SI Appendix

To accompany "Clique topology reveals intrinsic geometric structure in neural correlations."

Chad Giusti, Eva Pastalkova, Carina Curto*, and Vladimir Itskov*

Contents:

Supplementary Methods

Here we present further details for the analyses involving hippocampal place cell data, along with descriptions of the place field and scrambled place field models.

Selection criteria for cells and recordings. During each of the three behavioral conditions (spatial navigation, wheel running, and REM sleep), only putative pyramidal cells whose average firing rates were in the 0.2-7 Hz range were used. Putative interneurons, defined as having an average firing rate above 7 Hz over an entire recording session, were excluded. Recordings with at least N=60 neurons satisfying these criteria were selected for the analyses. A total of 18 recordings from 5 animals met the selection criteria. These consisted of 9 "open field" spatial navigation data sets from three animals, 5 wheel running data sets from two animals, and 4 REM sleep data sets from two animals. Periods of REM sleep were detected from the local field potential using the ratio of total delta power (0.1-3 Hz) to total theta power (5-10 Hz) in the spectrogram of the EEG. Periods with delta/theta ratio less than 1 were considered REM sleep.

The number of cells and length of each recording were: N = 76 (16 min), N=81 (16 min), N=72 (63 min), N=88 (63 min), N=68 (59 min), N=64 (50 min), N=67 (50 min), N=74 (59 min), and N=66 (60 min) for spatial navigation; N=77 (26 runs, 6 min), N=73 (25 runs, 5 min), N=83 (13 runs, 5 min), N=67 (7 runs, 2 min), and N=64 (10 runs, 3 min) for wheel running; and N=67 (30 s), N=69 (4 min), N=60 (2.5 min), and N=61 (3.5 min) for REM sleep. The N=88 spatial navigation data set was used in Figure 3.

Computation of the pairwise correlation matrices. The cross-correlograms were computed as $\mathrm{ccg}_{ij}(\tau) = \frac{1}{T}\int_0^T f_i(t) f_j(t+\tau)\mathrm{d}t$, where $f_i(t)$ is the firing rate of the i-th neuron and T is the total duration of the considered time period. The normalized cross-correlation on a timescale of $\tau_{\max}$ was computed as

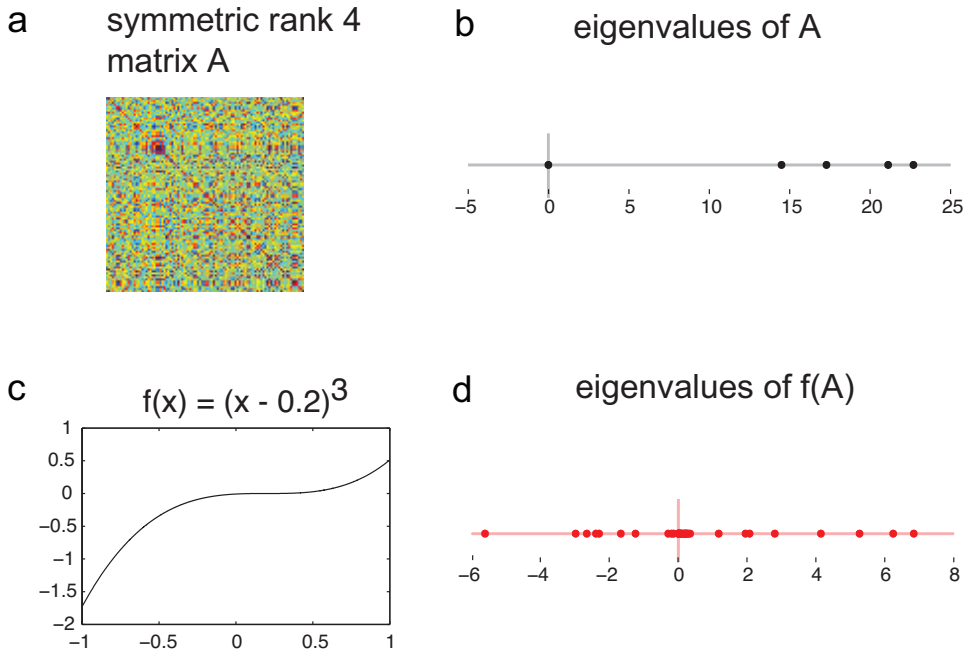
$$C_{ij} = \frac{1}{\tau_{\max} r_i r_j} \max\left(\int_0^{\tau_{\max}} \mathrm{ccg}_{ij}(\tau)d\tau, \int_0^{\tau_{\max}} \mathrm{ccg}_{ji}(\tau)d\tau\right),$$

where r_i is the average firing rate of the i-th neuron (see Supplementary Figure 3).

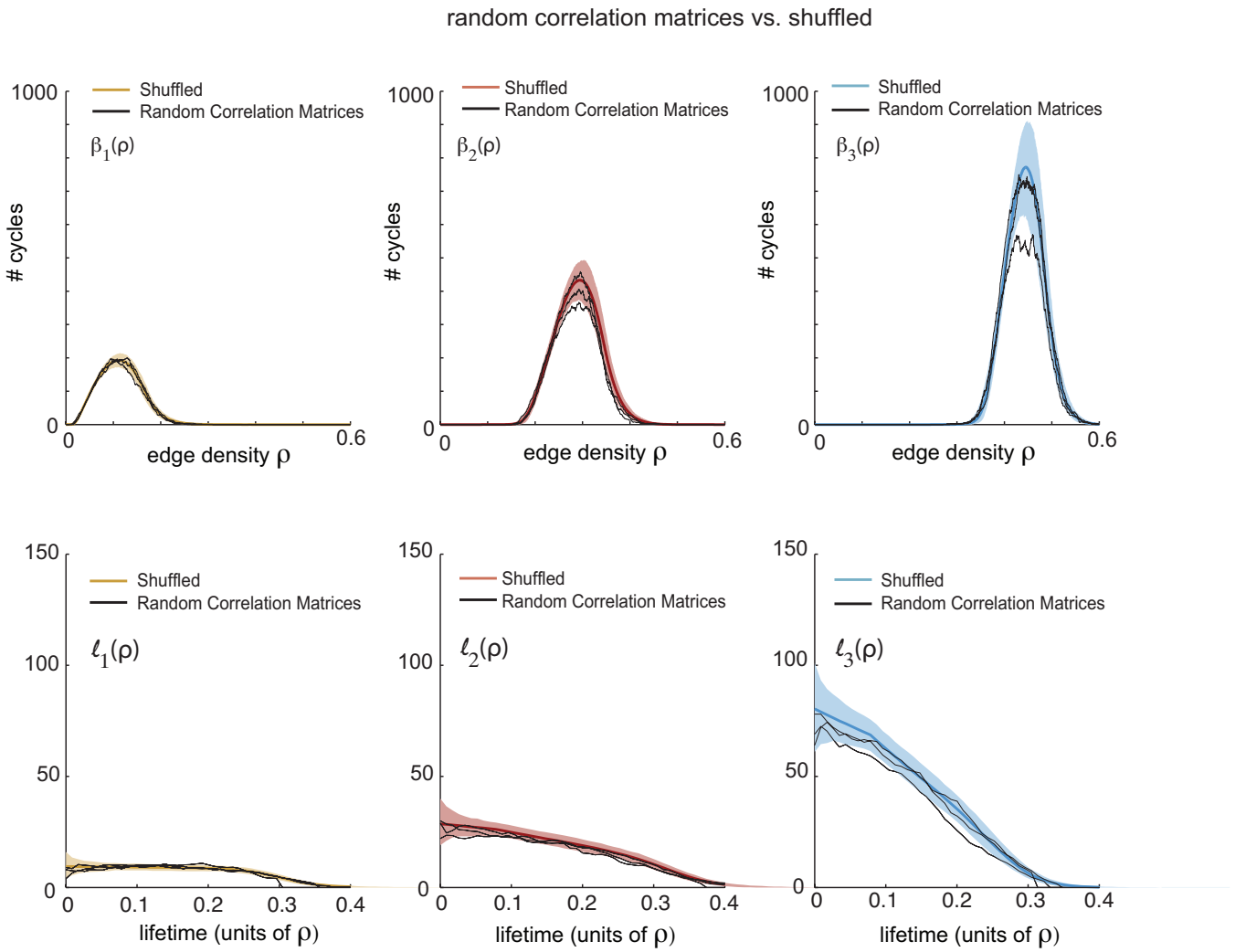
Simulated spike train data from PF and scrambled PF models. For each cell in the N=88 spatial navigation data set, place fields $F_i(\mathbf{x})$ were computed using a 100 × 100 grid of pixels. In each pixel, the number of spike events for each cell was normalized by the time spent at that location, and then smoothed with a 2-dimensional Gaussian ($\sigma = 5$ grid locations, see also Supplementary Figure 9). Simulated spike trains for the PF model were generated from the place fields as inhomogeneous Poisson processes with rate functions $r_i(t) = F_i(\mathbf{x}(t))$, using the animal's original trajectory, $\mathbf{x}(t)$.

Scrambled place fields $\widetilde{F}_i(\mathbf{x})$ were computed by randomly permuting the pixels in the 100 × 100 grid. A different permutation was used for each place field, in order to destroy the coherence of the spatial organization across the population. Similar spike trains for the scrambled PF model were again generated as inhomogeneous Poisson processes using the original trajectory, $\mathbf{x}(t)$, but with rate functions $r_i(t) = \widetilde{F}_i(\mathbf{x}(t))$ given by the scrambled place fields (see Supplementary Figure 9). The simulated data sets were analyzed in Figure 3 at the correlation timescale $\tau_{\max} = 1$sec.

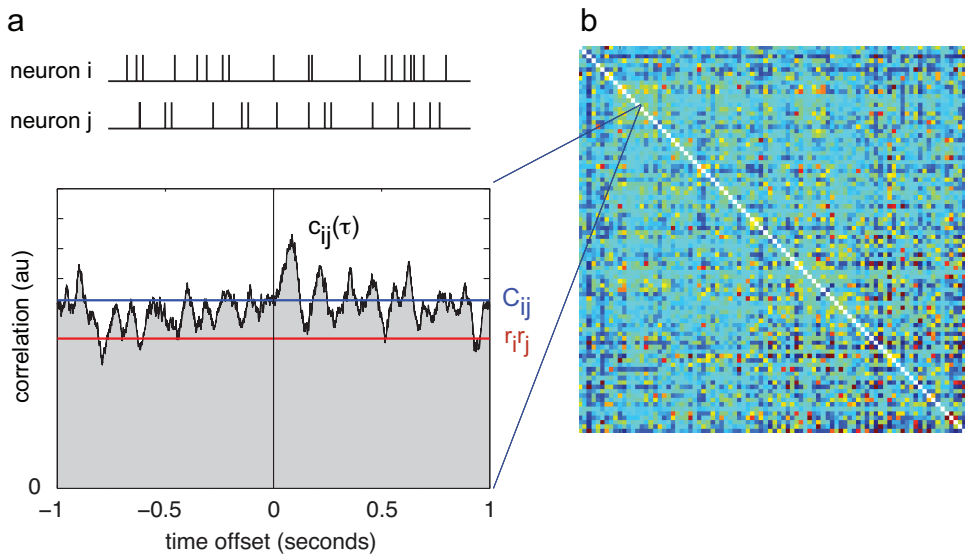
Supplementary Figures



Supplementary Figure 1: Spectral signatures of matrix structure are destroyed by nonlinear monotonically increasing transformations. (**a**) A 100×100 symmetric matrix A with $\text{rank}(A) = 4$. (**b**) The spectrum of A includes four nonzero eigenvalues that are all positive. This is the signature that A has rank 4 and is positive semidefinite. (**c**) The graph of the monotonically increasing function $f(x) = (x - 0.2)^3$. (**d**) The spectrum of the matrix $f(A)$ contains many nonzero eigenvalues. The spectral signature that A has low-rank structure has been destroyed by f.



Supplementary Figure 2: Topological properties of random correlation matrices are similar to those of random i.i.d. matrices. The $N \times N$ correlation matrix $C_{ij} = \mathrm{corr}(X_i, X_j)$ was computed using $10,000$ samples of $N = 88$ independent uniformly distributed random variables X_i. Each panel compares the Betti curves (top) and the persistence lifetimes (bottom) of the random correlation matrices to those of the random (shuffled) matrices. Each of the three black lines correspond to one instance of such a correlation matrix. Colored lines and shaded regions correspond to the mean curves and 95% confidence intervals for the random (shuffled) matrices.



