**Distribution Agreement**

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Elizabeth A. O'Gorman April 10, 2018

# Low-Dimensional Mapping of Corticostriatal Circuitry Dynamics Underlying Pair Bonding

by

Elizabeth A. O'Gorman

Dr. Gordon J. Berman
Adviser

Neuroscience and Behavioral Biology

Dr. Gordon J. Berman

Adviser

Dr. Robert C. Liu

Committee Member

Dr. Joseph R. Manns

Committee Member

Dr. Samuel J. Sober

Committee Member

2018

# Low-Dimensional Mapping of Corticostriatal Circuitry Dynamics Underlying Pair Bonding

By

Elizabeth A. O'Gorman

Dr. Gordon J. Berman

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Sciences with Honors

Neuroscience and Behavioral Biology

2018

Abstract

Low-Dimensional Mapping of Corticostriatal Circuitry Dynamics Underlying Pair Bonding
By Elizabeth A. O'Gorman

Extensive research has shed light on neurochemistry and neuroendocrinology contributing to the formation of socially monogamous relationships, known as pair bonds. However, until recently (Amadei et al., 2017), the dynamic neural circuitry underlying the formation of pair bonds has remained unstudied. By analyzing local field potential (LFP) recordings from brain regions in the "social brain network", namely the medial prefrontal cortex (mPFC), nucleus accumbens (NAcc), and bed nucleus stria terminalis (BNST) of prairie voles, the canonical model organism of pair bonding, we can assess whether the neural dynamics exhibited during pair bonding are stereotyped between individuals or behaviors. Here, an unsupervised machine learning method, *t*-distributed Stochastic Neighbor Embedding (*t*-SNE) (van der Maaten & Hinton, 2008) was used to map the structure of LFP recordings collected from the mPFC and NAcc (hit subjects), and mPFC and within or bordering the BNST (non-hit subjects) of female prairie voles during a six-hour cohabitation period with a male partner. The primary objective was to identify behavior-specific brain-states during pair bonding. Intra-behavior variability of hit subjects' neural dynamics was greater than the intra-animal variability, suggesting there may not be behavior-specific structure in the hit subjects' brain-state mappings. On the other hand, the intra-animal variability of non-hit subjects' neural dynamics was greater than the intra-behavior variability, suggesting there may be behavior-specific structure in the non-hit subjects' brain-state mappings. Furthermore, 36 stereotyped brain-states (i.e. specific pairings of peak oscillatory frequencies) were identified, which may be used for decoding neural signal. Overall, these results provide the basis for further analyses of stereotyped neural

dynamics across individuals and behaviors, as well as the temporal emergence of stereotyped neural dynamics over the course of pair bonding.

# Low-Dimensional Mapping of Corticostriatal Circuitry Dynamics Underlying Pair Bonding

By

Elizabeth A. O'Gorman

Dr. Gordon J. Berman

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Sciences with Honors

Neuroscience and Behavioral Biology

2018

Acknowledgements

I would like to thank Dr. Gordon Berman for taking me on, for his support and invaluable guidance, and of course, for providing me this project.

I would also like to thank Dr. Liz Ann Amadei for collecting these data used in this paper, and Dr. Robert Liu for allowing me to use these data.

I would like to thank the Rainnie Lab for my training throughout my undergraduate career, particularly Jeffrey Hsu for his mentorship and guidance for the past three years.

Lastly, I would like to thank my family and friends for their support, particularly my perfect person, Lauren Aulet. Thank you for your unending support throughout this entire process, especially when times were tough.

Table of Contents

**TABLES**

## INTRODUCTION

Affiliative social and socio-sexual behaviors are essential for facilitating social cohesion and bonding, which are fundamental for species survival (Goodson & Kabelik, 2010; Tinbergen, 1951). These natural social behaviors displayed across species (Getz, Carter, & Gavish, 1981; Stanley & Adolphs, 2013; Svendsend, 1989) are known to be endogenously modulated via individual neurochemistry (Lim et al., 2004) and neuroendocrinology (Aragona et al., 2006; Ross et al., 2009; Young & Wang, 2004), and exogenously modulated via influences from other conspecifics (Williams, Catania, & Carter, 1992). However, the detailed neural, specifically electrophysiological, mechanisms underlying the dynamic formation of a prosocial bond remains elusive.

