TOPOLOGICAL ANALYSIS OF THE CONNECTOME OF DIGITAL RECONSTRUCTIONS OF NEURAL MICROCIRCUITS

PAWEŁ DŁOTKO[*,1], KATHRYN HESS[*], RAN LEVI[*], MAX NOLTE[*], MICHAEL REIMANN, MARTINA SCOLAMIERO, KATHARINE TURNER, EILIF MULLER, AND HENRY MARKRAM

ABSTRACT. A recent publication provides the network graph for a neocortical microcircuit comprising 8 million connections between 31,000 neurons [7]. Since traditional graph-theoretical methods may not be sufficient to understand the immense complexity of such a biological network, we explored whether methods from algebraic topology could provide a new perspective on its structural and functional organization. Structural topological analysis revealed that directed graphs representing connectivity among neurons in the microcircuit deviated significantly from different varieties of randomized graph. In particular, the directed graphs contained in the order of 10^7 simplices groups of neurons with all-to-all directed connectivity. Some of these simplices contained up to 8 neurons, making them the most extreme neuronal clustering motif ever reported. Functional topological analysis of simulated neuronal activity in the microcircuit revealed novel spatio-temporal metrics that provide an effective classification of functional responses to qualitatively different stimuli. This study represents the first algebraic topological analysis of structural connectomics and connectomics-based spatio-temporal activity in a biologically realistic neural microcircuit. The methods used in the study show promise for more general applications in network science.

The Blue Brain Project (BBP) has recently generated the first draft digital reconstruction and simulation of a microcircuit of neurons in the neocortex of a two-week-old rat (Figure 1A) [7]. This reconstruction is made available through the Neocortical Microcircuit Portal (https://bbpnmc.epfl.ch) [11]. Based on sparse anatomical and physiological data for neurons and synapses and on a variety of biologically motivated organizing principles, the complete connectivity between neurons belonging to a neocortical microcircuit was digitally reconstructed – a microconnectome. The structural properties of the reconstruction have been extensively validated against independent data, and simulations of the reconstruction reproduced multiple in vitro and in vivo experiments without adjusting any parameter, further validating its biological accuracy.

In this article we apply methods from topology to the analysis of 42 variants of the digital reconstruction, grouped in six sets of seven microciruits each. The first five sets of microcircuits take into account biological variability in layer heights, proportions of cell types, and cell densities from five individual rats, while the sixth set is based on the average reconstruction across the five individuals. To form each

set of microcircuits, seven statistically varying instantiations of the microcircuit were reconstructed [12]. The 42 microcircuits are therefore all distinct, though the degree of resemblance within each set is higher than that between sets. The structural connectivity of each reconstructed microcircuit can be represented as a directed graph with approximately 3×10^4 vertices and 8×10^6 edges, while its functional connectivity can be represented as a time series of subgraphs formed by functionally effective connections.

Our topological analysis of the detailed structural and functional connectivity of these 42 neural microcircuits led to a number of surprising observations. Firstly, we found that the distribution of directed cliques (directed all-to-all connected subsets) of neurons by size is highly significantly different from both that in Erdős-Rényi random graphs with the same number of vertices and the same average connection probability and that in more sophisticated random graphs, constructed either by taking into account distance-dependent probabilities varying within and between cortical layers or morphological types of neurons, or according to Peters' Rule [9], [10] (Figure 1D). In particular, we found that directed cliques of up to eight neurons are highly prominent motifs in the reconstructed microcircuits: the average microcircuit incorporates approximately 10^8 3-cliques and 4-cliques, approximately 10^7 5-cliques, approximately 10^5 6-cliques, and approximately 10^3 7-cliques. Taking the alternating sum of the numbers of directed cliques of various sizes, we computed the *Euler characteristic (EC)* [5] of the 42 reconstructed microcircuits, obtaining in each case a value on the order of 10^7, indicating a preponderance of directed cliques consisting of an odd number of neurons (Figure 2).

Another topological metric that we considered in this analysis are the *Betti numbers* (SI, Supplementary Text, ST1.3) associated to a graph via its *directed flag complex* (Figure 1C). These are a sequence of natural numbers $\beta_0, \beta_1, \beta_2, ...$ that measure the higher-order organizational complexity of the network, detecting "cyclic" chains of intersecting directed cliques. For each graph considered here we determined its *homological dimension*, i.e., the maximum n such that $\beta_n \neq 0$. We showed that the reconstructed microcircuits have homological dimension 5 (Figure 2D), whereas the random graphs considered have homological dimension at most 4, strongly indicating that the microcircuits possess a higher degree of organizational complexity than the random graphs.

Topological methods also enabled us to distinguish functional responses to different input patterns fed into the microcircuit through thalamo-cortical connections. We ran simulations of neural activity in one of the reconstructed microcircuits during one second, over the course of which a given stimulus was applied every 50 ms (Figure 3). We then binned the output of the simulations by 5 ms timesteps and associated to each timestep a *transmission-response graph*, the vertices of which are all of the neurons in the microcircuit and the edges of which encode connections in the microcircuit whose activity in that time step leads to firing of the postsynaptic neuron (Figure 4). The size of the time bins and the precise rule for formation of the transmission-response graph for each time bin are biologically motivated, as explained in more detail in the Supplementary Methods section (SI, Supplementary Methods, SM1).

From the time series of transmission-response graphs for each of 20 trials of two different stimuli (called Circle and Point for geometric reasons (Figure 4A), we derived time series of two non-topological metrics (mean firing rate and number of
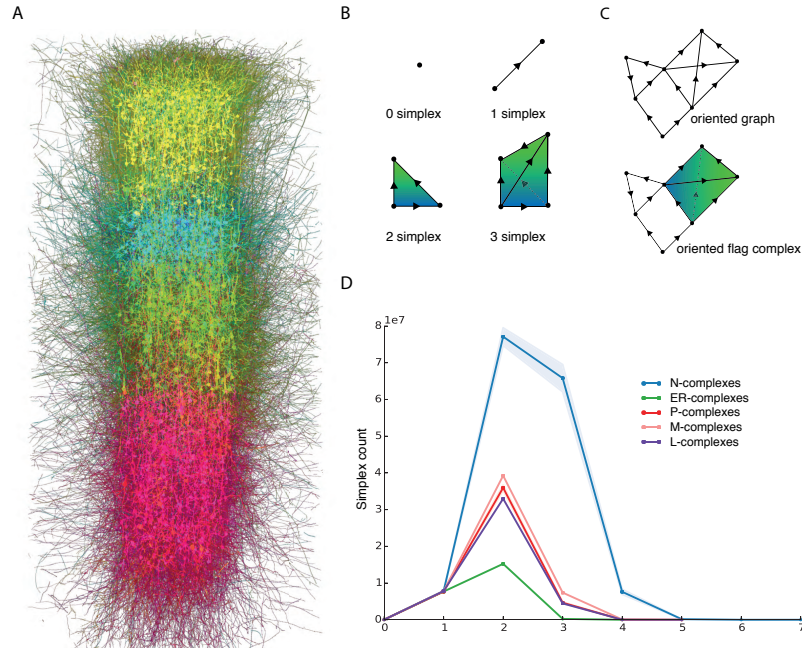
FIGURE 1. *(A)* A sparse visualization of the microcircuit (soma and dendrites only). Morphological types are color-coded, with m-types in the same layer having similar colors. *(B)* Examples of simplices in dimensions 0 through 3. *(C)* An example of a directed graph and its associated flag complex, in which there is one n-simplex for every directed $(n+1)$-clique in the graph. *(D)* A graph depicting the average number of simplices in each dimension for the flag complexes associated to the reconstructed microcircuit (N-complexes) and the four types of random graphs considered, each with the same number of vertices as the reconstructed microcircuit, where shading indicates standard deviation, which was very small for all except the N-complexes.

edges in the transmission-response graph) and five topological metrics (the number of 3-cliques, EC, β_0, β_1, and β_2) and applied a Gaussian Bayes classifier (SI, Supplementary Methods, SM2) to determine how successfully each of the metrics classified the 40 trials in the time bins corresponding to the first two stimulations and in the time bins immediately following those stimulations (Figure 5). In each of those crucial time bins, the metrics that were most successful at classification the number of 3-cliques (denoted 2D in the figure), β_2, and, in one case, the Euler characteristic (Figure 2A).

We expect the methods applied here will prove useful for studying networks in general.

A

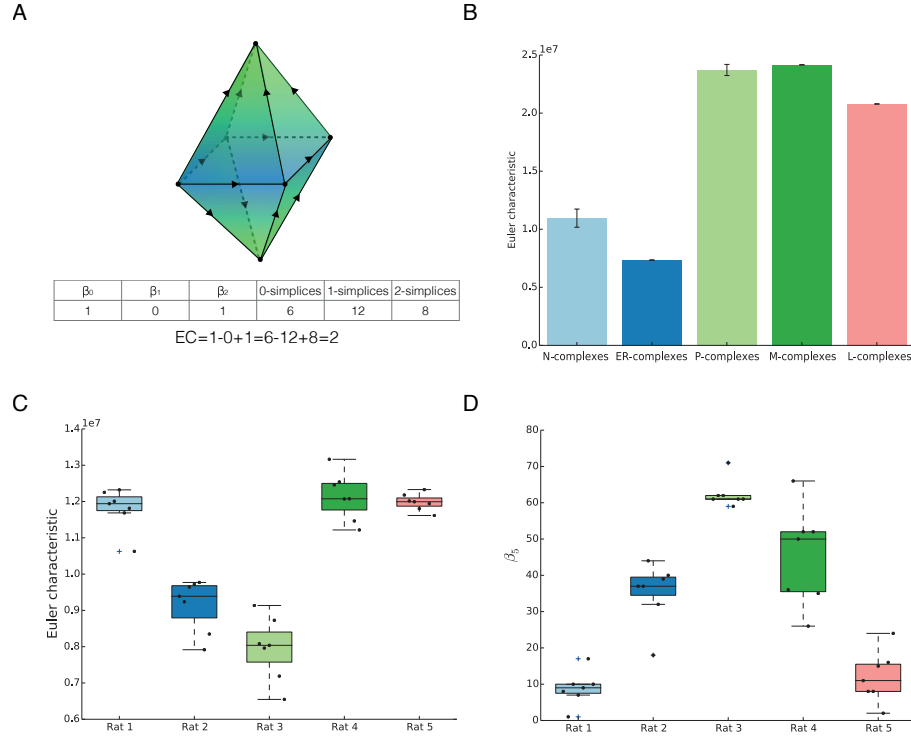| β_0	β_1	β_2	0-simplices	1-simplices	2-simplices
1	0	1	6	12	8

EC=1-0+1=6-12+8=2

B

C

D

FIGURE 2. *(A)* An oriented simplicial complex consisting of eight 2-simplices glued together along their 1-dimensional faces, together with a table of its Betti numbers and numbers of simplices in dimensions 0,1, and 2 and a computation illustrating that the Euler characteristic can be computed as the alternating sum of the Betti numbers or the simplex counts. *(B)* Graph depicting the average Euler characteristic of the reconstructed microcircuit (N-complexes) and of each of the types of random graph considered, where the whisker indicates standard deviation, which was very small, except for N-complexes and P-complexes. *(C)* Box-and-whisker plots depicting the Euler characteristics of 35 reconstructed microcircuits, seven for each individual rat. *(D)* Box-and-whisker plots depicting the 5th Betti number of 35 reconstructed microcircuits, seven for each individual rat.

