Dynamic correlations help prefrontal ensembles transmit information about social behavior

Nicholas A. Frost[1,3,4,5], Anna Haggart[1,3,4,5], and Vikaas S. Sohal[2,3,4,5*]

[1]Department of Neurology, [2]Department of Psychiatry, [3]Center for Integrative Neuroscience, [4]Weill Institute for Neurosciences, [5]Kavli Institute for Fundamental Neuroscience
University of California, San Francisco

*Correspondence to: vikaas.sohal@ucsf.edu

ABSTRACT

How neurons encode behavior is a fundamental question. Neuronal ensembles increase or decrease activity during specific behaviors. However, it is unclear whether ensembles encode information solely via changes in activity levels, or whether changes in correlations between neurons carry additional information. We used microendoscopic GCaMP imaging to measure prefrontal activity while mice were either alone or engaged in social interaction. Using neural network classifiers to measure how well prefrontal neurons transmit information about social behavior to downstream neurons, we find that surrogate datasets which preserve dynamic correlations outperform those which preserve ensemble activity but not correlations. Notably, this ability of correlations to enhance the information transmitted by neuronal ensembles is lost in mice lacking the autism-associated gene Shank3. These results show that dynamically modulated correlations create patterns of coactive neurons which are behaviorally-specific and enhance the information transmitted by neuronal ensembles. Furthermore, this process can be disrupted in pathological states.

INTRODUCTION

During behavior, the activity of neurons is organized with precise temporal relationships (deCharms and Merzenich, 1996; Cai *et al.*, 2016; Liang *et al.*, 2018). For example, during certain behaviors, subsets of neurons may exhibit correlated activity in which they become active at the same time or within short windows of time. However, it is unknown whether this sort of temporal organization is simply a byproduct of the interconnected nature of neuronal networks (Hebb, 1949), or contributes in a meaningful way to information encoding (Buzsáki, 2010). Groups of co-active neurons represent an attractive computational unit for information processing because they should optimize temporal summation in downstream targets. Thus, increases in correlations might further augment post-synaptic responses when pre-synaptic activity increases, or enhance post-synaptic responses even when the total level of pre-synaptic activity remains constant.

However, it is unclear whether behaviorally-driven changes in correlations actually encode additional behavioral information, beyond what is transmitted by changes in neuronal activity levels. In particular, with the advent of new technologies for simultaneously recording from large numbers of neurons in behaving animals, many studies have now shown that cortical ensembles encode behavioral information via increases or decreases in the activity of their constituent neurons. While correlations have been shown to contribute additional information for small groups (3-8 neurons) of cortical neurons (Averbeck and Lee, 2006), only a few studies have examined how correlations contribute to encoding within larger cortical ensembles. One study found that the identity of a conditioned stimulus was encoded in mean activity levels, but not in moment-to-moment patterns of co-activity (Ahmed *et al.*, 2020). Another study found that in hippocampal region CA1, disrupting correlations impairs the decoding of position, head direction and speed, but did not directly examine whether correlations themselves are dynamically modulated to encode these behavioral variables (Stefanini *et al.*, 2020). In particular, while multiple studies have shown that behavior can modulate correlations (Vaadia *et al.*, 1995; deCharms and Merzenich, 1996) the functional significance of this has remained unclear, because changes in correlations might simply reflect variation in activity levels (De La Rocha *et al.*, 2007) rather than contributing additional information.

To address these questions, we studied the mouse medial prefrontal cortex during simple social behaviors. The role of the medial prefrontal cortex in rodent social behavior is well-established (Yizhar *et al.*, 2011; Brumback *et al.*, 2017; Murugan *et al.*, 2017; Selimbeyoglu *et al.*, 2017). Many prefrontal neurons are recruited by social interaction (Brumback *et al.*, 2017; Murugan *et al.*, 2017; Liang *et al.*, 2018) as well as social stimuli such as odors (Levy *et al.*, 2019). These studies show that the activity levels of neuronal ensembles encode social behavior but have not examined whether changes in correlations between prefrontal neurons transmit additional information. Using microendoscopic GCAMP imaging in freely-moving mice, we identified prefrontal ensembles associated with social behavior. We used a neural network classifier to quantify how well these would transmit information about social behavior to downstream neurons. By examining the operation of this neural network and using surrogate datasets which preserve activity levels but either preserve or disrupt correlations, we find that changes in correlations enhance the information transmitted by neuronal ensembles. Notably, this was not

the case in a mouse model of autism (Shank3 knockout mice), demonstrating that this form of information transmission may be disrupted in pathological states.

RESULTS

Social interaction recruits prefrontal ensembles

We implanted microendoscopes (nVoke; Inscopix) into the medial prefrontal cortex (mPFC) of adult wildtype C57/B6 mice (WT) to image calcium transients using GCaMP6f expressed under control of the human synapsin promotor. We imaged freely moving mice during an assay which sequentially introduced 2 novel juvenile mice to the home cage of the subject mouse, first during an initial (novel) epoch and then again during a subsequent (familiar) epoch. These four epochs of social interaction were interleaved with epochs during which the subject mouse was alone in its home cage ('home cage' epochs). The first 5 minutes of each interaction epoch was scored by a blinded observer, and each wild-type mouse spent approximately 10 minutes interacting with the juvenile mice (393 +/- 25 s during the novel epochs and 235 +/- 18 s during the familiar epochs, $p = 0.00017$, paired t test, $n = 10$ WT mice).

We processed data using a modified PCA/ICA approach (Mukamel, Nimmerjahn and Schnitzer, 2009; Luongo, Horn and Sohal, 2016) to identify neurons which were active during the imaging session. To minimize the influence of the surrounding neuropil on neuronal signals, we calculated the mean signal within each ROI, then subtracted the mean signal calculated from a circular annulus surrounding each ROI (Supplemental Figure S1). Casual inspection of calcium traces revealed that some neurons were more active during epochs of social interaction (compared to periods of home cage exploration), whereas others exhibited the opposite pattern (**Figure 1A**). Correspondingly, aligning calcium traces to the onset of social interaction revealed many neurons that either increased or decreased activity at the onset of interaction (**Figure 1B**). Fluorescence traces were converted to binary event rasters (see Methods for details of event detection), in which most neurons were "active" in less than 5% of frames (**Figure 1C**). As a population, imaged neurons were more active during social interaction (**Figure 1C**, $n = 663$ neurons from 10 mice, percent time active in home cage: 1.8% +/- 0.1, percent time active during social interaction: 2.1 +/- 0.1%, $p = 0.00002$, paired t-test). There was a bimodal distribution of cells that were significantly more (>90th percentile, social: 152/663 neurons, home cage: 80/663 neurons; $p < 0.00001$, Chi-Squared Test) or less active (<10th percentile, social: 128/663 neurons, home cage: 119/663 neurons; $p = 0.5$) during either social interaction or matched periods when mice were alone in their home cage, as compared to circularly shuffled datasets (**Figure 1D**). These correspond to neuronal ensembles which are specifically recruited or inhibited during social interaction, respectively.

Using a neural network classifier to assess how well ensembles transmit information

Next, we sought to determine how well these prefrontal ensembles would transmit information about social behavior to downstream neurons. Since we and others have found that different prefrontal neurons are recruited during social vs. nonsocial behavior, we measured how well downstream neurons could decode whether a mouse was engaged in social behavior based on input from these prefrontal neurons. Later we will study how this was altered in Shank3 knockout mice. For this we used a simple neural network classifier that received input from the recorded neurons. Our rationale for using this kind of neural network classifier was threefold. First, a simple neural network measures information that is immediately and readily available to downstream neurons. Second, for a neural network with only one hidden layer, it is straightforward to examine the weights to determine how the network performs the

classification. This can provide insight into exactly how the neural network is able to decode behavior from the input activity. Third, examining how various parameters of the neural network affect its performance can provide additional clues about how information is represented within the input population.

Figure 2A shows the design of the neural network classifier. For clarity, we use the term 'neurons' specifically to refer to actual prefrontal neurons (which provide input to the neural network), and 'units' to refer to simulated elements within the network. The network consisted of a hidden layer containing 1000 units. We chose this number because it is both an order of magnitude larger than the number of input neurons and an order of magnitude smaller than the number of frames available for training (the latter helps ensure that there will be enough data to train the output weights). We simulated a different neural network for each mouse. Each hidden layer unit received input from a random subset of the prefrontal neurons from one mouse. I.e., each frame represents one timepoint and if neuron *i* is active in a frame then it provided an input of 1 to all the hidden units to which it is connected; otherwise it provides an input of 0. For each simulation, there was a fixed connection probability between each input neuron and each hidden layer unit. We tried different values for this connection probability in order to measure how classifier performance depends on the number of neurons that provide input to each hidden layer unit. Each hidden layer unit had an output weight which specifies how strongly that unit excites or inhibits a single output unit which classifies activity as belonging to periods in which a mouse was actively engaged in social interaction or alone in its home cage. E.g., output unit activity < 0.5 corresponds to the social condition, while output unit activity > 0.5 corresponds to the home cage condition. These output weights were adjusted during training (see Methods for details of the training rule) while the pattern of input connectivity was fixed. This models the situation in which prefrontal neurons transmit information to a downstream population of neurons (the hidden layer) that decode behavior via their output weights. We initially trained networks on 50% of the data (frames) and used the held-out data for testing. We trained and tested using intervals during which the mouse was actively engaged in social interaction or equivalent intervals when the mouse was alone in its home cage.

Classifier performance is optimal for intermediate connection probabilities

Classifier performance was strongly dependent on the probability that each input neuron was connected to each hidden unit. For the 8/10 datasets that could be classified above chance, classifier performance (measured on the 50% of data which was held-out during training) was near chance levels when the connection probability was < 0.1, but increased to a peak of 69 +/- 3% (**Figure 2B; Supplementary Table 1**) for a connection probability of 0.3. Accuracy decreased dramatically when the connection probability increased to 0.5 indicating that connection probabilities ~0.2 - 0.4 are optimal.

We also validated classifier performance by training and testing on surrogate datasets that were generated by 'swap shuffling' our original datasets. We created 'swap shuffled' surrogate datasets by randomly swapping blocks of activity between neurons (each block of activity = a set of consecutive frames during which the neuron was active). To understand this, think of the entire raster as a collection of blocks of activity. Each block occurs at a specific time, has a specific duration, and is associated with a particular neuron. Swap shuffling is equivalent to just shuffling the neurons associated with each block of activity (the start time and duration of each

block do not change). For example, if neuron *i* originally became active at time *t1* for *n1* frames and neuron *j* was active at time *t2* for *n2* frames, then in the surrogate dataset neuron *i* might become active at *t2* (but not at *t1*) while neuron *j* might become active at *t1* (but not *t2*). Swap shuffling preserves the number of neurons active at each point in time (because the timing of blocks of activity does not change). It also preserves the number of blocks of activity for each neuron, and this tends to preserve the overall level of activity of each neuron. Activity levels are not perfectly preserved, because blocks of activity can have different durations. Nevertheless, in practice, blocks of activity tend to have similar durations and the similarity between the mean activity level in each neuron before and after swap shuffling of entire datasets was 0.97 +/- 0.01. As expected, we found that neural network classifiers trained and tested on swap shuffled datasets performed near chance levels (Figure 2B).

Prefrontal neurons that drive classifier performance exhibit dramatic behaviorally-driven changes in their correlations

Next, we examined connections in trained networks to reveal factors which enable them to successfully classify social vs. home cage behavior (we analyzed networks with a connection probability = 0.3 since this maximized performance of the population). Each hidden layer unit has an output weight which measures how strongly it excites or inhibits the output unit that represents the 'decision' (social vs. home cage). Hidden units with output weights ~0 don't contribute to this decision. By contrast, hidden units with strong negative or positive weights promote the social or home cage decision, respectively (**Figure 3A**). Therefore, we hypothesized that there might be important differences in the pattern of input to hidden units, depending on whether those hidden units have large positive or negative output weights.