Socially monogamous relationships, known as pair bonds, occur in less than 5% of mammalian species; it is not a common phenomenon across species, but rather species-specific (Kleimen, 1977). The canonical model organism for examining pair bonding is the prairie vole (*Microtus ochrogaster*), known for co-parenting and being socially monogamous (McGraw & Young, 2010; Young & Wang, 2004). There exists an extensive body of literature concerning prairie vole's neuroendocrinology, namely the influence of oxytocin (OT) (Ross et al., 2009; Sofroniew, 1983), vasopressin (de Vries and Miller, 1998; de Vries and Panzica, 2006), and dopamine (DA) (Liu & Wang, 2003), on the formation and expression of a pair bond. In the prairie vole, the nucleus accumbens (NAcc), prelimbic cortex (PLC) located in the medial prefrontal cortex (mPFC), and bed nucleus stria terminalis (BNST) contain high densities of OT receptors (Insel & Shapiro, 1992). Blocking OT receptor activation in the NAcc and PLC with an OT antagonist prevents the formation of a partner preference (Young, Lim, Gingrich, & Insel, 2001). This indicates OT receptor activation in the NAcc and PLC are essential for the

development of a pair bond. Furthermore, the NAcc also contains inhibitory projection neurons, medium spiny neurons, expressing DA receptors, D1 or D2 (Lobo et al., 2010). Administering a D2 receptor antagonist in the NAcc prevents the formation of mating-induced partner preferences (Gingrich et al., 2000). Therefore, this indicates DA receptor activation in the NAcc is also necessary for the development of a pair bond. However, until recently, the neurophysiological underpinnings of pair bond formation have remained disproportionately under-examined.

The NAcc is well-characterized as a primary brain-region for reward-processing. Activity of this region facilitates goal- and particularly reward-directed behaviors (Nicola, 2007). It plays an essential role in encoding the reward of external stimuli (Stuber et al., 2011), which is at the forefront of affiliative social behavior and pair bonding (Floresco, 2015; Stuber et al., 2011). Also, the NAcc receives inputs from the mPFC. The mPFC is crucial for decision-making and biasing of social behavior, in that its activity facilitates the decision-making required to accomplish a goal or achieve reward (Block et al., 2007).

As the first to record neural activity in prairie voles during cohabitation, Amadei et al. (2017) examined the mPFC-NAcc corticostriatal circuitry supporting the formation of a pair bond. Socio-sexual interactions dynamically modulate corticostriatal circuity to drive changes in pair bond formation and expression. By recording local field potentials (LFPs) from the mPFC and NAcc, individual variation in the strength of the neural signal from the mPFC to the NAcc was shown to be dynamically modulated over the course of pair bond formation. Furthermore, by recording LFPs from the mPFC and within or bordering the BNST as an internal control, no individual variation in the strength of the neural signal from the mPFC to within or bordering the BNST was shown to be dynamically modulated over the course of pair bond formation.

Importantly, individual variation in the strength of this functional connectivity was measured by the mean Kullback-Liebler divergence ($D_{KL}$) of mPFC low-frequency phase activity modulating NAcc high-frequency amplitude and NAcc low-frequency phase activity modulating mPFC high-frequency amplitude. The mean $D_{KL}$ for one direction, NAcc to mPFC was subtracted from the other, mPFC to NAcc, to determine the net mean modulation index. It was concluded that the mean net modulation index, particularly after the first mating bout, predicts how quickly female subjects begin affiliative huddling with their male partner. If there are stereotyped brain-state-specific features during the formation and expression of a pair bond, it follows that those potential features change over the course of formation and expression of said pair bond. As such, there may be stereotyped neural dynamics underlying individual-specific variation in expression of affiliative behaviors and the formation of a pair bond.