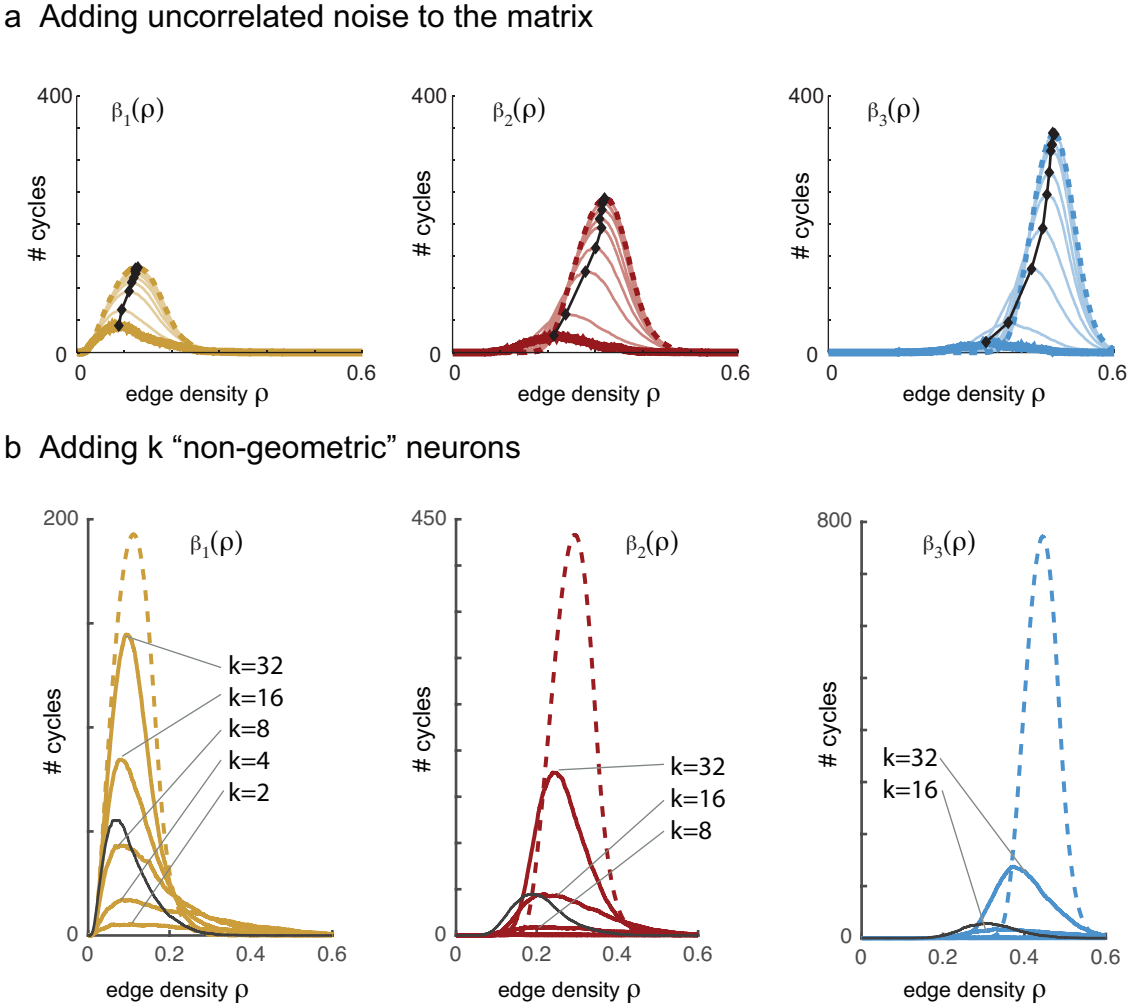
Supplementary Figure 3: Computation of pairwise correlation matrices from spike train data. **(a)** For a pair of spike trains $\{t_\ell^i\}_{\ell=1\ldots n_i}$ and $\{t_\ell^j\}_{\ell=1\ldots n_j}$ for neurons i and j (top), the cross-correlogram $\mathrm{ccg}_{ij}(\tau)$ is computed as

$$\mathrm{ccg}_{ij}(\tau) = \frac{1}{T}\int_0^T f_i(t) f_j(t+\tau)\mathrm{d}t,$$

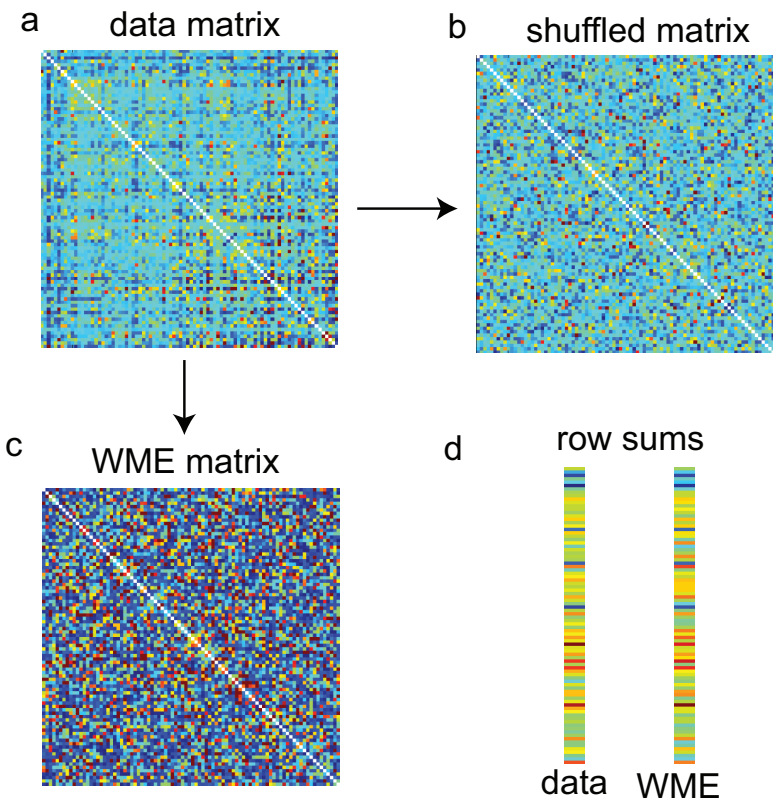
where $f_i(t) = \sum_{\ell=1}^{n_i} \delta(t - t_\ell^i)$ is the instantaneous firing rate of the i-th neuron. The graph of a smoothed $\mathrm{ccg}_{ij}(\tau)$ is displayed (black curve) along with the expected value of the cross-correlogram, $r_i r_j$ (red), for uncorrelated spike trains with matching firing rates, $r_i = n_i/T$. The pairwise correlations C_{ij} (blue line), with timescale $\tau_{\max}$, were computed as

$$C_{ij} = \frac{1}{\tau_{\max} r_i r_j} \max\left(\int_0^{\tau_{\max}} \mathrm{ccg}_{ij}(\tau)d\tau, \int_0^{\tau_{\max}} \mathrm{ccg}_{ji}(\tau)d\tau\right)$$

The timescale $\tau_{\max} = 1$ sec was used in all but one panel of Figures 3 and 4, while a range of timescales from 10 ms to 2 sec appears in Figure 3d. **(b)** The 88×88 matrix C for the spatial exploration data used in Figure 3a,b,c. The entry $C_{14,15}$ corresponds to the cross-correlogram in panel (a).



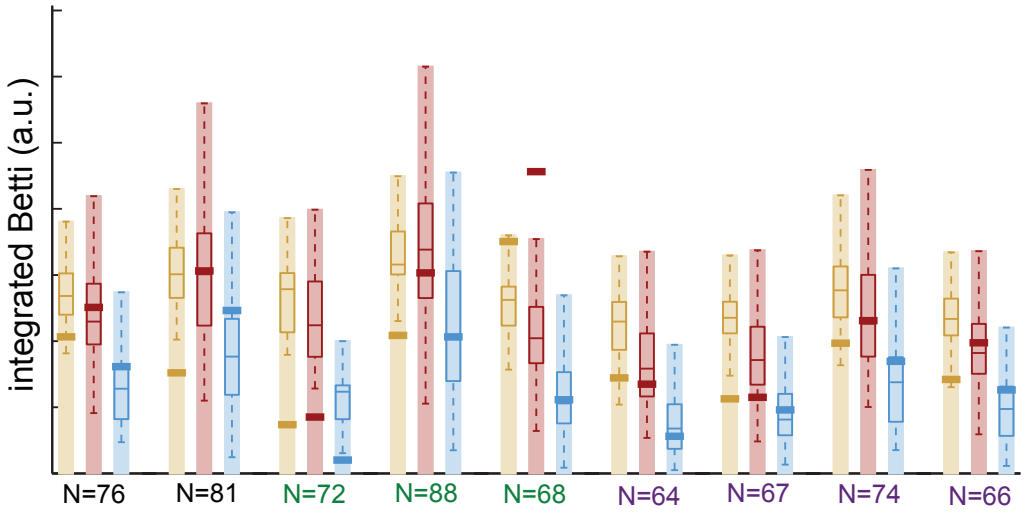
Supplementary Figure 4: The rightward and upward shift of the Betti curves can be explained by either adding noise to geometric matrices, or adding neurons that do not have any geometric organization to a geometrically organized population. (**a**) $N = 70$ random independent uniformly distributed points p_i in a d-dimensional unit cube were sampled in dimensions $d = N$. Noised geometric matrices were obtained as $A_{ij} = -\|p_i - p_j\| + (\nu\sqrt{d})g_{ij}$, where the terms g_{ij} were independent and normally distributed with zero mean and unit variance, and the noise strength ν took the following values: $0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.25, 0.5$. Each panel compares the Betti curves of the noised matrices (thin curves, stratified by the magnitude of the parameter ν) to those of the random matrices (dashed lines) and also the zero noise geometric matrices (thick curves). Black lines connect the maximum values of the mean curves for the different levels of noise. Small amounts of noise produce a rightward shift in the peak values of the Betti curves, while curves for $\nu = 0.5$ are indistinguishable from those of random matrices. (**b**) $N = 88$ random (uniformly distributed) points p_i were sampled in the 2-dimensional unit square. For each $k = 2, 4, 8, 16, 32$, the first k rows and columns of $A_{ij} = A_{ji} = -\|p_i - p_j\|$ were replaced with random values sampled independently from the uniform distribution on $[-\sqrt{2}, 0]$, thus yielding a matrix with k "non-geometric" neurons. Betti curves $\beta_1(\rho), \beta_2(\rho)$, and $\beta_3(\rho)$ (colored solid lines, stratified by k) represent averages over 35 trials. These Betti curves are shown superimposed over those of geometric matrices ($d = N$; solid black lines) and random matrices (dashed colored lines) of matching size. More than 10% of neurons must be non-geometric in order of the Betti curves to leave the geometric regime, but small numbers of non-geometric neurons increase the apparent dimension of the underlying space.



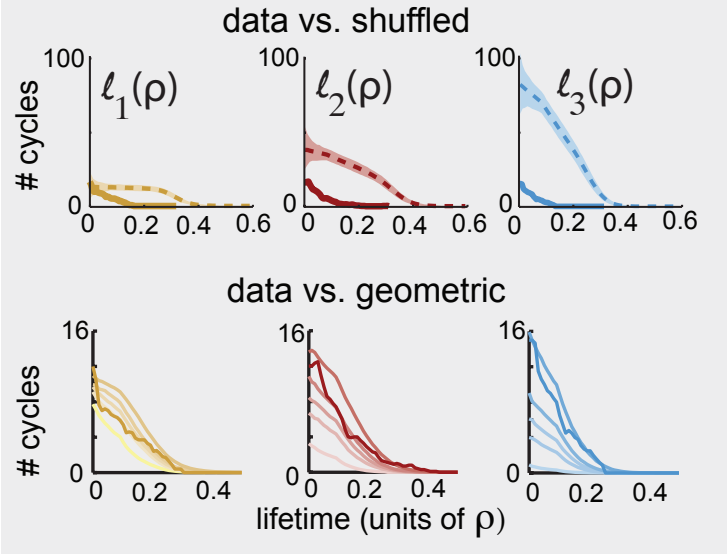
Supplementary Figure 5: Two types of random controls. (**a**) The matrix C_{ij} used in Figure 3a. (**b**) A shuffled matrix obtained by randomly permuting the $\binom{88}{2}$ off-diagonal elements of the symmetric matrix in panel (a). (**c**) A sample from the maximum entropy (WME) distribution on 88×88 symmetric matrices with prescribed expected values of row sums matching those of the matrix in panel (a). The distribution on each element (i, j) in the matrix is exponential with mean $\frac{1}{\theta_i+\theta_j}$ (Hillar & Wibisono, 2013 http://arxiv.org/abs/1301.3321). The parameters θ_i were obtained by solving the system of equations

$$\sum_{j \neq i} \frac{1}{\theta_i + \theta_j} = \sum_{j \neq i} C_{ij}, \qquad \text{for } i = 1, \ldots, N,$$