1. STRUCTURAL TOPOLOGY

We computed the binary adjacency matrices of all 42 digitally reconstructed microcircuits and then generated the associated *directed flag complexes* (SI, Supplementary Text, ST1.2), which are oriented simplicial complexes encoding the connectivity of all orders of the underlying directed graph: to each directed n-clique (SI, Supplementary Text, ST1.2) in the underlying graph corresponds an oriented $(n-1)$-simplex in the flag complex, and the faces of a simplex correspond

to the directed subcliques of its associated directed clique (Figure 1 B and C). For each neuron in the microcircuit, there is a vertex in the underlying directed graph that is labelled with the unique *global identification number (GID)* of the neuron. The (j, k)-coefficient of the structural adjacency matrix is 1 if and only if there is a directed connection in the microcircuit from the neuron with GID j to the neuron with GID k. We refer to this adjacency matrix as the *structural matrix* of the microcircuit and to its associated directed flag complex as a *neocortical microcircuit complex* or *N-complex*.

Having computed each of the 42 N-complexes, we counted the simplices in each dimension. For comparison with non-biological matrices, we generated five Erdős-Rényi random graphs [4] of a comparable size (31,000 vertices) and connection probability 0.8%, the same as the average arising from the structural matrices of the microcircuits (SI, Supplementary Methods SM3.1). We refer to the associated directed flag complexes as *ER-complexes*.

To have a more biological control, we also generated 20 adjacency matrices, given by partly randomizing the structural matrix of one of the average microcircuits, taking into account its biologically meaningful division into six layers in 10 cases and into 55 morphological neuron types (m-types) [7] in 10 cases. The randomization was carried out so that the distance-dependent connection probability for all pairs of layers (respectively, pairs of m-types) was identical to that of the original matrix, i.e., for each pair of layers (respectively, m-types) the number of connections between them was the same as that of the original and for each $25\,\mu$m distance bin the number of connections was identical. The matrices are completely random otherwise (SI, Supplementary Methods SM3.2, SM3.3). We call the associated directed flag complexes *L-complexes* (respectively, *M-complexes*). Note that since each m-type is restricted to a fixed layer, the M-complex should retain more of the structure of the original N-complex than the L-complex. Our final and most biological control consisted in the generation of 10 connectivity matrices for 31,000 neurons according to Peters' Rule [9], [10] for which the associated directed flag complexes are called *P-complexes* (SI, Supplementary Methods SM3.4). Having carried out the computations for 10 control matrices out of each randomized set of 20, the very small variance in the results convinced us that no further computations should be needed.

The resulting distribution of simplices displayed highly consistent behavior among the N-complexes, all of which we computed, with a small variation among the samples arising from different rats. Note that Figure 1 represents the analysis only of the seven N-complexes arising from the average reconstruction because the randomizations are based on those microcircuits. The ER-complexes showed almost identical behavior among the different instances, as did the L-complexes, M-complexes, and P-complexes. On the other hand, the N-complexes exhibited remarkably different distributions from the various random complexes (Figure 1 D), with much greater numbers of simplices and simplices of significantly higher dimension. We computed the Euler characteristic of all N-complexes, as well as that of the various random complexes, obtaining large positive values in all cases, due to the predominance of even-dimensional (particularly 2-dimensional) simplices.

The Betti numbers (SI, Supplementary Text, ST1.2) of a simplicial complex provide a much finer and more sophisticated measure of its organizational complexity than the dimension-wise simplex count or the Euler characteristic. The n-th Betti
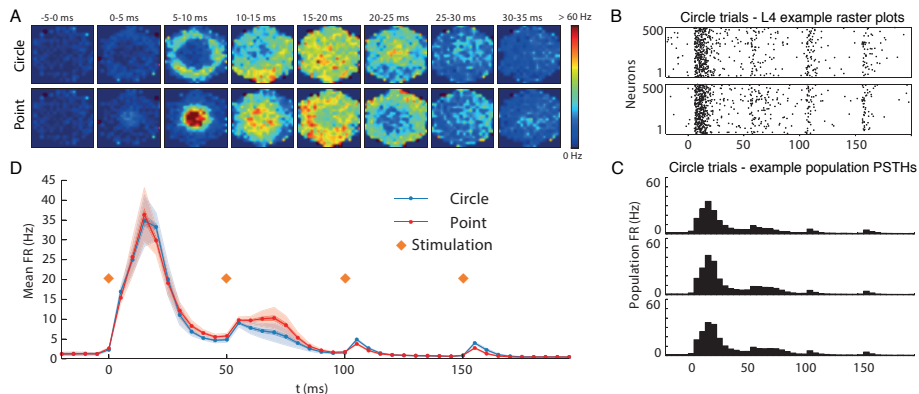
FIGURE 3. *(A)* Average firing rate (top-down projection) in the
stimulated microcircuit, plotted during the first 35 ms after the
first stimulation at t=0 ms in the Point vs. Circle experiment. *(B)*
Raster plots of the same 500 neurons randomly picked from layer
4, for two trials of the circle stimulus. *(C)* Population PSTH of all
neurons in the microcircuit for three trials of the Circle stimulus.
(D) Mean firing rate of the Circle and Point stimuli, between t and
$t + 5$ ms, where light shading indicates the standard deviation and
dark shading the error of the mean.

number, β_n, counts the number of chains of simplices intersecting along faces to
create an "n-dimensional hole" in the complex, which requires a certain degree of
organization among the simplices. On the other hand, computation of the Betti
numbers is much more expensive than that of the directed flag complex of a di-
rected graph or its Euler characteristic. In fact, the sheer size of the complexes we
considered here made it practically impossible to do so on a computer with 256 GB
of RAM. We succeeded in computing the highest nonzero Betti numbers of the N-
complexes, however, by restricting our attention to the 5-th and 6-th coskeleta (SI,
Supplementary Text, ST1.2). The top Betti number in all N-complexes appeared
in dimension 5, with β_5 varying between 1 and 80 (Figure 2D). By contrast, $\beta_n = 0$
for all $n > 3$ for all ER-complexes and P-complexes considered, while $\beta_n = 0$ for
all $n > 4$ for all L-complexes and M-complexes. Moreover β_4 varies between 0 and
6 for all L-complexes and M-complexes, so that these Betti numbers are almost
negligible.

2. FUNCTIONAL TOPOLOGY

We tested our methods on active microcircuits as well. In an experiment that we
call the *Point vs. Circle test*, we activated in a simulation the incoming thalamo-
cortical fibers of one of the average that the stimulated fibers formed first a point
shape, then a circle shape [7]. The size of the point shape was chosen such that the
average firing rate of the neurons was essentially the same as for the circle shape,
and in both cases the fibers were activated regularly and synchronously with a
frequency of 20 Hz for one second, similar to the whisker deflection approximation
in [7, Figure 17A]. We performed 20 trials of each stimulus (Figure 3). The trials of

each stimulus exhibit biological trial-to-trial variability in the neural response, due to the stochasticity of the synapse models and of some of the ion channel models. The aim of this experiment was to determine whether our topological methods were able to classify the two different stimuli, the point and the circle better than the firing rate, which is largely overlapping for the first two stimulations (see Figure 3D).

After a systematic analysis to determine the appropriate time bin size and conditions for probable spike transmission from one neuron to another (SI, Supplementary Methods, SM1.4), we divided the activity of the microcircuit into 5 ms time bins for 1 second after the initial stimulation and recorded for each $0 \leq n < 200$ a functional connectivity matrix $A(n)$ for the times between $5n$ ms and $5(n+1)$ ms. The (j, k)-coefficient of the binary matrix $A(n)$ is 1 if and only if the following three conditions are satisfied, where s_i^j denotes the time of the i-th spike of neuron j.

(1) The (j, k)-coefficient of the structural matrix is 1, i.e., there is a structural connection from the neuron with GID j to the neuron with GID k.
(2) There is some i such that $5n$ ms $\leq s_i^j < 5(n+1)$ ms, i.e., the neuron with GID j spikes in the n-th time bin.
(3) There is some l such that 0 ms $< s_l^k - s_i^j < 7.5$ ms, i.e., the neuron with GID k spikes after the neuron with GID j, within a 7.5 ms interval.

We call the matrices $A(n)$ *transmission-response matrices*, as it is reasonable to assume that the spiking of neuron k is influenced by the spiking of neuron j under conditions (1)–(3) above.

The goal of the *Point vs. Circle test* was to determine whether topological metrics, such as simplex counts, Betti numbers and Euler characteristic, could classify correctly two groups of stimuli of a similar nature and whether these metrics contain more information than the mean firing rate. In Figure 4C we provide plots of the time series of the average zeroth, first, and second Betti numbers, of the average numbers of 1- and 2-simplices, and of the average Euler characteristic for 20 trials of each stimulus. We applied a Gaussian Bayes classifier (SI, Supplementary Methods, SM2) to each metric in each time bin, to determine their success rate at classifying the various trials of the stimuli. To compare, we also classified the stimuli according to the mean firing rates. To allow for a fair comparison, we used three mean firing rates (between t to $t+5$, $t+5$ to $t+10$, and $t+10$ to $t+15$ ms) for the classification at each time step t, since the transmission-response edges for time step t are based on information from up to $t+12.5$ ms.