While behavioral expression of social bonding need not be stereotyped between or within species, it may be possible to predict an animal's behavior by assessing the underlying subtle neural dynamics as exemplified by Amadei et al.'s findings (2017). The extent to which stereotyped behavior-specific features can be extracted from complex neural data is reliant on simplifying the data – this is the main objective of dimensionality reduction techniques (Stephens, Osborne, & Bialek, 2010). Furthermore, by assessing the underlying neural dynamics of complex affiliative behaviors such as pair bonding, the temporal emergence of a social bond can be better understood.

Implementations of unsupervised machine learning for neural decoding best allow unbiased and robust analysis of large-scale and complex neural dynamics. Thus, applying unsupervised machine learning techniques allows for more subtle and unbiased analysis of continuous neural dynamics during prairie vole pair bonding. A particular unsupervised machine

learning method, *t*-distributed Stochastic Neighbor Embedding, embeds high-dimensional data into a lower dimensional space, minimizing the information lost when reducing the data to the lower dimension (van der Maaten & Hinton, 2008). Previous implementations of *t*-SNE have ranged from analyzing emotional states in response to stimuli (Cowen and Keltner, 2017), to mapping resting brain-states (Billings et al., 2017), to mapping the stereotyped behavior of fruit flies (Berman et al., 2014), to clustering large-scale neural recordings to identify spiking neurons (Dimitriadis, Neto, & Kampff, 2016). Here, *t*-SNE will be utilized to explore and map the structure of underlying neural dynamics during the development of pair bonding. Embedding this high-dimensional data into a lower dimension allows for visualization and straightforward extraction of brain-state-specific features during the cohabitation period. By mapping the underlying structure of neural activity, (i.e. brain-states during pair bonding), subtler neural dynamics and features can be extracted to classify affiliative behaviors. As such, the primary objective is to extract stereotyped neural features predictive of specific behaviors during prairie vole pair bonding.

This implementation relies on the LFP recordings collected from female prairie voles during a six-hour cohabitation period with a male partner to assess whether behavior-specific brain-states can be identified over the time course of pair bond formation. If specific brain-states are identified and stereotyped for individual behaviors across animals, then those identified patterns may be used to further decode neural signal. These identified features then may be used to assess the probability of a given behavior occurring during prairie vole pair bonding.

## METHODS

### Experiments

All procedures were approved by the Emory University Institutional Animal Care and Use Committee. To probe the underlying neural dynamics of pair bonding, LFP recordings from fifteen adult, sexually-naïve female prairie voles (laboratory-bred colony derived from wild-caught Illinois stock) 76-154 days of age at the start of experimentation were collected as previously described by Amadei et al. (2017).

All surgical procedures for electrode implantation were described by Amadei et al. (2017). However, it is important to note females were first ovariectomized to control for inter-animal hormonal variability, and then chronically implanted with electrodes 10-20 days later. Electrodes (tungsten microelectrodes, 1 MΩ, FHC) were stereotaxically targeted to the left mPFC, and NAcc for 9 hit subjects. For 6 non-hit subjects, electrodes were placed more posterior, in or bordering on the BNST, as verified by post-hoc histological analysis. These subjects are non-hit subjects and will be used as an internal control, as BNST is a part of the "social brain network" (Greenberg et al., 2010; Lee et al., 2008) and functionally connected to the mPFC (Lebow & Chen, 2016).

Electrodes were positioned in a fixed implant design that interfaced with the connector on the top of the skill, and in turn, the connector interfaced with a battery-powered Neurologger35 chip (1-GB model, New Behaviour) with eight channels (four neural data, two reference, one accelerometer, and one infrared synchronization) and the capability to sample data up to 500 Hz, to wirelessly collect LFP recordings during behavioral experiments. Before behavioral experiments, the Neurologger was programmed with sampling rate and data storage parameters and secured onto the connector on top of the animal's skull as pictured (Figure 1). The sampling

rate was 199.805 Hz for all subjects except hit subject 2 (489.075 Hz). For hit subject 2, linear interpolation was used to generate a time-series with matched sampling frequency as all other subjects.