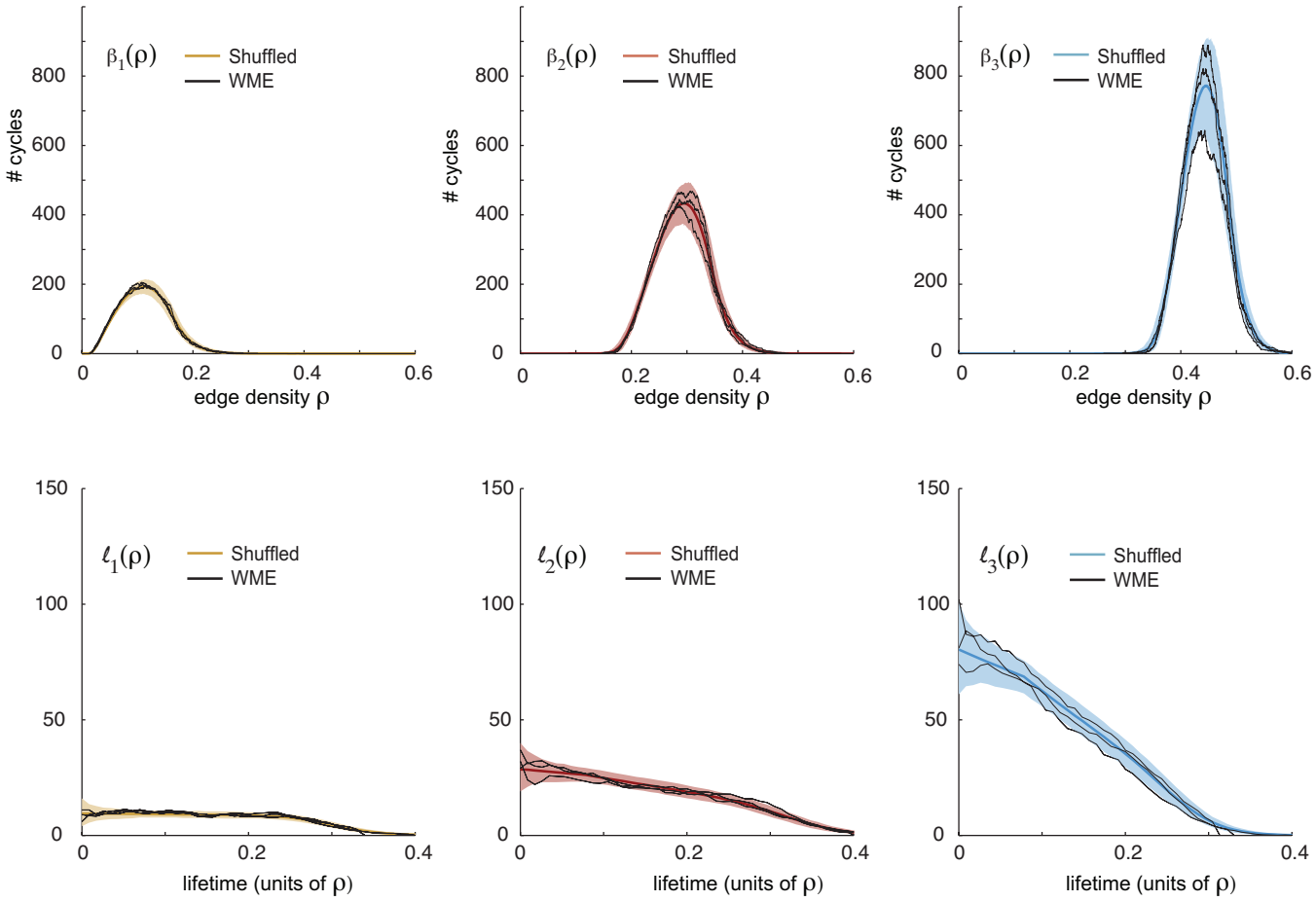
using standard gradient descent methods (Hillar & Wibisono, 2013). (**d**) (left) The row sums $\sum_{j \neq i} C_{ij}$ for the matrix in (a); (right) mean row sums for twenty samples from the distribution described in (c).



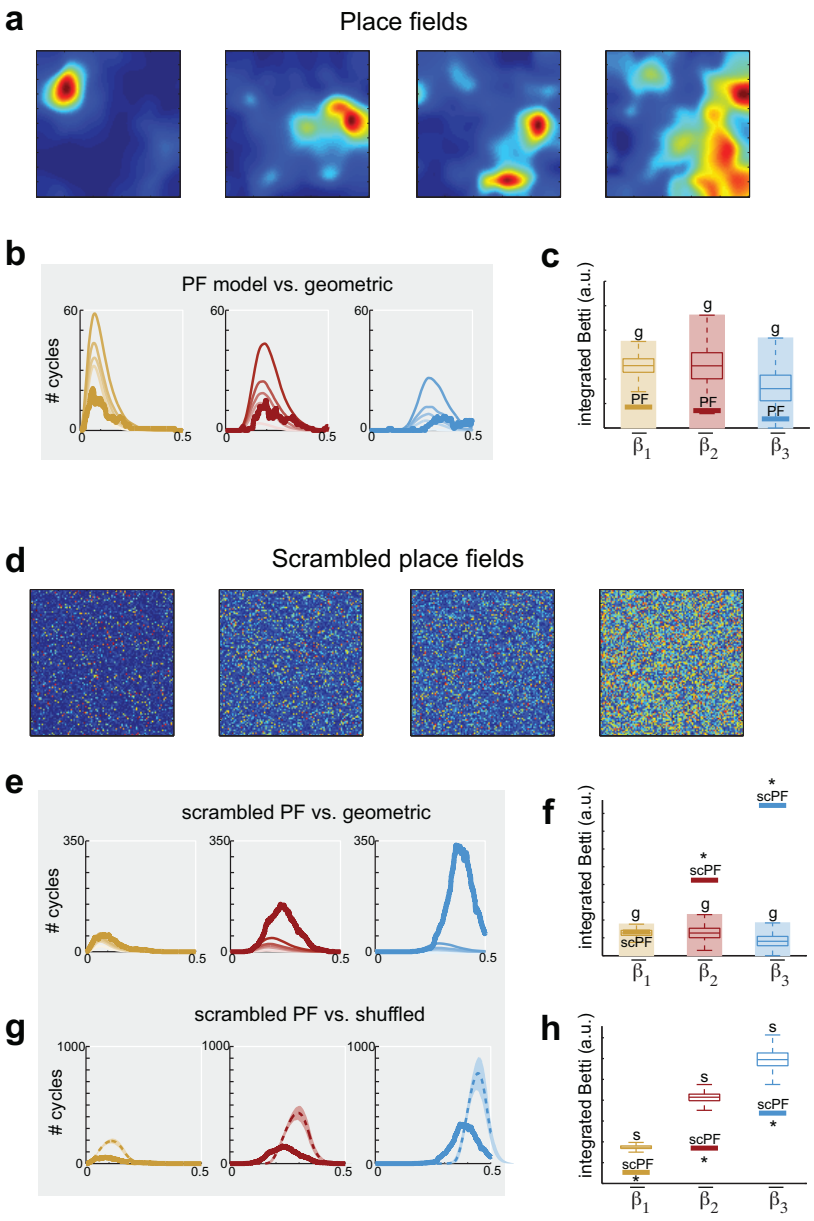
Supplementary Figure 6: Integrated Betti values obtained from neural activity during spatial navigation are consistent with those of geometric matrices. Integrated Betti values from place cell data are compared to geometric distributions with matching N across nine recordings of rat hippocampus during spatial navigation, obtained from three animals. The geometric box plots are shown for the dimension, $d = N$, while the shaded area indicates the confidence interval across the dimensions $d \leq N$. The number N of neurons is displayed in color (black, green, and purple) to indicate recordings from the same animal. Betti values for the place cell data are consistent with those of geometric matrices in all but one data set ($N = 68$).



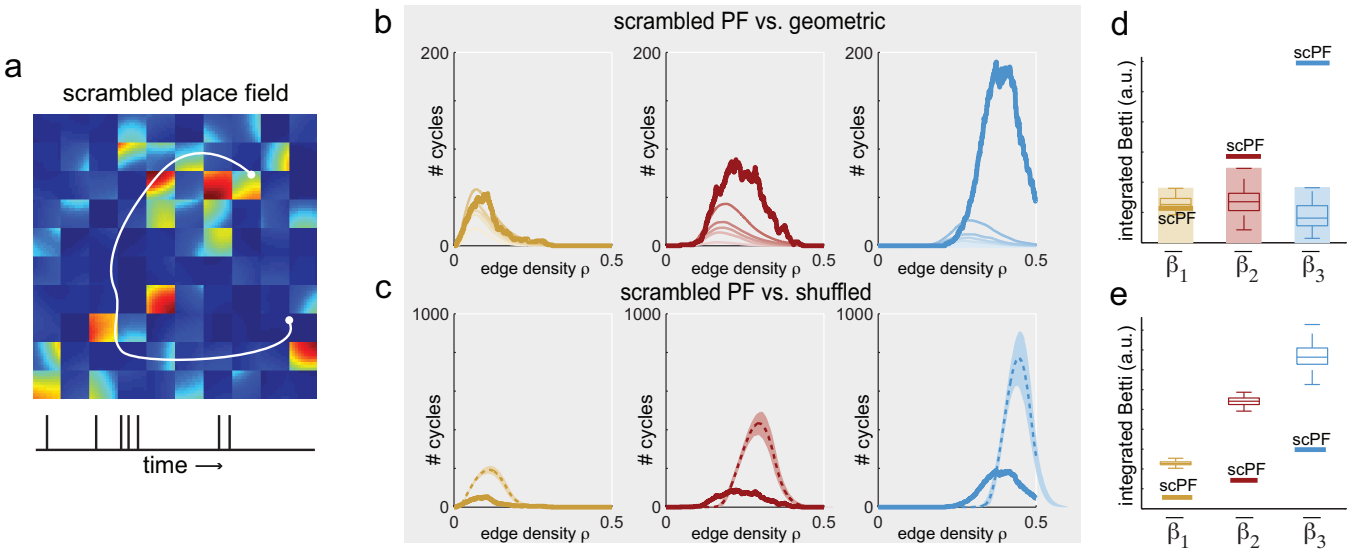
Supplementary Figure 7: The persistence lifetime distributions, computed from neural activity during spatial navigation resemble the geometric controls but are significantly below those of the shuffled matrices. Persistence lifetime distributions, $\ell_1(\rho)$ (yellow), $\ell_2(\rho)$ (red) and $\ell_3(\rho)$ (blue), computed from the same data set and controls as those used in Figure 3b. (Top) The lifetime distributions for the data (solid lines) fall off quickly, while those of the shuffled matrices are much broader (dashed lines are means over 1000 trials; shading shows the 95% confidence intervals). (Bottom) Here the data lifetime distributions (solid lines) are overlaid with the mean distributions (faint lines) for 1000 geometric matrices in each dimension $d = 5, 10, 16, 24$, and 88. As with the geometric Betti curves, the persistence lifetime distributions for geometric matrices are stratified by dimension, with the top curves corresponding to the highest dimension.



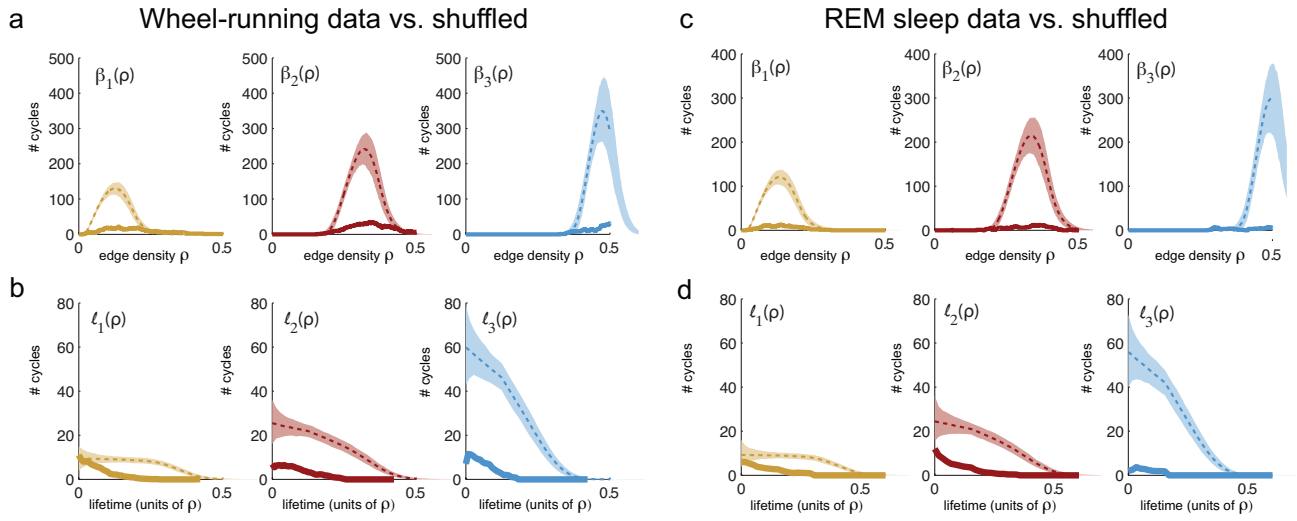
Supplementary Figure 8: Clique topology of WME matrices, sampled from the maximum entropy distribution with prescribed row sums, is similar to that of random (shuffled) matrices. Comparison of Betti curves (top) and persistence lifetimes (bottom) for WME matrices computed using the matrix C_{ij} used in Figure 3a. Each panel compares the Betti curves (top) and the persistence lifetimes (bottom) of the WME matrices, computed using the matrix C_{ij} used in Figure 3a, to those of the random (shuffled) obtained from the same C_{ij}. Each of the three black lines correspond to one sampling of a WME matrix. Colored lines and shaded regions correspond to the mean curves and 95% confidence interval for the shuffled matrices.



