bioRxiv preprint doi: https://doi.org/10.1101/2023.05.08.539868; this version posted June 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Revealing and reshaping attractor dynamics in large networks of cortical neurons

Chen Beer[1,2], Omri Barak[2,3*]

1 Andrew and Erna Viterbi Faculty of Electrical and Computer Engineering, Technion Israel Institute of Technology
2 Network Biology Research Laboratories, Technion Israel Institute of Technology
3 Rappaport Faculty of Medicine, Technion Israel Institute of Technology

* omri.barak@gmail.com

Abstract

Attractors play a key role in a wide range of processes including learning and memory. Due to recent innovations in recording methods, there is increasing evidence for the existence of attractor dynamics in the brain. Yet, our understanding of how these attractors emerge or disappear in a biological system is lacking.

By following the spontaneous network bursts of cultured cortical networks, we are able to define a vocabulary of spatiotemporal patterns and show that they function as discrete attractors in the network dynamics. We show that electrically stimulating specific attractors eliminates them from the spontaneous vocabulary, while they are still robustly evoked by the electrical stimulation. This seemingly paradoxical finding can be explained by a Hebbian-like strengthening of specific pathways into the attractors, at the expense of weakening non-evoked pathways into the same attractors. We verify this hypothesis and provide a mechanistic explanation for the underlying changes supporting this effect.

Author summary

There are many hints that could evoke the same memory. There are many chains of evidence that could lead to the same decision. The mathematical object describing such dynamics is called an attractor, and is believed to be the neural basis for many cognitive phenomena. In this study, we aimed to deepen our understanding of the existence and plasticity of attractors in the dynamics of a biological neural network. We explored the spontaneous activity of cultured neural networks and identified a set of patterns that function as discrete attractors in the network dynamics. To understand how these attractors evolve, we stimulated the network to repeatedly visit some of them. Surprisingly, we observed that the stimulated patterns became less common in the spontaneous activity, while still being reliably evoked by the stimulation. This paradoxical finding was explained by the strengthening of specific pathways leading to these attractors, alongside the weakening of other pathways. These findings provide valuable insights into the mechanisms underlying attractor plasticity in biological neural networks.

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.08.539868; this version posted June 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Introduction

Attractors are important elements in many cognitive processes such as memory formation and decision-making. These attractors are considered to arise from the dynamics of neuronal networks in the brain, which allow for the emergence of stable states that can persist over time. For instance, head-direction circuits need to integrate body motion over time, consistent with continuous attractor dynamics (1; 2). Working memory of discrete (3) or continuous (4) information was hypothesized to be supported by attractors (5). Decision-making can be interpreted as convergence to a discrete set of attractors (6), and many other examples exist (7). Nevertheless, despite their key role in brain function, the mechanisms underlying the generation of such attractors and their evolution over time remain largely unknown.

To address this challenge, we focus on the relationship between spontaneous and evoked activity (8). Attractors, as the name implies, attract neural activity from nearby starting points into a common trajectory. This set of initial conditions is known as a basin of attraction. If attractor dynamics are relevant for behavior, one would expect external stimuli to lead neural activity into one of these basins. Similarly, it is reasonable to expect spontaneous activity to occasionally land into one of the basins, and hence result in the activation of attractors. In line with these expectations, there have been reports of spontaneous reactivations that are similar to evoked activity (9; 10; 11).

We studied this question in a more controlled setting – using in-vitro cultured cortical neurons. These networks can sustain both spontaneous (12) and evoked (13) activity, and allow continuous monitoring over many hours. Furthermore, it was shown that structured stimulation can lead to learning in such networks (14).

In this paper, we show that the spontaneous activity of in-vitro cortical networks contains a vocabulary of spatiotemporal patterns that act as discrete transient attractors. Discreteness is manifested by the finite number of such patterns that repeat over time. We show that nearby initial conditions lead to the same pattern, consistent with basins of attraction. These attractors are transient, as these network bursts are of limited duration, and the network relaxes to a quiescent state following each burst. Furthermore, we demonstrate that specific localized stimulation can generate robust evoked responses from this vocabulary of attractors. We also show that prolonged stimulation of these specific attractors leads to their elimination from the spontaneous vocabulary, while still being robustly evoked by the stimulation.

This work provides the first direct evidence for the plasticity of multiple attractors in a biological neural network. In addition, the plasticity principles described in the paper improve our understanding of how attractors in a biological system evolve. This study sheds light on the mechanisms underlying attractor dynamics in the brain and offers a new perspective on how they can be manipulated.

Results

To study attractor dynamics, we use extracellular recordings of mature networks of cultured cortical neurons (18-21 DIV, see methods). Electrical activity is recorded from an array of 120 electrodes on which the neurons are plated (Fig 1A). Throughout the following sections, we will demonstrate the results using one example experiment, and show statistics across all experiments. Further details regarding all experiments are in the methods section.

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.08.539868; this version posted June 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Spontaneous vocabulary as attractor dynamics

One of the main characteristics of the activity of cultured neuronal networks is the presence of spontaneous synchronized bursts, in which a large fraction of the neurons fire almost simultaneously within a few hundred milliseconds (Fig 1B). We follow the spontaneous activity of matured cultured cortical neuronal networks, focusing on these bursting events. We define these events based on the overall activity across all electrodes, beginning with a threshold-crossing, and ending 100 msec later. A burst can be described as a spatiotemporal pattern in the high-dimensional space of the neural activity (Fig 1B, heat-map). For visualization purposes, we also project these events to a natural two-dimensional space (Fig 1A) – the physical location of the activity's center of mass (Fig 1B, rightmost plot). We noticed that each network has its own repertoire of such spatiotemporal patterns – a finite set of network bursts that repeat many times spontaneously (Fig 2C).