We arranged hidden layer units based on their output weights, i.e., the unit with the most negative weight was unit 1 and the unit with the most positive weight was unit 1000. Then we defined the 25 hidden layer units with the most negative weights as 'social units' and the 25 with the most positive weights as 'home cage units' (**Figure 3B**). For comparison we also defined the 25 hidden layer units with the weights closest to zero as 'neutral units.' For each pair of hidden units, we computed the similarity between their inputs (i.e., the correlation between their input vectors; **Figure 3C**). We then plotted the average input similarity of each hidden unit to either the social or home cage units (**Figure 3D**) or the neutral units (**Figure 3E**). Social and home cage units tended to receive input from the same prefrontal neurons as other hidden layer units with the same preference, i.e., which also had negative or positive output weights. By contrast neutral units did not exhibit any such relationship.

The preceding suggests that distinct ensembles of prefrontal neurons provide input to either social or home cage hidden units. We hypothesized that there might be important features of activity in these ensembles that support the classification of social vs. home cage behavior. For example, one possibility is that prefrontal neurons which provide input to social units might tend to increase activity during social behavior, whereas prefrontal neurons which provide input to home cage units do the opposite. Surprisingly, this was not the case. In fact, both ensembles of prefrontal neurons significantly increased their activity when mice were engaged in social interaction (**Figure 3F**; social ensemble: mean activity level 1.4 +/- 0.3% in home cage vs. 1.8 +/- 0.3% during social interaction, $p < 0.05$, sign-rank test; home cage ensemble: mean activity level 1.5 +/- 0.3% in home cage vs. 1.9 +/- 0.3% during social interaction, $p < 0.001$, sign-rank

test). Next, we examined pairwise correlations between the activity of prefrontal neurons within each ensemble. Strikingly, mean correlations within the social ensemble increased during social interaction (**Figure 3G**; (mean correlation coefficient between neurons in the social ensemble: 0.009 +/- 0.002 in home cage vs. 0.012 +/- 0.002 during social interaction, $p < 0.05$). By contrast, there was a non-significant decrease in correlations within the home cage ensemble (Figure 3G; home cage ensemble mean correlation coefficient 0.011 +/- 0.02 in home cage vs. 0.005 +/- 0.003 during social interaction, p=0.99, sign-rank).

Thus, the ensemble of prefrontal neurons which provide input to the social units form an assembly that collectively becomes more co-active (correlated) during social behavior. In contrast, the prefrontal neurons which provide input to the home cage units increase their activity, but not their co-activity, during social behavior. This suggests that behaviorally-driven changes in correlations may contribute to the encoding of social behavior.

Correlations enhance classifier performance

How can we quantitatively assess the contribution of these correlations, which are behaviorally-modulated, to classifier performance? Ideally we would first train a neural network on the original data. Then we would test this network's ability to classify data which maintained behaviorally-driven changes in activity levels, but either removed or preserved the correlations. Indeed, we have already developed methods for shuffling that achieve these goals. First, to shuffle the data in a manner that maintains behaviorally-driven changes in activity levels, but disrupts correlations, we can swap shuffle activity, but do so within each behavioral condition rather than across the entire testing dataset. In other words, we first divide up the raster into separate subrasters for each 5 minute behavior epoch (when the mouse was either engaged in social interaction or alone in its home cage). Then we performed swap shuffling (as described above) separately on each subraster, before recombining these swap shuffled subrasters to create the swap shuffled surrogate dataset for testing. Because swap shuffling tends to preserve activity levels, and because we swap shuffled activity within a behavioral condition, neurons that increase or decrease activity during periods of social interaction in the original dataset also tend to do so in the swap shuffled surrogate dataset.

To create surrogate datasets which preserve patterns of correlations as well as behaviorally-driven changes in activity, we used a method that we published previously: SHuffling Activity to Preserve Correlations, or SHARC (Luongo *et al.*, 2016). SHARC also re-assigns blocks of activity between neurons, but rather than doing so randomly, it instead follows an algorithm that achieves a target correlation matrix (in this case, the original correlation matrix) (**Figure 4B-C**). The full details of SHARC are presented in the Methods. Briefly: on each iteration, we randomly select one block of activity to be assigned to a new neuron. Instead of choosing the new neuron randomly, we first compute the difference between the target correlation matrix and the correlation matrix of the partially reconstructed surrogate dataset. Then we assign the block of activity to the neuron which will optimally reduce this difference. Finally, to maintain the mean activity level of each neuron, there is also an absolute limit on how many blocks of activity can be re-assigned to each neuron. We SHARC-shuffled each social or home cage subraster separately, then combined them to create a SHARC-shuffled surrogate dataset that preserves behaviorally-specific levels of activity and patterns of correlations.

We verified that both swap and SHARC shuffled surrogate datasets preserved levels of activity observed during both social interaction and periods when mice were alone in their home cages. Specifically, we computed the correlation between vectors in which each element represents the activity level of one neuron during one behavioral condition, and quantified the correlation between each real and surrogate dataset. For swap shuffled surrogate datasets, the similarity of activity levels (compared to real data) was 0.89 +/- 0.02 in the home cage and 0.82 +/- 0.04 during social interaction. For SHARC shuffled surrogate datasets, the similarity of activity levels (compared to real data) was 0.88 +/- 0.03 in the home cage and 0.86 +/- 0.03 during social interaction (n = 10 mice/datasets). We also computed the similarity of the pattern of correlations between each surrogate dataset and the corresponding real dataset. In this case, only SHARC shuffled surrogate datasets preserved patterns of correlations. For swap shuffled surrogate datasets, the similarity of correlations to the real data was 0.01 +/- 0.01 in the home cage, and 0.03 +/- 0.01 during social interaction. For SHARC shuffled surrogate datasets, the similarity was 0.50 +/- 0.05 in home cage and 0.55 +/- 0.03 during social interaction.

We then trained classifiers on each dataset and tested each classifier using either swap or SHARC shuffled surrogate datasets generated from the same dataset using for training (**Figure 4C**). Classifiers performed better than chance when tested with either surrogate dataset indicating that changes in activity levels encode behavioral information. However, performance was significantly higher for SHARC shuffled surrogates datasets than for swap shuffled ones (**Figure 4D**; classifier accuracy for SHARC shuffled surrogate datasets = 68 +/- 4%, classifier accuracy for swap shuffled surrogate datasets = 61 +/- 4%, $p < 0.05$, sign-rank test). This demonstrates that behaviorally-modulated patterns of correlations transmit additional information, beyond what is readily decodable from activity levels alone.

Combinations of coactive neurons occur in a behaviorally-specific manner

Interestingly, neural networks perform classification better for connection probabilities ~0.2 – 0.4 than for connection probabilities < 0.1. When the connection probability is low, each hidden unit receives input from individual prefrontal neurons or small groups of neurons. By contrast, when the input probability is higher, hidden units receive input from larger groups of prefrontal neurons. This suggests that training proceeds more efficiently when the network represents information about social vs. home cage behavior using multineuron combinations, instead of activity within individual neurons or small groups. Together with the fact that classifier performance was higher for SHARC shuffled datasets than swap shuffled ones, this indicates that multineuron patterns of coactivity, rather than just levels of activity within neuronal ensembles, transmit information about social behavior. Therefore as a proof-of-concept, we directly examined whether 3-neuron patterns of coactivity occur in a behaviorally-specific manner. We examined 3-neuron combinations because they measure network structure beyond pairwise correlations and are the building blocks of larger combinations. One could in principle analyze larger combinations, but because of the limited numbers of neurons and frames in our datasets, there is not always adequate statistical power to resolve larger combinations, i.e., to detect large numbers of combinations that occur more often in real datasets than expected by chance.

First, we quantified how often each possible 3-neuron combination occurred in real datasets. Then we calculated how often each of these combinations occurred in datasets that had been swap-shuffled (across the entirety of the dataset). For each real dataset we constructed 1000

swap-shuffled datasets, and identified 'enriched combinations,' which occurred more often in real datasets than in 95% of swap shuffled surrogate datasets. Enriched combinations are those which occur more often in real datasets than expected based on the chance overlap of activity between marginally independent neurons. Finally, we quantified how many of these enriched combinations were behaviorally-specific, i.e., occurred exclusively during social or home cage epochs. Combinations could appear to be behaviorally-specific simply because they only occurred at a single timepoint. Therefore we also restricted our analysis to enriched combinations which occurred during multiple distinct bouts of social interaction and/or matched sets of intervals during home cage epochs. Many of these repetitively-occurring enriched combinations were behaviorally-specific: 43.5% occurred during social interaction, 26.5% during home cage epochs, and 30% during both conditions.

The selective occurrence of enriched combinations either during social interaction or when a mouse is alone in its home cage may reflect changes in single neuron activity (i.e., neurons that form a social combination are only active during the social condition), and/or changes in correlations (i.e., neurons are active in both conditions but only *co-active* during social behavior). To test the hypothesis that changes in correlations underlie the behavioral specificity of significant combinations, we examined the 3-neuron combinations that were specifically enriched during either periods of home cage exploration or social interaction (**Figure 5**). We defined specific enrichment as those combinations which occurred more often in real data than in 95% of swap-shuffled surrogate datasets for one behavioral context, and less in real data than in 50% of swap-shuffled surrogate datasets for the other behavioral context. Based on these criteria, 12,408 3-neuron combinations were specifically enriched during social interaction, and 9,572 were specifically enriched during home cage exploration. There were 55,696 instances in which a social and nonsocial 3-neuron combination overlapped in 2 out of 3 neurons. In 97.0% of these cases, the neuron which was part of a social 3-neuron combination (triplet) but left out of the overlapping home cage triplet was part of a different 3-neuron combination that was enriched during homecage exploration (Figure 5, top right). Conversely, the neuron which was part of a nonsocial triplet but left out of the overlapping social 3-neuron combination was part of a different socially-enriched 3-neuron combination in 99.1% of cases (Figure 5, bottom right). Overall, an average of 71 enriched homecage combinations contained the neuron missing from the social triplet, and 85 enriched social combinations contained the neuron missing from homecage triplets. Thus, the specificity of a combination of co-active neurons for social vs. nonsocial behavior does not occur simply because some neurons were only active during one condition, but rather reflects the dynamic reorganization of patterns formed by neurons which are active in both conditions, i.e., changes in correlations. This – the behaviorally-specific occurrence of multineuron patterns of coactivity – represents the substrate through which correlations can add to the behavioral information transmitted by neuronal ensembles.

Socially-enriched combinations are deficient in Shank3 KO mice

We were curious whether there might be conditions under which these phenomena – the occurrence of multineuron combinations of coactivity during social behavior, and the ability of correlations to enhance the transmission of information about social behavior – might be impaired. To explore this, we performed microendoscopic GCaMP imaging in mice lacking the autism-associated gene Shank3 (Peça *et al.*, 2011; Duffney *et al.*, 2015; Chen *et al.*, 2020). These mice have been extensively studied as models of Phelan-McDermid syndrome, which often

includes autism as a clinical feature. Shank3$^{-/-}$ (KO) mice are known to have social deficits, and indeed, we found that compared to wild-type (WT) littermates, they spend significantly less time interacting with novel juvenile mice (**Figure 6A**).

We compared patterns of prefrontal activity in Shank3 KO mice and their WT littermates. As in WT mice, in Shank3 KO mice, many prefrontal neurons either increase or decrease activity during social interaction. However, compared to WT mice, the fraction of neurons whose activity increases during social interaction was significantly higher, whereas the fraction whose activity decreases was significantly lower (**Figure 6B-C;** 22% of 260 WT neurons vs. 39% of 290 KO neurons increased activity above the 90th percentile of shuffled data during social interaction, chi squared = 17.7, $p < 0.0001$; 25% of WT vs. 15% of KO neurons decreased activity below the 10th percentile of shuffled data during social interaction, chi squared = 8.2, $p < 0.0001$). Thus, Shank3 KO mice recruit abnormal neuronal ensembles during social behavior. We hypothesized that this might reflect a network-level disorganization that affects the normal patterning of activity during social behavior.

Indeed, we found that in KO mice a significantly smaller fraction of the 3-neuron combinations observed during social interaction were strongly enriched, i.e., occur more often in actual data than in 99.9% of swap-shuffled surrogate datasets (**Figure 6D**). This suggests that even though more neurons (i.e., larger ensembles), were recruited during social behavior in KO mice, these may have been less well-organized, such that the occurrence of socially-enriched patterns of activity is obscured by 'noise,' i.e., patterns formed by the chance overlap of activity between neurons that fire in a largely independent fashion. Notably, this deficiency was specific for social interaction. The fraction of 3-neuron combinations that were strongly enriched during home cage exploration (in comparison to swap-shuffled surrogate datasets) was similar in WT and KO mice (Figure 6D).