As illustrated by Figure 5 A, none of the metrics considered, topological or otherwise, succeeded very well at classifying the stimuli for times between 10 ms and 50 ms after the initial stimulation, which is not surprising given the strong similarity between the spatial propagation of activity of the two stimuli during this period (Figure 3). On the other hand, in the very first time bin, immediately after the initial stimulation, the 1- and 2-dimensional simplex counts and β_1 and β_2 all classify very well. In the second time bin the 2-dimensional simplex count and β_2 continue to classify very well, and the Euler characteristic classifies even better. Immediately after the second stimulation, from 50 ms to 55 ms after the initial stimulation, none of the metrics performs very well, but the 2-dimensional simplex count and β_2 still have the highest success rate. In the next time bin, from 55 ms to 60 ms after the initial stimulation, the 2-dimensional simplex count and β_2
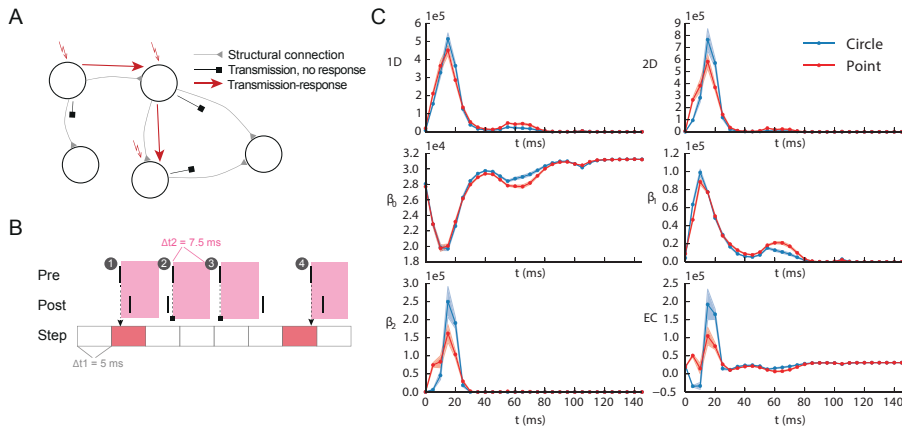
FIGURE 4. *(A)* Schematic representation of the transmission-response paradigm: there will be an edge from j to k in the graph associated to particular time bin if and only if there is a physical connection from neuron j to neuron k, neuron j fires in the time bin, and neuron k fires at most 7.5 ms after the firing of neuron j. Here, shading indicates the error of the mean.*(B)* Schematic representation of those firing patterns involving a presynaptic and a postsynaptic neuron that lead to an edge in the transmission-response graph, with a red block indicating successful transmission and a white block indicating lack of transmission. *(C)* Time series plots of the average value of the metrics 1D (number of 1-simplices), 2D (number of 2-simplices), β_0 (the zeroth Betti number, i.e., the number of connected components), β_1 (the first Betti number), β_2 (the second Betti number), and EC (the Euler characteristic) for the Circle and Point stimuli. Here, shading indicates the error of the mean.

again classify very well and are the only metrics to do so. In all of these cases, the topological metrics far outperform the metric based on firing rate.

3. DISCUSSION

We have introduced topological analysis of directed graphs encoding structural or functional connectivity of digital reconstructions of neural microcircuits. We showed in particular that these directed graphs differed significantly from random graphs of both Erdős-Rényi-type and types taking into account biologically constrained, distance-dependent connection probabilities. The topological analysis revealed not only the existence of high-dimensional simplices representing the most extreme form of circuit "motifs" - all-to-all connectivity within a set of neurons - that have so far been been reported for brain tissue, but also that there are a surprisingly huge number of these structures. We established moreover that topological methods effectively distinguish functional responses to distinct thalamic stimuli, introducing a new measure of the spatio-temporal activity responses generated by neural tissue. The results of our topological analysis of biologically realistic digital
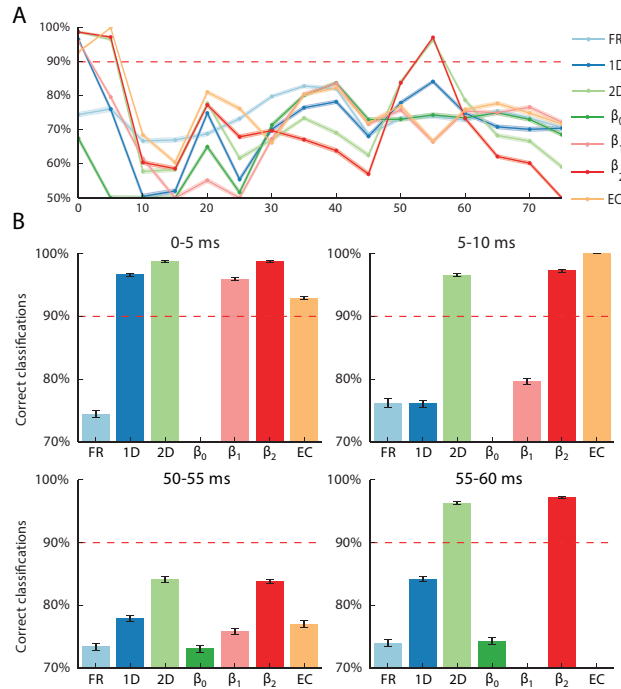
FIGURE 5. *(A)* Times series plot for the first 80 ms of the 40 trials of the percentage of correct classifications performed by a Gaussian Bayes classifier based on each of the metrics FR (sequences of mean firing rates over three consecutive time bins), 1D (number of 1-simplices), 2D (number of 2-simplices), β_0 (the zeroth Betti number, i.e., the number of connected components), β_1 (the first Betti number), β_2 (the second Betti number), and EC (the Euler characteristic). *(B)* Graphs depicting the percentage of correct classifications performed by a Gaussian Bayes classifier based on each of the metrics in four particularly important time bins: from 0 to 5 ms (immediately after the initial stimulation), from 5 to 10 ms, from 50 to 55 ms (the time bin immediately after the second stimulation), and from 55 to 60 ms.

reconstructions provide a convincing argument for considering topology as a useful mathematical tool for analyzing the structural and functional connectome of neural circuits.

Our results lead naturally to many new questions, most notably concerning the biological significance of the high-dimensional simplices and homology classes we have discovered in the digitally reconstructed neocortical microcircuits. We intend to explore these questions in future studies. In particular we hypothesize that the time series of different topological metrics could reveal an evolving spatio-temporal code that goes beyond either rate or timing information to one that incorporates the structural organization. Such metrics could yield a deeper understanding of how

the structural organization constrains emergent functional states. Age-dependent changes in such digital reconstructions may help reveal even more complex topological structures with development, and changes introduced by synaptic plasticity may reveal structures associated with learning and memory.

We expect the topological approach to studying directed graphs that we implement here will also prove useful in applications of network science outside of neuroscience, in the study of networks exhibiting intricate directed connectivity patterns, such as gene and protein networks, VLSI circuits, and electrical grids. The obvious utility of the directed flag complex in these applications may also encourage theorists to establish results analogous to those established by Kahle concerning Betti numbers of undirected flag complexes of random graphs [6].

4. Materials and methods

4.1. **Computation of flag complexes and their Betti numbers.** We represent the directed flag complex of a directed graph by a reference-based data structure, using vectors to store the references to the simplices in the simplicial complex. The required storage space grows linearly with the number of vertices and with the number of edges. A publicly available C++ implementation of the code will be available on http://neurotop.gforge.inria.fr/. All homology computations carried out for this paper were made with $\mathbb{F}_2$ coefficients, using the boundary matrix reduced by an algorithm from the PHAT [2] library. For further details, please see (SI, Supplementary Text, ST2).

4.2. **The Point vs. Circle experiment.** The stimulated reconstructed microcircuit is innervated by 310 VPM fibers, whose horizontal centers of innervation are evenly distributed over the microcircuit (one fiber per mini-column). It is therefore possible to activate the microcircuit with topographically different stimuli by selecting only a subset of these 310 fibers. Here we used two different stimuli, a point and a circle, which were calibrated by adjusting the respective number of fibers to evoke an overall similar mean firing rate (i.e., close enough to prevent clearly distinguishing between the two stimuli simply by the mean population firing rate). The microcircuit was stimulated by synchronous spikes, similar to the whisker deflection experiment described by Markram et al. (2015). The point stimulus consisted of synchronous spikes in the 46 neighboring fibers of the center of the microcircuit, whereas the circle stimulus involved 56 fibers near the periphery of the microcircuit. The stimulation was repeated every 50 ms, but only the firing rates after the first two stimulations (at 0 and 50 ms) are overlapping.

We used a Gaussian naïve Bayes classifier [8], where we performed 500 classification trials, randomly choosing 15 trials of each stimulus to be part of the training data, and five trials of each stimulus to be part of the test data. We then obtained the mean ratio of successfully classified test data points using 500 different training and test sets. The classification of the firing rate used the firing rates of three consecutive time bins, to make it a fairer comparison, since the edges may contain firing rate information of more than two time bins, over a range of 12.5 ms.

4.3. **Computation of transmission-response matrices.** Transmission-response matrices were calculated according to the specifications mentioned above, using a custom-written program in the Python programming language. It combined the

matrix of synaptic connections (structural matrix), constructed as part of the standard reconstruction process of the BBP, with the spiking output of a simulation run and user-defined values for time steps Δt_1 and Δt_2 (5 and 7.5 ms in our analyses). For further details, please see (SI, Supplementary Methods ,SM1).

4.4. **Gaussian Bayes classifiers.** The Gaussian Bayes classifier minimises the probability of misclassification under the assumption that the distributions are Gaussian. We randomly split the data into training and testing sets. Using the training set we model the distributions of the dot and circle classes by Gaussians $\mathcal{N}(\hat{\mu}_{\text{dot}}, \hat{\sigma}^2_{\text{dot}})$ and $\mathcal{N}(\hat{\mu}_{\text{circle}}, \hat{\sigma}^2_{\text{circle}})$ respectively. Assuming a uniform prior and Gaussian distributions, Bayes' theorem provides a classifier

$$\text{Class}(x) = \operatorname*{argmax}_{c \in \{\text{dot,circle}\}} \frac{1}{\sqrt{2\pi\hat{\sigma}^2_c}} \exp\left(\frac{-(x - \hat{\mu}_c)^2}{2\hat{\sigma}^2_c}\right).$$

5. Acknowledgments

References

[1] U. Bauer, M. Kerber, J. Reininghaus, *Clear and Compress: Computing Persistent Homology in Chunks*, *TopoInVis 2013*, in press.

[2] U. Bauer, M. Kerber, J. Reininghaus, Phat library, https://code.google.com/p/phat/.

[3] T. H. Cormen, C. E. Leiserson, R. L. Rivest and C. Stein, *Introduction to Algorithms*, *The MIT Press*, 2001.

[4] Erdős, P.; Rényi, A. *On random graphs*, I. Publ. Math. Debrecen 6 1959 290–297.

[5] A. Hatcher, *Algebraic Topology*, *Cambridge University Press* (Available Online), 2002.