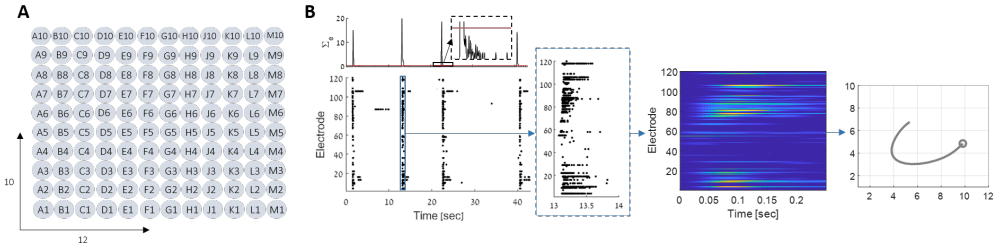

Fig 1. Burst extraction. (A) Electrodes layout in the MEA (10x12, 120 in total). See methods for more details. **(B)** The activity of all electrodes (bottom-left) is summed into a time series (top-left). A threshold is used to define a burst. For each burst, we bin and smooth the data to get a continuous time series for each of the electrodes (heatmap, see methods). For visualization only we use the center of mass representation (rightmost plot, circle denotes initial state. See methods).

Looking more closely at these bursts reveals attractor-like dynamics: Similar initial states lead to similar bursts. Specifically, we define the initial condition of a burst as the spatial activity pattern at the moment of threshold crossing. Examining all burst pairs allows us to look at the joint distribution of the similarity of initial conditions and the similarity of the overall burst patterns. We find that this distribution is bimodal, allowing us to define a threshold on the similarity of initial conditions that will lead to similar overall bursts (Fig 2A, note that θ is a network-specific threshold). Conversely, we see that most pairs of bursts are dissimilar – indicating the presence of more than one attractor. Further support to the attractor dynamics is given by the convergence of trajectories over time. If we consider all pairs of highly-correlated trajectories (above θ) and compute their instantaneous correlation, we see that the variability between them decreases over time (Fig 2B). We conclude that bursts can be described as distinct attractors, each with its basin of attraction.

Note that attractors in dynamical systems describe areas of phase space to which activity converges, and does not leave. In contrast, the bursts we describe are transient events. Nevertheless, we can think of them as attractors of the dynamical system before burst initiation. Once the burst is established, the dynamics change (probably due to adaptation), and the attractor destabilizes. Alternatively, one can consider a single global attractor – the quiescent state. The basin of attraction, however, is highly structured. Each of the bursts is a specific pathway within this basin, that is separated from the others. We are interested in the attraction phase of these dynamics, and not in the relaxation from them, and will thus refer to these events as discrete transient attractors.

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.08.539868; this version posted June 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

The correlation measure used here is one way out of many possibilities to characterize similarities between bursts. We use this measure and two others (See Fig 9 in methods) to construct a graph whose edges characterize similarity between pairs of bursts. We then use spectral clustering methods to create a vocabulary of spatiotemporal patterns, which act as attractors in the network dynamics. The resulting patterns for one network are shown as center-of-mass trajectories (Fig 2C). These two-dimensional projections do not capture the full phase space of neural activity. To provide another view of neural activity, we use a non-linear dimensionality reduction method (UMAP) to visualize all network bursts in a single projection (Fig 2D). This visualization emphasizes the existence of distinct pathways in the network dynamics, each corresponding to a cluster from the network vocabulary.

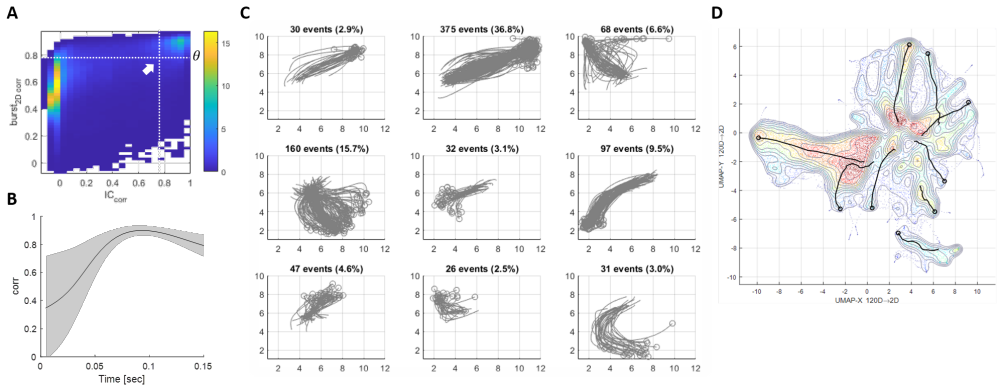

Fig 2. Attractor dynamics. (A) Similar initial states of bursts lead to similar bursts (white arrow). For every pair of bursts, we measure the correlation between their initial states and the overall spatiotemporal correlation between the full bursts. White dashed lines denote the thresholds used for clustering. θ is the 2D correlation threshold used as a similarity threshold between bursts. Initial states are defined as the first 5msec of a burst, after the threshold crossing. **(B)** Dynamics of convergence. For every pair of similar bursts (spatiotemporal correlation ¿θ), we measure the spatial correlation at every point in time. The black line denotes the mean and the grey shade represents the standard deviation (not SEM). Note the diminishing variability with time, indicating convergence from variable initial states. **(C)** Dynamics-based clustering (see methods). In this example, the vocabulary contains 9 main clusters, explaining about 85% of the spontaneous bursts. Each subtitle contains the number of bursts (events) in each cluster and the percentage out of all spontaneous bursts. Each center of mass trajectory is a single burst, the circles denote the initial state of each burst. **(D)** UMAP embedding representing all the 1017 spontaneous bursts recorded during 4 hours of activity. Contour denote the density of neural states. The black trajectories represent the median trajectory of each cluster in the spontaneous vocabulary.

Evoked responses

We showed the existence of attractors using spontaneous activity. The motivation to study attractors, however, stems from evoked activity. Attractors have been suggested to support memory of stimuli, to maintain a decision until it is carried out, or to support other computations related to evoked activity. Previous studies in-vivo showed conflicting accounts on the relationship between spontaneous and evoked activity (10; 11). We explored this question in our controlled settings. Namely, we asked whether spontaneous and evoked activity reside in the same dynamical landscape.

To this end, we divided the MEA into 20 stimulation sites: sets of 6 adjacent

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.08.539868; this version posted June 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

electrodes (with no overlap) that span the entire 2D space (Fig 3A). We then injected a voltage stimulation via these 6 electrodes simultaneously, for all the 20 sites, one after the other, with a 10 seconds delay. We repeated this sequence of 20 stimuli for 30 cycles and analyzed all the resulting evoked responses.