Correlations do not enhance the transmission of information about social behavior in Shank3 KO mice

The preceding shows that even though social behavior robustly recruits neuronal ensembles in Shank3 KO mice, the organization of these ensembles into multineuron combinations is disorganized. This suggests that the ability of patterns of co-activity to encode information about social behavior may be impaired in these mice. To test this, we directly examined whether correlations contribute to the transmission of information about social behavior in Shank3 KO mice. As before, we generated swap and SHARC shuffled surrogate datasets, then tested the ability of classifiers trained on the original datasets (from Shank3 KO mice) to classify activity associated with behavior during social interaction vs. in home cage. While we still observed above chance classification accuracy using a classifier with a connection probability of 0.3, there was no longer an increase in performance when correlations were preserved in SHARC shuffled surrogate datasets as compared to swap shuffled ones (**Figure 6E**; classifier accuracy: 62 +/ 4% for SHARC vs 63 +/- 2% for swap shuffled surrogate datasets, $p = 0.47$, sign-rank test). Thus, in Shank3 KO mice, multineuron patterns of coactivity during social behavior are disturbed, and correlations no longer add to the information about social behavior transmitted by prefrontal ensembles.

DISCUSSION

During complex behaviors, the brain can use many strategies to represent information about the external environment and internal state of the organism. The term 'ensemble' is often used to refer to a group of neurons whose activity is similarly modulated (either increased or decreased) during specific behaviors (Sakurai, 1999; Cai *et al.*, 2016; Sakurai *et al.*, 2018; Corder *et al.*, 2019; Ghandour *et al.*, 2019; Gründemann *et al.*, 2019). It is generally accepted that ensembles transmit behavioral information via changes in the activity levels of their constituent neurons. On the other hand, many studies have also shown that correlations between neurons can change during specific behaviors (Vaadia *et al.*, 1995; deCharms and Merzenich, 1996) or behavioral states (Abeles *et al.*, 1995; Pinto *et al.*, 2013; Dadarlat and Stryker, 2017). Importantly, correlations reflect changes in coactivity which exceed those expected to occur simply because of changes in the activity levels of the individual neurons (De La Rocha *et al.*, 2007). I.e., when an ensemble becomes more active, its correlations could go up, down, or remain unchanged. By optimizing synaptic interactions such as temporal summation, changes in correlated activity could potentially enhance the behavioral information transmitted by changes in ensemble activity, or transmit entirely different types of information, e.g., about internal states. Correlations have been studied extensively for the isolated retina responding to visual stimuli (Schneidman *et al.*, 2006). However, how correlations in recurrently connected cortical circuits such as the mPFC encode behavior has been more difficult to discern.

Here, we addressed this question using microendoscopic GCaMP imaging to measure activity from many (~40-100) prefrontal neurons during social behavior in mice. We used multiple approaches to disentangle the respective contributions of activity and correlations to the encoding of behavior. First, we used a simple neural network, in which prefrontal neurons provide input, there is one hidden layer, and a single output unit classifies social vs. nonsocial behavior, to quantify how well prefrontal ensembles would transmit information about social behavior to downstream neurons. Notably, classifier performance was at chance levels when hidden layer units only received input from one or a few prefrontal neurons, but was significantly higher when hidden units combined activity from 20-30% of prefrontal neurons. This suggests the network can be trained to discriminate social vs. nonsocial behavior most efficiently when each output weight corresponds to the activity in a large ensemble (representing ~20-30% of the network), rather than the activity of an individual prefrontal neuron or a small group.

Next, we extended a method we previously published, (Luongo *et al.*, 2016), to non-randomly shuffle datasets in order to preserve both behaviorally-modulated correlations and ensemble activity. This enabled us to compare the amount of information about social behavior transmitted by either SHARC-shuffled surrogate datasets or randomly-shuffled surrogates which preserved ensemble activity but not correlations. In this way, we found that correlations enhance the amount of information that prefrontal ensembles transmit about social behavior. Indeed, when we examined connections within neural network classifiers, we found that prefrontal neurons which serve to detect social behavior increase their correlations during social behavior (whereas neurons which detect nonsocial behavior do not).

Correlations measure neuronal coactivity that occurs more often than expected based on the chance overlap of activity between neurons. Thus, in accordance with our finding that behavior

modulates correlations, we found that multineuron patterns of coactivity which occur more often than expected by chance are behaviorally-specific. We then directly examined these behaviorally-specific and statistically-enriched combinations of coactive neurons. We found that they tend to be composed of neurons which are active in both conditions but only coactive in one, rather that neurons which are only active in one condition.

Interestingly, these statistically-enriched patterns of coactivity were specifically deficient during social behavior in mice lacking the autism-associated gene Shank3. Accordingly, in Shank3 KO mice, surrogate datasets which preserve behaviorally-modulated correlations failed to transmit more information about social behavior compared to randomly shuffled datasets which only preserved ensemble activity. This shows that the ability of correlations to enhance the transmission of information about social behavior is not automatic, and can in fact be disrupted in pathological states.

Similar to other recent studies (Ahmed *et al.*, 2020; Stefanini *et al.*, 2020), we have studied activity using binary activity rasters derived from GCaMP imaging. However, an important note is that any method of quantifying neural activity has limitations, such that there could be additional ways that neurons encode information which are not well resolved using this approach.

What is the meaningful size of ensembles in the cortex?

Complex behavior is possible because the brain reliably encodes features pertaining to the external environment as well as the internal state of the organism. These features may be encoded by the modulation of activity in neuronal ensembles (Sakurai, 1999; Cai *et al.*, 2016; Sakurai *et al.*, 2018; Corder *et al.*, 2019; Ghandour *et al.*, 2019; Gründemann *et al.*, 2019). How many neurons are needed to reliably encode an aspect of behavior? This is an important question because the capacity, robustness against noise, generalization ability, etc., of a network depend on how many neurons encode specific pieces of information.

We explored this question, not by measuring actual connections, but rather by asking what input connection probability would optimize the ability of a downstream network to classify behavior based on input from prefrontal ensembles. Note: input connections in neural network classifiers do not necessarily correspond to actual connections in the brain – rather they provide information about the size and nature of groups of neurons across which information should be combined to most efficiently decode behavior. Peak classifier performance occurred for connection probabilities ~0.2 - 0.3. Performance was markedly lower when the connection probability was 0.5. This is surprising because a connection probability of 0.5 would maximize the entropy of each connection; correspondingly, the number of distinct input combinations to a hidden unit is maximized when it receives connections input from half the input neurons. Thus, from the standpoint of encoding social behavior, combining activity from 20-30% of the input neurons may achieve some synergy that becomes degraded when the activity of additional neurons is included. This suggests that whatever mechanism normally generates behaviorally-meaningful patterns of coactivity in prefrontal neurons, the size of these patterns is limited to about 20-30% of the network. In the brain, nonrandom network connectivity (Alex M. Thomson, David C. West, Yun Wang, 2002; Jiang *et al.*, 2015) may similarly produce correlated activity / coactivity

within neurons that represent a specific subset of a larger population (Ko *et al.*, 2011; Litwin-Kumar and Doiron, 2012).

Combinatorial codes vs. sequential patterns of activity

Like many recent studies, we measured population-level activity in the mouse neocortex using genetically encoded calcium indicators. These indicators transduce neuronal spiking on timescales ~100 msec. Thus correlated activity / 'coactivity' imply that neurons jointly increase their activity within windows ~100 msec, and do not necessarily imply synchronous spiking on faster timescales (milliseconds or even tens of miliseconds). At the same time, correlated activity / coactivity on these timescales should be differentiated from sequential activity of neurons observed during the performance of sequential behaviors (i.e. spatial navigation or overtrained tasks) in which the activity of specific neurons corresponds to moving through a specific location or performing a specific portion of a complex task. As discussed above, in the neocortex correlations and coactivity likely reflect recurrent neural network connectivity (Ko *et al.*, 2011). By contrast, sequential patterns of neuronal activation can occur simply as a byproduct of the arrangement of spatial locations along a trajectory, the stereotyped order in which cues are encountered during a task, etc.

Relevance to disease states

Interestingly, in *Shank3* KO mice, which exhibit social deficits, the mPFC successfully recruits specific neuronal ensembles during social interaction. However these ensembles are enlarged, their organization into statistically-enriched patterns of coactivity is disrupted, and correlations between neurons fail to enhance the information that these ensembles transmit. Thus, the computational units by which information is processed in the mPFC appears to be inefficient, i.e., social behavioral recruits an abnormally large number of neurons at the expense of the precise temporal patterning of this activity. This central finding is similar to other findings in rodent models of autism at both the single neuron and network levels (Hamm *et al.*, 2017; Levy *et al.*, 2019; Chen *et al.*, 2020). In particular, we found an increase in the recruitment of prefrontal neurons during social interaction. This mirrors a recent study which found hyperdynamic response to whisker stimulation in the same mice (Chen *et al.*, 2020), possibly reflecting GABAergic circuit dysfunction and/or homeostatic compensations (Nelson and Valakh, 2015). (Note: these findings cannot be ascribed simply to the fact that Shank3 KO mice spend less time engaged in social interaction than their wild-type littermates; reduced interaction time would tend to reduce statistical power and thereby reduce the number of neurons that change their activity more than expected by chance.)

Increased excitatory activity causing decreased signal-to-noise ratio (SNR) has long been posited to contribute to the pathophysiology of autism (Rubenstein and Merzenich, 2003). However the exact nature of 'signal' and 'noise', and the specific mechanism through which excessive activity degrades the SNR have been unclear. Here, we show how enlarged neural ensemble recruitment by specific behavioral conditions disrupts information transmission by degrading the ratio between statistically meaningful patterns of coactivity (the signal) and the random overlap of activity between neurons (noise).

536 **METHODS**
537
538 Behavior: C57/B6 mice were obtained from Jackson Laboratories. We utilized adult mice of
539 either sex housed and bred in the UCSF animal facility. Adult mice were habituated to the room
540 and observer for 3 days prior to test day. All videos were subsequently scored by a blinded
541 observer. For imaging experiments, 5 WT and 4 KO littermates were generated through crosses
542 between Shank3 heterozygous parents and injected with AAV5.Syn.GCaMP6f.WPRE.SV40. We
543 included an additional 5 WT mice which were injected with
544 AAV5.Syn.GCaMP6m.WPRE.SV40 (Chen *et al.*, 2013). Viruses were obtained from Penn Viral
545 Core. Injections and 500 um GRIN lens (Inscopix) implantations were carried out in 8-12 week
546 old mice to express GCamp6f in prefrontal cortical neurons under control of the human Synapsin
547 promotor. Mice were anesthetized with 2% isoflurane and mounted in a stereotactic frame.
548 Craniotomies were made according to stereotaxic coordinates relative to Bregma. Coordinates
549 for injection into mPFC were (in mm relative to Bregma): +1.7 anterior–posterior (AP), –0.3
550 mediolateral (ML) and –2.75 dorsoventral (DV), and GRIN lenses were implanted at the same
551 AP and ML coordinates, to a depth of 2.25. We subsequently attached baseplates for attaching
552 the microendoscope, ~4 weeks later depending on GCamp expression. Mice were habituated
553 for three days with the scope attached, prior to test day. On test day, mice were habituated with
554 the scope turned on, then imaged in alternating home cage and social epochs. During social
555 epochs, one of 2 novel sex-matched juvenile mouse was introduced to the test mouse's
556 homecage, in sequential order so that there were two 'novel' epochs, followed by two 'familiar'
557 epochs interleaved with 'home cage' epochs during which the juvenile mice were removed and
558 the test mouse was free to explore its home cage. The first and last home cage epoch were 10
559 minutes in length; the others were 5 minutes in length. Each social epoch lasted 10 minutes but
560 only the first 5 minutes were recorded and scored. During each behavioral epoch, observer was
561 not in the room. Interaction epochs were defined from the moment test mouse first sniffed the
562 juvenile conspecific or object, until the test mouse turned away. Videos were recorded using
563 Anymaze, and scored by a blinded observer. For the bulk of analysis we pooled data across 10
564 WT mice. Shank3 KO mice were compared only to recordings from WT littermates.
565
566 Image acquisition and segmentation: Images were acquired using an Inscopix nVoke
567 microendoscope attached to a laptop computer and synced to a separate video acquisition
568 computer running Anymaze. Frame rate was 20 Hz and the laser power was 0.2 mW. Acquisition
569 was performed using 2x2 pixel binning, then subsequently downsampled again by 2.
570
571 We segmented neuronal signals using a modified PCA/ICA approach (Mukamel, Nimmerjahn
572 and Schnitzer, 2009; Luongo, Horn and Sohal, 2016), modified so that each segment was
573 expressed as a binary ROI consisting of pixels representing a single neuron. I.e., we used the
574 output from the PCA/ICA to identify a set of contiguous pixels which represent a neuron, then
575 averaged fluorescence signals across those pixels. To deconvolve neuronal signals from
576 background neuropil signals, we also subtracted the mean signal from each identified segment
577 from the mean value in pixels surrounding the edge of the segment (we excluded pixels that
578 belonged to another ROI). Signals were subsequently lowpass filtered to remove high frequency
579 noise using the Matlab command: designfilt('lowpassfir', 'PassbandFrequency', 0.5,
580 'StopbandFrequency', .65, 'PassbandRipple', 1, 'StopbandAttenuation', 25). All signal traces
581 shown represent normalized versions of the dF/F_0 trace, where F_0 *is estimated by the median*