[6] M. Kahle, *Sharp vanishing thresholds for cohomology of random flag complexes*, *Ann. of Math.* (2), **179** (2014), 1085–1107.

[7] H. Markram, et al., *Reconstruction and simulation of neocortical microcircuitry*, *Cell*, **163** (2015) no. 2, 456-492.

[8] Pedregosa, Fabian, et al., *Scikit-learn: Machine learning in Python*, The Journal of Machine Learning Research 12 (2011): 2825–2830.

[9] Peters, A., and Feldman, M.L. *The projection of the lateral geniculate nucleus to area 17 of the rat cerebral cortex. I. General description*, *J. Neurocytol.*, **5** (1976), 6384.

[10] Peters, A., Proskauer, C.C., Feldman, M.L., and Kimerer, L. *The projection of the lateral geniculate nucleus to area 17 of the rat cerebral cortex. V. Degenerating axon terminals synapsing with Golgi impregnated neurons*, *J. Neurocytol.*, **8** (1979), 331357.

[11] S. Ramaswamy, et al., *The neocortical microcircuit collaboration portal: a resource for rat somatosensory cortex*, *Frontiers in Neural Circuits*, **9** (2015), http://dx.doi.org/10.3389/fncir.2015.00044.

[12] M. Reimann, J. King, E. Muller, S. Ramaswamy, and H. Markram, *An algorithm to predict the connectome of neural microcircuits*, *Frontiers in Computational Neuroscience* (2015) 120, doi:10.3389/fncom.2015.00120.

[13] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, *Network Motifs: Simple Building Blocks of Complex Networks*, *Science*, 25 October 2002: 298 (5594), 824–827. [DOI:10.1126/science.298.5594.824]

SI Appendix

To accompany "Topological analysis of the connectome of digital reconstructions of neural microcircuits."

Paweł Dłotko, Kathryn Hess, Ran Levi, Max Nolte, Michael Reimann, Martina Scolamiero, Katharine Turner, Eilif Muller, Henry Markram

Contents

Supplementary Methods

SM1. Optimization of the parameters for the transmission-response matrices

The transmission-response matrices that allow us to analyze activity in an experiment (cf. the section on Functional Topology in the main body of the article) form a sequence depending on two parameters, Δt_1 and Δt_2. The number of matrices in the sequence is the duration of the experiment divided by Δt_1. In other words for a given experiment of duration T and fixed Δt_i, we obtain a sequence of matrices $S(\Delta t_1, \Delta t_2) = \{A(n) = A(n, \Delta t_1, \Delta t_2)\}_{n=1}^N$, where N is the integer value of $T/\Delta t_1$.

For fixed values of Δt_1 and Δt_2, the corresponding sequence $\{A(n)\}_{n=1}^N$ is obtained as follows. The spiking output of the simulation is first converted into lists of spike times, one for each neuron. Standard histogram methods, binning by Δt_1, are applied to each list to determine in which time steps a presynaptic neuron fired. For each time bin in which a particular neuron fired, the exact timing of its first spike in that bin is then compared to the full list of spike times of each neuron it innervates, to ascertain which of them had spiked at most Δt_2 ms after the presynaptic neuron. (Spiking of a pair of neurons within Δt_2 ms is ignored if they are not structurally connected.) For all pre-postsynaptic pairs satisfying this constraint on spike timing, the corresponding entry in the transmission-response matrix for that time step is set to 1 and all others to 0. More precisely, the (j, k)-coefficient of the binary transmission-response matrix $A(n)$ corresponding to the n-th time bin is 1 if and only if the following three conditions are satisfied, where s_i^j denotes the time of the i-th spike of neuron j.

(1) The (j, k)-coefficient of the structural matrix is 1, i.e., there is a structural connection from the neuron with GID j to the neuron with GID k, so that they form a pre-post synaptic pair.

(2) There is some i such that $n\Delta t_1$ ms $\leq s_i^j < (n + 1)\Delta t_1$ ms, i.e., the neuron with GID j spikes in the n-th time bin.

(3) There is some l such that 0 ms $< s_l^k - s_i^j < \Delta t_2$ ms, i.e., the neuron with GID k spikes after the neuron with GID j, within a Δt_2 ms interval.

Starting with firing data from spontaneous activity in the reconstructed microcircuit, we generated sequences of 20 transmission-response matrices for $\Delta t_i \in \{1, 2, 5, 10, 20, 50, 100\}$ ms, thus creating 49 such sequences corresponding to every possible choice of the pair $(\Delta t_1, \Delta t_2)$. We refer to each of these sequences as the *true transmission-response matrices* corresponding to the pair $(\Delta t_1, \Delta t_2)$.

In the rest of this section, we describe the procedure for optimizing the choice of the time intervals Δt_1 and Δt_2 so that the associated true transmission-response matrices best reflect the actual successful transmission of signals between neurons in the microcircuit.

SM1.1. Properties of the transmission-response matrix. The nonzero coefficients in a transmission-response matrix are a subset of those in the structural matrix. Due to the partly stochastic behavior of the *in silico* microcircuit, the subset will vary even for subsequent applications of the same stimulus. In fact, even an exact repetition of the same conditions will lead to different transmission-response matrices, if the random number generator is seeded differently. It follows that the generation of the transmission-response matrices for a given stimulus should be considered as a stochastic process. With the correct choice of the parameters Δt_i, the matrices should reflect how the microcircuit processes a stimulus and thus take into account parameters of neural processing, such as pre-post synaptic interaction.

To find parameters Δt_1 and Δt_2 that maximize the degree to which neural processing is captured by the transmission-response matrices, we first develop a stochastic model for synaptic firing that takes into account neural processing and that depends on Δt_1 and Δt_2. For the purpose of this analysis, we assume that the true transmission-response matrices are compatible with this model.

Based upon our model for synaptic firing, we formulate a simplified model that ignores neural processing. For this simplified model and for any choice of parameters Δt_1 and Δt_2, we explain how to obtain transmission-response matrices from actual firing data, by shuffling the firing data appropriately, then applying the algorithm for generating a transmission-response matrix of the previous section. Finally, for each choice of the parameters Δt_1 and Δt_2, we compare the true transmission-response matrices for spontaneous activity in the reconstructed microcircuit to those obtained by the simplified generation process. The parameters that we work with in the main body of the paper are the Δt_1 and Δt_2 that maximize the difference (measured by the ratio of the numbers of ones in the matrices) between the actual transmission-response matrices and those resulting from the simplified model.

SM1.2. Stochastic model with neural processing. Fix time intervals Δt_1 and Δt_2. Let $A = (a_{ij})$ denote the structural matrix of a reconstructed microcircuit, and let $A(n) = (a_{ij}^n)$ denote the transmission-response matrix of the n-th time bin, based on firing data from a trial of simulated activity in the microcircuit, for the given intervals Δt_1 and Δt_2. By Condition (1) above, if $a_{ij}^n = 1$ for any n, then $a_{ij} = 1$. It is reasonable to consider A to be static, at least over the time periods considered here.

We want to compute the probability that $a_{ij}^n = 1$, given that $a_{ij} = 1$, so we need to determine on which parameters and properties this probability depends. According to the definition of transmission-response matrices, a presynaptic and a postsynaptic spike are required for a_{ij}^n to be 1. To simplify the analysis somewhat, we assume that each neuron n_i has a *time-dependent, instantaneous firing rate* $F^i(t)$ that determines spiking probability at time t, i.e., spiking can be described as an inhomogeneous Poisson process. Under this assumption, the expected number $m_{\Delta t_1}^i(t_0)$ of spikes of neuron n_i in the interval $[t_0, t_0 + \Delta t_1]$ can be computed as

$$m_{\Delta t_1}^i(t_0) = \int_{t_0}^{t_0+\Delta t_1} F^i(u)du.$$

If $K^i_{\Delta t_1}(t_0)$ denotes the probability that neuron n_i spikes at least once in the interval $[t_0, t_0 + \Delta t_1]$, then

$$K^i_{\Delta t_1}(t_0) = 1 - \mathcal{P}\big(m^i_{\Delta t_1}(t_0)\big) = 1 - e^{-m^i_{\Delta t_1}(t_0)},$$

where $\mathcal{P}(\lambda)$ is the Poisson probability mass function with parameter λ at 0. (Recall that if X is a random variable that counts the number of spikes of neuron n_i in the interval $[t_0, t_0 + \Delta t_1]$, then $\mathcal{P}\big(m^i_{\Delta t_1}(t_0)\big)$ is the probability that $X = 0$.) If the change in $F^i(t)$ is slow compared to Δt_1, then $m^i_{\Delta t_1}(t) \approx F^i(t) \cdot \Delta t_1$. Moreover, $1 - \mathcal{P}(\lambda) \approx \lambda$ for small values of λ. For small enough Δt_1, the expected number $m^i_{\Delta t_1}(t_0)$ of spikes of neuron n_i will certainly be small, and change in $F^i(t)$ will be slow in compared to Δt_1, so that we may assume that

$$K^i_{\Delta t_1}(t_0) \approx F^i(t_0) \cdot \Delta t_1.$$

For the postsynaptic spike the situation is more complicated. As there is a causal relation between presynaptic and postsynaptic firing, mediated by synaptic transmission, we need to estimate the conditional probability of at least one postsynaptic spike, given that at least one presynaptic spike occured. Let n_i and n_j denote neurons such that $a_{ij} = 1$. Let $s_0 \in [t_0, t_0 + \Delta t_1]$ denote the time of the first presynaptic spike in this interval. Let $X^j_{\Delta t_2}(s_0)$ denote the random variable whose value is the number of times neuron n_j spiked in the time window $[s_0, s_0 + \Delta t_2]$. Let $Y^i_{\Delta t_1}(t_0)$ denote the random variable whose value is the number of times neuron n_i spiked in the time interval $[t_0, t_0 + \Delta t_1]$. We need to estimate the conditional probability

$$P\big(X^j_{\Delta t_2}(s_0) > 0 \,|\, Y^i_{\Delta t_1}(t_0) > 0\big).$$

The nature of this interaction is very intricate and depends on the identities of the presynaptic and postsynaptic neurons, the spiking history of the presynaptic neuron before s_0, and all other synaptic input the postsynaptic neuron received. It can be described as governed by some function G^{ij} modulating the spiking probability of the postsynaptic neuron n_j. This function takes as parameters the expected number of spikes of neuron n_j in the interval $[s_0, s_0 + \Delta t_2]$, the time t_0, and the "spiking history" of the presynaptic neuron n_i until s_0, which we write as a function $s^i_*(t)$ evaluated at s_0, giving rise to the expression

$$P\big(X^j_{\Delta t_2}(s_0) > 0 \,|\, Y^i_{\Delta t_1}(t_0) > 0\big) = 1 - e^{-G^{ij}(m^j_{\Delta t_2}(s_0), t_0, s^i_*(s_0))}.$$

Summarizing the analysis above, the following formula provides a good estimate of the probability that $a^n_{ij} = 1$ if $a_{ij} = 1$, for small enough Δt_1 and Δt_2, where s_0 denotes the time of the first presynaptic spike in the interval $[n\Delta t_1, (n+1)\Delta t_1]$ and $t_0 = n\Delta t_1$.

$$
(1) \quad
\begin{aligned}
P\left(a^n_{ij} = 1 | a_{ij} = 1\right) &= \left(1 - e^{-m^i_{\Delta t_1}(t_0)}\right) \cdot \left(1 - e^{-G^{ij}(m^j_{\Delta t_2}(s_0), t_0, s^i_*(s_0))}\right) \\
&\approx F^i(t_0) \cdot \Delta t_1 \cdot G^{ij}\big(F^j(s_0) \cdot \Delta t_2, t_0, s^i_*(s_0)\big).
\end{aligned}
$$

This conditional probability encodes not only the distinctive features of the structural connectivity (via a_{ij}) but also the potentially stimulus-dependant neuron-specific firing rates (via F^i and F^j) and their co-variation. Most crucially, it captures the stimulus-dependent functional modulation of postsynaptic firing by a presynaptic spike as well. We assume that the true transmission-reponse matrices capture the actual transmission of spikes according to the model of synaptic firing described by this formula.