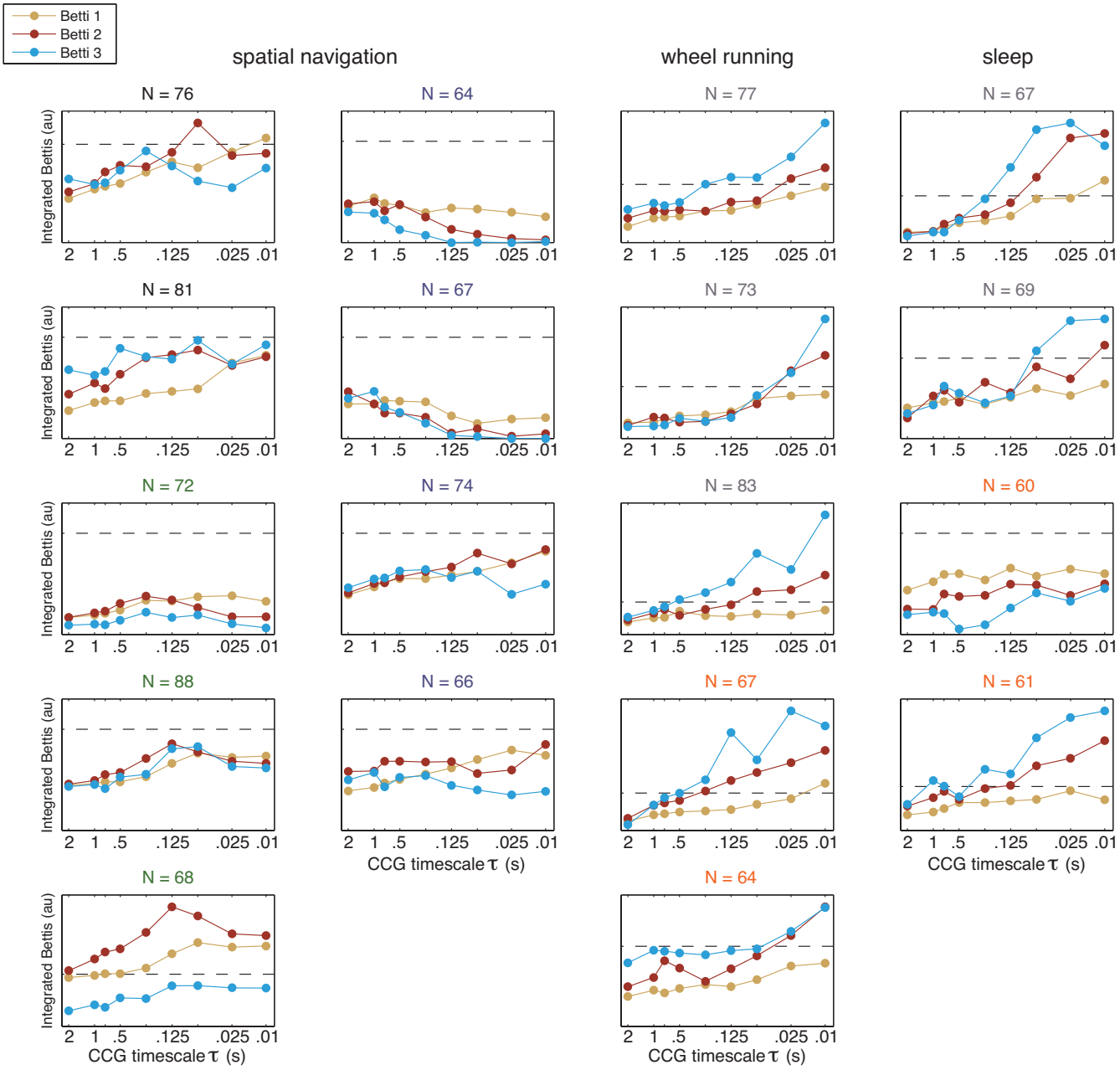
Supplementary Figure 9: (**a**) Example of several place fields computed from spike trains during spatial exploration in the N=88 data set. (**b**) Betti curves (bold lines) computed from the simulated spike trains ($\tau_{\max} = 1sec$) for the place field model versus the geometric Betti curves (thin lines stratified by dimensions). (**c**) Integrated Betti values (bold lines, labelled PF) for the curves in panel (b) lie in the geometric regime, in agreement with those of the original data. (**d**) The scrambled place fields corresponding to those in panel (a). These were obtained by sub-dividing the square into a 100×100 grid and randomly permuting each pixel. The permutations were independent for each cell. (**e**) Betti curves (bold lines) derived from the spike trains ($\tau_{\max} = 1sec$) generated using the scrambled PF model versus the geometric Betti curves (thin lines stratified by dimensions). (**f**) Integrated Betti values from the scrambled PF model (bold lines, labelled scPF) lie outside of the the significance threshold (see Supplementary Methods) for the geometric regime for $\bar{\beta}_2$ and $\bar{\beta}_3$. (**g**) Betti curves (bold lines) derived from the spike trains ($\tau_{\max} = 1sec$) generated using the Scrambled PF model are significantly smaller than those of shuffled controls. (**h**) Integrated Betti values derived from from the scrambled PF model (bold lines, labelled scPF) are outside the 99.9% confidence intervals for the shuffled matrices. Box plots for shuffled matrices are the same as in Figure 3c.



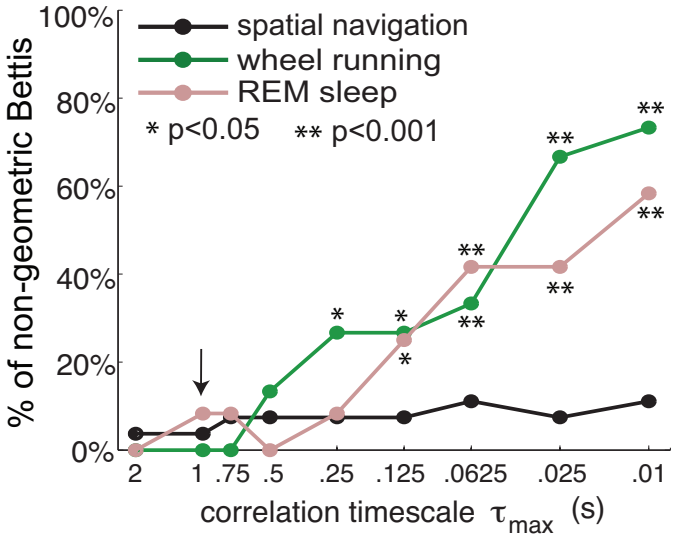
Supplementary Figure 10: Betti curves and summary stats for the 10x10 grid scrambled place fields. (**a**) A single scrambled place field corresponding to the place field in Figure 3e. The scrambling is performed on a 10×10 grid. A cartoon trajectory (white) is displayed together with the corresponding spike train (bottom). (**b**) Betti curves from the scrambled place field model (bold lines) lie outside the geometric regime for β_2 and β_3. (**c**) Betti curves for the scrambled place field model are significantly smaller than the shuffled controls. (**d**) Integrated Betti values (bold lines, labelled scPF) for the scrambled place field model also lie outside of the significance threshold (see Supplementary Methods) for the geometric regime for $\bar{\beta}_2$ and $\bar{\beta}_3$, while $\bar{\beta}_1$ is in the geometric regime. (**e**) Integrated Betti values (bold lines, labelled scPF) for the scrambled place field model are outside the 99.9% confidence intervals for the shuffled matrices. Box plots for geometric and shuffled matrices are the same as in Figure 3c.



Supplementary Figure 11: The clique topology of spike train correlations during wheel running and REM sleep is significantly non-random as compared to shuffled matrices. (**a**) Comparison of the data Betti curves (solid lines) for the spike trains during wheel running (N=73 in Figure 4a) to those of shuffled correlation matrices (dashed lines are means over 1000 trials and shading shows 95% confidence intervals, as in Figure 1e). For each $m = 1, 2, 3$, the data Betti curves are orders of magnitude smaller than those of the shuffled curves. (**b**) Persistence lifetime distributions (solid lines) for the same data as in (a), $\ell_1(\rho)$ (yellow), $\ell_2(\rho)$ (red) and $\ell_3(\rho)$ (blue). The lifetime distributions for the data fall off quickly, while those of the shuffled matrices are much broader (dashed lines are means over 1000 trials; shading shows the 95% confidence intervals). (**c**) Comparison of the Betti curves (solid lines) for spike train correlations during REM sleep (N= 67 in Figure 4b) to those of shuffled correlation matrices (dashed lines are means over 1000 trials and shading shows 95% confidence intervals, as in Figure 1e). For each $m = 1, 2, 3$, the data Betti curves are orders of magnitude smaller than those of the shuffled curves. (**d**) Persistence lifetime distributions (solid lines) for the same data as in (c). The lifetime distributions for the data fall off quickly, while those of the shuffled matrices are much broader (dashed lines are means over 1000 trials; shading shows the 95% confidence intervals).



Supplementary Figure 12: Integrated Betti values across a range of correlation timescales in spatial navigation, wheel running, and REM sleep data sets. For each of the datasets, and each $m = 1, 2, 3$, the integrated Betti number $\bar{\beta}_m = \int_0^1 \beta_m(\rho) d\rho$ was normalized by its significance threshold $b_m = Q_3 + 1.5 \times (Q_3 - Q_1)$ (see Supplementary Methods) that was obtained from the distribution of the integrated geometric Betti curves $\bar{\beta}_m^{\text{geom}} = \int_0^1 \beta_m^{\text{geom}}(\rho) d\rho$ in dimension $d = N$, where N was the number of cells. Each of the curves (yellow, red and blue) correspond to the values of $\bar{\beta}_m / b_m$ in dimensions $m = 1, 2, 3$ respectively. The dashed line marks the line $\bar{\beta}_m = b_m$; the appropriate integrated Betti numbers were deemed consistent with geometric distribution if they lay below this line. Note that the number N of neurons is displayed in color (black, green, purple, gray and orange) to indicate recordings from the same animal; there were a total of 5 animals.



Supplementary Figure 13: Percentage of integrated Betti values that were *not* consistent with geometric controls (i.e. above the significance threshold) across all considered data sets as a function of the correlation timescale, $\tau_{\max}$, used to compute the pairwise correlation matrix (see Methods). The arrow indicates the 1 s timescale used in the main figures. All three behavioral conditions, spatial navigation (black), wheel running (green) and REM sleep (pink), are consistent with geometric structure at timescales ranging from .5 s to 2 s. At finer timescales, however, the wheel running and REM sleep correlations are non-geometric, while the spatial navigation data remains consistent with geometric controls.

For each timescale, we computed integrated Betti values $\bar{\beta}_1$, $\bar{\beta}_2$, and $\bar{\beta}_3$ for all data sets under three different conditions: (i) spatial navigation, (ii) wheel running, and (iii) REM sleep (see Supplementary Figure 12). For each behavioral condition, we counted how many Betti values were above the significance threshold for rejecting the geometric hypothesis at each timescale. Because our significance threshold rejects the geometric hypothesis at a rate of less than 5%, the p-value for a given condition and timescale satisfies

$$p < \sum_{\ell=k}^{m} \binom{m}{\ell} (0.05)^{\ell} (1-0.05)^{m-\ell},$$

where k is the number of Betti values above the significance threshold, and m is the total number of Betti values. To obtain this upper bound on p-value, we used a binomial distribution with failure probability 0.05. Note that this assumes Betti values are independent. Although this is a reasonable assumption for Betti values from different data sets, Betti values from the same data set have statistical dependencies that are not well-understood.