value in the surround region. Threshold based event detection was performed on the traces by detecting increases in dF/F_0 exceeding 3σ over one second, then only keeping those events which exceeded a 15σ increase over two seconds, and a total area under the curve of 250σ. As there were occasional downward deflections due to surround subtraction, we instituted a final parameter requiring that the peak cross an absolute value of $dF/F_0 = 0.0125$. σ is the standard deviation of dF/F_0, calculated over the least active 50% of the movie. In some cases these parameters were adjusted slightly to optimize event detection to > 95% sensitivity and specificity, based on visual inspection, for each movie. After identifying these events in the GCaMP signal from a cell, the cell was considered "active" during the entire period from the beginning of an event until the GCaMP signal decreased 30% from the peak of the event, up to a maximum of 2 seconds. The peak of the event was defined as the local maximum of the entire event, from the beginning of the event until dF/F_0 returned to the pre-event baseline value. Calcium traces from segmented neurons were visually inspected and a small number of segments were removed if they did not appear to represent a single, unduplicated neuron. We restricted further analysis to those mice with 25 or more active neurons. We then created 2-dimensional event rasters consisting of detected events for each neuron over the course of the experiment.

Detection of behaviorally modulated neurons: To determine the response of individual neurons to behavioral context, we averaged the activity of each neuron during frames corresponding to periods of social interaction, or to a temporally matched set of frames during the preceding home cage epoch. We then created a 'null distribution' for each neuron that represents the percent of time active expected in each condition based on chance, by circularly shuffling the data 10,000 times. We then compared the activity of each neuron during either social interaction or home cage exploration to this null distribution. Neurons were considered positively modulated if they exceeded the 90th percentile of that observed in circularly-shuffled datasets, and negatively modulated if the percent of frames that a neuron was active during a given context was below the 10th percentile of observations from circularly-shuffled data.

SHARC: SHARC (SHuffling Activity to Rearrange Correlations) is an iterative method for generating surrogate datasets. SHARC nonrandomly shuffles blocks of activity within a raster to generate a new (surrogate) raster in which the pairwise correlations between neurons match a target correlation matrix (Luongo, Horn and Sohal, 2016). Here we apply this previously-published method, with modifications to also preserve the activity level in each neuron (Figure 4B).

To begin, note that each raster is equivalent to a collection of blocks of activity. Each block of activity is defined by the time at which it begins, its duration, and the neuron which is active. On each iteration one block of activity is randomly chosen and assigned to a new neuron as follows. Suppose block i has been chosen to be re-assigned. First, we find all the blocks of activity that overlap with block i. Next, we selected the subset of these blocks for which new cell identities had already been assigned. Call this set X. Let r_j represent the number of timepoints over which block $j \in X$ overlaps with block i, and let n_j represent the identity of the cell assigned to block $j \in X$. L_i and L_j are the lengths of blocks i and j, respectively. Then we constructed a vector,

$$\vec{P}_i = \sum_{j \in X} \frac{r_j}{\sqrt{L_i L_j}} \left(\vec{C}_{n_j} - \vec{C}'_{n_j} \right)$$

where $\vec{C}_{n_j}$ represents row j of the target correlation matrix, i.e. the target correlations between neuron n_j and the other neurons, and $\overrightarrow{C'}_{n_j}$ contains the current values of the correlations between neuron n_j and the other neurons based on the blocks of activity that have already been re-assigned. This step can be thought of as "guessing" which cell *should* be assigned to a particular block of activity by first figuring out what other cells are active at the same time, then choosing cells which are strongly correlated with these known active cells. Note that we assign values of $\vec{P}_i$ (i.e., construct "guesses" about which cell should be active), using the *difference* between the current correlation matrix ($\vec{C'}_{n_j}$) and the target correlation matrix ($\vec{C}_{n_j}$), in order to identify cell pairs for which the current correlation deviates from the target value, and force the new correlation matrix to progressively approximately the target correlation matrix.
We set elements of $\vec{P}_i$ to zero if the corresponding neuron had already been assigned to a block of activity that overlaps with block i, i.e. element n_j of $\vec{P}_i$ was set to zero $\forall\, j \in X$. Finally, we assigned block i to the neuron corresponding to the maximum value of $\vec{P}_i$. This can be thought of as choosing the cell that represents the "consensus" based on tallying up all of the "guesses" about which cells "should" be assigned to the block of activity being considered.

When all the elements of $\vec{P}_i$ were zero, e.g. because there no overlapping blocks of activity have had new cell identities assigned yet, then we chose a cell in order to match the originally observed level of activity. Specifically, after every iteration, we kept a log of the net number of blocks of activity that each neuron had donated or received. We used this vector to create a weighted probability whereby events from neurons which had received a net positive number of blocks were more likely to be chosen to be reassigned. To further ensure that the total number of active events for each neuron in the surrogate dataset was similar to the real dataset, if the difference between the number of blocks gained – lost in the reassignment process exceeded +10 for a particular neuron, then that neurons was no longer eligible to receive additional blocks of activity.

We extended this approach to generate surrogate datasets by shuffling data within shorter time windows (i.e., individual behavioral epochs). Here a discrete set of frames is chosen, corresponding to a subraster of the original raster. By repeating the process described above for each subrasters, then recombining the shuffled subrasters, we generate a complete shuffled dataset.

<u>Classifier</u>: We designed and trained a neural network to classify behavior (periods when a mouse was alone in its home cage vs. engaged in social interaction). This network contained 1000 units in a hidden layer, each of which received input from specific prefrontal neurons (from the real dataset). Thus, in each frame the activity of each hidden layer unit was just the summed activity of the connected prefrontal neurons. Each hidden layer unit had an output weight that represented the strength of its connection to a single output unit. On each frame the activity of the output unit was computed as:

$$y = \frac{1}{1 + e^{-\sum w_i x_i}}$$

where w_i is the output weight from hidden unit i and x_i is the activity of hidden unit i.

When we performed training and testing using the same dataset, we divided the dataset into alternating blocks of 500 frames for training vs. testing (in other cases we used the real dataset for training, then tested using a surrogate dataset). We restricted training or testing to frames in which mice were scored as actively engaged in social interaction (or matched frames during periods when the mouse was alone in its home cage). We also limited training / testing to frames with at least 3 active neurons.

We trained the output weights by performing 500 passes through the training data (each pass visited all of the training frames in a random order). On each training timestep, we calculated y, the activity of the output unit, and then adjusted each output weight based on:

$$\Delta w_i = \varepsilon y(1-y)(y-z)x_i$$

where z is the correct classification of the frame (0 for social behavior, 1 for home cage) and ε, the learning rate, was set to 0.05.

Following training, we examined the pattern of input connections and output weights. The distribution of output weights was roughly gaussian and centered near 0. We identified the selection of prefrontal neurons most likely to be connected to hidden layer units with large positive or negative weights. Hidden layer units with large negative or positive output weights bias classification towards the social or home cage condition, respectively. Therefore, we refer to the 25 hidden units with the most negative or positive weights as 'social' or 'home cage' units respectively. We calculated the number of input connections between each prefrontal neuron and the 25 home cage units or 25 social units. We then defined 'home cage' or 'social' ensembles as the 20% of prefrontal neurons with the most input connections to home cage or social units, respectively. As described in the main text, we then analyzed properties of these two ensembles.

<u>Quantification of multineuron combinations</u>: Estimating chance overlap between activity of largely independent neurons requires accounting for two factors. First, neurons with higher activity are more likely to overlap by chance with other neurons. Second, overall network activity is dynamic over time, creating a tendency for otherwise independent neurons to be recruited at similar times. Thus, it is necessary to identify combinations which occur more often than expected based on 1) the activity levels of the constituent neurons, and 2) the fact that activity in a network is not constant over time. We can do this by quantifying the occurrence of combinations in datasets which have been shuffled to preserve 1) the overall level of activity in each neuron, and 2) the total level of activity in the network at each point in time.

3 neuron combinations were quantified by identifying each combination present in frames in which 2 or more neurons were active. The number of frames each combination was active in *real data* was stored in a n-dimensional matrix. Surrogate datasets were then generated from event rasters by *swapping* the identity of neurons associated with detected events (periods of activity). As the timing of events themselves is unchanged, and only the identity of the participating neurons are exchanged, this preserves both the number of events per frame and the number of events that each neuron participates in. Therefore, the total number of combinations in

each frame and over the course of the experiment (i.e., the sum of occurrences across all combinations) is also preserved. The total number of combination occurrences in which a given neuron participates would also tend to be preserved in these swap-shuffled surrogate datasets.

We then quantified how often each combination occurred in real vs. swap-shuffled data. By comparing how often each combination occurred in real data vs. in 1,000 swap-shuffled surrogate dataset, we were able to quantify how 'enriched' each combination was, compared to the level of occurrence expected by chance based on the activity levels of its constituent neurons (and the overall temporal pattern of network activity). We expressed enrichment as a percentile, calculated relative to swap-shuffled surrogate data, e.g., the 100th percentile indicates that a particular combination occurred more often in real data than in all 1,000 surrogate datasets. Further analysis was restricted to 'enriched combinations', i.e., combinations that occurred more often in real datasets than in 95% of surrogate datasets.

Statistical analysis: Neurons and significant combinations from all animals and groups were pooled and counted as single units. Proportions were compared using chi-squared test. Activity levels were compared using paired *t*-tests (2-sided), unless otherwise noted. Where applicable, error bars denote standard error. Values of the classifier performance (accuracy) were generated by averaging after re-running the training / testing procedure at least 25 times.

735 **REFERENCES**
736

737 Abeles, M. *et al.* (1995) 'Cortical activity flips among quasi-stationary states', *Proceedings of*
738 *the National Academy of Sciences of the United States of America*. National Academy of
739 Sciences, 92(19), pp. 8616–8620. doi: 10.1073/pnas.92.19.8616.

740 Ahmed, M. S. *et al.* (2020) 'Hippocampal Network Reorganization Underlies the Formation of a
741 Temporal Association Memory', *Neuron*. Cell Press, 107(2), pp. 283-291.e6. doi:
742 10.1016/j.neuron.2020.04.013.

743 Alex M. Thomson, David C. West, Yun Wang, A. P. B. (2002) 'Synaptic Connections and Small
744 Circuits Involving Excitatory and Inhibitory Neurons in Layers 2–5 of Adult Rat and Cat
745 Neocortex: Triple Intracellular Recordings and Biocytin Labelling In Vitro', *Cerebral Cortex*,
746 12(9), pp. 936–953. Available at: https://academic.oup.com/cercor/article/12/9/936/383201
747 (Accessed: 29 July 2020).