SM.1.3. Null hypotheses: no neural processing. We introduce here a simplified model of synaptic spiking that is based upon formula [1] but that ignores pre-post synaptic interaction. We then explain how to obtain transmission-response matrices that correspond to this simplified model from firing data arising from simulated activity.

We begin by setting each G^{ij} to be the projection onto the first component, ignoring the pre-post synaptic interaction. After this simplification, the approximation obtained in the previous section now reads

$$P(a_{ij}^n = 1 | a_{ij} = 1) \approx F^i(t_0) \cdot F^j(s_0) \cdot \Delta t_1 \cdot \Delta t_2,$$

where s_0 denotes the time of the first presynaptic spike in the interval $[n\Delta t_1, (n+1)\Delta t_1]$ and $t_0 = n\Delta t_1$, as before. Since this drastic simplification neglects the central aspect of neural computation - pre-post synaptic interaction - it gives rise to control cases for each pair of parameters $(\Delta t_1, \Delta t_2)$ and each choice of firing rate functions $F^i(t)$. Comparison of the true transmission-response matrices for each pair of parameters to the corresponding control matrices for the same pair and a specific choice of the functions $F^i(t)$ will allow us to determine values for Δt_1 and Δt_2 for which the true transmission-response matrix optimally reflects neural processing.

We assume moreover that the individual firing rates consist of a neuron-dependent frequency that is up- or down-regulated by a global time series, i.e., that $F^i(t) = f(i) \cdot F(t)$, for some function $F(t)$ and some constant $f(i)$ for each neuron n_i. Transmission-response matrices corresponding to this simplified model for fixed Δt_1 and Δt_2, which we call *simplified transmission-response matrices*, can be generated by first shuffling all recorded spikes from simulated activity in the reconstructed microcircuit, while preserving both the number of spikes per neuron and per time bin, then applying the usual transmission-response matrix generation method.

SM.1.4. Optimization of parameters. The difference between the true transmission-response matrices and the control case described above is a consequence of the pre-post synaptic interaction. Comparison with the control case enables us therefore to measure how well that interaction is captured in the true transmission-response matrices. In particular, it is reasonable to optimize the parameters Δt_1 and Δt_2 so that the difference between the true transmission response matrices arising from actual simulation data and those arising in the control cases is maximized, as a maximal difference indicates that the effect of the pre-post synaptic interaction is captured optimally by the true transmission-response matrices.

The comparison between the true transmission-response matrices and the control cases was carried out by first producing 20 true transmission-response matrices and 20 simplified transmission-response matrices based on firing data obtained from spontaneous activity in the reconstructed microcircuit for every pair $(\Delta t_1, \Delta t_2)$, where $\Delta t_i \in \{1, 2, 5, 10, 20, 50, 100\}$ ms for $i = 1, 2$. The number of ones in each matrix was then computed and the average taken over each set of 20 matrices. Since no stimulus was applied to the microcircuit, the averages computed are meaningful, since the firing data should be fairly homogeneous across the time bins.

The average number of ones in the transmission-response matrix arising from simulated actitivity in the reconstructed microciruit, as a function of Δt_1 and Δt_2, is illustrated in Figure S1. Figure S2 shows the ratio of the average number of ones in the true transmission-response matrices to the average number of ones in the simplified transmission-response matrices, for various values of Δt_1 and Δt_2.

In all cases we find that the maximum lies between $\Delta t_2 = 5$ ms and $\Delta t_2 = 10$ ms, leading us to choose to work with $\Delta t_2 = 7.5$ ms. For Δt_1 we find a maximum at 50 ms, but we use $\Delta t_1 = 5$ ms (for which the maximum ratio is only slightly lower than for $\Delta t_1 = 50$ ms) instead to avoid more than one spike per neuron per bin.

SM2. GAUSSIAN BAYES CLASSIFIERS

Suppose there is a distribution ρ over $\mathbb{R} \times \{c_1, c_2, \ldots, c_k\}$, where $\{c_1, c_2, \ldots, c_k\}$ is a set of class labels. We can project ρ onto each of the coordinates to construct a real-valued random variable X and a class-label-valued random variable Y. We wish to build a classifier $C : \mathbb{R} \to \{c_1, c_2, \ldots, c_k\}$ which will, for any real number, choose the most likely class to which it might belong. That is,

$$C(x) = \underset{c \in \{c_1, c_2, \ldots, c_k\}}{\operatorname{argmax}} P(Y = c | X = x),$$

where $P(A|B)$ is the probability of A conditional on B and $\operatorname{argmax}_{a \in A} f(a)$ denotes the element $a \in A$ such that $f(a)$ is maximal. This element of A will in practice always be unique.

Bayes' theorem states that

$$P(Y = c | X = x)P(X = x) = P(X = x | Y = c)P(Y = c).$$

A Bayesian classifier picks the class with the highest conditional probability, which using Bayes' theorem is

$$C(x) = \underset{c \in \{c_1, c_2, \ldots c_k\}}{\operatorname{argmax}} \frac{P(X = x | Y = c)P(Y = c)}{P(X = x)}.$$

Usually ρ itself is unknown and must be infered from sample data. We then also assume some model distribution to estimate ρ from these samples. The Gaussian Bayes classifier is the Bayes' classifier under the assumption that the distribution of each separate class is Gaussian.

After calculating the means and variances of the sample data within each of the classes separately, we model their respective distributions by the Gaussians $N(\mu_{c_i}, \sigma_{c_i}^2)$. If $p(A)$ denotes the probability density function of A, then

$$\frac{P(X = x | Y = c)}{P}(Y = c)P(X = x) = \frac{p(X = x | Y = c)}{P}(Y = c)p(X = x)$$

$$= \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(\frac{-(x - \mu_c)^2}{2\sigma_c^2}\right) \frac{P(Y = c)}{p(X = x)}$$

A common situation, such as in our analysis, is a uniform prior. A uniform prior over $\{c_1, c_2, \ldots, c_k\}$ means $P(Y = c_i) = 1/k$ for all i. If we assume a uniform prior, then the factor $\frac{P(Y=c)}{p(X=x)}$ is common to all classes and thus does not affect which class achieves the maximum. Thus we get the formula

$$C(x) = \underset{c \in \{c_1, c_2, \ldots c_k\}}{\operatorname{argmax}} \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp\left(\frac{-(x - \mu_c)^2}{2\sigma_c^2}\right).$$

SM3. Randomization of connection matrices and other control cases

We created four types of random matrices of sizes and connection probabilities similar to the connectivity matrices of the BBP reconstruction.

SM3.1. Generation of Erdős-Rényi random matrices. For this basic control we first computed the overall connection probability in the reconstruction and found it to be 0.8%. We then generated random, binary square matrices of size 3.1×10^4, where 1's were placed at random off-diagonal in the matrix with probability 0.8%.

SM3.2. Randomization preserving the distance-dependent connectivity between layers. Input for this randomization method were the structural matrix and the matrix of pairwise soma distances, both generated as part of the standard BBP reconstruction process. The rows and columns of both matrices were first grouped into $N = 6$ groups according to the layer of the neuron they correspond to. This effectively partitioned both matrices into $N * N = 36$ submatrices each. For each pair of submatrices, the soma distances were grouped into bins of size $25\mu m$. Next, in the submatrix corresponding to each distance bin, we first replaced all 1's by 0's and then replaced randomly chosen 0's by 1's, so that the total number of 1's was preserved. Creation of autapses, i.e., a connection from a neuron to itself, was avoided by creating a separate bin for distances of $0\mu m$.

The result was a connection matrix with the same number of connections between each pair of layers and the same distance-dependent connection probability between pairs of layers, to within $25\mu m$, as the original matrix.

SM3.3. Randomization preserving the distance-dependent connectivity between m-types. This randomization method was identical to the preceding randomization, preserving connectivity between layers, except that the neurons were partitioned initially into $N = 55$ groups of morphological types instead of only six layers.

SM3.4. Generation of connection matrices according to Peters' Rule. For this control case, we started with a connection matrix that placed a connection not just where a synaptic connection was found in the reconstructed microcircuit, but between each pair of neurons whose arbors came within close proximity (closer than $3\mu m$). The resulting matrix had approximately 16 times more connections than the structural matrix. These connections were then pruned randomly with a uniform probability until the same number of connections as in the structural matrix was attained.

Supplementary Text

To accompany "Topological analysis of the connectome of digital reconstructions
of neural microcircuits."

Pawe l Dłotko, Kathryn Hess, Ran Levi, Max Nolte, Michael Reimann, Martina
Scolamiero, Katharine Turner, Eilif Muller, Henry Markram

ST1. THE TOPOLOGICAL TOOLBOX

Most of the mathematical methods we describe here are part of the basic toolbox
of algebraic topology, though perhaps not as well known in the directed variants
presented here. We give a brief account of these concepts for the benefit of the
non-expert, and refer to literature for the reader interested in further details.