Some of the stimulation sites generated a robust response (site 17, Fig 3B, bottom), while other sites did not (site 10, top). In order to quantify the robustness of a response to each site we calculated the pairwise 2D correlations between all its responses (Fig 3C). In the case of a robust response, we can ask whether it is part of the spontaneous vocabulary of the network or whether it represents an entirely different dynamics. For the robust response of site 17 shown in Fig 3B, we see that the correlation within different repetitions of the evoked response is as strong as the correlation between the evoked responses and one of the spontaneous attractors (cluster 5, Fig 3D). We can quantify the similarity between evoked and spontaneous activity by counting the number of spontaneous clusters required to explain most of the evoked repetitions from a given site (Fig 3E). If this number is small, it suggests that the external stimulus brought the network into the basin of attraction of a few clusters. Overall, we see that robust responses are mostly taken from the spontaneous vocabulary of the network (Fig 4).

It is important to note that comparing evoked responses to spontaneous bursts is not trivial. The signal recorded from the stimulating electrodes in the evoked responses is about an order of magnitude higher than all other electrodes. Therefore, when comparing a given evoked response to a spontaneous burst we exclude all 6 stimulating electrodes and only then calculate the correlation.

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.08.539868; this version posted June 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

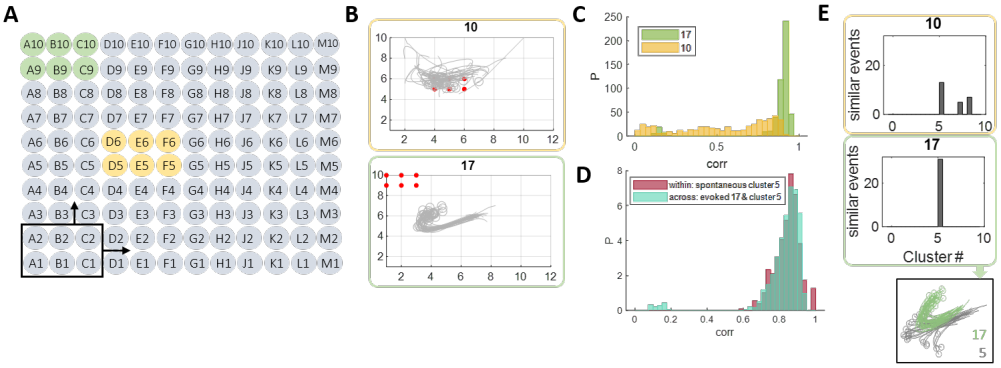

Fig 3. Evoked responses. **(A)** Electrode arrangement in the MEA (120 in total). Each stimulation site consists of 6 adjacent electrodes (black frame). There are 20 such stimulation sites with no overlapping electrodes. The set of 6 electrodes colored in green represents stimulation site number 17, and the set in yellow refers to site number 10. Probe stimulation: We stimulated all sites, one after the other, 30 times each. The time between stimuli is 10 seconds. This lasts 1.75 hours in total. **(B)** Visualization of the evoked responses to sites 10 and 17: center of mass representation for all of the 30 responses to each of these sites. The red dots denote the stimulating electrodes in each case. **(C)** Robustness: Pair-wise correlation values within each of the 2 sites. It is clear that the network response to stimulation at site 17 is much more robust and coherent compared to 10. **(D)** Existence: Comparing the evoked responses to the spontaneous vocabulary. Here we show the pair-wise correlation values within cluster 5 in the spontaneous activity and the pair-wise correlation values between the evoked responses to site 17 and the spontaneous bursts in cluster 5. They overlap almost completely, meaning that the evoked responses to 17 are indeed part of the spontaneous vocabulary of the network. **(E)** Existence in terms of spontaneous vocabulary: Which spontaneous clusters explain the 30 evoked responses for each of the 2 sites? In the case of site 10 – there is no specific cluster, also – a large part of the responses is not explained by any of the clusters. In the case of 17, cluster number 5 explains all of the evoked responses. The grey frame shows center of mass trajectories of the evoked responses to 17 (green) together with the spontaneous bursts in cluster 5 (grey).

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.08.539868; this version posted June 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

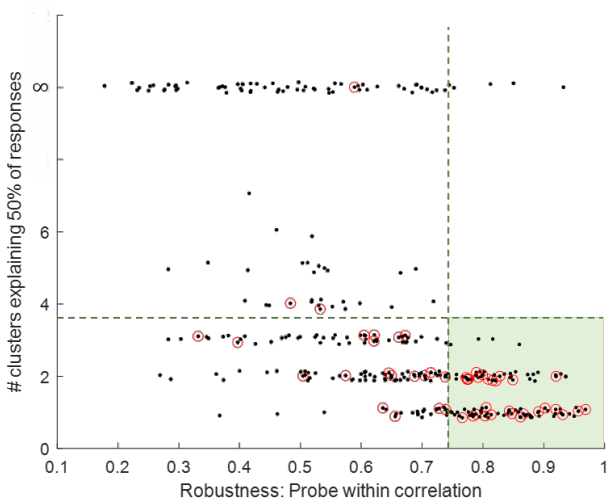

Fig 4. Existence as a function of robustness The x-axis denotes the median pair-wise correlation values for each probe response. The y-axis represents a measure for existence in the spontaneous vocabulary: the number of clusters are required to explain 50% of the responses. Small random noise is added to the number of clusters to aid in visualization. This plot shows data from 16 different experiments (see methods for more details). In general, the more robust the response is fewer clusters are required to explain it. Red circles denote the ones used for stimulation. The responses in the green area are used for further analysis.

In other words, the stimulation sites which generated a robust response can be used as switches to control the dynamics of the network – we can now "force" the network to visit specific areas in the dynamical space. This raises the following questions: What will happen to the network's evoked response to this stimulation? What will happen to the spontaneous dynamics? Will the spontaneous vocabulary change? What will happen to the stimulated attractors in comparison to the non-stimulated ones?

Strengthening and weakening specific pathways

In order to answer these questions we use the following protocol: We record the spontaneous activity of the network for four hours (during which the dynamics is stable), then we probe the network via 20 stimulation sites (Fig 3A). We then pick 3 stimulation sites to which the network responded in the most robust way and stimulate via those 3 sites for 10 hours (Fig 5). Finally, we record the spontaneous activity of the network again for an additional four hours.

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.08.539868; this version posted June 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

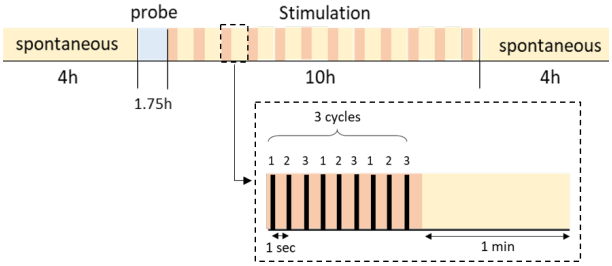

Fig 5. Experiment protocol. Each experiment starts with a 4-hour recording of spontaneous activity. We then probe the system in 20 different stimulation sites, one after the other, and analyze the evoked responses to each of the 20 sites. We choose the 3 stimulation sites which generated the most robust and distinct evoked responses and start a 10-hour stimulation period in which we alternate between stimulating these 3 sites and recording spontaneous activity (see inset). Following the 10-hour stimulation, we record the spontaneous activity for another 4 hours. Control experiments in which the 10-hour stimulation period had no stimulation but only spontaneous activity recordings, were also done.

We expected this protocol to strengthen the relevant attractors, in a Hebbian manner. Namely, the evoked responses will be more robust, and the corresponding spontaneous patterns will be more present in the spontaneous activity. Surprisingly, we observed two opposite effects: the spontaneous activity linked to stimulation weakened, while the evoked responses did become more robust.

To quantify the changes in the spontaneous vocabulary, we asked whether the stimulated patterns that were part of the spontaneous vocabulary before stimulation are still present afterwards. For instance, we can correlate the evoked responses to site 17 mentioned above (Fig 3E) to all 1017 spontaneous events that occurred before stimulation. The histogram in Fig 6A (blue) shows a large peak in high correlation values, consistent with the fact that this pattern is part of the vocabulary. Repeating the same analysis, but this time comparing to the 2092 spontaneous events from the period after the stimulation, results in a very different distribution (Fig 6A, purple). We can see that the stimulated attractor almost disappeared from the spontaneous vocabulary.

We quantified the changes in the existence of patterns using the cumulative probability distribution, exemplified in Fig 6B for the two distributions mentioned above. Intuitively, we care about changes in high values of correlation – as these indicate spontaneous events that are similar to the pattern of interest. The actual correlation values differ between networks, which is why we use a network-specific threshold as reference (θ, See Fig 2A). Using this threshold, we can calculate the change in the existence of high-correlation patterns – $\Delta CDF(\theta)$ (Fig 6B).

Is this change due to our stimulation or simply a result of drift over time? We repeated this analysis for 11 networks with stimulated patterns, and for 5 networks without stimulation. Importantly, for these 5 control networks, we also chose 3 robust patterns but simply did not stimulate them. There is some arbitrariness in our definition of the threshold θ, and we therefore scanned a short range of values relative to this threshold $\Delta CDF(\alpha\theta)$ (Fig 6C). We see that the stimulated patterns tend to disappear from the spontaneous vocabulary after stimulation (ΔCDF is negative), while the mean effect in the control experiments is roughly zero (Fig 6C).

The difference shown in Fig 6C could stem from two different effects – a larger drift in the spontaneous activity due to stimulation of the network, and a specific drift of the stimulated vs. the non-stimulated patterns within the stimulated networks. To dissociate the two, we now only considered the stimulated networks. For each network, we chose the clusters that were robustly evoked by stimulation (Fig 4, see methods) and

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.08.539868; this version posted June 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

calculated $\Delta CDF(\alpha\theta)$. We additionally chose the same number of non-stimulated clusters (see methods) for each network and repeated the same analysis. Networks are expected to differ not only in their correlation threshold, but also in their baseline drift rates. We therefore z-scored the ΔCDF values within each network before combining them across networks (Fig 6D). We can see that ΔCDF is negative, indicating an overall drift. For correlated patterns ($\alpha > 0.9$), we also see a trend towards larger drift in the stimulated clusters ($p = 0.08$ for $\alpha = 1$).

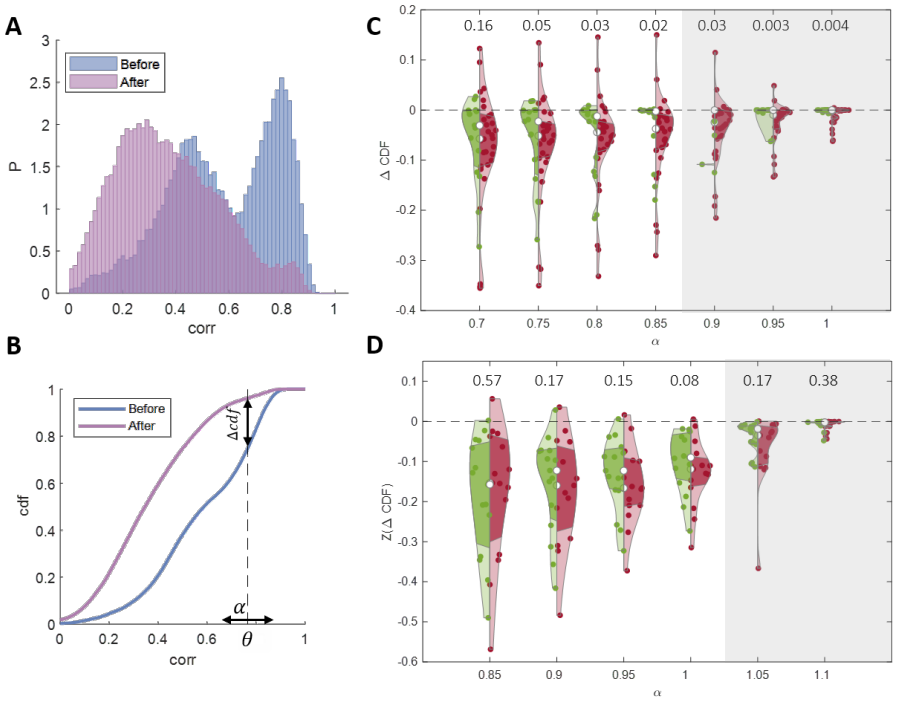

Fig 6. Changes in spontaneous activity. (A) 2D correlation values between all spontaneous events and the evoked responses to site 17, before (blue, high correlations) and after (purple, low correlations). **(B)** CDFs of the 2 distributions shown in (A). We quantify the effect by calculating the difference at $\alpha\theta$, were θ is the similarity threshold of the network and α is a value in the range [0.71.1]. **(C)** Existence of effect – stimulation vs. control. Statistics across 11 stimulation experiments and 5 control experiments. The violins represent $\Delta CDF(\alpha\theta)$ values in stimulation experiments (red) and in control experiments (green) for a range of α values (the fraction of θ at which ΔCDF was calculated). The numbers above each pair of violins represent the p-value of the hypothesis that the effect in the stimulation experiments is larger than in the control experiments. In the grey area ΔCDF was zero for some of the data points (the CDFs reached 1 for both before and after). **(D)** Specificity of the effect – measuring the effect in the spontaneous vocabulary. The violins represent the ΔCDF values for the stimulated clusters (red) and for the non-stimulated clusters (green). The numbers above each pair of violins represent the p-value of the hypothesis that the effect in the stimulated clusters is larger than in the non-stimulated clusters.

One simple possible explanation for this effect is that the stimulated pathways were "damaged" such that the network is no longer able to generate these patterns. Analyzing the evoked responses, however, shows the opposite is true. Throughout the 10-hour stimulation, not only that the network continues to generate these evoked responses, but they also get more robust with time. This can be visually appreciated by

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.08.539868; this version posted June 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

looking at the center-of-mass projections of one evoked response (site 17, Fig 7A), in which later responses are more tightly concentrated in space. We quantify this effect by measuring the variance between events across all networks (Fig 7B-C). In other words, these stimulated pathways remained accessible via stimulation but almost unreachable spontaneously. One can say that there is now a new association between the stimulated patterns and the specific stimulation that generates them.

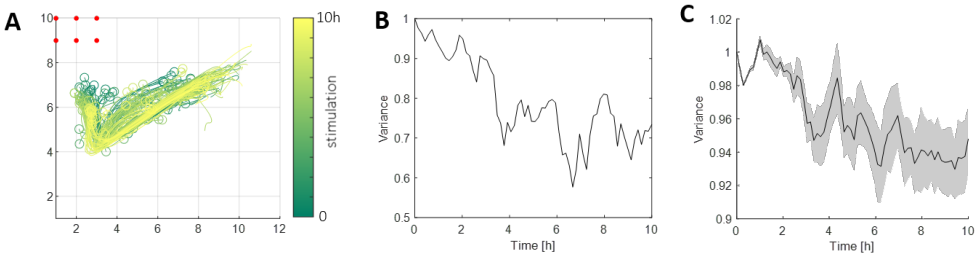

Fig 7. Evoked responses become more robust. (A) Center of mass trajectories of the responses to site 17 throughout stimulation (color: 0 (green) to 10 (yellow) hours). The 6 red dots denote the stimulating electrodes. **(B)** Variability between the responses in (A) throughout the 10-hour stimulation period. The variance is calculated in windows of 30 minutes throughout stimulation. In each window, we measure the deviation from the window's mean response (Euclidean distance). The plot is normalized by the variance at the first window. **(C)** Mean and variance of the variability between the responses for each stimulation site (statistics across all 11 experiments).

Mechanism

These effects raise many interesting questions – What causes this phenomenon? What is the mechanism behind these vocabulary changes? How do the background dynamics change to support such changes?

We imagine these two effects in the following way: Each network has a set of multiple discrete attractors that can be reached spontaneously, while some of them can also be reached through electrical stimulation. We show that throughout stimulation, the evoked responses become more robust with time – consistent with the basin of attraction becoming steeper on one side. On the other hand, the same attractors become much less accessible spontaneously – consistent with another side of the basin becoming flatter. One can imagine digging in the energy landscape and piling the dirt onto the other side.

In order to verify this hypothesis, we need to map the basin of attraction before and after stimulation. We do this via the set of initial states of bursts. Specifically, we ask whether the same set of initial states will lead to the same set of bursts. We define such a candidate set by considering the initial states of one of the stimulated clusters (see methods). We can now follow all the bursts that originate from this area. Before stimulation, these bursts are similar to one another (the peak at large correlation values in Fig 8A left, blue). After the stimulation, however, the bursts originating from the same area are much more variable (Fig 8A left, purple). To quantify this difference, we once again calculate $\Delta CDF(\alpha\theta)$ as shown in Fig 6B. Repeating the analysis on initial states stemming from a non-stimulated cluster shows a smaller effect (Fig 8A, right). We pool the data from all networks using z-scores of this value, showing a trend for the stimulated patterns to be more disrupted (Fig 8B).

We can visualize the change in the dynamics by looking at all the bursts associated with a single cluster – evoked and spontaneous, before and after the stimulation. Using nonlinear dimensionality reduction, we can see that for the non-stimulated patterns (Fig

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.08.539868; this version posted June 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

8C right, green frame), similar initial conditions lead to similar bursts. For the stimulated cluster (left, red frame), however, this is only true before stimulation (narrow distribution of blue trajectories), and not after (purple trajectories).

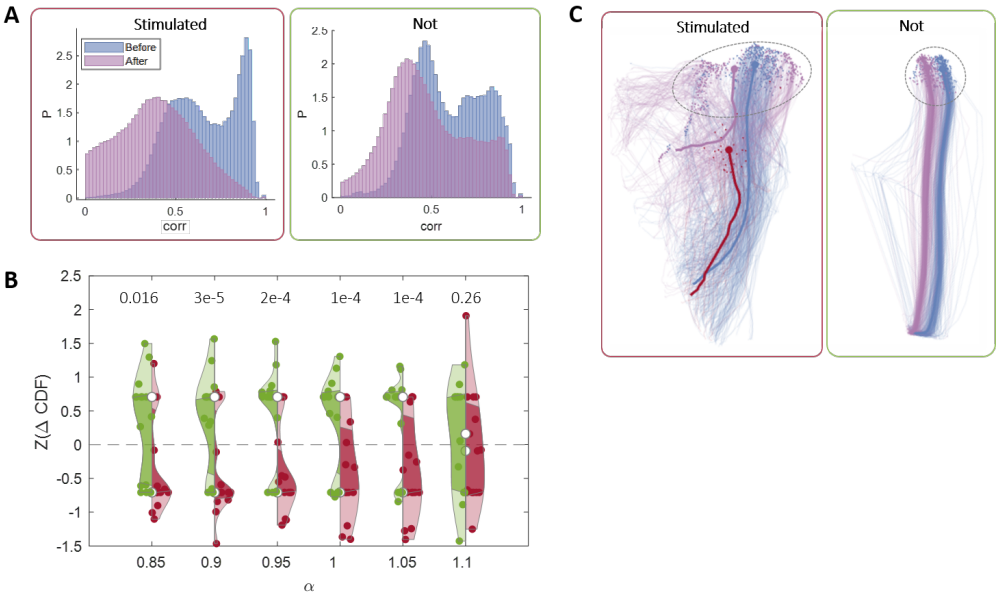

Fig 8. Mechanistic explanation. (A) The probability distribution of the spatiotemporal correlations between pairs of events with initial states similar to the ones of the stimulated patterns (red frame) and not similar to the stimulated patterns (green frame), before stimulation (blue), and after (purple). The difference between the two is measured by $\Delta CDF(\alpha\theta)$ as shown in Fig 6B. **(B)** Statistics across 11 stimulation experiments and 5 control experiments. The violins represent $\Delta CDF(\alpha\theta)$ values in stimulation clusters (red) and in non-stimulated clusters (green) for a range of α values (the fraction of θ at which ΔCDF was calculated). The numbers above each pair of violins represent the p-value of the hypothesis that the effect in the stimulation clusters is larger than in the non-stimulated clusters. **(C)** Trajectories in a non-linearly reduced 2D space (using UMAP) of one stimulated cluster (left, red frame) and one non-stimulated cluster (right, green frame), before (blue) and after (purple) stimulation. The dashed circles denote the area of initial states for each cluster.

Discussion

In this work, we analyzed the spontaneous activity of cultured neural networks. We showed that each such network has a finite repertoire of bursts that function as discrete attractors. Based on these dynamics, we were able to create a vocabulary of spatiotemporal patterns that describe the spontaneous dynamical space of the network. We showed that these attractors are accessible not only spontaneously, but also using electrical stimulation – we were able to find stimulation sites that generated robust and coherent evoked responses similar to the ones in the spontaneous vocabulary.

In order to answer questions regarding the plasticity of the vocabulary, we used electrical stimulation to force the network's dynamics to visit specific attractors repeatedly. We find that the targeted attractors are eliminated from the spontaneous vocabulary, while they are robustly evoked by the electrical stimulation. This seemingly paradoxical finding can be explained by a Hebbian-like strengthening of specific

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.08.539868; this version posted June 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

pathways into the attractors, at the expense of weakening non-evoked pathways into the same attractors.

Synchronized bursts are routinely observed in neural cultures and have been suggested to be a barrier to plasticity (15). Therefore, several attempts have been made to suppress them in order to allow plasticity (16; 17). Our work suggests that these synchronized bursts can also be informative, and serve as objects that advance the study of plasticity. In this work, we learned the network's dynamical structure and used it as a tool to shape the dynamics in specific directions. This is similar to the concept of learning within the "intrinsic manifold" presented in (18) which suggests that working within the constraints imposed by the underlying neural circuitry can make the learning process significantly easier and more accessible.

To our knowledge, this work provides the first direct evidence for the plasticity of multiple attractors in a biological neural network. The plasticity principles we describe improve our understanding of how attractors in a biological system evolve.

Methods and Materials

Cell culture

Cortical neurons were obtained from newborn rats within 24h after birth as described in (19). The neurons were plated directly onto multielectrode arrays (MEAs) and allowed to develop mature networks over a time period of 18-21 days. The number of neurons in a typical network is in the order of 10^6. The preparations were bathed in Minimal Essential Medium (MEM) supplemented with NuSerum (10%), L-Glutamine (2mM), glucose (20mM), and insulin (25mg/l), and maintained in an atmosphere of 37°C, 5% CO2 and 95% air in an incubator. Starting a week after preparation, half of the medium was replaced every 2 days with a fresh medium similar to the one described above excluding the NuSerum and with lower concentrations of L-Glutamine (0.5mM) and 2% B-27 supplement.

During recordings and stimulation, the cultures were removed from the incubator, but still maintained in an atmosphere of 37°C, 5% CO2, and 95% air. The dish was perfused at a constant ultra-slow rate of 2.5 ml/day by a custom-built perfusion system.

Experimental system

Network activity was recorded and stimulated through a commercial 120-channel headstage (MEA2100, MCS). The 120 30μm diameter electrodes are arranged in a 12x10 array, spaced 1mm vertically and 1.5mm horizontally. Data acquisition was performed using Multi Channel Suite. All data were stored as threshold crossing events, with the threshold set to 5σ, where σ is the standard deviation of the entire voltage trace.

Stimulation profile: As described in the text, 6 electrodes were selected for stimulation at each stimulation time. Biphasic voltage pulses of $+-700mV$ lasting $400\mu sec$, $200\mu sec$ for each phase were activated through all 6 electrodes simultaneously.

Data processing

Threshold crossings yield discrete time stamps of events from 120 extra-cellular electrodes. We smooth (using a Gaussian kernel, $\sigma = 2e-2$) and bin the data (bin size is 5msec) to get 120 continuous time series.

In order to avoid stimulation artifacts, we exclude the data from the stimulating electrodes when we analyze evoked responses. In addition, we ignore the first 5msec after stimulation offset from all electrodes.

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.08.539868; this version posted June 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Burst extraction

For each network, before and after stimulation, we detected all spontaneous bursting events using threshold crossing with the threshold set to 4σ, where σ is the standard deviation of the overall activity before and after stimulation respectively. We defined the starting and ending points of each such event as the crossing of the low threshold of 0.5σ. The typical duration of these spontaneous bursting events is 100 to 300 msec. In our analyses, we focus specifically on the first 100 msec of these events, as this time-frame tends to exhibit the greatest variability among them.

Clustering method

There are many possible metrics for comparing the spatiotemporal activity patterns. When clustering the spontaneous activity, we relied on the observation that similar initial conditions lead to similar patterns (bimodal distribution in Fig 2A). To capture different aspects of the patterns, we used 3 different metrics to measure the similarity between bursts. In each case, we consider two bursts $A, B \in R^{N \times T}$, where $N = 120$ and $T = 20$ (100msec).

1. **Spatiotemporal correlation**
   The correlation coefficient between 2 bursts in the following way:

$$corr2(A,B) = \frac{\Sigma_t \Sigma_n (A_{tn} - \bar{A})(B_{tn} - \bar{B})}{\sqrt{(\Sigma_t \Sigma_n (A_{tn} - \bar{A})^2)(\Sigma_t \Sigma_n (B_{tn} - \bar{B})^2)}}$$

2. **Euclidean distance between the center of mass of trajectories**
   We compute the center of mass (COM) of a burst as a weighted average of the activity from all 120 electrodes. Namely, each electrode n has coordinates $x_n \in R^2$ on the MEA. The 2D trajectory of the center of mass of pattern A, denoted $COMA_t \in R^{2xT}$ is then:

$$COMA_t = \sum_n x_i A_{tn}$$

   The Euclidean distance between 2 such trajectories is computed as the norm of the difference between the two across time: $\Sigma_t |COMA_t - COMB_t|$.

3. **Correlation of spatial profiles**
   The identity of the active electrodes is used to define this metric. The spatial profile of a burst $A_t n$ is defined as $SPA_t = \Sigma_t A$. The correlation coefficient between SPA_t and SPB_t is used.

The actual values for these 3 metrics vary between networks. In order to obtain measures that are more invariant, we rely on the bimodal distributions of the initial state and the full burst similarity shown in Fig 2A, but now extended to all three metrics in Fig 9A-C. For each network and each metric, we defined two thresholds (shown in dashed white lines) in order to distinguish between pairs of bursts that converge to the same attractor (close initial states and similar bursts) and pairs of bursts that converge to different attractors. Based on this distinction we can build a graph for each of the three metrics: each node is a burst; two nodes are connected if they cross both the initial condition threshold and the metric threshold. Then, we can sum these 3 non-directed graphs into a single graph with edges valued $0 - 3$ (Fig 9D) and perform spectral clustering (with the normalized symmetric Laplacian matrix, and k-medoids as the clustering method). We only consider clusters that capture at least 2%

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.08.539868; this version posted June 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

of spontaneous events, which accounts for the vast majority of events (see tables 1 and 2, "percent explained").

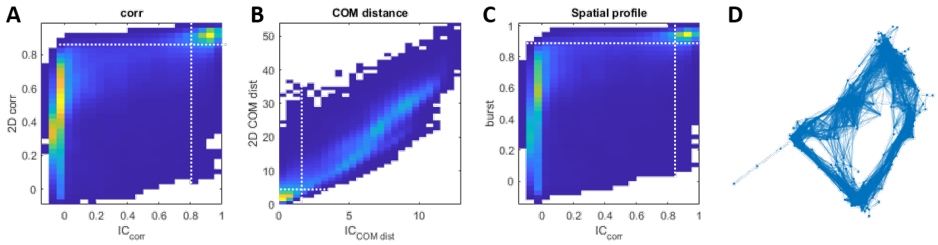

Fig 9. Dynamics-based clustering. (A) For each pair of spontaneous bursts we compute the Spatiotemporal correlation (2D corr) value and the correlation between the initial states, then we plot the probability density of the values of the first as a function of the second. The dashed white lines represent the thresholds that define the pairs of bursts that converge to the same attractor (right upper square). **(B)** Same as in A, using the COM distance metric. **(C)** Same as in A, using the spatiotemporal profile metric. **(D)** Connectivity graph representing the similarity between the spontaneous events, based on the three metrics bi-modal density.

Stimulated and non-stimulated clusters

In our analysis we define for each network a set of clusters that are similar to the stimulated patterns ("stimulated clusters") and a set of non-stimulated clusters. The definition of these two sets relies on the analysis shown in Fig 4. For each stimulated pattern in the green area, we defined a stimulated cluster as the one which explained most of the evoked responses to a given stimulation site. The non-stimulated clusters were all the clusters that explained *none* of the evoked responses to *all* 3 stimulation sites.

Comparing between evoked and spontaneous bursts

The comparison shown in Fig 6C was based on the *probe* evoked responses to the 3 stimulated patterns. As written in the main text, when comparing evoked responses to spontaneous activity, we exclude all 6 stimulated electrodes and then calculate the spatiotemporal correlation. The comparison shown in Fig 6D aimed to check whether the effect in C is specific to the stimulated patterns, and therefore was based solely on the spontaneous activity (stimulated and non-stimulated clusters, as described above). Here, we include all 120 electrodes.

Statistics over networks

Our data set consists of 11 stimulation experiments (table 1) and 5 control experiments (in which there was no stimulation during the 10 hours; table 2). The tables summarize some of the activity characteristics of each culture: MEA serial number, cell preparation date, number of spontaneous events (before and after stimulation), number of clusters (dictionary size; before and after stimulation), the percentage of spontaneous events that are part of the dictionary (this is because clusters smaller than 2% are discarded) before and after stimulation), and number of stimulated patterns (out of the selected 3) that are indeed part of the spontaneous vocabulary (according to the criterion in Fig 4).

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.08.539868; this version posted June 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

			Before stimulation				After stimulation		
MEA #	**Prep date**	**Age (DIV)**	**Events /hour**	**# of clusters**	**% Explained**	**Stimulated existence (out of 3)**	**Events /hour**	**# of clusters**	**% Explained**
26550	1.11	19	260	9	85	3	530	8	83
26549	8.11	20	292	14	98	1	651	11	97
38428	17.11	18	274	18	95	1	686	20	85
38427	1.11	20	694	13	87	2	739	14	92
26532	2.3	20	497	16	89	1	463	16	87
26550	3.5	19	617	17	93	1	563	18	89
38426	2.11	19	691	14	90	1	688	19	92
26549	11.11	19	679	7	66	1	605	8	98
26550	15.11	20	626	18	97	2	647	10	95
N/A	8.11	21	548	12	89	2	607	7	88
38428	20.2	21	301	7	96	3	392	8	94

Table 1. Stimulation experiments.

			Before stimulation				After stimulation		
MEA #	**Prep date**	**Age (DIV)**	**Events /hour**	**# of clusters**	**% Explained**	**Stimulated existence (out of 3)**	**Events /hour**	**# of clusters**	**% Explained**
26550	24.1	21	553	11	81	3	664	8	71
39740	24.4	18	609	17	88	1	659	10	79
38427	24.4	21	632	18	97	3	600	16	85
26536	7.2	20	234	15	94	3	453	17	83
38427	7.2	21	566	9	84	2	631	17	88

Table 2. Control experiments.

Success rate & probe as a criterion to proceed

The total number of cell preparations done in this study is about 100. A large number of them did not develop well enough (due to contamination events, low density of cells, and other reasons related to the maintenance atmosphere) and therefore were cleaned at early ages. The ones that matured successfully were transferred to the experimental system. We performed 17 stimulation experiments and 14 control experiments on cultures between the ages of 18-21 days. Some of these experiments are not part of the results presented in this paper due to low responsiveness to the 20 stimulation sites.

After the first 4-hour recording of spontaneous activity, there are 1.75 hours in which we probe the culture in 20 different sites, repeatedly. After this probing, we do a short analysis in which we pick the 3 stimulation sites which generated the most robust and coherent responses, then we continue the protocol as shown in Fig 5. In some of the cultures, there were no such responses at all; In these cases, we stopped the experiment right after the probing. The percentage of experiments (stimulation and control) that were completed (responded the at least 3 distinct stimulation sites robustly) is about 50%.

The decision of whether to continue an experiment after the probing stage was not based on a clear-cut condition, but on evaluation based on several figures aiming to evaluate the robustness of the responses. If there were less than 3 robust and distinct responses, or when there were very low activity levels (low number of participating electrodes), we stopped the protocol.

Acknowledgments

We thank Shimon Marom for many discussions and comments along the project. We thank Noam Ziv for useful comments on the manuscript, as well as Tamar Galateanu and Leonid Odesski for their technical support. OB is supported by the Israeli Science Foundation (grant 1442/21) and an HFSP research grant (RGP0017/2021).

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.08.539868; this version posted June 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

References

1. Aksay E, Gamkrelidze G, Seung HS, Baker R, Tank DW. In vivo intracellular recording and perturbation of persistent activity in a neural integrator. Nature Neuroscience. 2001;4(2):184–193. doi:10.1038/84023.

2. Kim SS, Hermundstad AM, Romani S, Abbott LF, Jayaraman V. Generation of stable heading representations in diverse visual scenes. Nature. 2019;576(7785):126–131. doi:10.1038/s41586-019-1767-1.

3. Miyashita Y, Chang HS. Neuronal correlate of pictorial short-term memory in the primate temporal cortexYasushi Miyashita. Nature 1988 331:6151. 1988;331(6151):68–70. doi:10.1038/331068a0.

4. Brody CD, Hernández A, Zainos A, Romo R. Timing and Neural Encoding of Somatosensory Parametric Working Memory in Macaque Prefrontal Cortex. Cerebral Cortex November. 2003;13:1196–1207. doi:10.1093/cercor/bhg100.

5. Compte A, Brunel N, Goldman-Rakic PS, Wang XJ. Synaptic mechanisms and network dynamics underlying spatial working memory in a cortical network model. Cerebral Cortex. 2000;10(9):910–923. doi:10.1093/cercor/10.9.910.

6. Piet AT, Erlich JC, Kopec CD, Brody CD, Piet A, Erlich J, et al. Communicated by Patrick Simen Rat Prefrontal Cortex Inactivations during Decision Making Are Explained by Bistable Attractor Dynamics memory model naturally accounts for optogenetic perturbations of FOF in the same task and correctly predicts a memory-dur. Neural Computation. 2017;29:2861–2886. doi:10.1162/NECO$_{a0}$1005.

7. Khona M, Fiete IR. Attractor and integrator networks in the brain. Nature Reviews Neuroscience. 2022;23(12):744–766. doi:10.1038/s41583-022-00642-0.

8. Avitan L, Stringer C. Not so spontaneous: Multi-dimensional representations of behaviors and context in sensory areas. Neuron. 2022;110(19):3064–3075. doi:10.1016/j.neuron.2022.06.019.

9. Grinvald A, Arieli A, Tsodyks M, Kenet T. Neuronal assemblies: Single cortical neurons are obedient members of a huge orchestra. Biopolymers. 2003;68(3):422–436. doi:10.1002/BIP.10273.

10. Berkes P, Orbán G, Lengyel M, Fiser J. enhanced actin depolymerization at the mDia1- bound barbed end. This inhibition occurs in the submillimolar range of P. Science. 2011;331(January):83–88.

11. Avitan L, Pujic Z, Mölter J, Zhu S, Sun B, Goodhill GJ. Spontaneous and evoked activity patterns diverge over development. eLife. 2021;10. doi:10.7554/ELIFE.61942.

12. Raichman N, Ben-Jacob E. Identifying repeating motifs in the activation of synchronized bursts in cultured neuronal networks. Journal of Neuroscience Methods. 2008;170(1):96–110. doi:10.1016/j.jneumeth.2007.12.020.

13. Eytan D, Brenner N, Marom S. Selective Adaptation in Networks of Cortical Neurons. Journal of Neuroscience. 2003;23(28):9349–9356. doi:10.1523/JNEUROSCI.23-28-09349.2003.

14. Shahaf G, Marom S. Learning in networks of cortical neurons. Journal of Neuroscience. 2001;21(22):8782–8788. doi:10.1523/jneurosci.21-22-08782.2001.

bioRxiv preprint doi: https://doi.org/10.1101/2023.05.08.539868; this version posted June 15, 2023. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

15. Madhavan R, Chao ZC, Wagenaar DA, Bakkum DJ, Potter SM. Multi-site stimulation quiets network-wide spontaneous bursts and enhances functional plasticity in cultured cortical networks. Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings. 2006; p. 1593–1596. doi:10.1109/IEMBS.2006.260571.

16. Wagenaar DA, Madhavan R, Pine J, Potter SM. Controlling Bursting in Cortical Cultures with Closed-Loop Multi-Electrode Stimulation. 2005;25(3):680–688. doi:10.1523/JNEUROSCI.4209-04.2005.

17. Kaufman M, Reinartz S, Ziv NE. Adaptation to prolonged neuromodulation in cortical cultures : an invariable return to network synchrony. 2014; p. 1–22.

18. Sadtler PT, Quick KM, Golub MD, Chase SM, Ryu SI, Tyler-Kabara EC, et al. Neural constraints on learning. Nature. 2014;512(7515):423–426. doi:10.1038/nature13665.

19. Marom S, Shahaf G. Development, learning and memory in large random networks of cortical neurons: Lessons beyond anatomy. Quarterly Reviews of Biophysics. 2002;35(1):63–87. doi:10.1017/S0033583501003742.