748 Averbeck, B. B. and Lee, D. (2006) 'Effects of noise correlations on information encoding and
749 decoding', *Journal of Neurophysiology*. American Physiological Society, 95(6), pp. 3633–3644.
750 doi: 10.1152/jn.00919.2005.

751 Brumback, A. C. *et al.* (2017) 'Identifying specific prefrontal neurons that contribute to autism-
752 associated abnormalities in physiology and social behavior', *Molecular Psychiatry*. doi:
753 10.1038/mp.2017.213.

754 Buzsáki, G. (2010) 'Neural Syntax: Cell Assemblies, Synapsembles, and Readers', *Neuron*,
755 68(3), pp. 362–385. doi: 10.1016/j.neuron.2010.09.023.

756 Cai, D. J. *et al.* (2016) 'A shared neural ensemble links distinct contextual memories encoded
757 close in time', *Nature*. doi: 10.1038/nature17955.

758 Chen, Q. *et al.* (2020) 'Dysfunction of cortical GABAergic neurons leads to sensory hyper-
759 reactivity in a Shank3 mouse model of ASD.', *Nature neuroscience*. Nature Publishing Group,
760 pp. 1–13. doi: 10.1038/s41593-020-0598-6.

761 Chen, T.-W. *et al.* (2013) 'Ultrasensitive fluorescent proteins for imaging neuronal activity',
762 *Nature*, 499(7458), pp. 295–300. doi: 10.1038/nature12354.

763 Corder, G. *et al.* (2019) 'An amygdalar neural ensemble that encodes the unpleasantness of
764 pain', *Science*. American Association for the Advancement of Science, 363(6424), pp. 276–281.
765 doi: 10.1126/science.aap8586.

766 Dadarlat, M. C. and Stryker, M. P. (2017) 'Locomotion enhances neural encoding of visual
767 stimuli in mouse V1', *Journal of Neuroscience*. Society for Neuroscience, 37(14), pp. 3764–
768 3775. doi: 10.1523/JNEUROSCI.2728-16.2017.

769 deCharms, R. C. and Merzenich, M. M. (1996) 'Primary cortical representation of sounds by the
770 coordination of action-potential timing', *Nature*, 381(6583), pp. 610–613. doi:
771 10.1038/381610a0.

772 Duffney, L. J. *et al.* (2015) 'Autism-like Deficits in Shank3-Deficient Mice Are Rescued by
773 Targeting Actin Regulators', *Cell Reports*. doi: 10.1016/j.celrep.2015.04.064.

774 Ghandour, K. *et al.* (2019) 'Orchestrated ensemble activities constitute a hippocampal memory

engram', *Nature Communications*. Nature Publishing Group, 10(1), p. 2637. doi: 10.1038/s41467-019-10683-2.

Gründemann, J. *et al.* (2019) 'Amygdala ensembles encode behavioral states.', *Science (New York, N.Y.)*. American Association for the Advancement of Science, 364(6437), p. eaav8736. doi: 10.1126/science.aav8736.

Hamm, J. P. *et al.* (2017) 'Altered Cortical Ensembles in Mouse Models of Schizophrenia', *Neuron*, 94(1), pp. 153-167.e8. doi: 10.1016/j.neuron.2017.03.019.

Hebb, D. O. (1949) *The organization of behavior: A neuropsychological theory*. New York: Wiley.

Jiang, X. *et al.* (2015) 'Principles of connectivity among morphologically defined cell types in adult neocortex', *Science*. American Association for the Advancement of Science, 350(6264). doi: 10.1126/science.aac9462.

Ko, H. *et al.* (2011) 'Functional specificity of local synaptic connections in neocortical networks', *Nature*, 473(7345), pp. 87–91. doi: 10.1038/nature09880.

De La Rocha, J. *et al.* (2007) 'Correlation between neural spike trains increases with firing rate', *Nature*. doi: 10.1038/nature06028.

Levy, D. R. *et al.* (2019) 'Dynamics of social representation in the mouse prefrontal cortex.', *Nature neuroscience*. doi: 10.1038/s41593-019-0531-z.

Liang, B. *et al.* (2018) 'Distinct and Dynamic ON and OFF Neural Ensembles in the Prefrontal Cortex Code Social Exploration', *Neuron*. Cell Press, 100(3), pp. 700-714.e9. doi: 10.1016/J.NEURON.2018.08.043.

Litwin-Kumar, A. and Doiron, B. (2012) 'Slow dynamics and high variability in balanced cortical networks with clustered connections.', *Nature neuroscience*. NIH Public Access, 15(11), pp. 1498–505. doi: 10.1038/nn.3220.

Luongo, F. J. *et al.* (2016) 'Correlations between prefrontal neurons form a small-world network that optimizes the generation of multineuron sequences of activity', *Journal of Neurophysiology*, 115(5), pp. 2359–2375. doi: 10.1152/jn.01043.2015.

Luongo, F. J., Horn, M. E. and Sohal, V. S. (2016) 'Putative microcircuit-level substrates for attention are disrupted in mouse models of autism', *Biological Psychiatry*. doi: 10.1016/j.biopsych.2015.04.014.

Mukamel, E. A., Nimmerjahn, A. and Schnitzer, M. J. (2009) 'Automated analysis of cellular signals from large-scale calcium imaging data.', *Neuron*, 63(6), pp. 747–60. doi: 10.1016/j.neuron.2009.08.009.

Murugan, M. *et al.* (2017) 'Combined Social and Spatial Coding in a Descending Projection from the Prefrontal Cortex', *Cell*, 171(7), pp. 1663-1677.e16. doi: 10.1016/j.cell.2017.11.002.

Nelson, S. B. and Valakh, V. (2015) 'Excitatory/Inhibitory Balance and Circuit Homeostasis in Autism Spectrum Disorders', *Neuron*, 87(4), pp. 684–698. doi: 10.1016/j.neuron.2015.07.033.

Peça, J. *et al.* (2011) 'Shank3 mutant mice display autistic-like behaviours and striatal dysfunction HHS Public Access', *Nature*, 472(7344), pp. 437–442. doi: 10.1038/nature09965.

Pinto, L. *et al.* (2013) 'Fast modulation of visual perception by basal forebrain cholinergic

neurons', 16(12). doi: 10.1038/nn.3552.

Rubenstein, J. L. R. and Merzenich, M. M. (2003) 'Model of autism: increased ratio of excitation/inhibition in key neural systems', *Genes, Brain and Behavior*. doi: 10.1034/j.1601-183X.2003.00037.x.

Sakurai, Y. (1999) 'How do cell assemblies encode information in the brain?', *Neuroscience and Biobehavioral Reviews*. doi: 10.1016/S0149-7634(99)00017-2.

Sakurai, Y. *et al.* (2018) 'Multiple approaches to the investigation of cell assembly in memory research—present and future', *Frontiers in Systems Neuroscience*. Frontiers Media S.A. doi: 10.3389/fnsys.2018.00021.

Schneidman, E. *et al.* (2006) 'Weak pairwise correlations imply strongly correlated network states in a neural population', *Nature*. Nature Publishing Group, 440(7087), pp. 1007–1012. doi: 10.1038/nature04701.

Selimbeyoglu, A. *et al.* (2017) 'Modulation of prefrontal cortex excitation/inhibition balance rescues social behavior in CNTNAP2-deficient mice.', *Science translational medicine*. American Association for the Advancement of Science, 9(401), p. eaah6733. doi: 10.1126/scitranslmed.aah6733.

Stefanini, F. *et al.* (2020) 'A Distributed Neural Code in the Dentate Gyrus and in CA1', *Neuron*. Cell Press. doi: 10.1016/j.neuron.2020.05.022.

Vaadia, E. *et al.* (1995) 'Dynamics of neuronal interactions in monkey cortex in relation to behavioural events', *Nature*. Springer, 373(6514), pp. 515–518. doi: 10.1038/373515a0.

Yizhar, O. *et al.* (2011) 'Neocortical excitation/inhibition balance in information processing and social dysfunction', *Nature*. doi: 10.1038/nature10360.

Acknowledgments: We thank Josh Berke and Ruchi Malik for helpful comments on the manuscript. NAF is funded by K08NS105938 from NINDS and 5R25NS070680-05 from NINDS.

Author contributions: Experiments were performed by NAF. Analysis was designed by NAF and VS and performed by NAF; AH assisted with collecting data; the manuscript was written by NAF and VS**.**

Declaration of interests: Authors declare no competing interests.

Data and materials availability: All code used in the analysis will be deposited on GITHUB for any researcher for purposes of reproducing or extending the analysis.

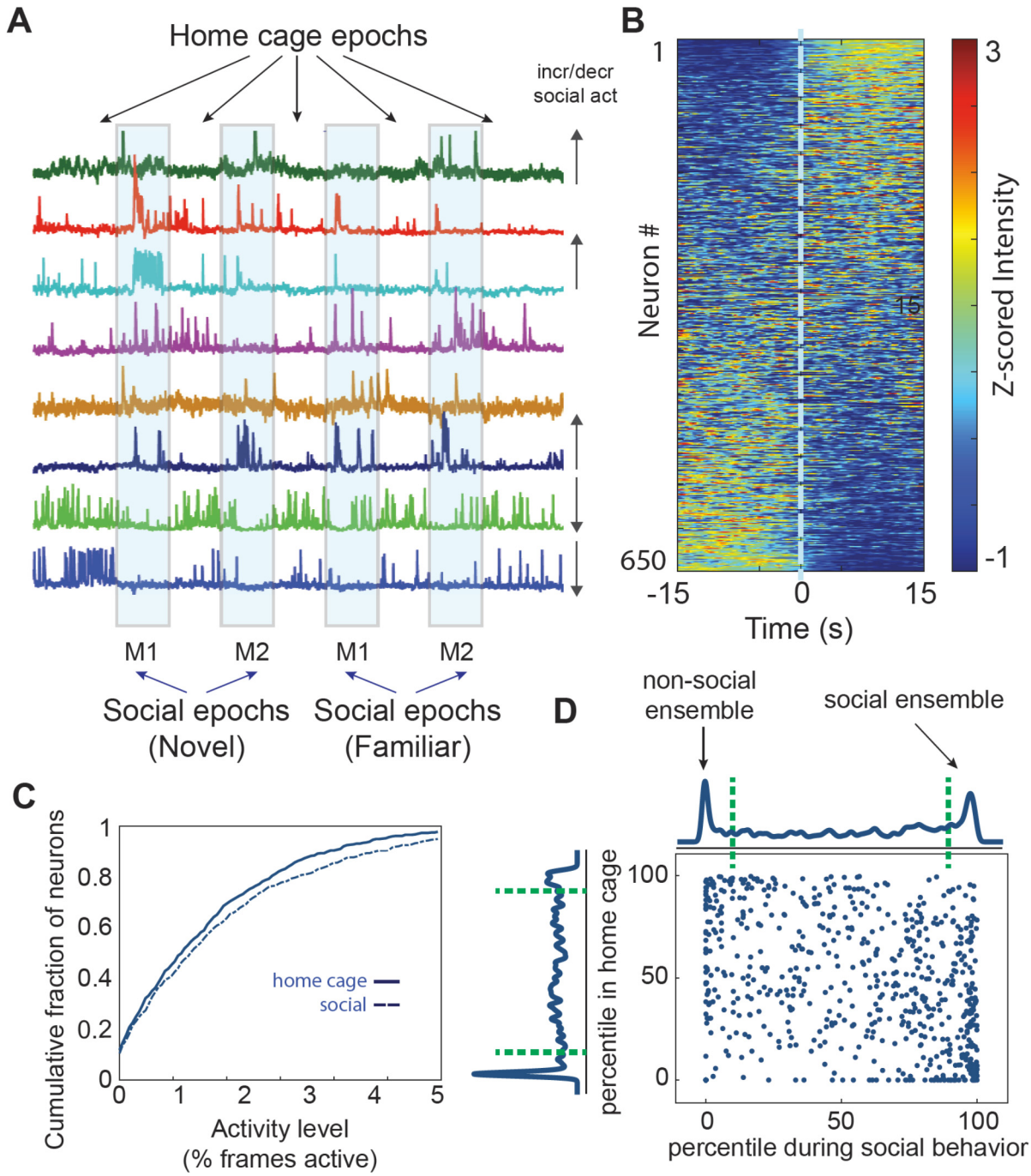


Figure 1. Social interaction modulates activity levels within prefrontal ensembles.