We explain first how to associate to any directed graph a simplicial complex
known as its *directed flag complex*, then recall two types of important invariants of
simplicial complexes, which turn out to be very useful for analyzing the digitally
reconstructed microcircuits: the Euler characteristic and Betti numbers. We then
describe the data structures and algorithms that we implemented in order to con-
struct the flag complexes of the directed graphs representing the microcircuits and
to compute their Euler characteristics and Betti numbers.

ST1.1. Directed graphs. A *directed graph* $\mathcal{G}$ consists of a pair of finite sets
(V, E) and a function $\tau \colon E \to V \times V$. The elements of the set V are the *vertices*
of $\mathcal{G}$, the elements of E are the *edges* of $\mathcal{G}$, and the function τ associates with each
edge an ordered pair of vertices. The *direction* of an edge e with $\tau(e) = (v_1, v_2)$
is taken to be from $\tau_1(e) = v_1$, the *source vertex*, to $\tau_2(v) = v_2$, the *target vertex*.
The function τ is required to satisfy the following two conditions.

(1) For each $e \in E$, if $\tau(e) = (v_1, v_2)$, then $v_1 \neq v_2$, i.e., there are no loops in
the graph.
(2) The function τ is injective, i.e., for any pair of vertices (v_1, v_2), there is at
most one edge directed from v_1 to v_2.

A vertex $v \in \mathcal{G}$ is said to be a *sink* if $\tau_1(e) \neq v$ for all $e \in E$, and a *source* is if
$\tau_2(e) \neq v$ for all $e \in E$.

To compare two graphs, we require the following notion. A *morphism of directed
graphs* from a directed graph $\mathcal{G} = (V, E, \tau)$ to a directed graph $\mathcal{G}' = (V', E', \tau')$
consists of a pair of set maps $\alpha : V \to V'$ and $\beta : E \to E'$ such that β takes an edge
in $\mathcal{G}$ with source v_1 and target v_2 to an edge in $\mathcal{G}'$ with source $\alpha(v_1)$ and target
$\alpha(v_2)$, i.e., $\tau' \circ \beta = (\alpha, \alpha) \circ \tau$. Two graphs $\mathcal{G}$ and $\mathcal{G}'$ are *isomorphic* if there is
morphism of graphs $(\alpha, \beta) : \mathcal{G} \to \mathcal{G}'$ such that both α and β are bijections, which
we call an *isomorphism of directed graphs* (Figure S3).

A *path* in a directed graph $\mathcal{G}$ consists of a sequence of edges $(e_1, ..., e_n)$ such that
for all $1 \leq k < n$, the target of e_k is the source of e_{k+1}, i.e., $\tau_2(e_k) = \tau_1(e_{k+1})$.
The *length* of the path $(e_1, ..., e_n)$ is n, the number of edges of which the path is
composed. If, in addition, target of e_n is the source of e_1, i.e., $\tau_2(e_n) = \tau_1(e_1)$, then
$(e_1, ..., e_n)$ is an *oriented cycle*.

ST1.2. Simplicial complexes. An *abstract oriented simplicial complex* is a collection $\mathcal{S}$ of finite, ordered sets with the property that if $\sigma \in \mathcal{S}$, then every subset τ of σ is also a member of $\mathcal{S}$. A *subcomplex* of an abstract oriented simplicial complex is a sub-collection $\mathcal{S}' \subseteq \mathcal{S}$ that is itself an abstract oriented simplicial complex. Henceforth, we simplify terminology and usually refer to abstract oriented simplicial complexes merely as simplicial complexes.

The elements of a simplicial complex $\mathcal{S}$ are called its *simplices*. A simplicial complex is said to be *finite* if it has only finitely many simplices. If $\sigma \in \mathcal{S}$, we define the *dimension* of σ, denoted $\dim(\sigma)$, to be $|\sigma| - 1$, the cardinality of the set σ minus one. If σ is a simplex of dimension n, then we refer to σ as an *n-simplex* of $\mathcal{S}$. The set of all n-simplices of $\mathcal{S}$ is denoted $\mathcal{S}_n$. A simplex τ is said to be a *face* of σ if τ is a subset of σ of a strictly smaller cardinality. A *front face* of an n-simplex $\sigma = (v_0, ..., v_n)$ is a face $\tau = (v_0, ..., v_m)$ for some $m < n$. Similarly, a *back face* of σ is a face $\tau' = (v_i, ..., v_n)$ for some $0 < i < n$. If $\sigma = (v_0, ..., v_n) \in \mathcal{S}_n$, then the ith *face* of σ is the $(n-1)$-simplex σ^i obtained from σ by removing the vertex v_i.

A simplicial complex gives rise to a topological space by means of the construction known as *geometric realization*. In brief, one associates a point (a standard geometric 0-simplex) with each 0-simplex, a line segment (a standard geometric 1-simplex) with each 1-simplex, a filled-in triangle (a standard geometric 2-simplex) with each 2-simplex, etc., glued together along common faces. The intersection of two simplices in $\mathcal{S}$, neither of which is a face of the other, is a proper subset, and hence a face, of both of them. In the geometric realization this means that the geometric simplices that realize the abstract simplices intersect on common faces, and hence give rise to a well-defined geometric object. A geometric n-simplex is nothing but a $(n+1)$-clique, canonically realized as a geometric object. An n-simplex is said to be *oriented* if there is a linear ordering on its vertices. In this case the corresponding $(n+1)$-clique is said to be a *directed $(n+1)$-clique*.

If $\mathcal{S}$ is a simplicial complex, then the union $\mathcal{S}^{(n)} = \mathcal{S}_n \cup \cdots \cup \mathcal{S}_0$, which is called the *n-skeleton* of $\mathcal{S}$, is a subcomplex of $\mathcal{S}$. We say that $\mathcal{S}$ is *n-dimensional* if $\mathcal{S} = \mathcal{S}^{(n)}$, and n is minimal with this property. If $\mathcal{S}$ is n-dimensional, and $k \leq n$, then the collection $\mathcal{S}_k \cup \ldots \cup \mathcal{S}_n$ is not a subcomplex of $\mathcal{S}$ because it is not closed under taking subsets. However if one adds to that collection all the faces of all simplices in $\mathcal{S}_k \cup \ldots \cup \mathcal{S}_n$, one obtains a subcomplex of $\mathcal{S}$ called the *k-coskeleton* of $\mathcal{S}$, which we will denote by $\mathcal{S}_{(k)}$. The computational usefulness of coskeleta will become clear when we discuss homology computation (ST1.3).

Directed graphs give rise to abstract oriented simplicial complexes in a natural way. Let $\mathcal{G} = (V, E, \tau)$ be a directed graph. The *directed flag complex* associated to $\mathcal{G}$ is the abstract simplicial complex $\mathcal{S} = \mathcal{S}(\mathcal{G})$, with $\mathcal{S}_0 = V$ and whose n-simplices $\mathcal{S}_n$ for $n \geq 1$ are $(n+1)$-tuples $(v_0, \ldots, v_n)$, of vertices such that for each $0 \leq i < j \leq n$, there is an edge in $\mathcal{G}$ from v_i to v_j. Notice that because of the assumptions on τ, an n-simplex in $\mathcal{S}$ is characterised by the (ordered) sequence $(v_0, \ldots, v_n)$, but not by the underlying set of vertices. For instance (v_1, v_2, v_3) and (v_2, v_1, v_3) are distinct 2-simplices with the same set of vertices.

ST1.3. Homology, Betti numbers, and Euler characteristic. We now recall certain well known invariants of simplicial complexes arising in algebraic topology, which are preserved under a class of morphisms that is relevant in algebraic topology and that includes isomorphisms. These invariants serve to measure the

"complexity" of simplicial complexes, from various topological perspectives, leading us to refer to them as *metrics*.

Homology is an important algebraic invariant of topological spaces. In this paper we use only *mod-2 simplicial homology*, computationally the simplest variant of homology, which is why we choose to work with it in applications, though other types of simplicial homology may provide deeper information. We do not give a complete account of homology here, but rather an elementary description of what it is and its basic properties.

Let $\mathbb{F}_2$ denote the field of two elements, which we denote by 0 and 1. Let $\mathcal{S}$ be a finite simplicial complex. Define the *chain complex* $C_*(\mathcal{S}, \mathbb{F}_2)$ to be the sequence $\{C_n = C_n(\mathcal{S}, \mathbb{F}_2)\}_{n \geq 0}$, such that C_n is the $\mathbb{F}_2$-vector space whose basis elements are the n-simplices $\sigma \in \mathcal{S}_n$, for each $n \geq 0$. In other words, the elements of C_n are formal linear combinations of n-simplices in $\mathcal{S}$ with coefficients in $\mathbb{F}_2$. For each $n \geq 0$, there is a linear transformation called a *differential*

$$\partial_n \colon C_{n+1} \to C_n$$

defined by $\partial_n(\sigma) = \sigma^0 + \sigma^1 + \cdots + \sigma^n$ for every n-simplex σ, where σ^i is the i-th face of σ, as defined above. Having defined ∂_n on the basis, one extends the definition linearly to the entire vector space C_n.

The n-th Betti number $\beta_n(\mathcal{S})$ of a simplicial complex $\mathcal{S}$ is the $\mathbb{F}_2$-vector space dimension of its n-th mod 2 *homology group*, which is defined by

$$H_n(\mathcal{S}, \mathbb{F}_2) = \mathrm{Ker}(\partial_{n-1}) / \mathrm{Im}(\partial_n).$$

Computing the Betti numbers is conceptually very easy. Let $|\mathcal{S}_n|$ denote the number of n-simplices in the simplicial complex $\mathcal{S}$. If one encodes the differential ∂_n as a $\left(|\mathcal{S}_n| \times |\mathcal{S}_{n+1}|\right)$-matrix D_n with coefficients in $\mathbb{F}_2$, then one can easily compute its *nullity*, $\mathrm{null}(\partial_n)$, and its *rank*, $\mathrm{rk}(\partial_n)$, which are the $\mathbb{F}_2$-dimensions of the null-space and the column space of D_n, respectively. The *Betti numbers* of $\mathcal{S}$ are then a sequence of natural numbers defined by

$$\beta_0(\mathcal{S}) = \dim_{\mathbb{F}_2}(C_0) - \mathrm{rk}(\partial_0), \quad \text{and} \quad \beta_n(\mathcal{S}) = \mathrm{null}(\partial_{n-1}) - \mathrm{rk}(\partial_n).$$

The n-the Betti number β_n counts the number of "n-dimensional holes" in the geometric realization of $\mathcal{S}$. When $\mathcal{S} = \mathcal{S}(\mathcal{G})$ is the directed flag complex of a directed graph $\mathcal{G}$, both the simplices of $\mathcal{S}$ and these "n-dimensional holes" can be regarded as particularly important "metamotifs" [13] in the graph $\mathcal{G}$.