Supplementary Text

To accompany "Clique topology reveals intrinsic geometric structure in neural correlations." Chad Giusti, Eva Pastalkova, Carina Curto*, and Vladimir Itskov*

Contents

1 Introduction

The purpose of this supplement is to provide a more complete account of the mathematics underlying our analyses in the main text. In particular, the *order complex* and *clique topology* are described more precisely here. The order complex of a matrix is analogous to its Jordan Form, in that it captures features that are invariant under a certain type of matrix transformation. Likewise, the clique topology of a matrix is analogous to its eigenvalue spectrum, in that it provides a set of invariants that can be used to detect structure. While the Jordan Form and eigenvalue spectrum are invariant under linear change of variables, the order complex and clique topology are invariant under monotonic transformations of the matrix entries.

Seeking quantities that are invariant under linear coordinate transformations is natural in physical applications, where measurements are often performed with respect to an arbitrary basis, such as the choice of x, y and z directions in physical space. In contrast, measurements in biological settings are often obtained as nonlinear (but monotonic) transformations of the underlying "real" variables, while the choice of basis is meaningful

and fixed. For example, basis elements might represent particular neurons or genes, and measurements (matrix elements) could consist of pairwise correlations in neural activity, or the co-expression of pairs of genes. Unlike change of basis, these transformations are of the form

$$L_{ij} = f(M_{ij}),$$

where f is a nonlinear, but monotonically increasing function that is applied to each entry of M. The Jordan Form of a matrix, and its spectrum, may be badly distorted by such transformations; it also discards basis information which may be meaningful and should be preserved.

Given a symmetric, $N \times N$ matrix that reflects correlations or similarities between N entities (such as neurons, imaging voxels, etc.), we have two basic questions:

Q1. Is the matrix a monotonic transformation of a random or geometric[1] matrix?

Q2. Can we distinguish between these two possibilities, without knowing f?

Perhaps surprisingly, information sufficient to answer these questions is contained in the ordering of matrix entries, and is encoded in its *order complex*, to be described in the next section. To extract the relevant features, we compute certain topological invariants of the order complex, which we refer to as the *clique topology* of the matrix. The motivation for this choice stems from recent mathematical results by M. Kahle [1], describing the clique topology of random symmetric matrices asymptotically (for large N); and our own computational results, showing that random and "generic" Euclidean distance matrices can be readily distinguished using clique topology for $N \sim 100$.

We have made an effort to keep these explanations self-contained, but details of how certain computations are performed have been left to the references for the sake of brevity. Standard material from algebraic topology [2] is described in a minimal fashion, with an emphasis on homology of clique complexes. The reader is expected to be familiar with linear algebra.

Comparison to prior applications in biology

Topological data analysis has previously been used in biological applications to identify individual persistent cycles that may have meaningful interpretation [3, 4, 5, 6, 7, 8]. In contrast, our approach relies on the *statistical properties* of cycles, as captured by Betti curves, in order to detect geometric structure (or randomness) in symmetric matrices. In particular, the relevant space from which the data points are sampled may not possess any meaningful persistent cycles, as in the square box environment covered by place fields. The background Euclidean geometry, however, has a strong effect on the statistics of cycles, enabling detection of geometric structure and providing a sharp contrast to Betti curves of random matrices with i.i.d. entries.

[1] Recall from the main text that a *geometric* matrix refers to a matrix of (negative) Euclidean distances among random points in $\mathbb{R}^d$.

2 The order complex

Recall that a function $f : \mathbb{R} \to \mathbb{R}$ is said to be *monotonically increasing* if $f(x) > f(y)$ whenever $x > y$. Let $f : \mathbb{R} \to \mathbb{R}$ be a monotonically increasing function. For any real-valued matrix M, we define the matrix $f \cdot M$ by

$$(f \cdot M)_{ij} = f(M_{ij}).$$

Note that this action preserves the ordering of matrix entries. That is, if $L = f \cdot M$, then all pairs of off-diagonal entries, (i, j) and (k, ℓ), satisfy:

$$L_{ij} < L_{k\ell} \;\Leftrightarrow\; M_{ij} < M_{k\ell}.$$

Equivalence classes of matrices can thus be represented by integer-valued matrices that record the ordering of off-diagonal entries (and carry no information on the diagonal). Figure 1 shows three matrix orderings for $N = 5$. For a given symmetric matrix M, we

Figure 1: Three matrix orderings, reproduced from Figure 2a in the main text.

denote the representative matrix ordering by $\widehat{M}$, where

$$\widehat{M}_{ij} = |\{(k, \ell) \mid 0 < k < \ell \leq N \text{ and } M_{k\ell} < M_{ij}\}|$$

simply counts the number of upper-triangular entries of M that are smaller than M_{ij} for $i \neq j$, while the diagonal entries of $\widehat{M}$ are left undefined ($|\cdot|$ denotes the size of the set). If M_{ij} is the smallest off-diagonal entry, then $\widehat{M}_{ij} = 0$; if M_{ij} is the largest matrix entry, and all upper-triangular entries are distinct, then $\widehat{M}_{ij} = \binom{N}{2} - 1$. With this notation, we have:

Lemma 2.1. *$\widehat{L} = \widehat{M}$ if and only if there exists a monotonically increasing function f such that $L = f \cdot M$.*

Proof. ($\Leftarrow$) is obvious, since the action of f preserves the ordering of matrix entries. ($\Rightarrow$) One can construct $f : \mathbb{R} \to \mathbb{R}$ by setting $f(M_{ij}) = L_{ij}$ for each off-diagonal entry, and interpolating monotonically (e.g., linearly). Since we assume $\widehat{L} = \widehat{M}$, this function is monotonically increasing and well-defined. □

In order to analyze the information present in the ordering of entries for an $N \times N$ symmetric matrix, it is useful to represent it as a sequence of nested simple graphs. Recall that a *simple graph* G is a pair $([N], E)$, where $[N] = \{1, 2, \ldots, N\}$ is the ordered set of *vertices*, and E is the set of *edges*. Each edge is undirected and connects a unique pair of distinct vertices (no self-loops). We will use the notation $(ij) \in G$ to indicate that the edge corresponding to vertices i, j is in the graph.

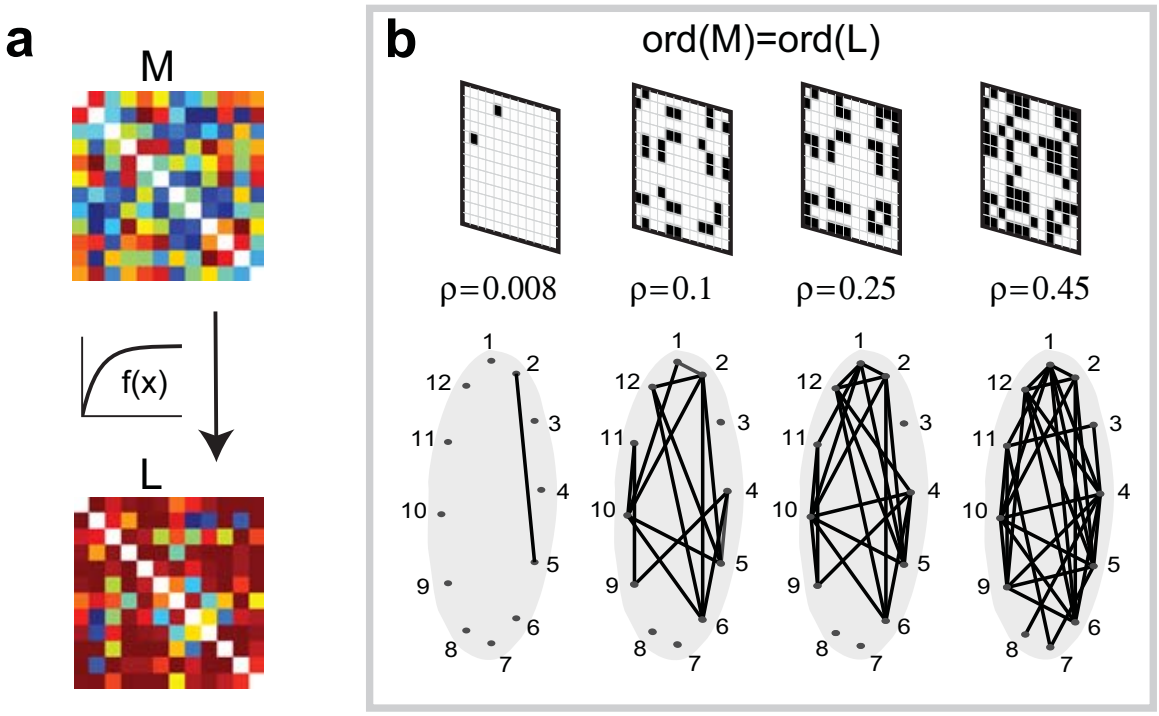


Figure 2: Selected graphs in an order complex, adapted from Figure 1 in the main text.

Definition 2.2. Let M be a real symmetric matrix with matrix ordering $\widehat{M}$, and let $p = \max_{i<j} \widehat{M}_{ij}$. The *order complex* of M, denoted $\mathrm{ord}(M)$, is the sequence of graphs

$$G_0 \subset G_1 \subset \cdots \subset G_{p+1},$$

such that

$$(ij) \in G_r \quad \Leftrightarrow \quad \widehat{M}_{ij} > p - r \ \text{ for each } \ r = 0, \ldots, p+1.$$

Note that G_0 has no edges, G_1 contains only the edge (ij) corresponding to the largest off-diagonal entry of M, and subsequent graphs are each obtained from the previous one by adding an additional edge for each next-largest entry until we reach the complete graph, G_{p+1}. A portion of an order complex is illustrated in Figure 2. It is clear from the definition that:

$$\mathrm{ord}(L) = \mathrm{ord}(M) \ \Leftrightarrow \ \widehat{L} = \widehat{M}.$$

Because of Lemma 2.1, the order complex ord(M) captures all features of M that are preserved under the action of monotonically increasing functions.

3 Clique topology

We are now ready to introduce *clique topology*, a tool for extracting invariant features of a matrix from the ordering of matrix entries. We begin by describing the clique topology of a single graph G, by which we simply mean the homology of its *clique complex*:

$$H_i(X(G), \mathbf{k}),$$

where $\mathbf{k}$ is a field (more on the field in section 3.2). The clique complex, $X(G)$, is defined in section 3.1; while the simplicial homology groups, $H_i(X(G), \mathbf{k})$, will be defined in section 3.3. We refer to these invariants as *clique* topology in order to indicate that we are measuring topological features of the organization of cliques in the graph, rather than the usual topology of the graph.

We summarize the information present in clique topology via a set of *Betti numbers*, $\beta_i(X(G))$, which are the ranks of the corresponding homology groups:

$$\beta_i(X(G)) \stackrel{\text{def}}{=} \text{rank } H_i(X(G), \mathbf{k}).$$

The clique topology of a symmetric matrix M, with order complex $G_0 \subset G_1 \subset \cdots \subset G_{p+1}$, is reflected in the sequences of Betti numbers $\beta_i(X(G_r))$, computed for various dimensions $i = 0, 1, 2, \ldots$, and for each graph G_r in ord(M) (see section 3.4).

The reader familiar with homology of simplicial complexes, including clique complexes, should feel free to skip the next few sections and proceed directly to section 3.4, where we define *Betti curves*.

3.1 The clique complex of a graph

Recall that a *clique* in a graph G is an all-to-all connected collection of vertices in G. An *m-clique* is a clique consisting of m vertices. Note that if σ is a clique of G, then all subsets of σ are also cliques.

Definition 3.1. Let G be a graph with N vertices. The *clique complex* of G, denoted $X(G)$, is the set of all cliques of G:

$$X(G) = \{\sigma \subset [N] \mid \sigma \text{ is a clique of } G\}.$$

We write $X_m(G)$ for the set of $(m+1)$-cliques of G.

The shift in index reflects the "dimension" of a clique, when the clique complex is represented geometrically. If we think of the vertices of the graph G as embedded generically in a high-dimensional space, each clique represents the simplex given by the convex hull of its vertices. For example, the convex hull of two vertices is a 1-dimensional edge, for three vertices we obtain a 2-dimensional triangle, and four vertices yields a 3-dimensional tetrahedron. Thus, cliques in $X_m(G)$ consist of $m+1$ vertices, but represent m-dimensional simplices.

The *boundary* of a clique $\sigma \subseteq G$ is the collection of subcliques $\tau \subset \sigma$ which have one fewer vertex. This corresponds to the set of lower-dimensional simplices that comprise the boundary of the simplex defined by σ (Figure 3b).

The *homology* of a clique complex $X(G)$, to be defined in section 3.3, is a measurement of relationships among the cliques in G. Intuitively, homology counts *cycles* in the clique complex, a higher-dimensional generalization of the notion of cycles in a graph (Figure 3a). A collection of cliques forms a cycle if their boundaries overlap so as to "cancel" one another (Figure 3b).

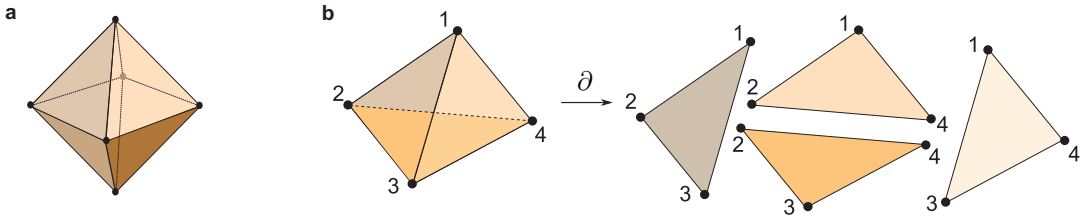


Figure 3: Illustration of homology cycles in clique complexes.

3.2 Chains and boundaries

To make the above notion of "cancellation" of boundaries precise (and computable), one introduces linear combinations of cliques, called *chains*. Given a set of cliques $\sigma_1, \ldots, \sigma_\ell \in X(G)$, one can form a vector space consisting of formal linear combinations of cliques with coefficients in a field $\mathbf{k}$:

$$\sum_{i=1}^{\ell} a_i c_{\sigma_i}, \quad \text{where } a_i \in \mathbf{k},$$

and c_{σ_i} denotes the basis element corresponding to the clique σ_i. To define chain groups, one considers linear combinations of cliques of the same size. Recall that $X_m(G)$ denotes the set of $(m+1)$-cliques of G.

Definition 3.2. The *m-th chain group* of $X(G)$, with coefficients in $\mathbf{k}$, is the $\mathbf{k}$-vector

space:

$$C_m(X(G);\mathbf{k}) \stackrel{\text{def}}{=} \left\{ \sum_{i=1}^{\ell} a_i c_{\sigma_i} \mid \sigma_i \in X_m(G) \text{ and } a_i \in \mathbf{k} \text{ for each } i = 1, \ldots, \ell \right\}.$$

As we will always be working with coefficients in an aribitrary field $\mathbf{k}$,[2] we will omit it from the notation and write $C_m(X(G))$ instead of $C_m(X(G);\mathbf{k})$.Note that $C_0(X(G))$ consists of formal linear combinations of 1-cliques (vertices), $C_1(X(G))$ of 2-cliques (edges), $C_2(X(G))$ of 3-cliques (triangles), and so on.

The boundaries of cliques can also be described algebraically, allowing this notion to be extended to chains. If $\sigma = \{v_i\}_{i=0}^m$ is an m-clique of G, we use the notation

$$c_\sigma = c_{v_0 v_1 \ldots v_m},$$

where $v_0 < v_1 < \ldots < v_m$ (recall that each $v_i \in [N]$). Consistent ordering is important because it affects the signs in the boundary map. Given a sequence of vertices $v_0 v_1 \ldots v_m$, we denote by $v_0 v_1 \ldots \hat{v}_i \ldots v_m$ the sequence obtained by omitting the element v_i. Note that for each $\sigma \in X_m(G)$, the element $c_\sigma = c_{v_0 v_1 \ldots v_m}$ is a basis element of the vector space $C_m(X(G))$.

Definition 3.3. The *boundary* map $\partial_m : C_m(X(G)) \to C_{m-1}(X(G))$, for $m > 0$, is given on basis elements $c_{v_0 v_1 \ldots v_m}$ by

$$\partial_m(c_{v_0 v_1 \ldots v_m}) = \sum_{i=0}^{m} (-1)^i c_{v_0 v_1 \ldots \hat{v_i} \ldots v_m},$$

and is extended via linearity to general chains; i.e. $\partial_m(\sum_j a_j c_{\sigma_j}) = \sum a_j \partial_m(c_{\sigma_j})$. The map ∂_0 is defined to be the zero map.

Recall that in the geometric picture, an $(m+1)$-clique corresponds to a m-dimensional simplex, and the boundary of this simplex is the set of m-cliques comprising its $(m-1)$-dimensional *facets* – that is, all subcliques on one fewer vertex. We have thus defined the boundary of a chain in $C_m(X(G))$ in a fashion consistent with our geometric understanding: as a formal sum of chains in $C_{m-1}(X(G))$, corresponding to simplices that are one dimension lower (see Figure 3b). Note that signs are assigned to the elements of this formal sum to indicate the *orientation* of cliques, which will be critical for obtaining the desired "cancellation" of boundaries (see Remark 3.6 for details).

[2]For readers uncomfortable with the notion of a general field $\mathbf{k}$, it is relatively harmless to substitute $\mathbb{R}$ or $\mathbb{Q}$ for $\mathbf{k}$ for the remainder of the discussion. One should keep in mind, however, that the actual computations typically take place with $\mathbf{k}$ a finite field, $\mathbb{Z}/p\mathbb{Z}$. This can have an effect on the result: in such a field, one can add a boundary to itself a finite number of times and get zero, creating "extra" cycles – called *torsion* cycles – that would not be present over $\mathbb{R}$. These extra cycles measure aspects of the clique complex that are not relevant to our purposes. In our software we have chosen the field to be $\mathbb{Z}/2\mathbb{Z}$, but this choice is somewhat arbitrary and not important. So long as all computations are done using the same field, comparing the resulting homology groups across different graphs is entirely valid.

Example 3.4. Suppose $\sigma, \tau \in X_2(G)$ are cliques on vertices $\{1,2,3\}$ and $\{1,2,4\}$ respectively. The boundary of the 2-chain $c_\sigma - c_\tau \in C_2(X(G))$ is

$$\begin{aligned}\partial_2(c_{123} - c_{124}) &= \partial_2(c_{123}) - \partial_2(c_{124}) \\ &= (c_{23} - c_{13} + c_{12}) - (c_{24} - c_{14} + c_{12}) \\ &= c_{23} - c_{13} - c_{24} + c_{14}.\end{aligned}$$

The cancellation of c_{12} reflects the fact that the clique $\{1,2\}$ appears twice in the boundary of $c_\sigma - c_\tau$, with *opposite* orientation. Note also that applying ∂_1 to the resulting 1-chain yields

$$\begin{aligned}\partial_1(\partial_2(c_{123} - c_{124})) &= \partial_1(c_{23} - c_{13} - c_{24} + c_{14}) \\ &= (c_3 - c_2) - (c_3 - c_1) - (c_4 - c_2) + (c_4 - c_1) \\ &= 0.\end{aligned}$$

In fact, it is straightforward to check from the definition that the composition of two subsequent boundary maps always yields 0. In other words,

Lemma 3.5. *For any $m > 0$, $\partial_m \circ \partial_{m+1} = 0$.*

Remark 3.6. The *orientation* of cliques can be *positive* or *negative*. The vertices of a clique $c_{v_0 v_1 \ldots v_m} \in X_m(G)$ have a *canonical* ordering induced by the usual ordering of the vertices $[N]$ of G. We define the canonical ordering to have positive orientation for each clique. Any other ordering can be obtained as a permutation of the canonical ordering, and the resulting ordering is positive or negative according to the sign of the permutation. For example, c_{124} has a positive orientation, while c_{214} is negatively oriented. When we compute the boundary of a clique c_σ in Definition 3.3, the signs arise as a result of the *induced orientation* on the boundary cliques. The result of taking all cliques on the boundary is the signed sum we obtain in Definition 3.3.

3.3 Homology of a clique complex

For a given graph G, the chain groups $C_m(X(G))$ can be strung together to form a *chain complex*:

$$0 \xrightarrow{\partial_{k+1}=0} C_k(X(G)) \xrightarrow{\partial_k} C_{k-1}(X(G)) \xrightarrow{\partial_{k-1}} \ldots \xrightarrow{\partial_2} C_1(X(G)) \xrightarrow{\partial_1} C_0(X(G)) \xrightarrow{\partial_0=0} 0,$$

The zeroes at either end of the complex represent the zero-dimensional $\mathbf{k}$-vector space, and the maps at each end are necessarily the zero map.

If a chain is in the kernel of the boundary map, it is because the (oriented) boundaries of its constituent cliques cancel one another. This is precisely the desired notion of a *cycle*, so the set of m-cycles is exactly $\ker(\partial_m)$; in particular, 1-cycles correspond to the

usual notion of cycles in a graph. Note also that any chain in $C_m(X(G))$ which forms the boundary of a clique in $X_{m+1}(G)$ is itself a cycle, so its own boundary should be zero. This is reflected in the fact that $\partial_m \circ \partial_{m+1} = 0$ (Lemma 3.5). In particular,

$$\operatorname{im} \partial_{m+1} \subset \ker \partial_m.$$

When we are counting cycles for homology, we do not want to consider those which arise as boundaries of chains, as these are "filled in." For example, the two clique complexes in Figure 4 should have the same number of homology 1-cycles. In Figure 4b, we do not wish to count the chain $c_{23} + c_{35} - c_{25} \in C_1(X(G))$ as a 1-cycle because it is the boundary of a clique, $c_{235} \in C_2(X(G))$.

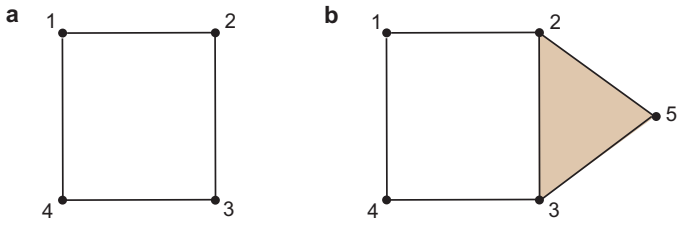


Figure 4: Two clique complexes for graphs on 4 and 5 vertices.

In order to eliminate cycles that are boundaries of higher-dimensional cliques, one computes quotient vector spaces, $\ker(\partial_m)/\operatorname{im}(\partial_{m+1})$.

Definition 3.7. The m-th *homology group* of $X(G)$ with coefficients in $\mathbf{k}$ is the quotient space

$$H_m(X(G); \mathbf{k}) \stackrel{\text{def}}{=} \frac{\ker(\partial_m)}{\operatorname{im}(\partial_{m+1})}.$$

As with chain groups, we will omit the field from our notation and write simply $H_m(X(G))$.

Observe that the zeroth homology group is special: since $\partial_0 = 0$, its kernel is always $C_0(X(G))$. The quotient $\ker(\partial_0)/\operatorname{im}(\partial_1)$ thus identifies vertices which are connected to one another, so that $H_0(X(G))$ is a vector space whose basis can be chosen to correspond to the *connected components* of G.

Example 3.8. Let G be the graph on four vertices in Figure 4a. The kernel of the boundary map $\partial_1 : C_1(X(G)) \to C_0(X(G))$ is the one-dimensional space spanned by $\sigma = c_{12} + c_{23} + c_{34} - c_{14}$. Indeed, $\partial_1(\sigma) = (c_2 - c_1) + (c_3 - c_2) + (c_4 - c_3) - (c_4 - c_1) = 0$. Since there are no cliques of size greater than 2, $C_2(X(G)) = 0$ and hence $\partial_2 = 0$. It follows that $H_1(X(G))$ is precisely the one-dimensional vector space spanned by σ. Furthermore, since $C_1(X(G))$ has dimension 4 and $\ker \partial_1$ has dimension 1, it follows that $\operatorname{im} \partial_1$ has dimension 3. We can thus deduce that $H_0(X(G))$ is also one-dimensional, consistent with the fact that G has just one connected component.

Next, consider the graph G' on five vertices in Figure 4b. This graph has been obtained from G by "attaching" the clique $\{2, 3, 5\}$. The kernel of ∂_1 is now 2-dimensional, and is spanned by both σ and a new cycle, $\tau = c_{23} + c_{35} - c_{25}$. However, $\tau \in \operatorname{im} \partial_2$, so we find that $H_1(X(G'))$ continues to be one-dimensional, consistent with our intuition that G and G' both have just one cycle that has not been "filled in" by cliques.

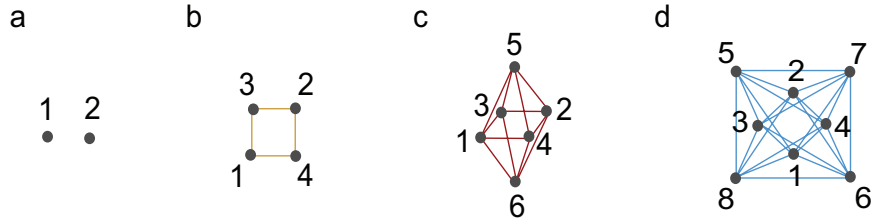


Figure 5: Cross-polytopes generate the minimal clique complexes which produce homology in each dimension. Adapted from Figure 1 of the main text.

Example 3.9. The smallest example of a graph G_m whose clique complex has non-trivial m-th homology group is the 1-skeleton of the $(m+1)$-dimensional cross-polytope (Figure 5). Such a graph can be built inductively starting from the graph G_0 (Figure 5a), having just two vertices and no edges. To obtain G_1 from G_0, we attach two new vertices and include all edges between the new vertices and the vertices of G_0 (Figure 5b). More generally, to obtain G_i from G_{i-1} we attach two new vertices and all edges between these new vertices and those of G_{i-1}. Thus, we obtain G_2 (Figure 5c) and G_3 (Figure 5d), which give minimal examples of graphs whose clique complexes have a non-trivial homology 2-cycle and 3-cycle, respectively.

A useful characterization of the clique topology of a graph is obtained by simply tracking the dimensions of the homology groups. This is done via the so-called Betti numbers.

Definition 3.10. The *m-th Betti number* of $X(G)$, denoted $\beta_m(X(G))$, is the rank of $H_m(X(G); \mathbf{k})$ as a $\mathbf{k}$-vector space.

While this information discards the identities of individual cycles, it is well-suited to statistical methods as it reduces the clique topology of a graph to a sequence of integers.

3.4 Clique topology across the order complex

We now turn our attention to the clique topology of all graphs in the order complex at once. For a matrix M, the Betti numbers of the graphs in $\operatorname{ord}(M)$ are collected as follows.

Definition 3.11. Let M be a real symmetric matrix and $\operatorname{ord}(M) = (G_0 \subset G_1 \subset G_2 \subset \cdots \subset G_{p+1})$ its order complex, where $p = \max_{i<j} \widehat{M}_{ij}$. The *$m$-th Betti curve* of M is the

sequence of numbers $\{\beta_m(\rho_r)\}_{r=1}^{p+1}$, where ρ_r is the edge density of the graph G_r, and

$$\beta_m(\rho_r) \stackrel{\text{def}}{=} \operatorname{rank} H_m(X(G_r)).$$

As the matrix M will be clear from context, we omit it from the notation.

While each Betti curve is a discrete sequence, we can think of it as being a piecewise constant function. To simplify comparison, we consider as a summary statistic the integral of the entire Betti curve. We call this the *m-th total Betti number* of the matrix M, given by

$$\bar{\beta}_m(M) \stackrel{\text{def}}{=} \sum_{r=1}^{p+1} \beta_m(\rho_r)\Delta\rho_r = \int_0^1 \beta_m(\rho)\, \mathrm{d}\rho,$$

where $\Delta\rho_r$ is the change in edge density between G_r and G_{r-1}.[3] Typically, $\Delta\rho_r = 1/\binom{N}{2}$, which is the change in density after adding a single edge. As we will see, the $\bar{\beta}_m$ alone can distinguish between a random symmetric matrix, drawn from a distribution with i.i.d. entries, and a geometric matrix, which arises from distances between a set of randomly-distributed points in Euclidean space. Thus, we can use the total Betti number to test the hypotheses that a matrix is random or geometric.

3.5 Clique topology of random and geometric matrices

In order to interpret the results of computing clique topology for matrices of interest, we need suitable null models for comparison. This brings us back to our motivating questions Q1 and Q2 from section 1. Can we use clique topology to reject the hypothesis that a given matrix is random or geometric? This will be possible if matrices in these categories have stereotyped Betti curves. In this case, it can be shown that a matrix with a substantially different Betti curve is unlikely to have come from the given null model distribution, and a p-value can be assigned to quantify the significance.

Because clique topology depends only on $\operatorname{ord}(M)$, it suffices to describe the distributions of order complexes we obtain for random and geometric matrices. In both families, the details of the Betti curves change with N; however, we find that their large-scale features are robust once $N > 50$. This means Betti curves can indeed be used to reject these models.

The distribution of *random order complexes* arises by sampling a matrix ordering $\widehat{M}$ from the uniform distribution on all such orderings. For $N \times N$ symmetric matrices with distinct entries $\{M_{ij}\}_{i<j}$, this can be achieved by sampling permutations of $\{0, \ldots, \binom{N}{2}-1\}$ uniformly at random. Equivalently, the matrix can be chosen with i.i.d. entries drawn from any continuous distribution, or by shuffling the elements of a given matrix with

[3] This measurement, $\bar{\beta}_m(M)$, also appears as the first element in the basis for the ring of algebraic functions on the collection of all persistence structures described in [9].

distinct off-diagonal entries. Thus, in a graph G_ρ of ord(M), each edge has independent probability ρ of appearing. In other words, the graphs in the order complex are a nested family of Erdös-Rényi random graphs. The clique topology of such complexes is relatively well understood from a theoretical perspective [1], with highly stereotyped, unimodal Betti curves as illustrated in Figure 6a.

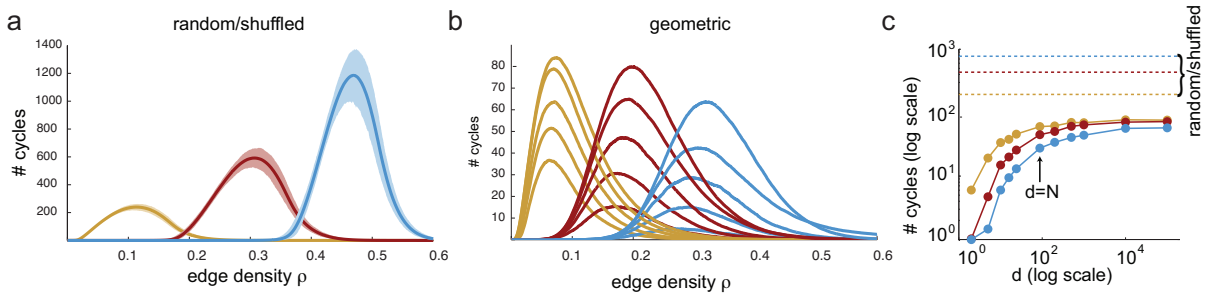


Figure 6: Betti curves for random and geometric matrices. **(a)** Betti curves of random matrices with $N = 100$. The means for the Betti curves $\beta_1(\rho)$ (yellow), $\beta_2(\rho)$ (red), and $\beta_3(\rho)$ (blue) are displayed with bold lines, while shading indicates 99.5% confidence intervals. **(b)** Betti curves of geometric matrices with $N = 100$. Average Betti curves are displayed for dimensions $d = 10, 50, 100, 1000, 10000$, in increasing order (i.e., higher curves correspond to larger dimensions). **(c)** The maxima of Betti curves for $N = 88$ geometric matrices as a function of dimension. Beyond $d = N$, peak values increase only slightly and appear to saturate. The $d = 10^4$ and $d = 10^5$ values are nearly identical, and far from the peak values for random/shuffled matrices with matching N (dashed lines). Adapted from Figure 2b of the main text.

A *geometric order complex* is one arising from the negative distance matrix of a collection of points embedded in some Euclidean space. We choose *negative* distance matrices so that the highest matrix values correspond to the nearest distances; this is consistent with the intuition that correlations should decrease with distance, as described in the main text. Sampling such a complex consists of sampling N i.i.d. points, $\{p_i\}$, from some distribution on $\mathbb{R}^d$. The associated sequence of clique complexes, $X(G_0) \subset X(G_1) \subset \ldots \subset X(G_{p+1})$, corresponding to a geometric order complex is also referred to as the *Vietoris-Rips complex* of the underlying points. These complexes have been heavily studied in cases where the points are presumably sampled from an underlying manifold [10].

In our setting, we sample points from the uniform distribution on the unit cube in $\mathbb{R}^d$. To our knowledge, the Betti curves of geometric order complexes are largely unstudied. Our numerical experiments show that they are highly stereotyped (Figure 6b), irrespective of d for a large range of dimensions. Moreover, the peaks of the geometric Betti curves are roughly an order of magnitude smaller than those of random order complexes with matching N, and the peak values *decrease* rather than increase as we move between $\beta_1(\rho)$

to $\beta_2(\rho)$ and $\beta_3(\rho)$. We observed similar Betti curves to those in Figure 6b for values of d that were orders of magnitude larger than N, testing dimensions up to $d = 10^5$ (for $N = 88$). Beyond $d = N$, peak values of Betti curves increase only slightly and appear to saturate, while remaining far from random/shuffled Betti curves (Figure 6c). The differences between the Betti curves of random and geometric matrices can also be understood through the lens of *persistence lifetimes*, which we will describe in section 4.2.

3.6 Some remarks on geometric order complexes

Any matrix ordering $\widehat{M}$ appears with equal probability in the distribution of random symmetric matrices with i.i.d. entries. The consistency of the Betti curves in Figure 6a indicates that "most" of these matrix orderings have a similar organization of cliques. For geometric matrices, the possible matrix orderings are sampled in a highly non-uniform manner, leading to dramatically different Betti curves. Despite this, it is worth noting that any matrix ordering can in fact arise from a geometric matrix.

Definition 3.12. A set of points $p_1, \ldots, p_N \in \mathbb{R}^d$ is called a *geometric realization* of the matrix ordering $\widehat{M}$ if the distance matrix $D_{ij} = ||p_i - p_j||$ has $\widehat{D} = \widehat{M}$.

Note that for each collection of three or more points, the (higher) triangle inequalities implied by the metric impose strong constraints on $\widehat{M}$. This means that for most matrix orderings, the probability of sampling a point configuration in the unit cube that yields a geometric realization of $\widehat{M}$ is vanishingly small. This is why geometric Betti curves are, on average, so different from those of random matrices. Nevertheless, geometric realizations do always exist, provided $d \geq N - 1$.

Lemma 3.13. *Every $N \times N$ matrix ordering $\widehat{M}$ that has $\binom{N}{2}$ distinct off-diagonal entries possesses a geometric realization in $(N-1)$-dimensional Euclidean space. Moreover, this realization can be chosen as*

$$p_i = \frac{1}{\sqrt{2}} \left(\vec{e}_i - \frac{\varepsilon}{2} \sum_{j=1}^{N} M_{ij} \vec{e}_j \right),$$

for small enough $\varepsilon > 0$, where M is any symmetric matrix with ordering $\widehat{M}$ and zeroes on the diagonal, and $\{\vec{e}_i\}_{i=1}^N$ is the standard orthonormal basis in $\mathbb{R}^N$.

Proof. With the choice above, $||p_i - p_j||^2 = ||p_i||^2 + ||p_j||^2 - 2p_i \cdot p_j = 1 + \varepsilon M_{ij} + \mathcal{O}(\varepsilon^2)$. □

Despite this fact, when we constrain the dimension d of the Euclidean space we do find matrix orderings that cannot be geometrically realized at all. This was the basis for our examples in Figure 2a of the main text.

Figure 2a examples from the main text. Here we prove that the $d \geq 2$ and $d \geq 3$ matrices (reproduced in Figure 1) cannot be geometrically realized in lower dimensions.

To see why the $d \geq 2$ matrix cannot arise from an arrangement of points on a line, observe that the three smallest matrix entries are M_{12}, M_{13}, and M_{14}. This implies the three shortest distances in a corresponding point arrangement must all involve the point p_1, which is not possible for points on a line.

To see why the $d \geq 3$ matrix cannot arise from an arrangement of points on a plane, notice that the six smallest matrix entries are $M_{i\alpha}$, for $i = 1, 2, 3$ and $\alpha = 4, 5$. This means the six smallest distances are those of the form $\|p_i - p_\alpha\|$, for $i = 1, 2, 3$ and $\alpha = 4, 5$. Without loss of generality we can assume $\|p_i - p_\alpha\| < 1$, and all other distances are greater than one. Now suppose the points $p_1, \ldots, p_5$ all lie in a plane. Then $p_4, p_5 \in D(p_1) \cap D(p_2) \cap D(p_3)$, where $D(p_i)$ is a disk of radius 1 centered at p_i. Since none of the disk centers in contained in any of the other two disks, the largest distance between two points in the intersection $D(p_1) \cap D(p_2) \cap D(p_3)$ is less than one, and thus $\|p_4 - p_5\| < 1$, which is a contradiction. We conclude that the matrix cannot arise from points in the plane. We thank Anton Petrunin for this example.

4 Computational aspects and persistence

Each graph in an order complex, $G_0 \subset G_1 \subset \cdots \subset G_{p+1}$, is a subgraph of its successor. Intuitively, this means that the clique topology of any G_r is closely related to the clique topology of the previous graph, G_{r-1}. Exploiting this structure dramatically reduces the computational complexity of finding Betti curves (defined in section 3.4), and also provides us with finer matrix invariants in the form of *persistence lifetimes* of cycles. This is achieved via *persistent homology*, an approach that enables homology cycles to be tracked as we move from one graph in the order complex to the next.

4.1 A brief history of persistent homology

The mathematics underlying persistent homology has existed since the middle of the twentieth century, in the guise of Morse theory and spectral sequences for the homology of filtered spaces. Its interpretation as a tool for data analysis, however, is a much more recent development. One can trace the origins of these applications to work on size theory in computer vision [11, 12, 13] and alpha shapes in computational geometry [14, 15, 16, 17]. The use of persistent homology as a tool for the study of data sets relies on two fundamental and recent developments: computabilty and robustness.

Computability arose from the *persistence algorithm*, developed first for subsets of three-dimensional complexes in [17] and then extended to work with general simplicial complexes in [18]. In addition to the algorithm, these papers introduced the notions of persistence diagrams and modules. Several software packages [19, 20, 21] have been developed based on the persistence algorithm, and recent work using discrete Morse theory has led to further improvements in speed and memory efficiency [22].

Robustness to perturbations of the underlying simplicial complexes, on the other hand, was first explicitly shown through the bottleneck stability theorem of [23]. Further work has broadened this result by developing more complete theoretical tools for the comparison of persistence structures, divorcing their stability from any underlying geometry [24, 25, 26]. It is this interpretation of stability that most clearly applies to our study of order complexes.

Although persistent homology has only recently emerged a tool for studying features of data, it has already found a broad range of applications [27, 28, 29, 30].

4.2 Persistent homology of order complexes

Here we present the basic ideas in persistent homology, restricted to the special case of computing clique topology for order complexes. This means we need to apply the persistence algorithm to filtered families of clique complexes,

$$X(G_0) \subset X(G_1) \subset \cdots \subset X(G_p) \subset X(G_{p+1}),$$

where the graphs $\{G_r\}$ comprise the order complex of a symmetric matrix. In order to track homology cycles from one clique complex to the next, we need to understand how the natural inclusion maps on the graphs, $\iota_r : G_r \hookrightarrow G_{r+1}$, translate to maps on the corresponding cliques, chains, and homology groups, $H_m(X(G_r))$. This turns out to be straightforward, as there is an obvious extension to maps on clique complexes, $\iota_r : X(G_r) \hookrightarrow X(G_{r+1})$, and these in turn can be extended linearly to maps between chain groups.

Lemma 4.1. *Consider the order complex $G_0 \subset G_1 \subset \cdots \subset G_{p+1}$. The standard inclusion maps, $\iota_r : G_r \hookrightarrow G_{r+1}$, induce maps on homology $(\iota_r)_m : H_m(X(G_r)) \to H_m(X(G_{r+1}))$.*

Using these maps, one can follow individual cycles and understand their evolution as we move from one graph to the next in the order complex. Of particular interest are the edge densities at which cycles appear and disappear (Figure 7).

Definition 4.2. Let $\omega \in H_m(X(G_r))$ be a non-zero cycle which is not in the image of ι_{r-1}, and let $s > r$ be the smallest integer such that $\iota_{s-1} \circ \iota_{s-2} \circ \ldots \circ \iota_r(\omega) = 0$. We say that ω is *born* at r and *dies* at s, and has *persistence lifetime* $\ell(\omega) = s - r$.

For a given order complex, the distribution of persistence lifetimes provides a measure of matrix structure that is complementary to the Betti curves defined in section 3.4.

4.3 Persistence lifetimes of random and geometric order complexes

Recall that there is a sharp qualitative difference in the Betti curves of random order complexes and those of geometric order complexes (Figure 6). These differences are also

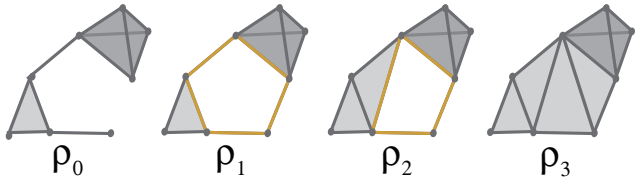


Figure 7: Illustration of persistence lifetime. The 1-cycle (yellow) appears at edge density ρ_1 and disappears at ρ_3, so it's lifetime is $\ell = \rho_3 - \rho_1$.

reflected in the distributions of their persistence lifetimes. While random complexes have relatively broad distributions (Figure 8a), the geometric complexes are heavily weighted toward shorter lifetimes (Figure 8b). The shapes of these distributions are a direct consequence of the order in which edges are added in the order complex.

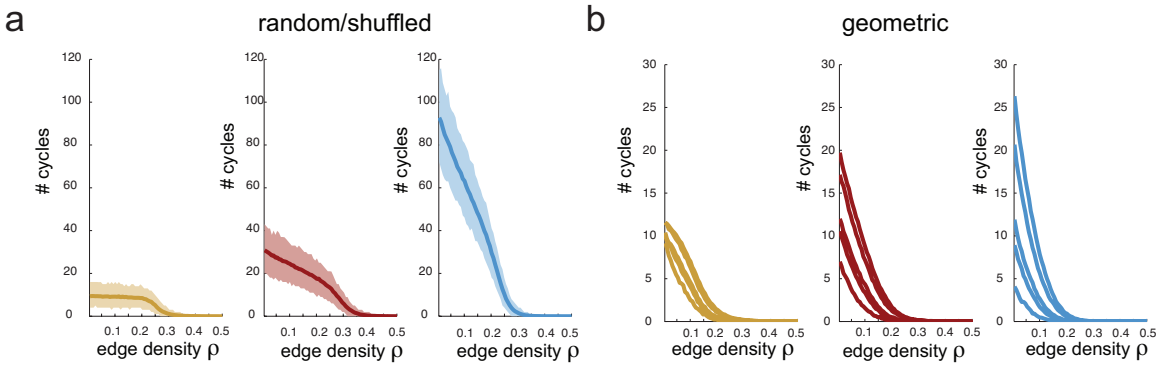


Figure 8: Persistence lifetimes for random and geometric matrices. (a) N = 100, and mean lifetime distributions for 1-cycles (yellow), 2-cycles (red), and 3-cycles (blue) are displayed with bold lines, while shading indicates 99.5% confidence intervals. (b) N=100, and average lifetime distributions are displayed for dimensions $d = 10, 50, 100, 1000, 10000$, in increasing order (i.e., higher curves correspond to larger dimensions).

The qualitative differences in these distributions can be understood by thinking about dependencies in edge orderings in the order complex. Minimal cycles, represented by cross-polytopes (Figure 5), are known to constitute the large majority of cycles in random order complexes [31], and can thus be used to understand the shape of the distribution. Such a cycle's lifetime is governed by the density at which the first additional edge appears, since the extra edge destroys the cycle by creating new cliques. Since the ordering of the edges is completely random, the lifetimes will be broadly distributed. In contrast, geometric order complexes are constrained by triangle inequalities (and higher-dimensional

analogues); these produce dependencies in the edge ordering which imposes an upper limit on the lifetime of small cycles, like the cross-polytopes.[4] Persistence lifetimes in geometric complexes are thus concentrated at short lifetimes.

4.4 CliqueTop software

To compute clique topology for symmetric matrices, we developed the CliqueTop Matlab package. This software is maintained by Chad Giusti (one of the authors), and is available on GitHub at `https://github.com/nebneuron/clique-top`. At the time of this writing, CliqueTop makes use of one other package: Perseus [22], by Vidit Nanda. Perseus provides an implementation of the persistence algorithm, and is available at `http://www.sas.upenn.edu/~vnanda/perseus/index.html`. Previous versions of CliqueTop also used the Cliquer software package [32].

References

[1] Matthew Kahle. Topology of random clique complexes. *Discrete Math.*, 309(6):1658–1671, 2009.

[2] Allen Hatcher. *Algebraic topology*. Cambridge University Press, Cambridge, 2002.

[3] Carina Curto and Vladimir Itskov. Cell groups reveal structure of stimulus space. *PLoS Comput Biol*, 4(10), 2008.

[4] Gurjeet Singh, Facundo Memoli, Tigran Ishkhanov, Guillermo Sapiro, Gunnar Carlsson, and Dario L Ringach. Topological analysis of population activity in visual cortex. *Journal of vision*, 8(8):11, 2008.

[5] Monica Nicolau, Arnold J. Levine, and Gunnar Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270, 2011.

[6] Y. Dabaghian, F. Memoli, L. Frank, and G. Carlsson. A topological paradigm for hippocampal spatial map formation using persistent homology. *PLoS Comput Biol*, 8(8), 08 2012.

[7] Joseph Minhow Chan, Gunnar Carlsson, and Raul Rabadan. Topology of viral evolution. *Proceedings of the National Academy of Sciences*, 2013.

[4] This is true statistically. It is, of course, possible for two distances to be close in absolute terms and still be separated by many edges in the order complex, but this is rare enough that the intuition about Betti curves still holds.

[8] Zhe Chen, Stephen N. Gomperts, Jun Yamamoto, and Matthew A. Wilson. Neural representation of spatial topology in the rodent hippocampus. *Neural computation*, 26(1):1–39, Jan 2014.

[9] Aaron Adcock, Erik Carlsson, and Gunnar Carlsson. The ring of algebraic functions on persistence bar codes, 2013.

[10] Anne D. Collins, Afra Zomorodian, Gunnar E. Carlsson, and Leonidas J. Guibas. A barcode shape descriptor for curve point cloud data. *Computers & Graphics*, 28(6):881–894, 2004.

[11] Patrizio Frosini and Claudia Landi. Persistent Betti numbers for a noise tolerant shape-based approach to image retrieval. In *Computer analysis of images and patterns. Part I*, volume 6854 of *Lecture Notes in Comput. Sci.*, pages 294–301. Springer, Heidelberg, 2011.

[12] Francesca Cagliari, Massimo Ferri, and Paola Pozzi. Size functions from a categorical viewpoint. *Acta Appl. Math.*, 67(3):225–235, 2001.

[13] Patrizio Frosini and Michele Mulazzani. Size homotopy groups for computation of natural size distances. *Bull. Belg. Math. Soc. Simon Stevin*, 6(3):455–464, 1999.

[14] V. Robins. Towards computing homology from finite approximations. In *Proceedings of the 14th Summer Conference on General Topology and its Applications (Brookville, NY, 1999)*, volume 24, pages 503–532 (2001), 1999.

[15] H. Edelsbrunner. The union of balls and its dual shape. *Discrete Comput. Geom.*, 13(3-4):415–440, 1995.

[16] H. Edelsbrunner and E. P. Mucke. Three-dimensional alpha shapes. *ACM Trans. Graphics*, 13:43–72, 1994.

[17] Herbert Edelsbrunner, David Letscher, and Afra Zomorodian. Topological persistence and simplification. In *41st Annual Symposium on Foundations of Computer Science (Redondo Beach, CA, 2000)*, pages 454–463. IEEE Comput. Soc. Press, Los Alamitos, CA, 2000.

[18] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. *Discrete Comput. Geom.*, 33(2):249–274, 2005.

[19] Andrew Tausz, Mikael Vejdemo-Johansson, and Henry Adams. Javaplex: A research software package for persistent (co)homology. `http://javaplex.github.io/javaplex/`, 2011.

[20] Dmitriy Morozov. Dionysis, c++ library for computing persistent homology. `http://www.mrzv.org/software/dionysus/`, 2011–2014.

[21] Andrew Tausz. phom: Persistent homology in r. `http://cran.r-project.org`, 2011.

[22] Vidit Nanda. Perseus, the persistent homology software. `http://www.sas.upenn.edu/~vnanda/perseus` Accessed 06/14, 2012–2014.

[23] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete Comput. Geom.*, 37(1):103–120, 2007.

[24] F. Chazal, D. Cohen-Steiner, M. Glisse, L. J. Guibas, and S. Y. Oudot. Proximity of persistence modules and their diagrams. In *Proc. 25th Annu. Symposium on Computational Geometry*, pages 237–246, 2009.

[25] Frederic Chazal, Vin de Silva, Marc Glisse, and Steve Oudot. The structure and stability of persistence modules, 2012.

[26] Frederic Chazal, Vin de Silva, and Steve Oudot. Persistence stability for geometric complexes, 2012.

[27] Joseph Minhow Chan, Gunnar Carlsson, and Raul Rabadan. Topology of viral evolution. *Proc. Natl. Acad. Sci. USA*, 110(46):18566–18571, 2013.

[28] Gurjeet Singh, Facundo Memoli, Tigran Ishkhanov, Guillermo Sapiro, Gunnar Carlsson, and Dario L Ringach. Topological analysis of population activity in visual cortex. *Journal of vision*, 8(8):11, 2008.

[29] Jennifer Gamble and Giseon Heo. Exploring uses of persistent homology for statistical analysis of landmark-based shape data. *Journal of Multivariate Analysis*, 101(9):2184 – 2199, 2010.

[30] Gunnar Carlsson, Tigran Ishkhanov, Vin de Silva, and Afra Zomorodian. On the local behavior of spaces of natural images. *Int. J. Comput. Vision*, 76(1):1–12, January 2008.

[31] Matthew Kahle and Elizabeth Meckes. Limit theorems for Betti numbers of random simplicial complexes. *Homology Homotopy Appl.*, 15(1):343–374, 2013.

[32] Sampo Niskanen and Patric Östergård. Cliquer. `http://users.tkk.fi/pat/cliquer.html`, 2002.