A. Mice were imaged across 9 consecutive behavioral epochs (each lasting 5 min) during which they were either alone in their homecage or interacted with one of two novel sex-matched juvenile mice introduced to the homecage ('M1' or '(M2'). Each novel mouse was subsequently re-introduced to the home cage during a familiar epoch. GCaMP traces during show examples of neurons that appear to increase or decrease activity during social epochs (see arrows at the right of each trace).

B. Mean z-scored GCaMP traces for all neurons recorded from wild-type mice (663 neurons from 10 mice) aligned to the onset of social interaction during the first bout of interaction within each social epoch.

C. Cumulative plot showing the distribution of activity levels for individual neurons during homecage epochs or periods of social interaction (percent time active in homecage: 1.8% +/- 0.1, percent time active during social interaction: 2.1 +/- 0.1%, $p = 0.00002$, paired t-test; n = 663 neurons from 10 WT mice).

D. Scatter-plot showing the activity of each neuron during each behavioral condition, expressed as a percentile relative to a null distribution generated by circularly shuffling that neuron's activity. Activity levels during social interaction or while the mouse was alone in its home cage are plotted on the x and y axis, respectively. Kernel density plots along the axes indicate the fraction of neurons whose activity was at a given percentile of the null distribution. Neurons with activity > 90th percentile of shuffled datasets (green dotted line) were considered to be positively modulated, whereas neurons with activity < 10th percentile (green dotted line) were considered to be negatively modulated during each behavior (>90th percentile, social: 152/663 neurons, home cage: 80/663 neurons; $p < 0.00001$, chi-squared test; <10th percentile, social: 128/663 neurons, home cage: 119/663 neurons; $p = 0.5$, chi-squared test).

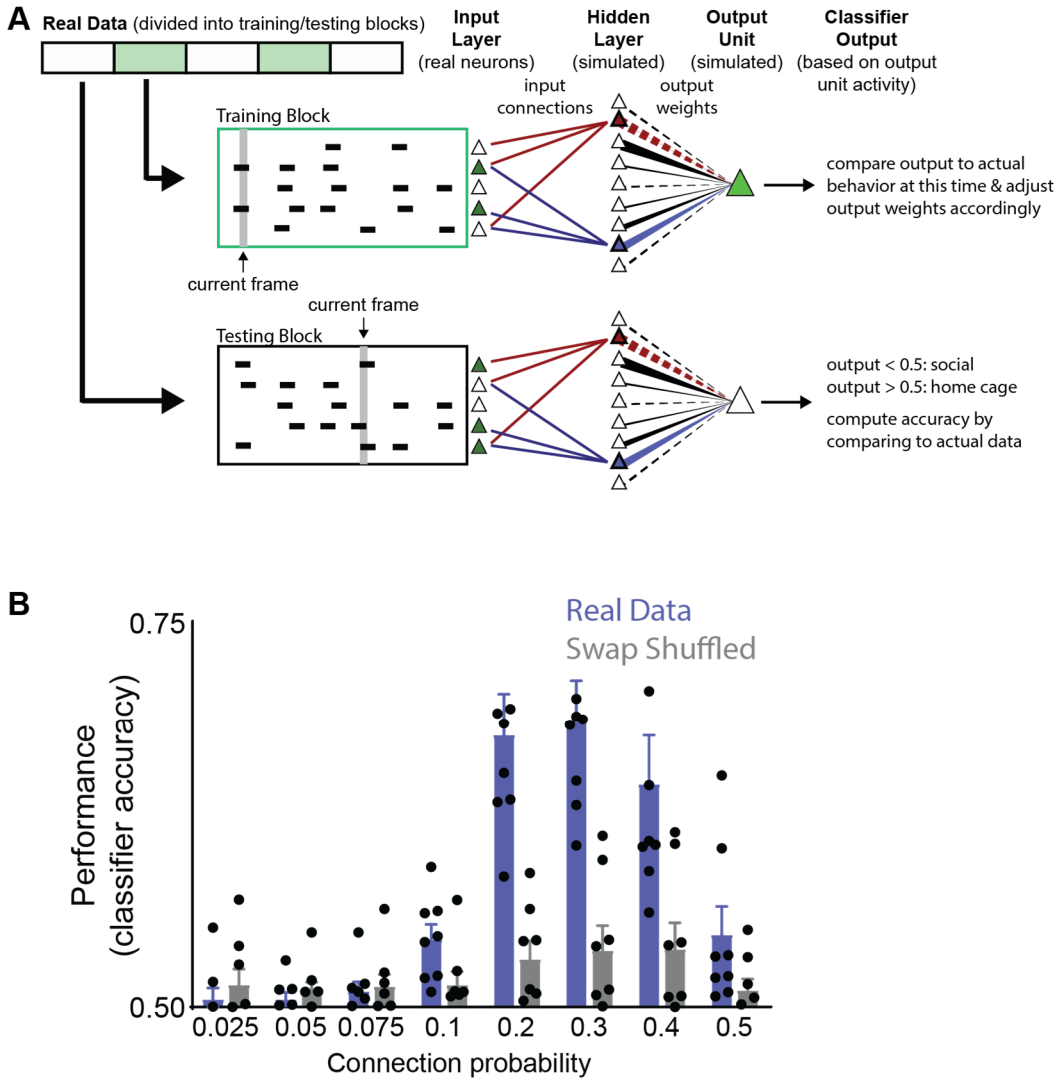


Figure 2. Classifying behavior from prefrontal ensembles using a simple neural network.

A. We constructed a neural network consisting of a single hidden layer (containing 1000 units) which were connected to a single output unit. The thickness of lines between each hidden layer unit and the output unit reflects the magnitude of the output weight. Positive and negative weights are indicated by solid and dashed lines, respectively. Each hidden layer unit received input from a random subset of prefrontal neurons from one real dataset. For clarity, we have only shown input connections to two hidden layer units (which are differentiated by their blue and red colors) – output weights from other hidden units are shown in black. The output weight from each hidden layer neuron was iteratively updated during training. We trained the classifier to distinguish periods marked as home cage exploration or social interaction by dividing a dataset into 500-frame blocks, and then using alternating blocks for training or testing.

B. The classifier performed poorly (near chance) when the input connection probability (governing the number of prefrontal neurons that provided input to each hidden layer unit) was <10%. Classification accuracy was above chance in 8/10 mice and increased to a peak of 69 +/- 3% in these mice, before decreasing again for connection probabilities >30%. The

classifier performed near chance levels when we trained and tested using data that had been randomly swap-shuffled.

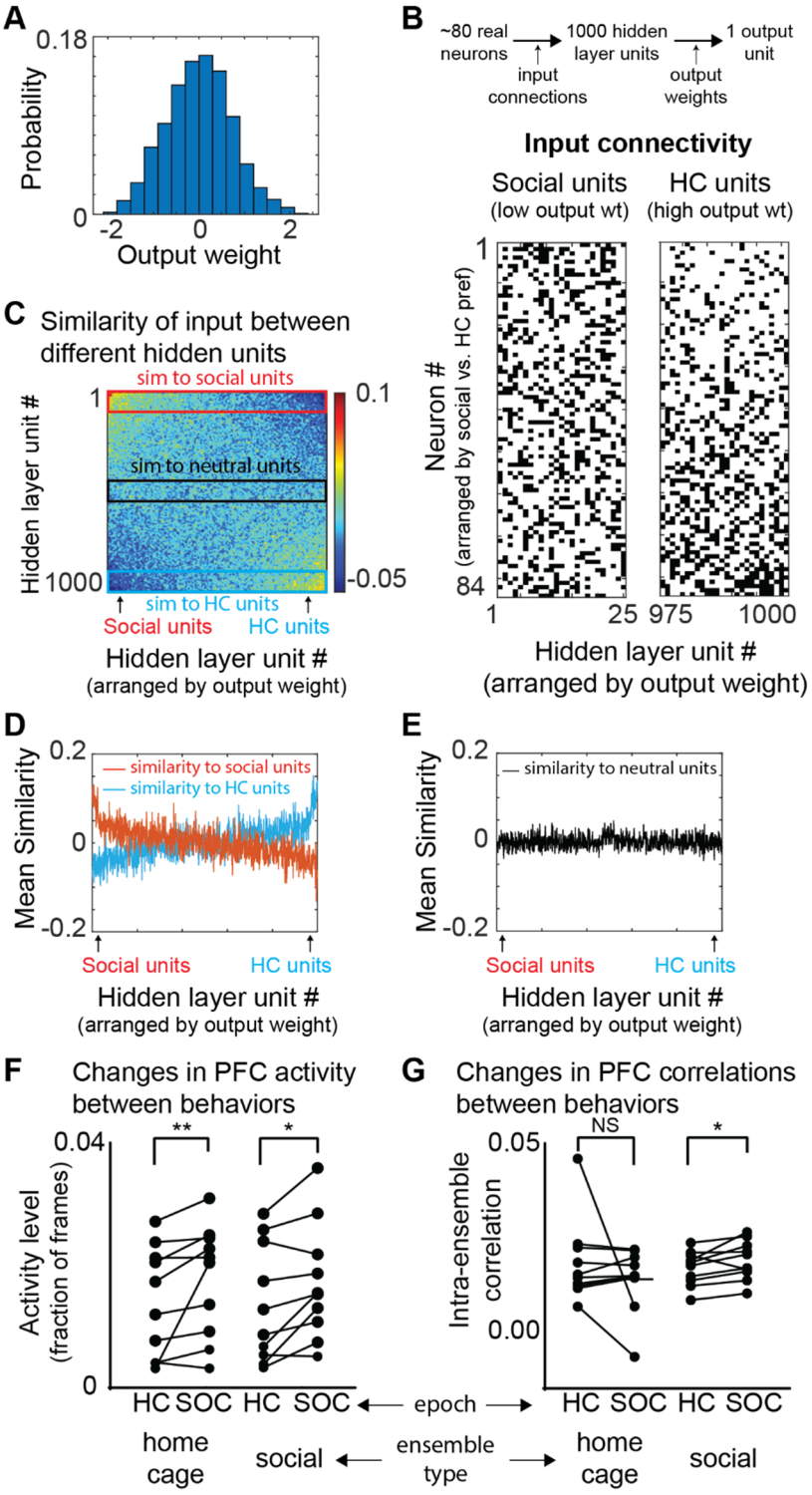


Figure 3. Classifier weights reveals an ensemble that increases correlations during social behavior and detects social behavior.

A. Example histogram depicting the distribution of output weights assigned to connections between hidden layer units and the output unit over the course of training.

B. Matrix of input connections for hidden units which detect the social (left) or home cage condition (right). The hidden layer units (x-axis) have been arranged in order of increasing output weights to identify 'social units' (25 most negative output weights) and 'home cage units' (25 largest positive output weights). Prefrontal neurons (y-axis) have been arranged in order of their preference for social interaction vs. home cage, i.e. the difference between their activity levels in the two conditions.

C. Correlation matrix showing the input similarity, i.e., the pairwise correlation between binary vectors representing the input connections to each pair of hidden layer units. Hidden layer units are arranged in order of increasing output weight. Red and blue rectangles indicate correlations with social or home cage units, respectively. A gaussian filter with a standard deviation of 3 was applied to the 1000x1000 matrix to improve visualization.

D. For each hidden layer unit, we plotted its average input similarity to either the 25 social units (red) or the 25 home cage units (blue). Hidden layer units (x-axis) are again arranged by output weight. Social units had similar patterns of input compared to each other but not to home cage units and vice-versa.

E. The average input similarity of each hidden layer unit to 25 hidden layer units with near-zero output weights ('neutral units'; black rectangle in C).

F. We defined social and home-cage (HC) ensembles as the 20% of prefrontal neurons most likely to provide input to the social or home cage units, respectively. The mean activity of both home cage and social ensembles increased during social interaction compared to the home cage condition (social ensemble: mean activity level 1.4 +/- 0.3% in home cage vs. 1.8 +/- 0.3% during social interaction, $p < 0.05$, sign-rank test; home cage ensemble: mean activity level 1.5 +/- 0.30% in home cage vs. 1.9 +/- 0.3% during interaction, $p < 0.001$, sign-rank test).

G. Correlations between neurons in the same ensemble increased during social interaction for the social ensemble but for the home cage ensemble (mean correlation coefficient between neurons in the social ensemble: 0.009 +/- 0.002 in home cage vs. 0.012 +/- 0.002 during social interaction, $p < 0.05$; home cage ensemble mean correlation coefficient 0.011 +/- 0.02 in home cage vs. 0.005+/- 0.003 during social interaction, $p=0.99$, sign-rank).

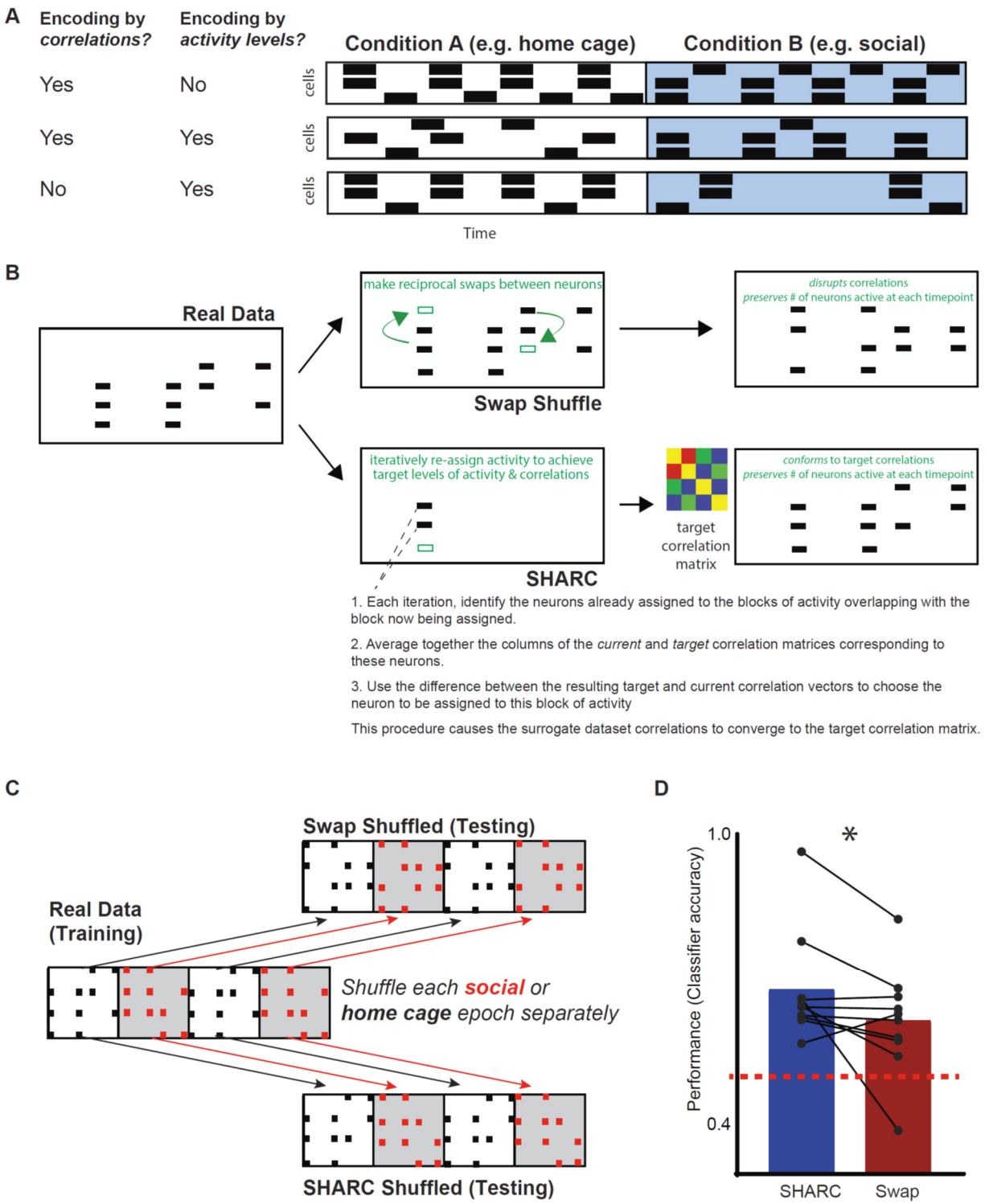


Figure 4. Correlations transmit additional information that is not efficiently conveyed by changes in activity levels alone.

A. Cartoon illustrating that information about behavior may be encoded through changes in activity levels, correlations between neurons, or both. When behavior modulates activity levels, correlations in two behavioral conditions may differ or be the same, and vice-versa.

B. To disentangle the roles of activity levels and correlations in transmitting information we used two different methods to create shuffled (surrogate) datasets which preserve changes in activity levels, but either do or do not preserve patterns of correlations. We made random, reciprocal swaps of activity between neurons to generate surrogate datasets which maintained network activity in each frame as well as the number of blocks of activity for each neuron. However, these datasets destroyed the correlation structure. In a second set of surrogate datasets we used SHARC to iteratively generate surrogates in which the correlation structure was also maintained.

C. To maintain dynamic changes in activity levels and correlations that are associated with the two behavioral conditions we swap-shuffled or performed SHARC separately for each behavioral epoch, then concatenated the 9 resulting surrogate subrasters to create each surrogate dataset.

D. We trained a classifier (with a connection probability = 0.3) on each real dataset, then tested that classifier on swap or SHARC-shuffled surrogate datasets generated from that real dataset. Accuracy was significantly higher for the SHARC-shuffled surrogates, which maintain the behaviorally-modulated correlations found in the original dataset (accuracy for SHARC shuffled surrogate datasets = 68 +/- 4%, classifier accuracy for swap shuffled surrogate datasets = 61 +/- 4%, $p < 0.05$, sign-rank test, n = 10 mice).

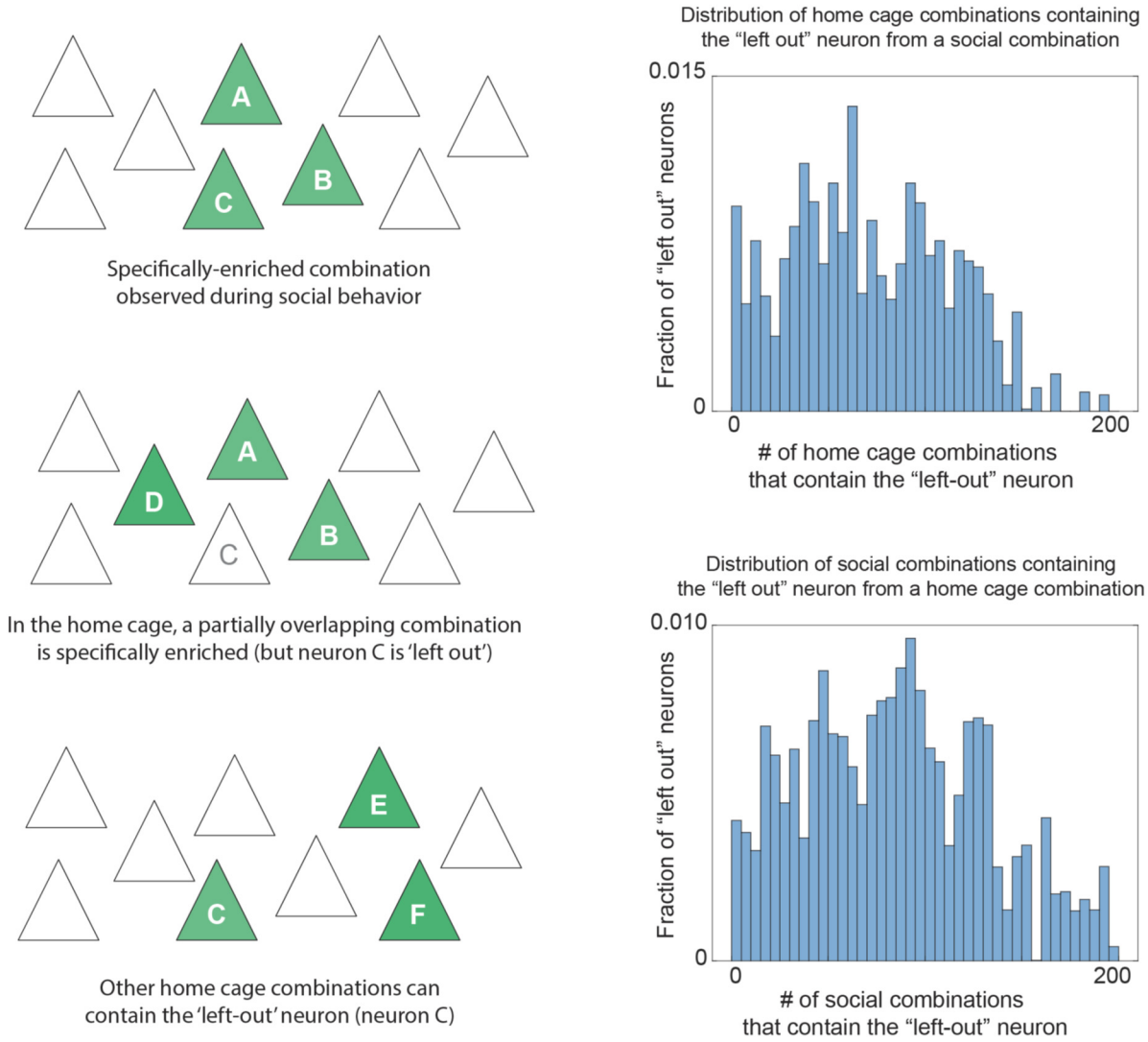


Figure 5. Behaviorally-specific patterns of coactivity are formed by neurons that are active in both conditions, but coactive with different partners in each condition.
Left: We identified combinations of 3 neurons that are specifically enriched during one behavioral condition (occurring more often during social interaction than in 95% of surrogate datasets, and occurring less often during home cage exploration than in 50% of surrogate datasets, or vice-versa). We then identified overlapping combinations occurring during the opposite behavioral condition in which a single neuron was 'left out.' In other words, we identified combinations from the two conditions that overlapped in exactly two neurons. **Right, top:** Histogram showing the number of distinct 3 neuron home cage combinations that contain the neuron which participates in a social combination but is 'left out' during home cage behavior. **Right, bottom:** Histogram showing the number of distinct 3 neuron social combinations that contain the neuron which participates in a home cage combination but is 'left

970 out' during social interaction. In the vast majority of cases, neurons that are 'left out' in one
971 condition are still active during that condition and participate in other combinations.