It is easy to show that the n-th Betti number of a simplicial complex $\mathcal{S}$ is equal to that of its $(n-1)$-st coskeleton $\mathcal{S}_{(n-1)}$, i.e., $\beta_n(\mathcal{S}) = \beta_n(\mathcal{S}_{(n-1)})$, for all n. This observation turns out to be computationally very useful, since there is no need to store the simplices of dimension less than $n - 1$ that are not faces of higher dimensional simplices in order to compute $\beta_n(\mathcal{S})$. In this paper it was exactly this trick that allowed us to compute the top dimensional homology of the 42 N-complexes we worked with.

Homology actually encodes far more information than what is intimated here, which can potentially be used for analyzing networks, but for the purposes of this article the description above will suffice.

If $\mathcal{S}$ is a simplicial complex and $|\mathcal{S}_n|$ denotes the cardinality of the set of n-simplices in $\mathcal{S}$, then the Euler characteristic of $\mathcal{S}$ is defined to be

$$\chi(\mathcal{S}) = \sum_{n \geq 0} (-1)^n |\mathcal{S}_n|.$$

There is a well known, close relationship between Euler characterstic and Betti numbers [5], which is expressed as follows. If $\{\beta_n\}_{n\geq 0}$ is the sequence of Betti numbers for $\mathcal{S}$, then

$$\chi(\mathcal{S}) = \sum_{n\geq 0} (-1)^n \beta_n(\mathcal{S}).$$

See Figure 2A for a specific example.

ST1.4. Hasse Diagrams. A *Hasse diagram*, otherwise known as a directed acyclic graph, is a directed graph $\mathcal{H} = (V, E, \tau)$ with no oriented cycles. Hasse diagrams can be used to encode various combinatorial, geometric, and topological structures, such as posets and cubical complexes. Below we explain in detail how Hasse diagrams encode simplicial complexes. We include this discussion here because our computational algorithm (Algorithm 1) is based on this idea.

A Hasse diagram $\mathcal{H}$ is said to be *stratified* if for each $v \in V$, every path from v to any sink has the same length. Thus in a stratified Hasse diagram the vertices are naturally partitioned into disjoint strata, where every directed path from a vertex in the k-th stratum V_k to any sink is of length k. In particular, the 0-th stratrum V_0 is the set of sinks of $\mathcal{H}$. Moreover, for all $e \in E$, there exists $k > 0$ such that $\tau_1(e) \in V_k$ and $\tau_2(e) \in V_{k-1}$. Note that if $\mathcal{H}$ and $\mathcal{H}'$ are isomorphic Hasse diagrams, and $\mathcal{H}$ is stratified, then so is $\mathcal{H}'$.

An *orientation* ς on a Hasse diagram $\mathcal{H}$ consists of a linear ordering $<_{\varsigma,v}$ of the set E_v of edges with source v, for every vertex v of $\mathcal{H}$. If $\mathcal{H} = (V, E, \tau)$ and $\mathcal{H}' = (V', E', \tau')$ are Hasse digrams equipped with orientations ς and ς', respectively, then a *morphism of oriented Hasse diagrams* from $(\mathcal{H}, \varsigma)$ to $(\mathcal{H}', \varsigma')$ is a morphism of directed graphs $(\alpha, \beta) : \mathcal{H} \to \mathcal{H}'$ such that for every $v \in V$, the restriction of β to a set map $E_v \to E_{\alpha(v)}$ preserves the orientation, i.e, if $e <_{\varsigma,v} e'$ for some $e, e' \in E_v$, then $\beta(e) <_{\varsigma',\alpha(v)} \beta(e')$. A morphism (α, β) of oriented Hasse diagrams is an *isomorphism* if α and β are bijections. A stratified Hasse diagram equipped with an orientation is called *admissible*.

Vertices in the k-th stratum of a stratified Hasse diagram $\mathcal{H}$ are said to be *of level k*. If $k < n$, and v, u are vertices of levels k and n respectively, then we say that v is a *face* of u if there is a path in $\mathcal{H}$ from u to v. If $\mathcal{H}$ is also oriented and therefore admissible, and there is a path $(e_1, ..., e_{n-k})$ from u to v such that $e_i = \min E_{\tau_1(e_i)}$ for all $1 \leq i \leq n - k$, we say that v is a *front face* of u. Similarly, v is a *back face* of u if there is a path $(e_1, ..., e_{n-k})$ from u to v such that $e_i = \max E_{\tau_1(e_i)}$ for all $1 \leq i \leq n - k$. We let Face(u) denote the set of all faces of u and Face(v)$_k$ the set of those that are of level k, while Front(u) and Back(u) denote its sets of front and back faces, respectively. See Figure S4 for an illustration of the concepts introduced above.

Example 1. *If $\mathcal{G} = (V, E, \tau)$ is a directed graph, then $\mathcal{G}$ can be equivalently represented by an admissible Hasse diagram with level 0 vertices V, level 1 vertices E, and directed edges from each $e \in E$ to its source and target. The ordering on the edges in the Hasse diagram is determined by the orientation of each edge e in $\mathcal{G}$.*

Every simplicial complex $\mathcal{S}$ gives rise to an admissible Hasse diagram $\mathcal{H}_\mathcal{S}$ as follows. The level d vertices of $\mathcal{H}_\mathcal{S}$ are the d-simplices of $\mathcal{S}$. There is a directed edge from each d-simplex to each of its $(d-1)$-faces. The stratification on $\mathcal{H}_\mathcal{S}$ is thus given by dimension, and the orientation is given by the natural ordering of the faces of a simplex from front to back. See Figure S5.

The *Euler characteristic* of a stratified Hasse diagram $\mathcal{H} = (V, E, \tau)$ is defined to be the integer

$$\chi(\mathcal{H}) = \sum_{k \geq 0} (-1)^k |V_k|.$$

It is easy to see that isomorphic stratified Hasse diagrams have the same Euler characteristic. It is also straight forward to show that if $\mathcal{H}$ is a stratified Hasse diagram associated to a simplicial complex $\mathcal{S}$, then the Euler characteristic of $\mathcal{H}$ coincides with that of $\mathcal{S}$.

ST2. Data structures and algorithms

In this section we describe our basic data structures and provide a detailed overview of the algorithm that constructs the directed flag complex associated to a directed graph. We also indicate briefly how our homology computations were performed. A publicly available C++ implementation of the code will be available on http://neurotop.gforge.inria.fr/.

ST2.1. Data structures. We represent an admissible Hasse diagram $\mathcal{H}$ corresponding to the directed flag complex of a directed graph $\mathcal{G} = (V, E, \tau)$ by a reference-based data structure, using vectors to store the references to the vertices of the diagram. Each vertex $v \in \mathcal{H}$ stores the following information.

(1) $\mathrm{Ver}(v)$: A vector of the vertices of $\mathcal{G}$ determining the simplex of the flag complex to which v corresponds.
(2) $\mathrm{Tar}(v)$: A vector of references to the vertices that are targets of edges with source v.
(3) $\mathrm{Src}(v)$: A vector of references to the vertices that are sources of edges with target v.

The admissible Hasse diagram $\mathcal{H}$ is thus represented by an ordered set of d vectors, where d is the maximal level in $\mathcal{H}$, and where the i-th vector contains the references to all level i vertices.

Let S_{int} denote the size of integer data types, and for a given graph $\mathcal{G} = (V, E, \tau)$, let $|V|$ and $|E|$ denote the cardinalities of the corresponding sets. Each edge of the Hasse diagram is stored in two vertices of the diagram. If each reference requires S_{int} storage, then we require $O(|E| \cdot S_{\mathrm{int}})$ space to store all references. In addition, each vertex stores the vector of vertices in V of the simplex in the flag complex of $\mathcal{G}$ to which it corresponds, which requires an additional $O(S_{int} \cdot d)$ of space per vertex. The total size of a Hasse diagram is thus bounded by $O((S_{\mathrm{int}} \cdot d) \cdot |V| + |E| \cdot S_{\mathrm{int}})$. In particular, the required storage space grows linearly with the number of vertices and with the number of edges. For our complexity analysis below we assume that accessing any vertex, using Tar or Src, takes $O(1)$ time.

ST2.2. Creation of the directed flag complex associated to a directed graph. We describe our algorithm that creates a directed simplicial complex given a directed graph $\mathcal{G}$. The output is a Hasse diagram $\mathcal{H}$, stored as the data structure described above. The identifier $\mathrm{Ver}(v)$ of a vertex v in $\mathcal{H}$, corresponding to a simplex σ in the directed flag complex, is the vector of vertices in $\mathcal{G}$ that represents σ.

For every level $n \geq 1$ vertex v in $\mathcal{H}$ such that $\mathrm{Ver}(v) = [v_0, \ldots, v_n]$, the algorithm additionally records a vector U_v of references to level 0 vertices u satisfying the following properties:

Algorithm 1 Directed flag complex generation.

Input: A directed graph $\mathcal{G} = (V, E, \tau)$.
Output: A Hasse diagram $\mathcal{H}$ representing the directed flag complex associated to $\mathcal{G}$.
 1: Convert $\mathcal{G}$ to level 0 and level 1 vertices of $\mathcal{H}$ (cf. Example 1).
 2: **for** every level 1 vertex $e \in \mathcal{H}$ **do**
 3:    **if** exist e_1, e_2 such that $\tau_1(e_1) = \tau_1(e)$, $\tau_1(e_2) = \tau_2(e)$ and $\tau_2(e_1) = \tau_2(e_2) = u$ **then**
 4:       Add u to U_e;
 5: $dim = 2$;
 6: **repeat**
 7:    $next\_level\_nodes$ – empty vector of references to nodes;
 8:    **for** top–level vertex $e \in \mathcal{H}$ **do**
 9:       **for** Every $u \in U_e$ **do**
10:          Create a node t of a Hasse diagram;
11:          $\text{Ver}(t) = [\text{Ver}(e), u]$;
12:          $U_t = U_e$;
13:          Add e to $\text{Tar}(t)$;
14:          Add t to $\text{Src}(e)$;
15:          **for** Every $bd \in \text{Tar}(e)$ **do**
16:             **for** Every $cbd \in \text{Src}(bd)$ **do**
17:                **if** $u \in \text{Ver}(cbd)$ **then**
18:                   Add cbd to $\text{Tar}(t)$;
19:                   Add t to $\text{Src}(cbd)$;
20:                   $U_t = U_t \cap U_{cbd}$;
21:          Add t to $next\_level\_nodes$;
22:    Add $next\_level\_nodes$ to $\mathcal{H}$;
23:    $dim = dim + 1$;
24: **until** $next\_level\_nodes = \emptyset$
25: Return $\mathcal{H}$;

---

   (1) $u \neq v_i$ for all $0 \leq i \leq n$, and
   (2) for every $u \in U_v$ and every $0 \leq i \leq n$, there exists an edge in $\mathcal{G}$ from v_i to u.