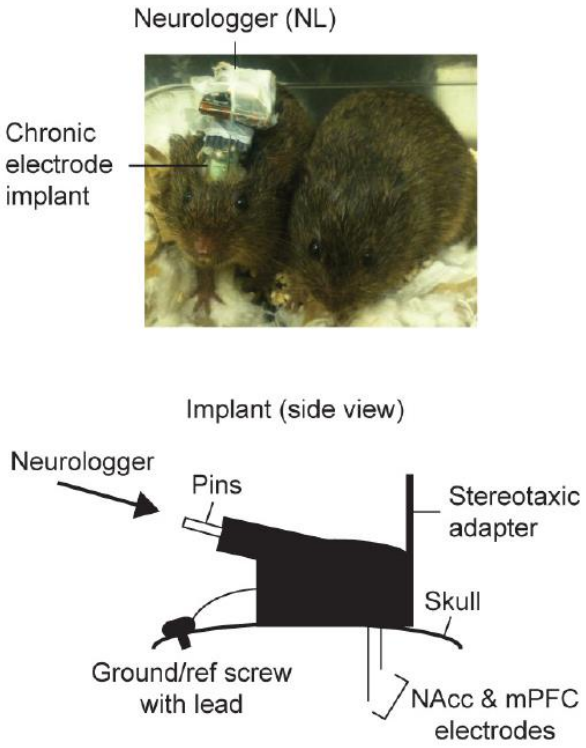



**Figure 1.** Neurologger recording device on a female subject animal during cohabitation with a male. Neurologger interfaces with chronic electrode implantations targeting mPFC and NAcc in hits subjects of mPFC and BNST in non-hits subjects to record LFP signal. Subjects were habituated to this device for at least 1 hour on the day before experiments. Adapted from Amadei et al. (2017).

**LFP Data Collection During Cohabitation**

Prior to behavioral experiments, all female subjects were primed with estradiol benzoate (17-β -estradiol-3-benzoate, Fisher Scientific, daily injections of 1–2 μ g dissolved in sesame oil starting 3-4 days before experiments) to induce socio-sexual interest in males (Donaldson, Spiegel, & Young, 2010). For hit subjects, LFPs were then recorded from the mPFC and NAcc of awake, behaving female subjects during a 6-hour cohabitation period with an adult, sexually experienced male partner animal under 1.5 years of age. For non-hit subjects, LFPs were recorded from the mPFC and within or bordering the BNST of awake, behaving female subjects

during the same behavioral paradigm. Male partner animals were matched by age (within 61 days) and weight (within approximately 5 g) for each female subject.

Subjects with electrodes within or on the medial border of the NAcc were included as hit subjects (n = 9), while subjects with electrodes posterior to the NAcc (within or bordering BNST) were included as non-hit subjects (n = 6). For the purposes of this paper, subjects are ordered from hits 1-9 and non-hits 1-6 by their latency to huddle for a cumulative 5 minutes.

Neural and video recording were performed throughout the baseline, 10–15 minute solo habituation period, and 6-hour cohabitation period, and were synchronized using periodic timestamps delivered every 100 frames (3.3 s) from a Cleversys Topscan system running on a 32-bit Dell Precision T3500 computer. Notably, to account for 60-Hz electrical noise, experiments were performed under a Faraday cage.

Cohabitation videos were behaviorally scored by developing an ethogram to define mating, self-grooming, and huddling behaviors occurring in these experiments (Figure 2). Mating accelerates pair bond formation (Williams, Catania, & Carter 1992), side-by-side huddling is an index of bond expression (Ahern et al. 2009; Lim, et al., 2004), and self-grooming is a self-directed, high-motion control behavior (Figure 2). Notably, behaviors were variable from individual to individual, but not significantly different between hit and non-hit subjects (Amadei et al., 2017).


**Figure 2.** Ethogram definitions of mating, self-grooming, and huddling used to score experimental videos. Adapted from Amadei et al. (2017).

**Overview of Analyses**

The general framework of analyses is depicted in Figure 3. While Amadei et al. (2017) assessed coherence, Granger causality, and cross-frequency coupling between brain regions during mating, self-grooming, and huddling, the spectral features of the LFP data were not assessed on a continuous time scale. As such, previous analyses may not have provided a complete description of behavioral-specific brain-states or spectral features. LFP data from the mPFC and NAcc was first wavelet transformed to generate wavelet spectrograms. Next, the spectrograms were used to construct spectral feature vectors, which were embedded into two dimensions using *t*-SNE (van der Maaten & Hinton, 2008). The probability distribution over this two-dimensional space was estimated, and resolvable peaks were identified in the distribution.

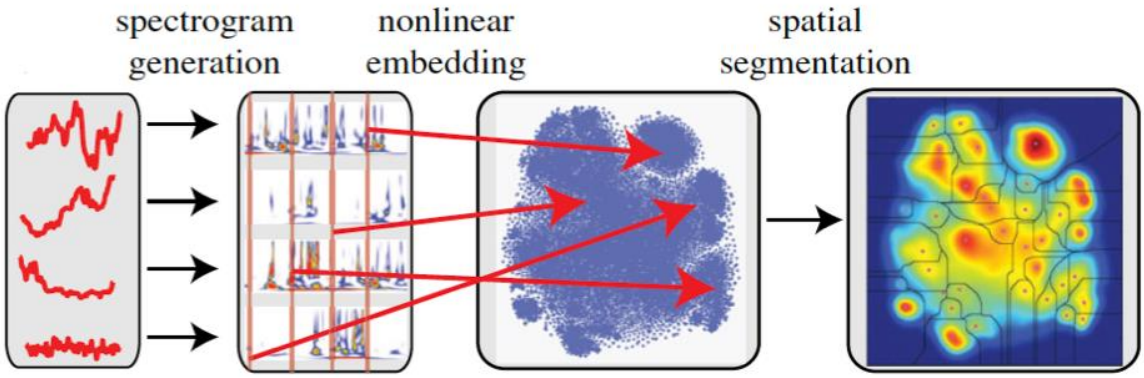



**Figure 3.** Overview of the data analysis pipeline. Raw LFP time-series collected from the mPFC, NAcc (hits), and BNST (non-hits) are collected. A Morlet continuous wavelet transform is applied to each time series for each animal, creating a spectrogram representation of the LFP time-series collected from each brain region. After normalization, each point in time is mapped into a two-dimensional space via *t*-SNE (van der Maaten & Hinton, 2008). Lastly, a watershed transform is applied to the Gaussian-smoothed probability density functions, isolating peaks of highest density to determine regions with unique spectral properties. Adapted from Berman et al. (2014).

**Spectrogram Generation**

As an alternative to Amadei et al.'s approach (2017), we used a spectrogram representation of the neural dynamics, measuring the power, $S(k, f; \tau)$, at a set of frequencies, $f$, for each brain region, $k$, over an interval of time, $\tau$. Each LFP time-series, $y(t)$ was assessed, such that

$$\boldsymbol{Y} = \{y_{mPFC}(t), y_{NAcc}(t)\} \quad (1)$$

where $y(t)$ was approximately 4,300,000 time-points long, (i.e. 6 hours sampled at approximately 200 Hz). The Morlet continuous wavelet transform was then used to provide a multiple time-scale representation of brain-state dynamics. LFP signal was decomposed in terms of time and frequency information. Although a Fourier transform or fast-Fourier transform could have been similarly used to decompose the LFP signal, a Morlet continuous wavelet transform can uniquely extract both frequency and time information to better assess temporal changes in power at multiple frequencies. This allows for more complete analysis of the LFP signal, as the

power at different frequencies at one time-point reflect changes in power at neighboring time points. For this reason, the Morlet wavelet is better suited for isolating short instances or changes of power at specific frequencies (Daubechies, 1992). By measuring the amplitudes of the transform at high temporal and frequency resolution, the need to assess the signal using moving time-windows is eliminated. Details of these calculations are included in Appendix B.

For the analysis, 100 frequency channels, dyadically spaced between 1 and 100 Hz, the latter being the Nyquist frequency, were used. These frequencies were used to ensure analysis of power at all previously identified frequency bands of interest – 5 Hz and 80 Hz (Amadei et al. 2017).

**Spatial Embedding**

$S(k, f; \tau)$ comprises 100 frequency channels for each of the 2 brain regions. We would like to find a low-dimensional representation that captures the important features of the dataset, such that subtle neural dynamics throughout cohabitation can be assessed. The aim of dimensionality reduction of the feature vectors was to construct a space, $\boldsymbol{B}$ of behavioral-specific brain-states at each point in time. We aimed to maximally preserve the local structure of the data, namely, spectral features on small time scales, while simultaneously preserving the general, global structure of the data. An ideal embedding reduces dimensionality by altering the distances between more distant points on the manifold. This was used to assess clustering of similar brain-states and behavioral-specific brain-states over time.

One method that does possess this property is *t*-SNE (van der Maaten & Hinton, 2008). Like other embedding algorithms, *t*-SNE aims to take data from high dimensional space and embed it into a space of much smaller dimensionality, preserving some set of invariants as best as possible. For *t*-SNE, the conserved invariants are related to the Markov transition probabilities

if a random walk is performed on the dataset. Thus, this is a stochastic modeling process. Specifically, we defined the transition probability from time-point $t_i$ to time-point $t_j$, $p_{j|i}$, to be proportional to a Gaussian kernel of the distance (as of yet, undefined) between them

$$p_{j|i} = \frac{exp(-d(t_i,t_j)^2/2\sigma_i^2)}{\sum_{k\neq 1} exp(-d(t_i,t_k)^2/2\sigma_i^2)'} \quad (2)$$

All self-transitions (i.e. $p_{i|i}$) were assumed to be zero. Each of the $\sigma_i$ were set such that all points had the same transition entropy, $H_i = \sum jp_{j|i} \log p_{j|i} = 5$. This can be interpreted as restricting transitions to roughly 32 neighbors.

The *t*-SNE algorithm was then used to embed the data points in the lower-dimensional space while keeping the new low-dimensional set of transition probabilities, $q_{j|i}$ as similar to the high-dimensional set of transition probabilities, $p_{j|i}$ as possible. The $q_{j|i}$ were defined similarly to the high-dimension transition probabilities but were proportional to a Cauchy (or Student-*t*) kernel of the points' Euclidean distances in the embedded space. This algorithm results in an embedding that minimizes local distortions. If $p_{j|i}$ is initially very small or zero, it will place little to no constraint on the relative positions of the two points, but if the original transition probability is large, it will factor significantly into the cost function.

Because this is computationally expensive, it is impossible to incorporate the entire dataset into the embedding, as that would mean incorporating approximately 40,000,000,000 points. Therefore, we used an importance sampling technique to select a training set of approximately 35,000 data points, build the space from these data, and then re-embed the remaining points into the space as best as possible (Appendix C). Roughly 4,380 data points from 8 hit subjects, out of 4,380,000 data points per subject, were used to create a representative set of data. *t*-SNE was performed on 20,000 randomly selected data points from 8 hit subjects,

and the resultant embedding was then used to estimate a probability density by convolving each point with a two-dimensional Gaussian whose width is equal to the distance from the point to its 5 nearest-neighbors. This space was segmented by applying a watershed transform to the inverse of the PDF. Points were grouped by region, and the number of points selected out of each region was proportional to the integral over the PDF in that region. This was performed for 8 hit subject data sets, yielding a total of approximately 35,000 data points in the training set.