972

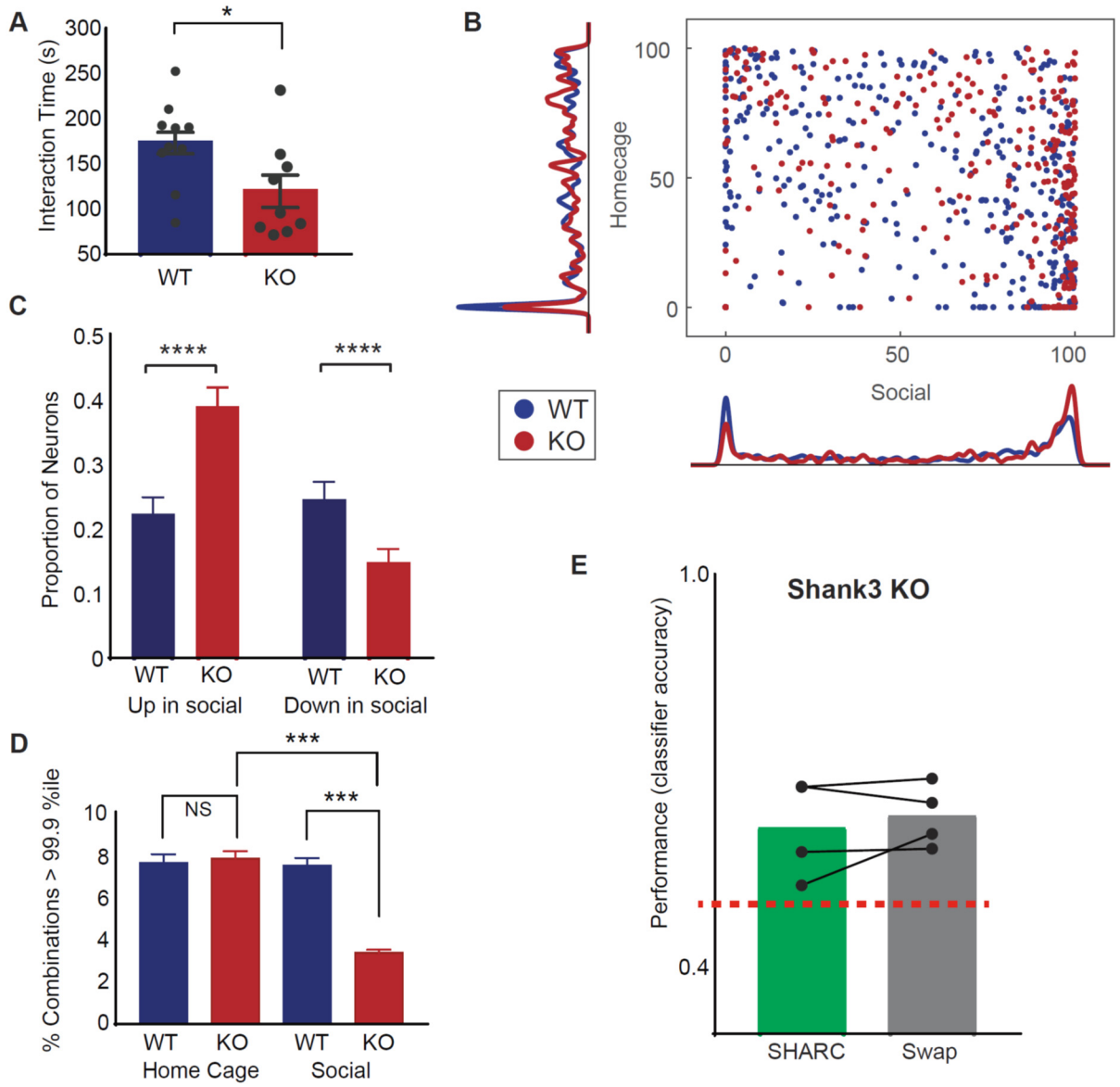


Figure 6. Shank3 KO mice have disorganized ensembles for which correlations fail to enhance the transmission of information about social behavior.

A. The mean time that *Shank3* KO mice or wild-type (WT) littermates spend interacting with a novel juvenile mouse of the same sex during a 5 min assay. Data has been pooled from 8 unimplanted WT mice as well as the 5 implanted WT mice used for microendoscopic imaging, and 5 unimplanted KO mice in addition to the 4 implanted mice used for imaging. For implanted mice we used the average of interaction time for the 2 novel mice. Pooled data showed decreased interaction in KO mice (173 +/- 12 s vs. 120 +/- 18 s for WT and KO respectively, $p < 0.05$, t-test). The un-implanted cohort alone shows a similar significant decrease in interaction time for KO mice (165 +/- 15 s vs 110 +/- 16 s for WT and KO respectively, $p < 0.05$, t-test). In the implanted cohort there was a similar trend toward

986 decreased interaction for KO mice (186 +/- 20 s vs 133 +/- 37 s, for WT and KO
987 respectively, p = 0.21, t-test).

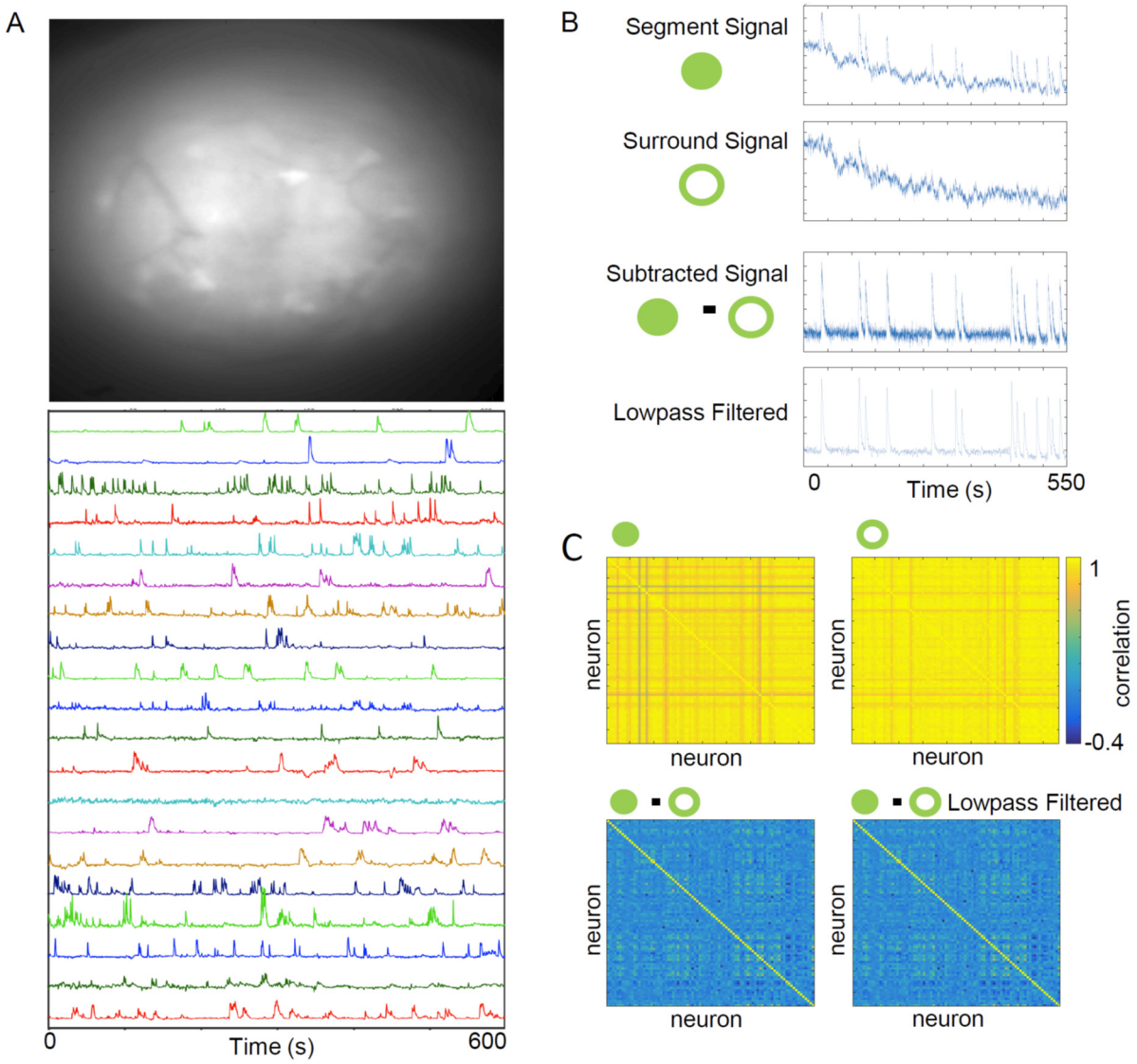
988 **B.** (Similar to Fig. 1D). Scatter-plot showing the activity of each neuron during each behavioral
989 condition, expressed as a percentile relative to a null distribution generated by circularly
990 shuffling that neuron's activity. Activity levels during social interaction or while the mouse
991 was alone in its home cage are plotted on the x and y axis, respectively. Kernel density plots
992 along the axes indicate the fraction of neurons whose activity was at a given percentile of the
993 null distribution. Neurons with activity > 90th percentile of shuffled datasets (green dotted
994 line) were considered to be positively modulated, whereas neurons with activity < 10th
995 percentile (green dotted line) were considered to be negatively modulated during each
996 behavior**.** Data is plotted for Shank3 KO mice (red) and WT littermates (blue) (mean
997 percentile of activity during social interaction, WT: 50 +/- 2 percentile; KO: 64 +/- 2
998 percentile; p < 0.0001 by 2-sample *t*-test; mean percentile of activity during home cage, WT:
999 47 +/- 2 percentile; KO: 51 +/- 2 percentile; p = 0.1, *t*-test).

1000 **C.** Bar graph showing the fraction of neurons whose activity was positively or negatively
1001 modulated (>90th percentile or <10th percentile) during social interaction. The proportion of
1002 neurons which increased activity during social interaction was significantly greater in KO
1003 mice (22% in WT vs. 39% in KO, chi-squared = 17.7, p < 0.0001), whereas the
1004 downregulated ensemble was significantly smaller in KO mice (25% in WT vs. 15% in KO,
1005 chi-squared test, 8.2, p < 0.005). Error bars denote the binomial S.E.M. algebraically derived
1006 from total number of neurons and the proportion that were modulated in the specified
1007 direction.

1008 **D.** The proportion of 3 neuron combinations occurring during home cage exploration that are
1009 enriched > the 99.9^{th} percentile compared to swap-shuffled datasets was similar across WT
1010 (Blue; 7.6%) and KO (Red; 7.8%) mice. By contrast, the proportion of 3 neuron
1011 combinations occurring during social interaction that are enriched > 99.9^{th} percentile
1012 compared to swap-shuffled datasets was 7.5% in WT compared to only 3.4% in KO mice
1013 (total number of home cage combinations: 4,187 in 5 WT mice, and 5,878 in 4 KO mice;
1014 total number of social combinations: 5,487 in 5 WT mice, 16,326 in 4 KO mice). The top
1015 two plots show histograms of enrichment for the home cage (upper) or social conditions
1016 (middle); the lower panel is a bar graph showing the fraction of these combinations that were
1017 specifically enriched above the 99.9^{th} percentile (chi-squared = 165, p < 0.0001). Error bars
1018 denote the S.E.M. algebraically derived from the binomial distribution, the number of 3
1019 neuron combinations in each condition, and the proportion of those combinations that were
1020 enriched.

1021 **E.** Performance of classifier trained on real datasets and tested on surrogate datasets.
1022 Performance was not better (and was non-significantly worse) when correlation structure was
1023 maintained using SHARC (classifier accuracy: 62 +/ 4% for SHARC vs 63 +/- 2% for swap
1024 shuffled surrogate datasets, p = 0.47, sign-rank test).

1025
1026
1027
1028
1029
1030



Supplemental Figure S1. Spatial Decorrelation of Neuronal Signals.

A. Example image (top) and individual neuron GCaMP traces (bottom) from prefrontal cortex imaged with implanted endoscope.

B. The average GCaMP signal from a region of interest (ROI), corresponding to one neuron, was corrected by subtracting the average GCaMP signal from the surrounding pixels, in order to spatially deconvolve signals from each ROI vs. the surrounding neuropil. Examples traces from a single neuron are shown.

C. The pairwise correlation matrix between signals from different neurons is shown (calculated from 550 seconds of activity from a single wildtype mouse), for the original GCaMP signals (top left), the surround signals (top right), the surround-subtracted signals (bottom left), and the surround-subtracted signals after lowpass filtering (bottom right).

Mouse	Sex	Neurons	Peak Accuracy
WT			
1	f	84	0.687
2	f	39	0.705
#3	m	25	0.53
4	f	75	0.605
#5	m	37	0.459
6	*m*	*86*	*0.647*
7	*m*	*74*	*0.848*
8	*m*	*73*	*0.691*
9	*m*	*71*	*0.694*
10	*f*	*99*	*0.633*
KO			
1	f	83	0.772
2	m	77	0.832
3	f	45	0.34
4	f	85	0.627

Supplemental Table S1. Details for each mouse included in this study.

The table shows the genotype, sex, number of imaged neurons, and peak classifier accuracy (performance when half the data was used for training and half for testing). Figure 2 showed how classifier accuracy depends on the input connection probability; because their performance was not >50% for multiple input connection probabilities, WT mice 3 & 5 (marked with #) were not included in this illustrative plot. However, we did not exclude data from these mice in any analyses. WT mice 1-5 were wild-type littermates of Shank3 KO mice. WT mice 6-10 (indicated by italics) were not littermates of Shank3 KO mice and were therefore not included as WT controls for the analyses shown in Figure 6.