Finally, we assume that the graph $\mathcal{G}$ itself is given as an admissible Hasse diagram, as described in Example 1. Under these assumptions Algorithm 1 below is used to create the directed flag complex associated to $\mathcal{G}$.

ST2.3. Discussion of Algorithm 1. At the start of the algorithm (Line 1) only levels 0 and 1 of the Hasse diagram $\mathcal{H}$, which are the same as those of the Hasse diagram representation of $\mathcal{G}$ itself, have been created (cf. Example 1). The *for* loop in the line 2 initialises the creation of the vectors U_v for level 1 vertices. For every level 1 vertex e, the vector U_e stores the references to all the 0-simplices that, together with e, will form a level 2 vertex t. The construction of level 2 vertices in $\mathcal{H}$ is performed during the first iteration of the *repeat-until* loop starting in Line 6.

We analyze the generation of level 2 vertices as a generic case, since the arguments may clearly be generalised to higher levels. The *if* condition in Line 3 ensures that the vertex u will be the terminal vertex of the 2-dimensional simplex corresponding to the level 2 vertex t, created in the first iteration of the *repeat-until* loop (Line 6). Moreover the level 1 vertex e will correspond to a front face of the

2-simplex associated to t. Therefore, the ordering of $\mathrm{Ver}(e)$ can be extended to ordering of $\mathrm{Ver}(t)$, as in Line 11. Thus all level 2 vertices corresponding to 2-simplices in the directed flag complex of $\mathcal{G}$ will be created by the algorithm. Also, since every simplex has a unique 1-dimensional front face, every 2-simplex will be created only once by this process.

Notice also that the *if* condition in Line 3 ensures that only triangles in $\mathcal{G}$ consisting of three edges oriented as (v_1, v_2), (v_2, v_3), and (v_1, v_3) will give rise to level 2 vertices in $\mathcal{H}$. It follows by induction that the analogous condition on orientations is then automatically satisfied for simplices of dimension greater than 2. To see this, fix $n \geq 2$, and suppose that all simplices of dimension less than or equal to n have the desired property. Fix an n-simplex $S = [v_0, \ldots, v_n]$ and $u \in U_S$. By definition of the set U_S, there is an edge from v_i to u for every $i \in \{0, \ldots, n\}$. Note that $u \in U_{S'}$ for any $S' \in \mathrm{Tar}(S)$. The previous iteration of the *repeat-until* loop (Line 6) created an oriented simplex from S' together with u, of which u is the last vertex. Since the ordering of elements in S' is a restriction of the ordering of elements in S, the ordering of a $n+1$ dimensional simplex $[v_0, \ldots, v_n, u]$ restricted to any face yields the orientation of that face. It follows that Algorithm 1 does indeed construct a directed flag complex.

We now discuss the termination of Algorithm 1. If a level n vertex v is a face of a level $(n+1)$ vertex w, then the last vertex u in $\mathrm{Ver}(w)$ is not present in $\mathrm{Ver}(v)$, but is listed in U_v. From Lines 12 and 21 of the algorithm it is clear that $U_w \subset U_v$ and moreover that $u \notin U_v$. The cardinalities of the vectors $U_{(-)}$ are therefore decreasing for the newly created vertices. More precisely, for a vertex t and its faces s_i, there exist i such that $|U_t| \lneq |U_{s_i}|$. Level $n+1$ vertices are created only if there exist a level n vertex t such that $U_t \neq \emptyset$. Since the cardinality of the $U_{(-)}$ decreases with each iteration of the repeat loop, the algorithm will terminate.

We remark finally that the size of the directed flag complex corresponding to a given directed graph $\mathcal{G}$ may be exponential in the size of $\mathcal{G}$. In that case, the process of creation of a complex is usually stopped at some fixed dimension n. The time complexity of Algorithm 1 is proportional to the size of the output complex $\mathcal{H}$, multiplied by maximal level of a vertex in $\mathcal{H}$ (due to the target-source search performed in Line 15) of the algorithm.

ST2.4. Homology and Betti numbers. All homology computations carried out for this paper were made with $\mathbb{F}_2$ coefficients, using the boundary matrix reduced by an algorithm from the PHAT [2] library.
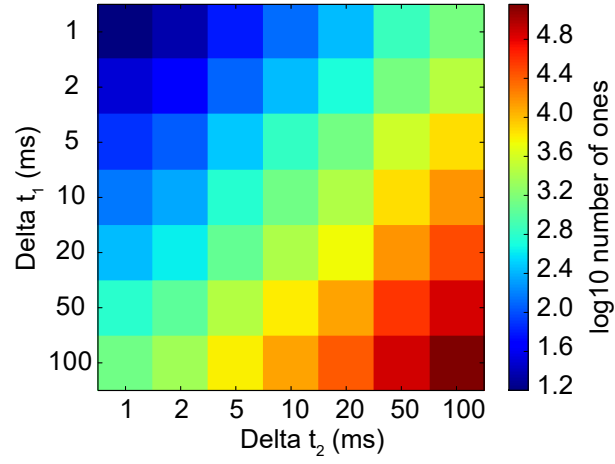
6. Supplementary Figures

Figure S1. Average number of ones in the true transmission-response matrices for different pairs of parameters $(\Delta t_1, \Delta t_2)$ in a simulation of spontaneous, in-vivo-like activity (Ca 1.2)
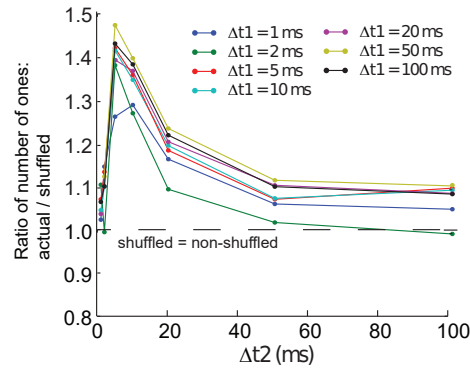
FIGURE   S2. Comparing   randomized   and   non-randomized
transmission-response matrices: average number of ones in a true
transmission-response (t-r) matrix divided by the average number
of ones obtained when the recorded spikes were randomized
before calculating the t-r matrix. Matrices were calculated from
simulated spontaneous, ongoing activity with different values for
Δt_1 (in different colors) and Δt_2 (along the x-axis). For each
pair $(\Delta t_1, \Delta t_2)$, matrices for 20 time steps were calculated, and
the mean ratio is shown. Spikes were randomized by shuffling the
identities of the firing neurons, thus conserving the number of
spikes in any given time step and the total number of spikes fired
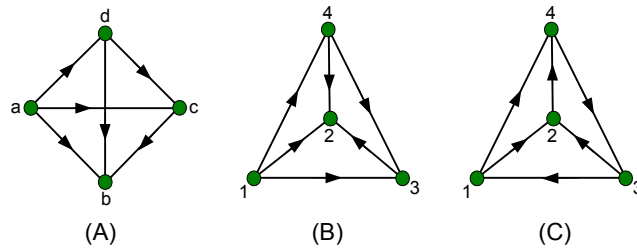by each neuron.

FIGURE S3. (A-C) Examples of directed graphs. Graphs (A) and
(B) are isomorphic, where the isomorphism is given by the map
sending vertex a to 1, b to 2, c to 3, and d to 4. Graphs (A) and
(B) are not isomorphic to graph (C). Vertex b in graph (A) is a
sink, vertex a in the same graph is a source. Graph (C) has no
sources or sinks, which explains the lack of isomorphism to graphs
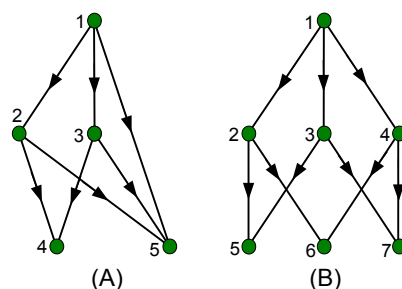(A) and (B).

FIGURE S4. (A) A Hasse diagram that is not stratified, due to the edge from the vertex 1 to 5. (B) A stratified Hasse diagram, where vertices 5, 6, and 7 are the vertices of level 0, vertices 2, 3, and 4 are of level 1, and vertex 1 is of level 2. This is also an admissible Hasse diagram, where the outgoing edges are ordered from left to right. Vertex 2 is a front face of vertex 1, while vertex 3 is neither a front nor a back face of a vertex 1, and vertex 4 is back face of a vertex 1.
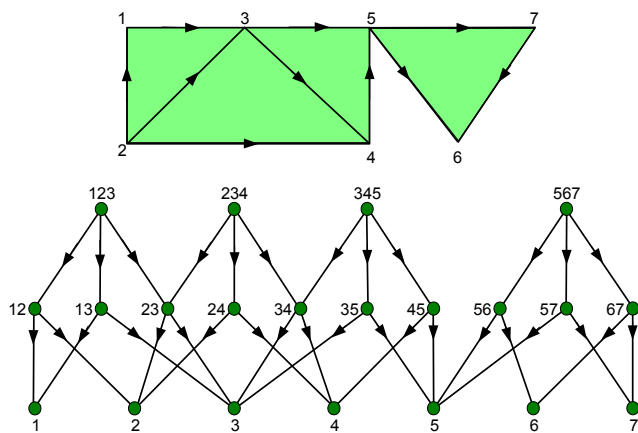

FIGURE S5. Top: The geometric realization of a simplicial complex consisting of seven 0-simplices (labeled 1,...,7), ten 1-simplices, and four 2-simplices. The orientation on the edges is denoted by arrows, i.e., the tail of an arrow is its source vertex, while the head of an arrow is its target. Bottom: The Hasse diagram corresponding to the simplicial complex above. Level k vertices correspond to k-simplices of the complex and are labeled by the ordered sets of vertices that constitute the corresponding simplex. Note that, e.g., vertex 23 is a back face of a vertex 123 and a front face of a vertex 234.

Geometrica, Inria, Saclay, France

Laboratory for Topology and Neuroscience, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Institute of Mathematics, University of Aberdeen, Aberdeen, UK

Blue Brain Project, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Blue Brain Project, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Laboratory for Topology and Neuroscience, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Laboratory for Topology and Neuroscience, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Blue Brain Project, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland

Blue Brain Project, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland