.*

Proof. To prove the claim, it is sufficient to construct a strong deformation retraction

$$r : |H| \to |S(H)|.$$

Suppose $V = [n] = \{\tilde{0}, \tilde{1}, \ldots, \tilde{n}\}$, where we identify $\tilde{i} \in [n]$ with the point $(0, \ldots, 1, \ldots, 0) \in \mathbb{R}^{n+1}$ with all coordinates 0 except the $(i+1)$-th coordinate which is 1. So, the hypergraph H is contained in the simplex Δ with vertices V. Furthermore, its realization $|\Delta|$ is the convex hull of the set of points $\{\tilde{0}, \tilde{1}, \ldots, \tilde{n}\} \in \mathbb{R}^{n+1}$.

Define $F = \Delta \setminus H$ and $F^{(i)}$ as the subset of F consistent of those elements of size i. Furthermore, define L_i the subcomplex of Δ' (the first baricentric subdivision of the simplex Δ) generated be the following vertices of Δ' (remember that the vertices of Δ' are subsets of V):

$$2^V \setminus \cup_{j>i} F^{(i)}.$$

Finally, define Z_i, the finite collection of vertices of the simplicial complex Δ', as follows: $Z_i = \{\hat{\sigma} \mid \sigma \in F^{(i)}\}$, $1 \leq i \leq m$.

The following is easy to verify:

1. $L_0 = H'$,
2. If Δ is an $(m-1)$-simplex, then $L_m = \Delta'$
3. L_{i-1} is a subsimplex of L_i, and $|L_i| \setminus |L_{i-1}|$ is a pairwise disjoint union of a finite (possible empty) collection of open cones with apices in Z_i, $1 \leq i \leq m$.

Define the strong deformation retraction along the lines of the cones

$$r_i : |L_i| \setminus Z_i \to |L_{i-1}|$$

In fact,

$$r_i(|L_i| \setminus Z_i) = \cup_{\sigma \in F^{(i)}} \partial\sigma * \tau_\sigma$$

where $\partial\sigma$ is the subcomplex of Δ' which is the boundary of σ and τ_σ is a simplex of Δ' with vertices $\{v_{\sigma_1}, \ldots, v_{\sigma_p}\}$, where $\sigma < v_{\sigma_1} < \cdots < v_{\sigma_p}$ and $v_{\sigma_j} \in L_{i-1}$, $1 \leq j \leq p$.

Note that $v_{\sigma_j} \in \Delta$ and its size is greater than i, and since it is not in L_{i-1}, then $v_{\sigma_j} \notin F$. Consequently,

$$r_i(|L_i| \setminus Z_i) \subset r_i(|L_i| \setminus \cup_{\sigma \in F^{(i)}} \text{relint } |\sigma|) \subset \cup_{\sigma \in F^{(i)}} \cup_1^p \text{ relint } |\tau_{\sigma_j}| \subset |L_{i-1}| \cap |H|$$

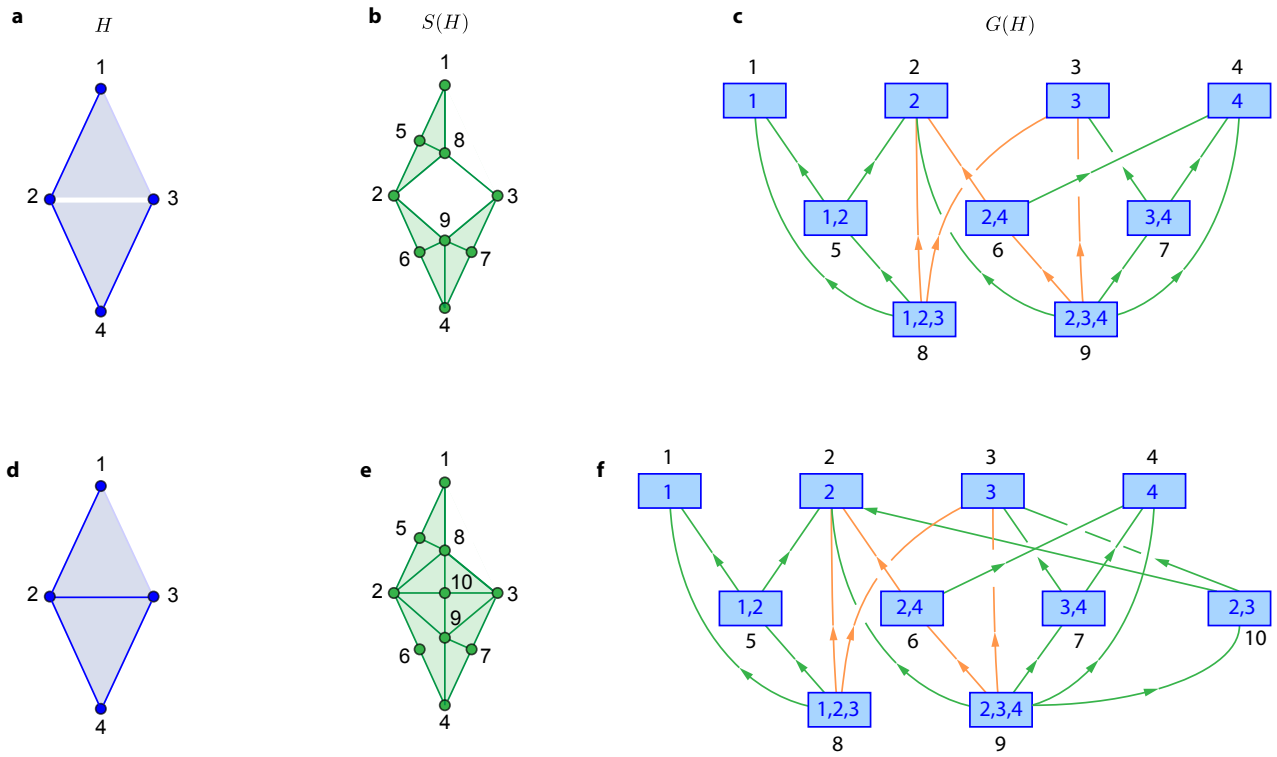
and then $r_i(|L_i| \cap |H|) \subset |L_{i-1}| \cap |H|$.

This allows to define $r : (|L_m| \cap |H|) = |H| \to (|L_0| \cap |H|) = |H'|$ as the composition $r_m \circ \cdots \circ r_1$. Note that r_i |: $|L_i| \cap |H| \to |L_{i-1}| \cap |H|$ is defined along the lines of the cones, then r is also a strong deformation retraction.

□

3.3 An algorithm for calculating the homology of arbitrary hypergraphs

Theorem 1 guarantees that that we can calculate the homology of any hypergraph H by calculating the homology of its subdivision $S(H)$, which is a simplicial complex. Here we use this fact to implement an



Supplementary Figure 4 | Analyzing the homology of hypergraphs. **a.** The hypergraph $H = [[1],[2],[3],[4],[1,2],[2,4],[3,4],[1,2,3],[2,3,4]]$. **b.** Its subdivision $S(H)$. Note this is a simplicial complex. **c.** Inclusion graph $G(H)$ of H, used to compute the subdivision of the graph. **d.** Another hypergraph without holes. **e.** Subdivision for the hypergraph of panel d. **f.** Inclusion graph for the hypergraph of panel d.

efficient algorithm to calculate the homology of arbitrary hypergraph leveraging on existing algorithms for calculating the homology of simplicial complexes via their Vietoris-Rips complexes.

First, note that the notion of hypergraph subdivision in Definition 12 can be understood abstractly as replacing each hyperedge by a vertex in a new hypergraph, and then connecting these new vertices according to the inclusion of hyperedges they represent. We illustrate this fact in the elementary hypergraph H of Fig. 4a. The corresponding subdivision $S(H)$ is shown in Fig. 4b. This subdivision contains nine vertices, one per hyperedge of H. For example, the vertex 9 in $S(H)$ is associated to the hyperedge $[2,3,4]$ in H, and so on. The hyperedges of $S(H)$ are obtained by inclusion. For example, the hyperedge $[1,8] \in S(H)$ exists because $[1] \subseteq [1,2,3]$ in the hyperdges of H.

The above observation is useful because it allows to implement the subdivision in two steps:

1. **Build the inclusion graph**. This is a directed graph $G(H)$ containing the same vertex set V' as $S(H)$ (i.e., one per hyperedge of H). A directed edge $(v_i' \to v_j')$ is included if the hyperedges $h_i, h_j \in H$ associated to $v_i', v_j' \in \mathsf{V}'$ satisfy $h_j \subset h_i$. In Fig. 4c, we illustrate the construction of the inclusion graph for the hypergraph of Fig. 4a.

2. **Calculate the Vietoris-Rips complex associated to the undirected inclusion graph**. Namely, let $G_u(H)$ denote the inclusion graph $G(H)$ with the direction of edges removed. Then, the Vietoris-Rips complex is a simplicial complex with vertices V' and with one simplex σ for each complete subgraph of $G_u(H)$.

For the inclusion graph of Fig. 4c, note that its Vietoris-Rips complex is precisely the hypergraph subdivision of Fig. 4b. This is not a coincidence, as we can prove the following:

Proposition 1. *Let $K(H)$ denote the Vietoris-Rips complex built from $G(H)$. Then $S(H) = K(H)$.*

Proof. The proof follows by observing that, by construction, each simplex of $\sigma \in S(H)$ is a complete subgraph of $G(H)$. Namely, let $\sigma \in S(H)$ be a simplex in the subdivision. Let $[h_1, h_2, \cdots, h_n]$ be the

hyperedges of H associated to this simplex. By Definition 12, it follows that $h_1 \subseteq h_2 \subseteq \cdots \subseteq h_n$. Therefore, the associated vertices $h_1, h_2, \cdots, h_2$ form a complete subgraph in $G(H)$. On the other hand, if $h_1, h_2, \cdots, h_2$ is a complete subgraph of $G(H)$, then they are ordered by inclusion as $h_1 \subseteq h_2 \subseteq \cdots \subseteq h_n$, and hence they should be a simplex in $S(H)$. □

Proposition 1 guarantees that our subdivision algorithm is correct in the sense that it calculates the subdivision of any hypergraph. We provide the Julia function

`hypergraph_subdivision(H, simplification = false)`,

returning the subdivision of the hypergraph `H`. The options for `simplification` are `true` or `false`. Choosing `simplification = true` will return the simplicial complex $S(H)$ by calculating the Vietoris-Rips complex of the inclusion graph. Choosing `simplification = false` will return the inclusion graph $G(H)$ only. Furthermore, the relevance of calculating the hypergraph subdivision through a Vietoris-Rips complex is that it allows us to efficiently calculate the homology of the subdivision, as we discuss in the next subsection.

We emphasize that, despite both an hypergraph H and its inclusion graph $G(H)$ contain exactly the same information, it is not easy or immediate for humans to identify the holes by looking only at $G(H)$. Indeed, it is the geometric information that is added to hypergraph when it is (geometrically) realized that allows us to detect and most easily conceptualize the presence of holes. We illustrate this fact in the following example:

Example 10. *For the particular hypergraph H of this example (Fig. 4a), there is a one-dimensional hole between species $[2, 3]$, which is made evident in its subdivision $S(H)$ (Fig. 4b). The corresponding inclusion graph $G(H)$ is shown in Fig. 4c, where vertices are hyperedges (e.g., species collections that coexist), and edges connect two vertices if they can be ordered by inclusion (i.e., one is contained in the other). From this inclusion hypergraph, one may be tempted to conclude that the hole exists because there is a cycle $\{[2, 9], [9, 3], [3, 8], [8, 2]\}$, which is precisely the boundary of the hole (orange cycle in Fig. 4c, note the direction of the edges is not considered). The challenge here is then to conclude if such bounding cycle is "empty" or "not". To understand the difficulty of this challenge, in Fig. 4f we show the inclusion graph in the case the bounding cycle is actually filled. Despite the bounding cycle remains in the inclusion graph (orange), it is not obvious by looking only to the graph what changed from Fig. 4c to Fig. 4f. By contrast, it the additional geometric structure added to the hypergraph and its subdivision allow us to immediately conclude that the bounding cycle is filled in the latter case (Fig. 4d,e). In this sense, the inclusion graph $G(H)$ is a more "primitive" representation compared to the (geometrical realization) of a hypergraph H.*

4. Constructing assembly hypergraphs from population dynamics models

Here we start considering ecological systems with population dynamics described by the Lotka Volterra (LV) model. The LV is the classical population dynamics model that has been successfully applied to model a very diverse class of ecological systems, from terrestrial, lake and marine foodwebs, to plant ecological systems [9, 10] and microbial communities [11, 12] in the human body. Compared to other population dynamics models, calculating the assembly hypergraph for LV models is computationally more efficient because analytical conditions for coexistence exist. In Section 4.3, we discuss using more general population dynamics to construct assembly hypergraphs.

4.1 Estimating assembly and disassembly hypergraphs using the Lotka-Volterra model.

In vector form, the Lotka-Volterra model for S species is

$$\frac{\mathrm{d}x(t)}{\mathrm{d}t} = x(t) \odot [Ax(t) + r], \quad x(0) = x_0, \tag{S1}$$

where $x(t) = (x_1(t), \cdots, x_N(t))^\intercal \in \mathbb{R}^S_{\geq 0}$ and $x_i(t)$ is the abundance of the i-th species at time $t \geq 0$. Above, $x \odot v$ denotes the entry wise multiplication of vectors $x, v \in \mathbb{R}^S$. The parameters of the GLV model are $A \in \mathbb{R}^{S\times S}$ and $r \in \mathbb{R}^S$, representing the interaction matrix and the intrinsic growth rate of species, respectively.

We assume we are given the parameters (A, r) capturing the population dynamics of the ecological system. These parameters can be inferred from experimental data using system identification techniques. Then, the objective is to build the assembly hypergraph by using the Eq. (S1) to predict if certain species collection $\Sigma \in 2^V$ coexists or not. To test for coexistence, we considered the criterion of **permanence** [13–15]. In the following two sub-subsection we describe such criterion in details. Overall, we our Julia package contains all functionalities needed to build the assembly hypergraph from a LV model using the following function:

`assembly_hypergraphGLV(A, r; method, regularization, z_tolerance, iterations)`.

This function returns an array of arrays `H` containing the hyperedges of the assembly hypergraph (i.e., species collections that coexist). Above, the parameters:

1. `A, r` are the interaction matrix A and intrinsic growth rate vector r of the ecological system.
2. `method` specifies the criterion for testing coexistence. The function accepts two criteria: "permanence" or "localstabilitiy".
3. The remaining three parameters, are accepted only for the permanence criterion:
   - `regularization = reg` adds the negative random values Uniform(-reg, 0) to the diagonal of A. This is useful to obtain bounded abundances when the A matrix has zero diagonal and some species are autotrophs in the environment (i.e., $r_i > 0$). By default, the function uses `regularization = 0`.
   - `z_tolerance` sets the tolerance for solving the linear program that characterizes permanence. By default, the function uses `z_tolerance = -1e-60`.
   - `iterations` is the number of iterations used to solve the linear program that characterizes permanence. By default, the function uses `iterations= 5e4`.

In our results we used the default parameters. From the assembly hypergraph H, our Julia package allows constructing the disassembly hypergraph D using the function

`disassembly_hypergraph(H)`.

4.2 Permanence.

Permanence has been used as a criterion for coexistence in various seminal studies [13–15]. Mathematically, permanence is defined by the existence of an attractive set in the interior of the state space. To introduce this notion, consider an ecological system of S species and let $\mathsf{V} = \{1, 2, \cdots, S\}$ the set of all species.

Definition 13. *[16, pp. 160]. The species collection $\Sigma \in 2^{\mathsf{V}}$ is* **permanent** *if there exists a compact set $\Omega \subset \mathbb{R}^{|\Sigma|}_{>0}$ in the interior of the state space such that all orbits in the interior end up in Ω.*

Let $x_i(t)$ denote the abundance of the i-th species at time $t \geq 0$. Permanence means that, whenever $x_i(0) > 0$ for all $i \in \Sigma$ and $x_j(0) = 0$ for all $j \notin \Sigma$, there exists constants $0 < \delta \leq D < \infty$ such that

$$\delta < \liminf_{t\to\infty} x_i(t) \leq \limsup_{t\to\infty} x_i(t) \leq D, \quad \forall i \in \Sigma.$$

In words, permanence means that if some species are initially present, then the population dynamics will not lead to their extinction.

Consider a species collection $\Sigma \in 2^{\mathsf{V}}$ and let $A_\Sigma \in \mathbb{R}^{|\Sigma|\times|\Sigma|}$ and $r_\Sigma \in \mathbb{R}^{|\Sigma|}$ denote the submatrix of A and subvector of r obtained by considering only the species in Σ. In our work, we used the following conditions to determine the permanence of a species collection Σ:

1. A **necessary condition for permanence** is the existence of a feasible interior equilibrium [16, pp. 169]. Namely, the existence of a point $x^*_\Sigma \in \mathbb{R}^{|\Sigma|}_{>0}$ such that $A_\Sigma x^*_\Sigma + r_\Sigma = 0$.

2. A **sufficient condition for permanence** is the existence of vector $h_\Sigma \in \mathbb{R}^{|S|}_{\geq 0}$ such that

$$h^\intercal_\Sigma\,(r_\Sigma + A_\Sigma x^*_\Sigma) > 0, \quad \forall x^*_\Sigma \in \mathcal{E}(\Sigma).$$

   Above, $\mathcal{E}(\Sigma)$ is the set of non-negative boundary equilibria of the species collection Σ, i.e., $\mathcal{E}(S) = \{x^* \in \mathbb{R}^{|\Sigma|}_{\geq 0} | A_\Sigma x^* + r_\Sigma = 0\}$. The above conditions is usually referred as Jansen's criterion of permanence [17] (see also [16, pp. 176]).

   We note that Jansen's criterion is a necessary and sufficient for permanence when $|\Sigma| \leq 3$, and it is only sufficient for $|\Sigma| \geq 4$.

Jansen's criterion has been a popular in applications because it can be written as a linear program [13, 14]. Namely, Let $n^*_\Sigma = |\mathcal{E}(\Sigma)|$ and denote by $X^*_\Sigma = [x^*_1|\cdots|x^*_{n^*}] \in \mathbb{R}^{n\times n^*}$ the matrix obtained by concatenating the equilibria of $\mathcal{E}(\Sigma)$ by columns, with $x^*_i \in \mathbb{R}^{|\Sigma|}$. Then, Jansen's criterion can be rewritten as the linear program:

$$\text{solve } h^\intercal A_\Sigma X^*_\Sigma + h^\intercal r_S \Sigma \mathbf{1}^\intercal > 0, \quad \text{subject to } h \in \mathbb{R}^{|\Sigma|}_{\geq 0}.$$

Above, $\mathbf{1}^\intercal = (1, 1, \cdots, 1) \in \mathbb{R}^{n^*_S}$. For implementation purposes, we rewrite the above linear program as

$$\text{solve } h^\intercal A_\Sigma X^*_\Sigma + h^\intercal r_\Sigma \mathbf{1}^\intercal + z\mathbf{1}^\intercal \geq 0, \quad \text{subject to } h \in \mathbb{R}^{|\Sigma|}_{\geq 0}, \tag{S2}$$

where $z < 0$ is a small constant.

We built a Julia function to calculate the permanence of a species collection Σ with parameters (A_Σ, r_Σ):

`is_GLVpermanent(Asigma, rsigma; regularization, z_tolerance, iterations)`.

This function returns `true` if conditions 1 and 2 above are satisfied, and returns `false` if at least one of those two conditions is not satisfied. The parameters of this function are:

- `regularization = reg` adds the negative random values Uniform(-`reg`, 0) to the diagonal of A. This is useful to obtain bounded abundances when the A matrix has zero diagonal and some species are autotrophs in the environment (i.e., $r_i > 0$). By default, the function uses `regularization = 0`.

- `z_tolerance` sets the value for the variable z in Eq. (S2). By default, the function uses `z_tolerance = -1e-60`.
- `iterations` is the maximum number of iterations used to solve the linear program of Eq. (S2). By default, the function uses `iterations= 5e4`.

4.3 Estimating assembly hypergraphs from general population dynamics models.

In principle, more general population dynamics model can better capture the coexistence of ecological systems. Compared to the LV equations of Eq. (S1), other models may use different functional responses [18, 19] to describe the interactions between species (such as replacing the term $x_i x_j$ by the Holling Type II term $x_i x_j/(\theta + x_j)$), and may contain higher order interactions [20] (term like $x_i x_j x_k$). Similarly, in addition to the species dynamics, models may incorporate the dynamics of available resources (i.e., abiotic conditions), such as in the classical MacArthur's consumer-resource model [21]. In this sense, a rather general class of population dynamics modeling the abundance of S species $x(t) \in \mathbb{R}^S_{\geq 0}$ and M resources $R(t) \in \mathbb{R}^M_{\geq 0}$ correspond to the differential equations:

$$\begin{aligned} \frac{\mathrm{d}x(t)}{\mathrm{d}t} &= x(t) \odot f_\theta\big(x(t), R(t)\big), \quad x(0) = x_0; \\ \frac{\mathrm{d}R(t)}{\mathrm{d}t} &= R(t) \odot g_\gamma\big(x(t), R(t)\big), \quad R(0) = R_0. \end{aligned} \tag{S3}$$

Above $f_\theta : \mathbb{R}^N_{\geq 0} \times \mathbb{R}^M_{\geq 0} \to \mathbb{R}^N$ and $g_\gamma : \mathbb{R}^N_{\geq 0} \times \mathbb{R}^M_{\geq 0} \to \mathbb{R}^M$ are functions parametrized by the parameters (θ, γ). The pair $\{f_\theta, g_\gamma\}$ characterize the population dynamics of the ecological system. For example, MacArthur's model is obtained by using the affine functions

$$f_\theta(x, R) = \Theta_1 R + \theta_2, \quad g_\gamma(x, R) = \Gamma_1 R + \Gamma_2 x + \gamma_3.$$

Here we assume we are given such a pair $\{f_\theta, g_\gamma\}$ with their parameters adjusted to represent an empirical ecological system. For certain pairs and under additional hypothesis, the coexistence of species can be analytically derived. One example is MacArthur's model with time-scale separation between resources and species, where the existence of a unique interior equilibrium of species implies its global stability [22]. In general, however, analytically characterizing coexistence of a species collection from an arbitrary pair $\{f_\theta, g_\gamma\}$ is very challenging. Therefore, for such general models, permanence needs to be numerically estimated. The basic idea is to numerically solve Eqs. (S3) for an ensemble of initial conditions (x_0, R_0). Then, we calculate the fraction of the resulting trajectories that enter and remain in an interior compact set for a long enough simulation time interval.

5. Assembly and disassembly holes in systems with Lotka-Volterra dynamics

5.1 Increasing the "complexity" of an interaction matrix generates more skeletons

In Fig. 3 of the Main Text, we studied how the assembly and disassembly skeletons are generated by a pair (A, r). In particular, Fig. 3 contains the skeletons generated by the following hypothetical interaction matrix:

$$A = \begin{pmatrix} -1. & 0.250537 & 0. & -0.198409 & 0.327808 & 0. \\ 0. & -1. & -0.0825464 & 0. & 0. & 0. \\ 0. & 0. & -1. & 0. & -0.294949 & 0. \\ 0.210027 & 0.0230585 & 0.0537077 & -1. & 0.608267 & 0.100475 \\ 0. & 0.502158 & -0.156564 & 0. & -1. & 0.287838 \\ 0.394669 & 0. & -0.257042 & 0. & 0. & -1. \end{pmatrix} \tag{S4}$$

We can increase the "complexity" of this interaction matrix in two ways: (1) increasing the strength of the interspecific interactions already present, or (2) increasing the number of non-zero interspecific interaction. For example, we can increase the complexity of Eq. (S4) by increasing its interspecific strengths as follows:

$$A = \begin{pmatrix} -1. & 1.24247 & 0. & -0.74856 & 4.05051 & 0. \\ 0. & -1. & -0.506785 & 0. & 0. & 0. \\ 0. & 0. & -1. & 0. & -1.59741 & 0. \\ 1.61705 & 0.605579 & 0.483881 & -1. & 10.0131 & 0.6016 \\ 0. & 2.92263 & -0.794511 & 0. & -1. & 2.11763 \\ 4.54829 & 0. & -2.83429 & 0. & 0. & -1. \end{pmatrix}.$$

Note this interaction matrix has the same zeros as the original one (Supplementary Fig. 5a). Increasing the complexity in this way increases the number of possible assembly and disassembly skeletons (Supplementary Fig. 5b).

As mentioned above, we can also increase the complexity of the interaction matrix by increasing the number of non-zero interspecific interactions between species. For example, we can increase the complexity of Eq. (S4) as follows:

$$A = \begin{pmatrix} -1. & 0.250537 & -0.257042 & -0.198409 & 0.327808 & 0.250537 \\ 0.0825464 & -1. & -0.0825464 & 0.100475 & 0.608267 & -0.210027 \\ 0.0825464 & -0.198409 & -1. & -0.100475 & -0.294949 & -0.327808 \\ 0.210027 & 0.0230585 & 0.0537077 & -1. & 0.608267 & 0.100475 \\ 0.327808 & 0.502158 & -0.156564 & 0.210027 & -1. & 0.287838 \\ 0.394669 & -0.394669 & -0.257042 & 0.250537 & 0.198409 & -1. \end{pmatrix}.$$

Note this interaction matrix "fills" the zero entries of the original one (Supplementary Fig. 5c). Increasing the complexity in this way also increases the number of possible assembly and disassembly holes (Supplementary Fig. 5d).

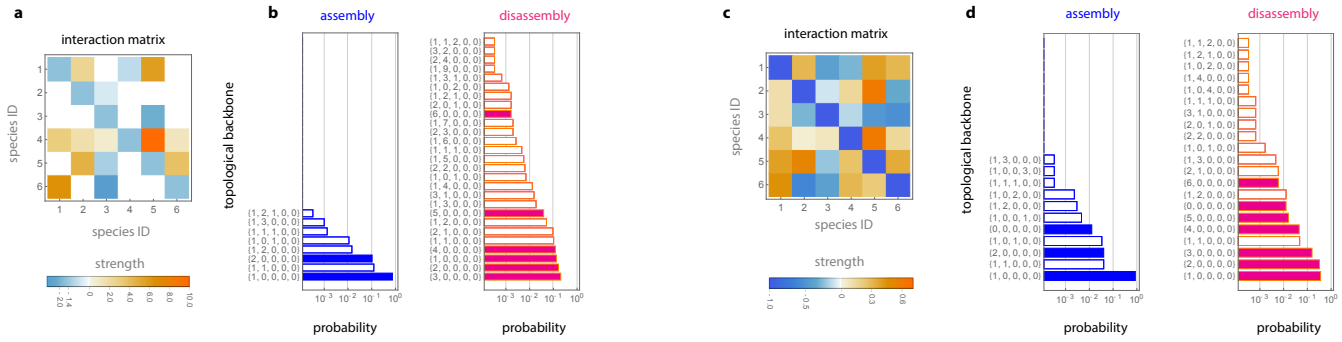
5.2 Constructing an ensemble of random LV models

To systematically analyze how coexistence holes emerge in ecological systems with the LV dynamics, we generated random LV models using the methodology developed in previous studies (see, e.g., [23, 24]). More precisely:

1. For the interaction matrix $A = (a_{ij})$, its non-diagonal entries a_{ij} are constructed by sampling as

$$a_{ij} \sim \text{Bernoulli}(C_A)\ \text{Normal}(0, \sigma_A), \quad i \neq j.$$

Above, $C_A \in [0, 1]$ characterizes the **connectance** of the interaction matrix —controlling the frac-



Supplementary Figure 5 | Changing the complexity of the interaction matrix generates more assembly and disassembly skeletons. a. An interaction matrix with identical zero-pattern as in Eq. (S4), but with higher inter-species interaction strengths. **b.** Increasing the interaction strength generated a larger diversity of assembly and disassembly skeletons. **c.** An interaction matrix with identical interspecific interaction strengths as in panel a, but with more non-zero interactions. **d.** Increasing the connectivity generated a larger diversity of assembly and disassembly skeletons.

tion of zero entries— and $\sigma_A \geq 0$ is the **characteristic interspecific interaction strength** between species. Diagonal entries are set to $a_{ii} = -1$ for $i = 1, \cdots, S$.

2. For the intrinsic growth rate vector r, we considered a randomization where its 2-norm is maintained but its direction is chosen uniformly at random over a sphere. More precisely, given a vector $r_0 \in \mathbb{R}^S$, we first generate a vector $v \in \mathbb{R}^S$ with $v_i \sim \text{Normal}(0, 1)$. The randomized version of r_0 is then

$$r = \|r_0\|_2 \frac{v}{\|v\|_2}.$$

Recall that, by sampling $v_i \sim \text{Normal}(0, 1)$, the vector $v/\|v\|_2$ is uniformly distributed over S-dimensional unit sphere.

3. For the intrinsic growth rate vector r, we also considered a randomization where growth rates are samples uniformly at random over the positive sector of the units sphere. This was done by independently sampling $v_i \sim \text{Normal}(0, 1)$ for $i = 1, \cdots, N$, and rejecting the sample if at least one $v_i < 0$. Then, we built

$$r = \|r_0\|_2 \frac{v}{\|v\|_2}.$$

5.3 Random species interactions very likely generate assembly and disassembly holes

To analyze how likely to observe coexistence holes in in LV ecological systems, we quantified the probability of observing assembly/diassembly holes in our ensemble of random LV ecological systems. For this, we considered ecological systems of S species. We constructed $m > 0$ random matrices A for each set of parameters $(C_A, \sigma_A) \in [0, 1] \times [0, 3]$. For each random matrix, we generated $\ell > 0$ vectors r uniformly at random with unit norm. Finally, we calculated the corresponding assembly and disassembly holes. We say that matrix A has assembly holes (or disassembly holes) if, with at least one of those r vectors, the interaction matrix generates assembly holes of some dimension (or disassembly holes of some dimension). Then, for a pair (C_A, σ_A), the expected probability of finding assembly holes (or disassembly holes) is the fraction of the m random matrices A for those parameters that have at least one assembly hole (or disassembly holes).

Figure 3d of the Main Text shows the results of the above methodology using $S = 8$ species, $m = 20$ random matrices, and $\ell = 100$ random vectors. For assembly holes, we find that the expected probability of finding assembly holes is higher than 0.8 if the connectivity satisfies $C_A > 0.4$ and the typical interspecific interaction strength satisfies $\sigma_A > 0.5$. This region where assembly holes are likely occupies a large portion of the parameter space, indicating that assembly holes are rather common. We find a similar result for disassembly holes, but in this case the region where disassembly holes are very likely is even larger. Overall, these results show that the presence of assembly and disassembly holes is the norm rather than the exception in random LV models.

5.4 Disassembly strongly influences assembly, but not vice versa

Because assembly and disassembly skeletons occur simultaneously, it is important to understand their co-occurrence. For this purpose, we calculated the joint probability of observing a pair of assembly and disassembly skeletons. The joint probability can be represented by a bipartite graph where vertices correspond to skeletons. Supplementary Fig. 6a illustrates this graph for the system with the interaction matrix of Eq. (S4) when choosing the intrinsic growth rate vector uniformly at random over the unit sphere. Here, the size of a node corresponds to the marginal probability of observing the given skeleton, and the strength of edges between nodes corresponds to the joint probability of observing a given pair of assembly and a disassembly skeletons. Supplementary Fig. 6a shows that by looking at the size of nodes in the bipartite graph generated by our hypothetical interaction matrix $\mathbf{A}$, there is less uncertainty about which assembly skeleton will occur than in the disassembly skeleton. This result is expected because the probability distribution of assembly skeletons is heavily concentrated compared to the more uniform distribution observed for disassembly skeletons.

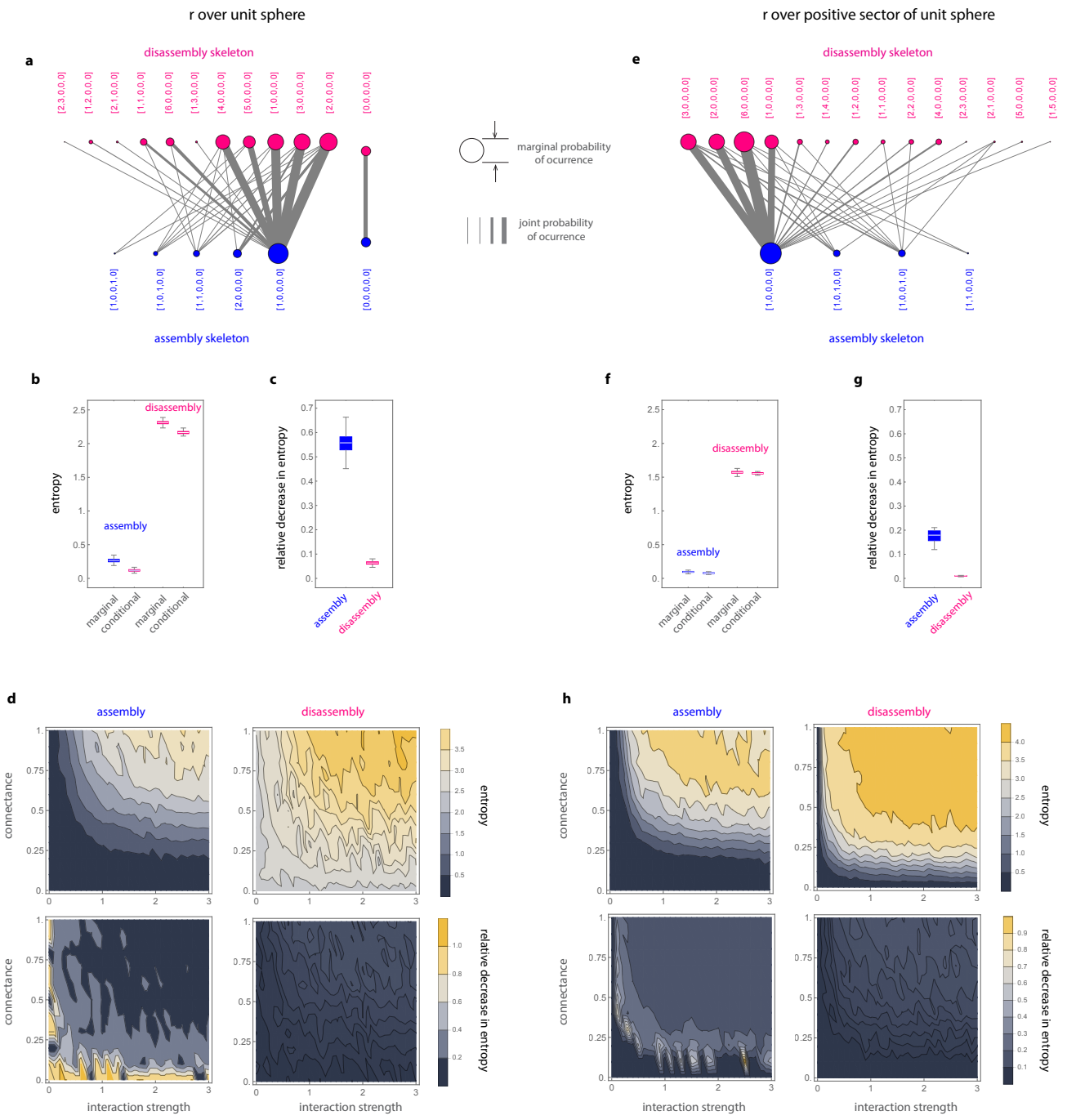
Moreover, the links of the illustrative bipartite graph generated above show that if we are given information about the observed disassembly skeleton, the uncertainty of occurrence for the assembly skeleton strongly decreases. This can be confirmed because most disassembly skeletons have few strong links (Supplementary Fig. 6a). Recall that the stronger the link, the higher the conditional probability. However, these patters are not present in the other direction. In other words, in this example, the disassembly skeleton strongly determines the assembly skeleton, but not vice versa. Note that, despite the assembly hypergraph determines the disassembly hypergraph (i.e., $D = D(H)$), there is an infinite number of assembly hypergraphs with the same assembly skeleton. This means that there is also an infinite number of disassembly hypergraphs associated to same assembly skeleton. Consequently, this implies that the observation that disassembly skeletons determine assembly skeletons is not a simple byproduct of the construction of the assembly/disassembly hypergraphs.

We can further quantify the findings above by calculating the marginal and conditional entropies of assembly and disassembly skeletons—recall that higher entropy means larger uncertainty. As expected from our hypothetical example, Supplementary Fig. 6b shows that the marginal entropy of the assembly skeletons is much lower than the marginal entropy of the disassembly skeletons. By conditioning assembly on disassembly or disassembly on assembly, the entropy of the conditional probability distribution decreases—uncertainty decrease because we are given additional information. We can then calculate the relative change in entropy due to conditioning as: 1−(marginal entropy)/(conditional entropy). In our hypothetical example, we found that the entropy in the assembly reduces about 55% when we know the disassembly, whereas knowing the assembly reduces the entropy in the disassembly only about 6% (Supplementary Fig. 6c). This result is aligned with the previous observations derived from the bipartite graph in Supplementary Fig. 6a, confirming that disassembly strongly determines assembly, but not vice versa. We found this same behavior when systematically analyzing a random ensemble of LV models (Supplementary Fig. 6d).

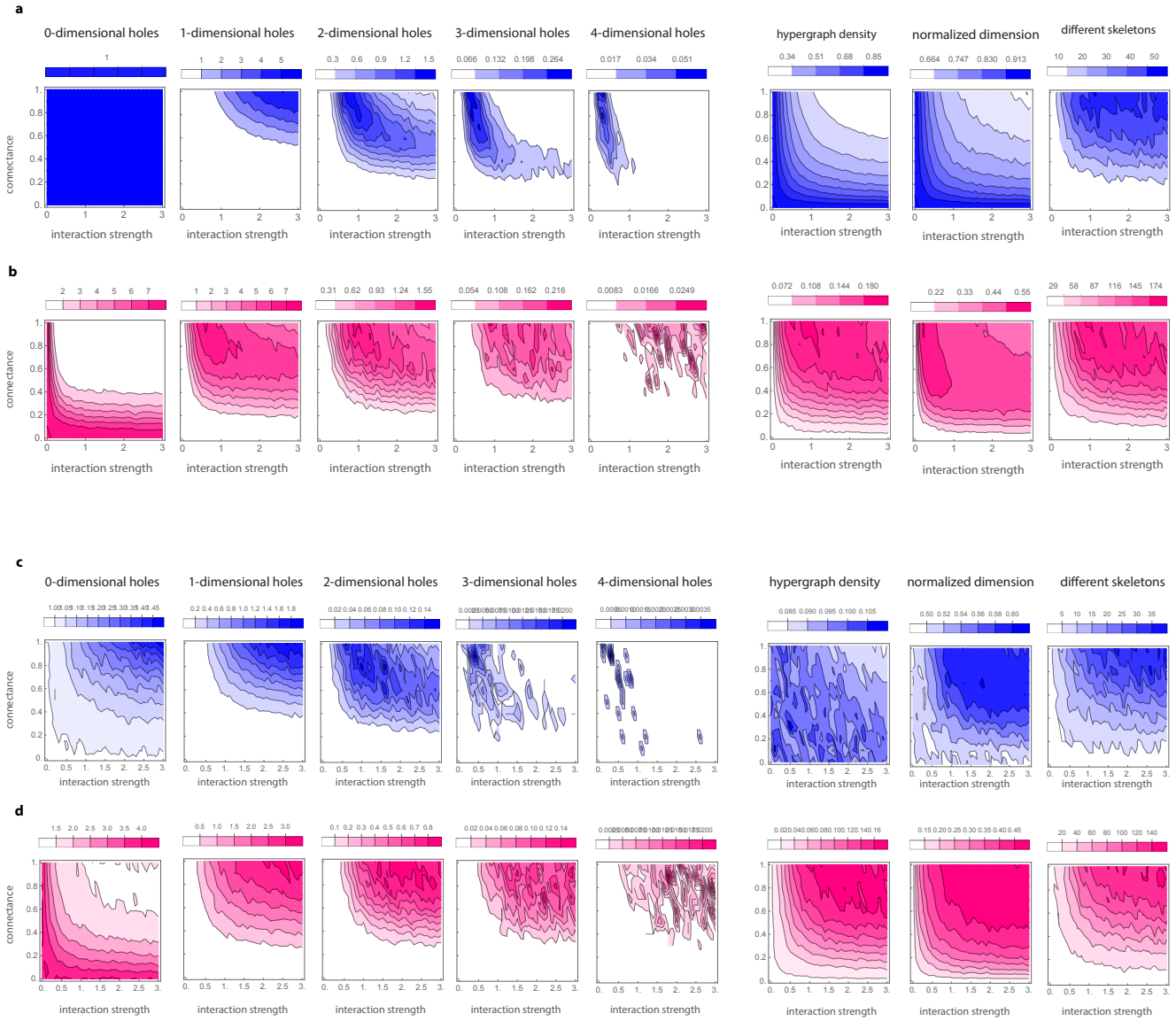
In our hypothetical example, when the intrinsic growth rate vector is restricted to the positive sector of the unit sphere, we qualitatively find the same behavior: disassembly strongly influence assembly but not viceversa (Supplementary Fig. 6e). Yet, we find lower marginal and conditional entropies for both assembly and disassembly (Supplementary Fig. 6f). This is expected, since the domain of values for the intrinsic growth rates is smaller. In this case, the entropy in the assembly reduces about 17% when we know the disassembly, whereas knowing the assembly reduces the entropy in the disassembly only about 0.1% (Supplementary Fig. 6g). We found this same behavior when systematically analyzing a random ensemble of LV models (Supplementary Fig. 6h).

5.5 The probability of observing holes changes with their dimension, and the ensemble parameters

Here we study how the probability of observing assembly/disassembly holes of different dimension change with the parameters of the ensemble of random LV models. Specifically, we generate an en-



Supplementary Figure 6 | Disassembly strongly influences assembly, but not vice versa. Panels a-d consider a growth rate vector uniformly chosen over the unit sphere. **a.** Bipartite graph between assembly and disassembly skeletons generated by the interaction matrix of Eq. (S4). Here, the size of the node is proportional to the marginal probability, and the weight of a link is proportional to the joint probability. **b.** Marginal and conditional entropies obtained for the interaction matrix of Eq. (S4). **c.** Relative decrease in the entropy due to conditioning obtained for the interaction matrix of Eq. (S4). **d.** Relative decrease in entropy for an ensemble of $20,000$ random LV models of $S = 8$ species. For each ecological system, we calculate its assembly (H). and disassembly (D) skeletons. Left panel shows the conditional entropy $\mathcal{H}(H|D)$ of an assembly skeleton given a disassembly skeleton, for all ecological systems with the same combination of parameters (C, σ_A). Right panel shows the corresponding conditional entropy $\mathcal{H}(D|H)$. The conditional entropy $\mathcal{H}(H|D)$ is higher than $\mathcal{H}(D|H)$. A low conditional entropy $\mathcal{H}(D|H)$ indicates that the disassembly skeleton is determined by the assembly skeleton. Similarly, the large conditional entropy $\mathcal{H}(D|H)$ indicates that the disassembly skeleton is NOT strongly determined by the assembly skeleton. **e. to h.** Same as panels a to c, except the growth rate vector is uniformly chosen over the positive sector of the unit sphere.



Supplementary Figure 7 | Coexistence holes in the Lotka-Volterra model with random parameters. Results are for a system of $S = 8$ species. Panels a and b are for positive growth rates over the unit sphere. Panels c and d for arbitrary growth over the unit sphere. In all panels, colors correspond to the expected number of holes of different dimension, obtained from 10,000 pairs (A, r) generated with the same value of (σ_A, C_A). For a hypergraph H, its density is $|H|/(2^S - 1)$, namely the proportion of hyperedges it has compared the maximum number of hyperedges that it can have.

semble random LV models for $S = 8$ species with parameters $(C_A, \sigma_A) \in [0 : 0.1 : 1] \times [0 : 0.1 : 3]$, sampling 10,000 random matrices at each step. For each of those matrices, we generate a random intrinsic growth vector either chosen uniformly at random over the positive sector of the unit sphere (Supplementary Fig. 7a,b), or uniformly at random over the whole unit sphere (Supplementary Fig. 7c,d). Then, for each generated pair (A, r) we calculate their assembly/disassembly hypergraphs, and the number of assembly/disassembly holes of different dimensions. Here we also calculate three additional statistics of the hypergraphs: their density (number of hyperedges they have divided by $2^S - 1$), their dimension (maximum number of elements in their hyperedges), and the different skeletons generated when A is fixed and r randomly changes. Note the hypergraph dimension characterizes the limits of coexistence.

As expected, the limits of coexistence (i.e., the dimension of the assembly hypegraph) increase when σ_A or C_A are decrease. Below these limits, additional structure exists as characterized by assembly and disassembly holes. For a wide region in the parameter plane (σ_A, C_A), we find only one 0-dimensional assembly hole, meaning that there are no species groups where coexistence is totally mutually exclusive. Also, 0-dimensional disassembly holes are more likely at low values of σ_A or C_A, indicating that a low "complexity" $\sigma_A C_A$ is more likely to produce the simple closed-under-inclusion assembly rule. For higher dimension, the expected number of assembly and disassembly holes depends on the σ_A, C_A, and the hole's dimension. In general, this expected number of holes decrease with their dimension. The maximum number of low-dimensional assembly holes occurs at higher values of σ_A and C_A compared to high-dimensional assembly holes. This result is reasonable because higher-dimensional assembly holes require that bigger species collections can coexist. We also find that the assembly hypergraph tends to be denser than the disassembly hypergraph, and that the systems tends to adopt a larger number of different disassembly skeletons compared to assembly skeletons. Results are similar when the intrinsic growth rate vector r can take positive and negative values, except that more 0-dimensional assembly holes appear and fewer high-dimensional coexistence holes occur.

6. Analysis of empirical ecological systems.

6.1 Description of the empirical datasets

We analyzed five experimental microbial communities to identify the presence of coexistence holes. The list of species in each ecological system are shown in Supplementary Tables 1 to 5. For each community, we used the empirical values of the interaction matrix $A \in \mathbb{R}^{S\times S}$ and intrinsic growth-rate vectors $r \in \mathbb{R}^{S}$ previously inferred and validated in each study.

ID	species name
1	Paramecium blirsaria
2	Paramecium aurelia
3	Paramecium caudatum
4	Blepharisnia sp

Supplementary Table 1 | Species list for Vandermeer system [25].

ID	species name
1	Enterobacter aerogenes
2	Pseudomonas aurantiaca
3	Pseudomonas chlororaphis
4	Pseudomonas citronellolis
5	Pseudomonas fluorescens
6	Pseudomonas putida
7	Pseudomonas veronii
8	Serratia marcescens

Supplementary Table 2 | Species list for Friedman system [26]

ID	species name
1	Barnesiella
2	undefined_genus_of_Lachnospiraceae
3	unclassified_Lachnospiraceae
4	Other
5	Blautia
6	undefined_genus_of_unclassified_Mollicutes
7	Akkermansia
8	Coprobacillus
9	Clostridium_difficile
10	Enterococcus
11	undefined_genus_of_Enterobacteriaceae

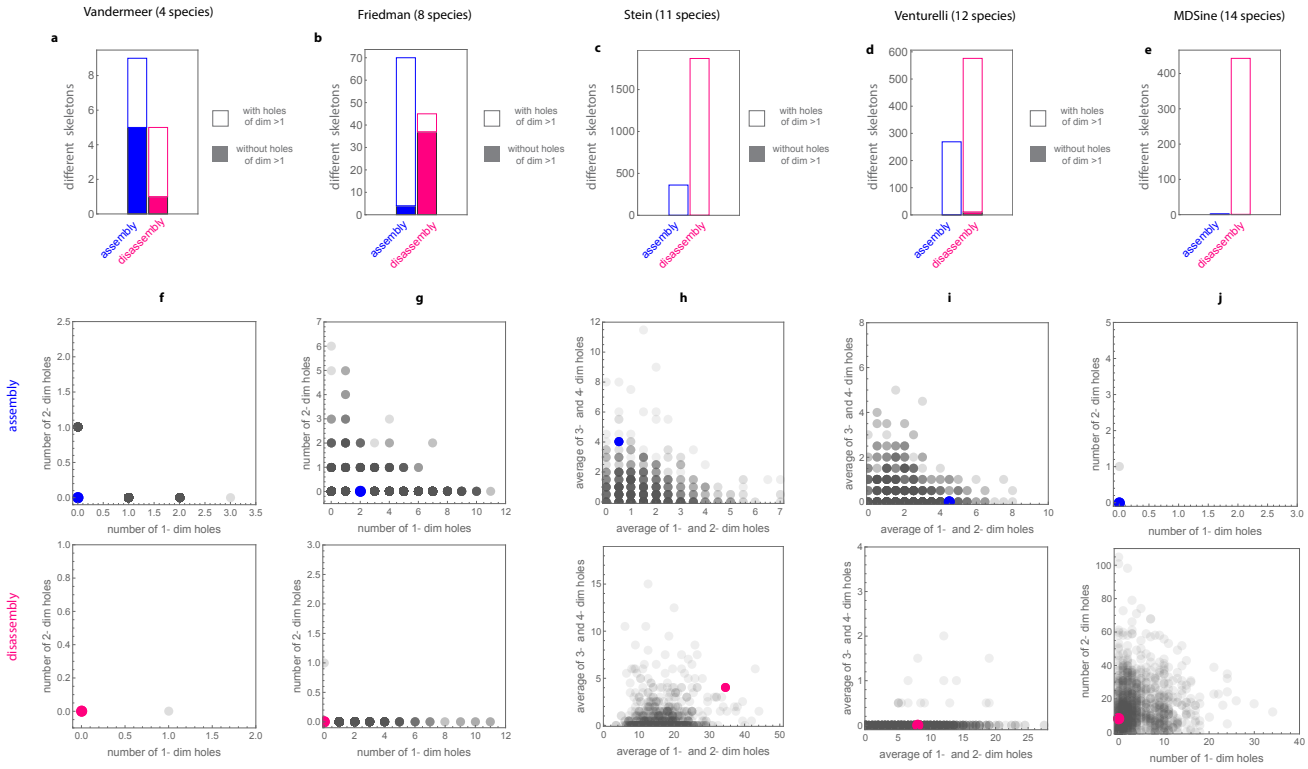
Supplementary Table 3 | Species list for Stein system [11].

ID	species name
1	Blautia hydrogenotrophica
2	Collinsella aerofaciens
3	Bacteroides uniformis
4	Prevotella copri
5	Bacteroides ovatus
6	Bacteroides vulgatus
7	Bacteroides thetaiotaomicron
8	Eggerthella lenta
9	Faecalibacterium prausnitzii
10	Clostridium hiranonis
11	Desulfovibrio piger
12	Eubacterium rectale

Supplementary Table 4 | Species list for Venturelli system [12].

ID	species name
1	Clostridium hiranonis
2	Clostridium difficile
3	Proteus mirabilis
4	Clostridium scindens
5	Ruminococcus obeum
6	Clostridium ramosum
7	Bacteroides ovatus
8	Akkermansia muciniphila
9	Parabacteroides distasonis
10	Bacteroides fragilis
11	Bacteroides vulgatus
12	Klebsiella oxytoca
13	Roseburia hominis
14	Escherichia coli

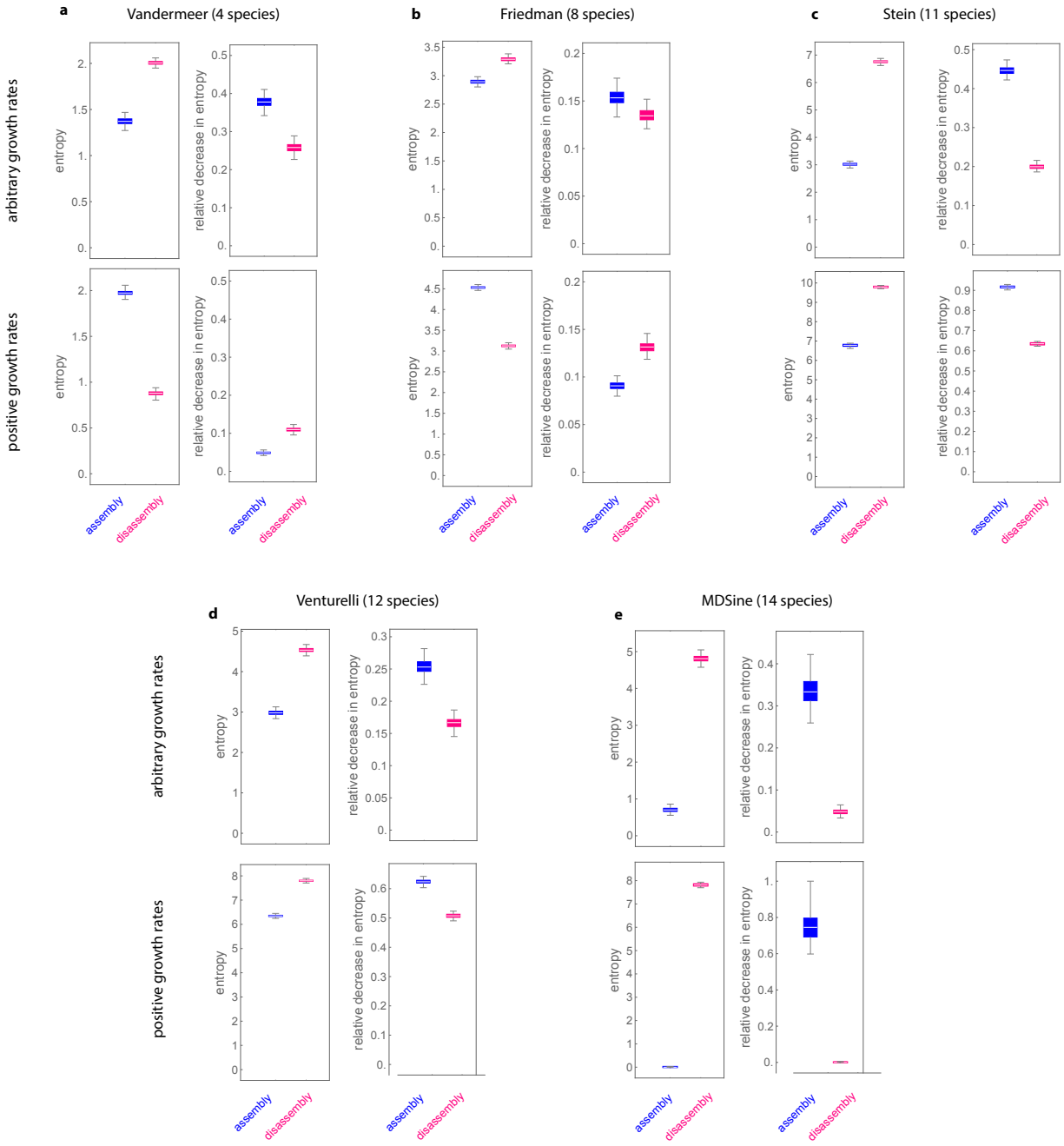
Supplementary Table 5 | Species list for the MDSine system [27].



Supplementary Figure 8 | Empirical species interactions generate assembly and disassembly holes. Results are for each empirical interaction matrix and $5,000$ skeletons generated by choosing the intrinsic growth-rate vector uniformly at random over the positive section of the unit sphere. bf a. to e. Number of assembly and disassembly skeletons generated by each empirical interaction matrix. **f. to j.** Grey dots show the number of low- and high-dimensional holes observed in the 5,000 random skeletons obtained by using the empirical interaction matrices (strength of color is proportional to the probability of occurrence). Color dots show the empirical skeletons obtained by the empirical interaction matrix and the empirical intrinsic growth-rate vector.

6.2 Disassembly strongly determines assembly in empirical systems, but not vice versa

For each of the five empirical interaction matrices A, we followed the methodology of Supplementary Note S4.5 to calculate the entropy of the disassembly and assembly skeletons, and the relative decrease in the entropy obtained after conditioning. The results are shown in Supplementary Fig. 9. As in the case of unstructured ecological systems, for all five empirical ecological systems we find that conditioning on the disassembly reduces more strongly the entropy (i.e., the uncertainty about the assembly) compared to conditioning on the assembly.



Supplementary Figure 9 | Relation between assembly and disassembly in empirical ecological systems.

7. Calculating assembly hypergraphs using only co-culture experiments.

Here we consider available an experimental dataset $\mathfrak{D}$ containing the co-culture experiments for *all* local communities with different species collections. More precisely, we consider that each experiment has associated an initial **species collection** $z \in \{0,1\}^S$ as well as the corresponding final **abundance** $x \in \mathbb{R}^S$ of species (e.g., steady-state abundance of species). Each pair (z,x) is called a **sample**, and the dataset $\mathfrak{D}$ contains those samples $\mathfrak{D} = \{(z,x)\}$. We assume that multiple experiments may exists for the same species collection, which we call **repetitions**.

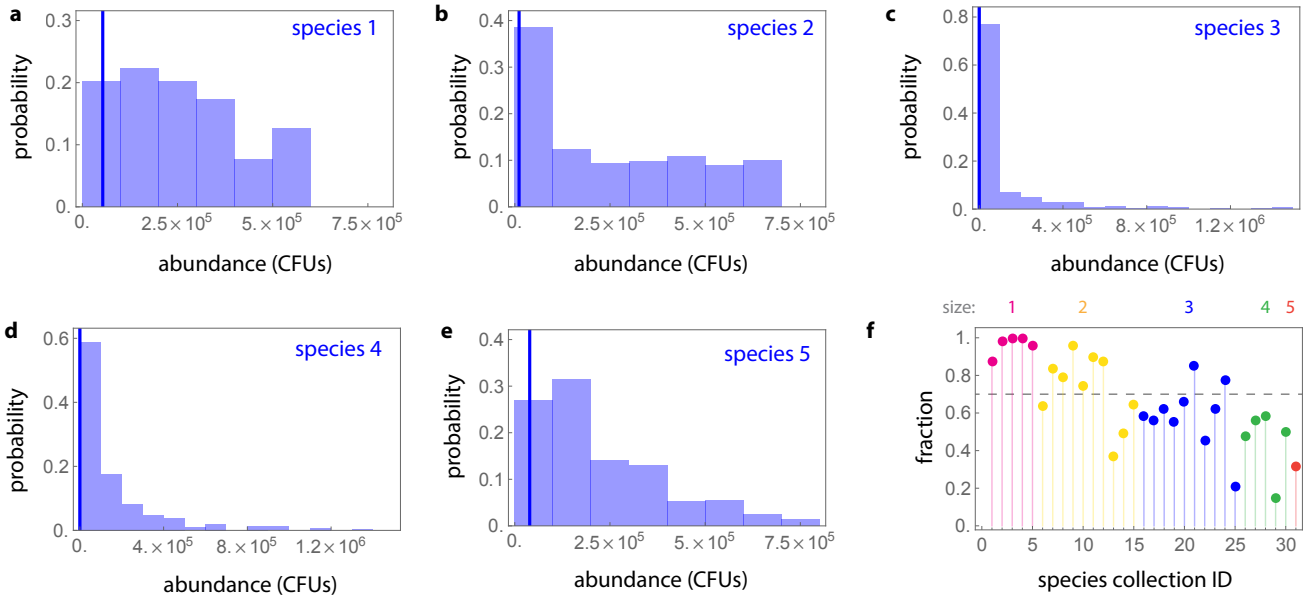
Under the above conditions, the problem is to introduce a suitable coexistence function. Note that solving this problem is challenging because of multiple sources of variability that are intrinsic to co-culture experiments. These source include:

a. Measurement errors. These include cases when a species that is present in the local community is measured as not present (e.g., because its abundance is below the detection threshold of the measurement devices). Here, it may also happen that a species that is absent from a local community is measured as present (e.g., misidentifications, such as those produced with 16S rRNA gene sequencing for bacteria identification [28]).

b. Differences in the environment between experiments. The same species collection may coexist in some environment, and not coexist in other environments. Here, we understand environment as everything necessary to calculate the vital rates of species's individuals [29], such as the abundance of different species, the density of resources or predators, or other abiotic factors. The coexistance of a species collection depends on those conditions, but they cannot be identical between co-culture experiments, affecting their reproducibility.

To address the above limitations, given experimental data $\mathfrak{D}$ is given, we consider the following method to determine species coexistence:

1. Choose: a **noise level** $\varepsilon \in [0,1]$, and a **minimal probability of coexistence** $p^* \in (0,1)$.
2. Let X_i be the vector containing the abundance x_i of species i in all samples $x \in \mathfrak{D}$ with $z_i = 1$. Calculate the ε quantile $q_i(\varepsilon)$ for X_i. Abundances $x_i \leq q_i(\varepsilon)$ are considered to be zero, but measured as non-zero due to experimental errors. Here, choosing $\varepsilon = 0$ corresponds to assuming that there exist no experimental errors.
3. Construct the thresholded dataset $\mathfrak{D}_\varepsilon$ as follows. For each sample $x \in \mathfrak{D}$, calculate the thresholded sample $x_\varepsilon \in \mathbb{R}^S_{\geq 0}$ using
$$x_{\varepsilon,i} = \begin{cases} 0 & \text{if } |x_i| \leq q_i(\varepsilon), \\ x_i & \text{otherwise,} \end{cases}$$
for $i = 1, \cdots, N$. Build $\mathfrak{D}_\varepsilon$ by collecting the pairs $\{(z, x_\varepsilon)\}$.
4. For each different initial species collection $\Sigma \in \{0,1\}^S$, do:
   i) Collect in $\mathcal{X}_\varepsilon^\Sigma \subset \mathbb{R}^N$ all $x_\varepsilon \in \mathfrak{D}_\varepsilon$ for which $z = \Sigma$. The set $\mathcal{X}_\varepsilon^\Sigma$ contains the thresholded abundances for all the repetitions of species collection Σ.
   ii) Compute the species collection $\hat{z} = \text{sign} \circ \text{abs}(x_\varepsilon) \in \{0,1\}^N$ of each $x_\varepsilon \in \mathcal{X}_\varepsilon^\Sigma$, and build the set $\mathcal{Z}_\varepsilon^\Sigma = \{\hat{z}\} \subset \{0,1\}^N$.
   iii) Let
$$\mathbf{1}(\hat{z} = \Sigma) = \begin{cases} 1 & \text{if } \hat{z} = \Sigma, \\ 0 & \text{otherwise,} \end{cases},$$
   be the indicator function. Compute
$$p(\Sigma) = \frac{1}{|\mathcal{Z}_\varepsilon^\Sigma|} \sum_{\hat{z} \in \mathcal{Z}_\varepsilon^\Sigma} \mathbf{1}(\hat{z} = \Sigma),$$



Supplementary Figure 10 | Analyzing coexistence in co-culture experiments of the core *Drosophila melangoster* gut microbiota. a-e. Histograms for the abundance of each species across all samples. Blue vertical lines correspond to the quantiles $q_i(\varepsilon)$ generated for $\varepsilon = 0.1$. **f.** Fraction of repetitions in which each species collection coexists. Dashed line indicates $p^* = 0.7$.

where $|\mathcal{Z}_\varepsilon^\Sigma|$ is the cardinality of $\mathcal{Z}_\varepsilon^\Sigma$. Here, $p(\Sigma) \in [0,1]$ quantifies the probability (over repetitions of the dataset) that the species collection Σ coexists.

iii) Finally, for each species collection Σ, **define the coexistence function** as $c(\Sigma) = 1$ if $p(\Sigma) \geq p^*$, and $c(\Sigma) = 0$ otherwise.

7.1 Calculating the assembly of Drosophila melanogaster gut microbiota.

We illustrate the above method using an experimental dataset with the assemblage of the $2^5 - 1 = 31$ local communities of the $S = 5$ "core" species of *Drosophila melanogaster* fruit fly gut microbiota [30]. In this dataset, species are numbered as:

1. *Lactobacillus plantarum*; 2. *Lactobacillus brevis*; 3. *Acetobacter pasteurianus*; 4. *Acetobacter tropicalis*; and 5. *Acetobacter orientalis*.

Abundance of species was measured by counting colony-forming units (CFUs). In total, this dataset contains 47 repetitions for the abundance of each possible species collection, resulting in 1449 samples.

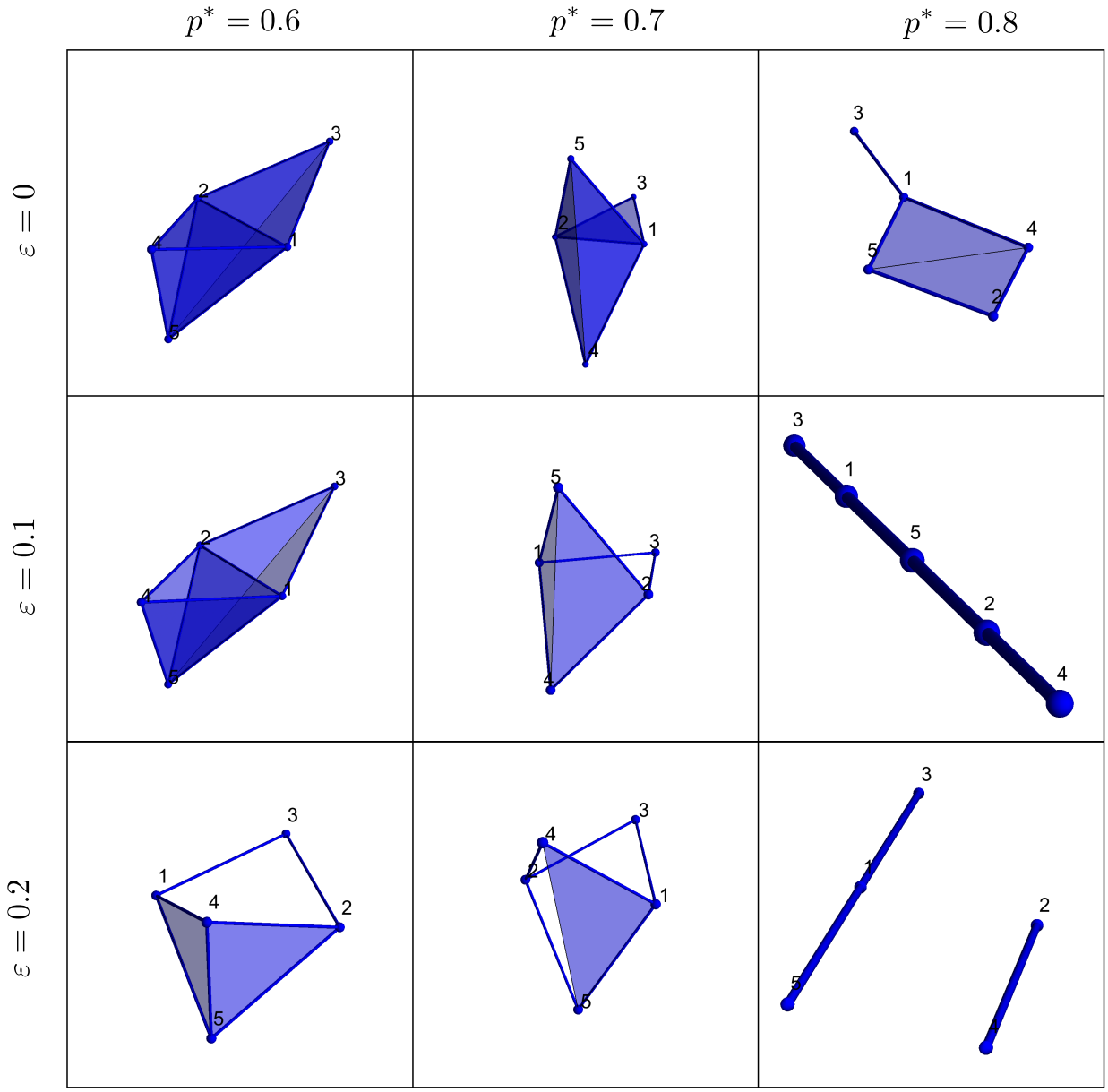
First, we compute the vectors X_i, $i = 1, \cdots, N$, containing the abundance of the i-th species in all samples that originally contained it. These five vectors can be visualized as histograms displaying the distribution of abundances that each species takes, see Fig. 10a-e. Next we choose $\varepsilon = 0.1$, leading to the quantiles

$$q(\varepsilon) = (53657,\ 11133,\ 0,\ 0,\ 39210).$$

Namely, for this choice, 90% of the abundances in X_i are larger than $q_i(\varepsilon)$ (blue vertical lines in Fig. 10). Actually, these quantiles can be compared to the maximum abundances of each species: (537000, 668000 ,1.47×10^6, 1.31×10^6, 706000). Next we compute the probabilities $p(\Sigma)$, $\Sigma \in 2^V$ for each species collection (Fig. 10f). Finally, choosing $p^* = 0.7$, we select those species collections S that for which $p(S) \geq p^*$ to obtain the assembly hypergraph:

$$H = \big\{\{1\},\{2\},\{3\},\{4\},\{5\},\{1,3\},\{1,4\},\{1,5\},\{2,3\},\{2,4\},\{2,5\},\{1,4,5\},\{2,4,5\}\big\}.$$

In Fig. 11 we visualize the assembly hypergraph H of the core *Drosophila melanogaster* gut microbiota for different combinations of parameters (ε, p^*).



Supplementary Figure 11 | Assembly hypergraphs of the core *Drosophila melanogaster* gut microbiota. Coexistence hypergraphs obtained by choosing different combinations of noise level ε parameter and minimal probability of coexistence across repetitions p^*.

References

[1] V. De Risi. Analysis situs, the foundations of mathematics and a geometry of space. In *The Oxford Handbook of Leibniz* (Oxford University Press, 2019).

[2] C. Berge. *Hypergraphs: combinatorics of finite sets*, vol. 45 (Elsevier, 1984).

[3] S. Klamt, U.-U. Haus and F. Theis. Hypergraphs and cellular networks. *PLoS computational biology* **5** no. 5, p. e1000 385 (2009).

[4] A. Ritz et al. Signaling hypergraphs. *Trends in biotechnology* **32** no. 7, pp. 356–362 (2014).

[5] A. J. Golubski et al. Ecological networks over the edge: hypergraph trait-mediated indirect interaction (tmii) structure. *Trends in ecology & evolution* **31** no. 5, pp. 344–354 (2016).

[6] A. R. Benson, D. F. Gleich and J. Leskovec. Higher-order organization of complex networks. *Science* **353** no. 6295, pp. 163–166 (2016).

[7] R. Lambiotte, M. Rosvall and I. Scholtes. From networks to optimal higher-order models of complex systems. *Nature physics* p. 1.

[8] J. Hocking and G. Young. Topology addison wesley. *Reading, Massachusetts* .

[9] V. D. Picasso et al. Crop species diversity affects productivity and weed suppression in perennial polycultures under two management strategies. *Crop Science* **48** no. 1, pp. 331–342 (2008).

[10] V. Halty et al. Modeling plant interspecific interactions from experiments with perennial crop mixtures to predict optimal combinations. *Ecological Applications* **27** no. 8, pp. 2277–2289 (2017).

[11] R. R. Stein et al. Ecological modeling from time-series inference: insight into dynamics and stability of intestinal microbiota. *PLoS computational biology* **9** no. 12.

[12] O. S. Venturelli et al. Deciphering microbial interactions in synthetic human gut microbiome communities. *Molecular systems biology* **14** no. 6.

[13] R. Law and J. C. Blackford. Self-assembling food webs: a global viewpoint of coexistence of species in lotka-volterra communities. *Ecology* **73** no. 2, pp. 567–578 (1992).

[14] R. Law and R. D. Morton. Permanence and the assembly of ecological communities. *Ecology* **77** no. 3, pp. 762–775 (1996).

[15] K. Sigmuiud. Darwin's circles of complexity: Assembling ecological communities. *Complexity* **1** no. 1, pp. 40–44 (1995).

[16] J. Hofbauer, K. Sigmund et al. *The theory of evolution and dynamical systems: mathematical aspects of selection* (Cambridge University Press, 1988).

[17] W. Jansen. A permanence theorem for replicator and lotka-volterra systems. *Journal of Mathematical Biology* **25** no. 4, pp. 411–422 (1987).

[18] C. Jost and S. P. Ellner. Testing for predator dependence in predator-prey dynamics: a non-parametric approach. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **267** no. 1453, pp. 1611–1620 (2000).

[19] Y. Chen, M. T. Angulo and Y.-Y. Liu. Revealing complex ecological dynamics via symbolic regression. *BioEssays* **41** no. 12, p. 1900 069 (2019).

[20] J. Grilli et al. Higher-order interactions stabilize dynamics in competitive network models. *Nature* **548** no. 7666, p. 210 (2017).

[21] R. MacArthur. Species packing and competitive equilibrium for many species. *Theoretical population biology* **1** no. 1, pp. 1–11 (1970).

[22] P. Chesson. Macarthur's consumer-resource model. *Theoretical Population Biology* **37** no. 1, pp. 26–38 (1990).

[23] P. J. Taylor. Consistent scaling and parameter choice for linear and generalized lotka-volterra models used in community ecology. *Journal of theoretical biology* **135** no. 4, pp. 543–568 (1988).

[24] C. Song, R. P. Rohr and S. Saavedra. A guideline to study the feasibility domain of multi-trophic and changing ecological communities. *J. of Theoretical Biology* **450**, pp. 30–36 (2018).

[25] J. H. Vandermeer. The competitive structure of communities: an experimental approach with protozoa. *Ecology* **50** no. 3, pp. 362–371 (1969).

[26] J. Friedman, L. M. Higgins and J. Gore. Community structure follows simple assembly rules in microbial microcosms. *Nature ecology & evolution* **1** no. 5, p. 0109 (2017).

[27] V. Bucci et al. Mdsine: Microbial dynamical systems inference engine for microbiome time-series analyses. *Genome biology* **17** no. 1, pp. 1–17 (2016).
[28] J. M. Janda and S. L. Abbott. 16s rrna gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *Journal of clinical microbiology* **45** no. 9, pp. 2761–2764 (2007).
[29] S. Lion. Theoretical approaches in evolutionary ecology: environmental feedback as a unifying perspective. *The American Naturalist* **191** no. 1, pp. 21–44 (2018).
[30] A. L. Gould et al. Microbiome interactions shape host fitness. *Proceedings of the National Academy of Sciences* **115** no. 51, pp. E11 951–E11 960 (2018).