Lastly, we defined a distance function, $d(t_i, t_j)$ between the feature vectors. This function should accurately measure the difference between the shapes of two frequency spectra. Simply measuring the Euclidean distance between two spectra will be greatly affected by amplitude modulations. However, because $S(k, f; t)$ is composed of a set of wavelet amplitudes, it must be positive semi-definite. As such, if we define

$$\hat{S}(k, f; t) = \frac{S(k,f;t)}{\sum_{k',f'} S(k',f';t)'} \quad (3)$$

then we can treat this normalized feature vector as a probability distribution over all frequency channels for each brain region at a given point in time. Thus, an appropriate distance function is the $D_{KL}$ (Cover & Thomas, 2006) between two feature vectors in time,

$$\begin{aligned} d(t_1, t_2) &= D_{KL}(t_1 || t_2) \\ &= \sum_{f,k} \hat{S}(k, f; t) \log_2 \left[\frac{\hat{S}(k,f;t_1)}{\hat{S}(k,f;t_2)}\right] \end{aligned} \quad (4)$$

**Jensen-Shannon Divergence**

To assess the extent to which two probability density functions are dissimilar, the Jensen-Shannon divergence (JSD) was used. This is a symmetric and smoothed version of the $D_{KL}$, such that values from 0 to 1 indicate similarity to dissimilarity of the probability density functions, respectively. Specifically, for two probability density functions, $P$ and $Q$, the JSD is defined by

$$JSD(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M), \quad (5)$$

where $D_{KL}(P||M)$ is the $D_{KL}$ from $M$ to $P$, $D_{KL}(Q||M)$ is the $D_{KL}$ from $M$ to $Q$, and $M = \frac{1}{2}(P + Q)$ (Fuglede & Topsøe, 2004; Lin, 1991). This was used to assess intra-animal and intra-behavior variability for hit and non-hit subjects over the full cohabitation period, during the first hour of cohabitation after the first bout of mating, and during the last hour of cohabitation. To determine statistical differences between the intra-animal and intra-behavioral JSD, a Wilcoxon rank-sum test was used (Wilcoxon, 1945). A Wilcoxon signed rank-sum test was then used to determine statistical differences between the intra-animal JSD for the first hour of cohabitation after the first mating bout and for the last hour of cohabitation for hit and non-hit subjects. This was repeated to determine statistical differences between the intra-behavior JSD for the first hour after the first mating bout and for the last hour of cohabitation for hit and non-hit subjects.

## RESULTS

### Structure and Dynamics of the Low-Dimensional Embedded Space

To assess the neural dynamics underlying affiliative social behavior and the development of a pair bond, LFP signals from the mPFC and NAcc were embedded in a low-dimensional space as previously described. First, spectral feature vectors, (e.g. amplitudes of the wavelet-transformed LFPs from mPFC and NAcc), were embedded into a two-dimensional space $(z_1, z_2)$ for eight of nine hit subjects (subjects 1-6, 8, 9) to generate an optimized training set embedding. Neighboring optimized training set points do not exhibit similar log-normalized amplitudes - there is no obvious clustering of nearby points by similar power ($\sum_{k,f} S(k, f; t)$) (Figure 4. A.). The lack of clustering by amplitude indicates nearest-neighbors, and thus structure of the embedded space, are not determined by the amplitude information alone. When

re-embedding all data points from all 9 hit subjects, ~93.6% of points remain in the training set embedding space. The re-embedding of all 9 hit subjects show clear local maxima and minima – peaks and valley – to be assessed for spectral features (Figure 4. B.). These peaks may indicate stereotyped brain-states between or within animals, or stereotyped behavior-specific brain-states between or within animals.

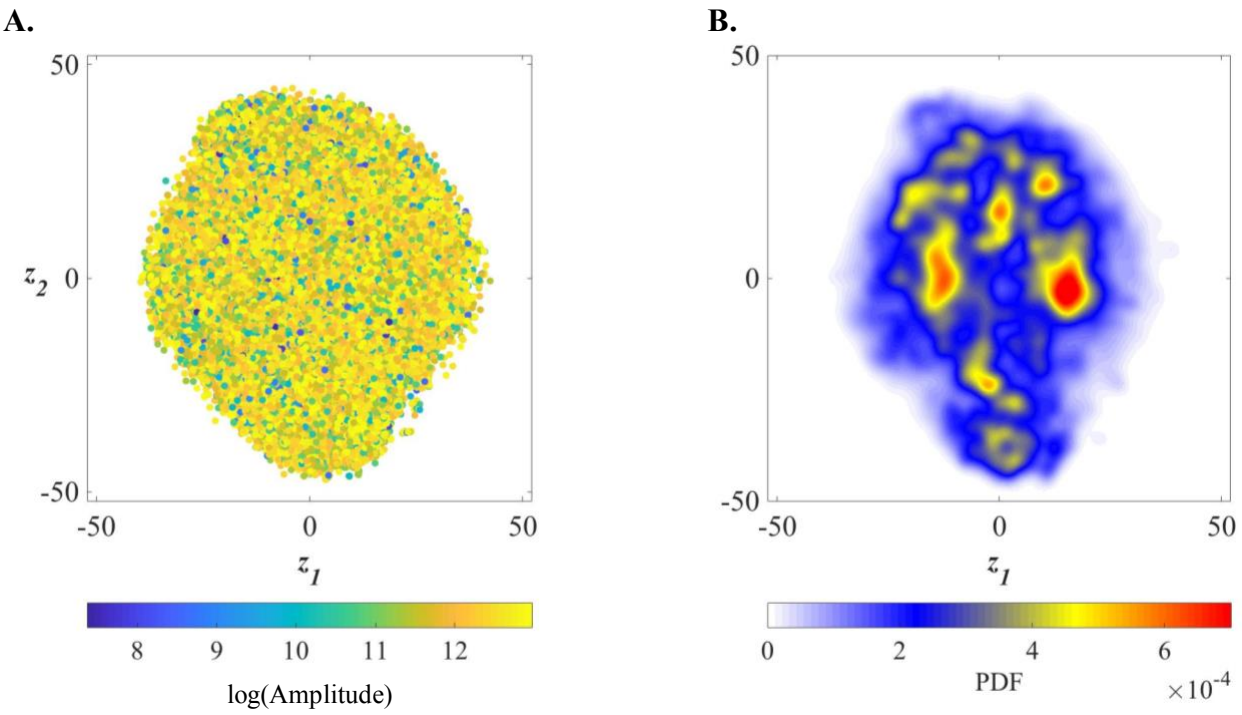



**Figure 4.** Low-dimensional embedding of wavelet-transformed LFP signal from the mPFC and NAcc. **A.** ~35,000 training set points subsampled from 8 hit subjects embedded into two dimensions via *t*-SNE. Coloring is proportional to the logarithm of the training set amplitudes, $\sum_{k,f} S(k, f; t)$. **B.** Probability density function (PDF) generated from embedding all 9 hit subject points via *t*-SNE and convolving all points with a Gaussian ($\sigma$ = 1.3), which represents local maxima of embedded points.

Likewise, when embedding the same spectral feature data into three dimensions using the same parameters, training set points are also not clustered by their amplitude information (Figure 4. A.). The general topology of embedded points in three dimensions remains the same, as the data are distributed along the same axes and exhibit a similar shape (Figure 5. B.). When embedding the data into this higher dimension, the embedding cost function (C1) is reduced by

0.6166 bits and improved by 3% (6.429 bits for two dimensions versus 5.812 for three dimensions compared to the total Shannon entropy calculated for the sparse transition matrix resulting from embedding in two dimensions, 20.717 bits) (Figure 5). This means a similar amount of total information, total entropy, is preserved in both a two- and three-dimensional mapping. In other words, approximately 97% of the data can be explained by the mapping in two-dimensions, and there is no dramatic improvement or unique three-dimensional structure that suggests we should analyze the data in three-dimensions.

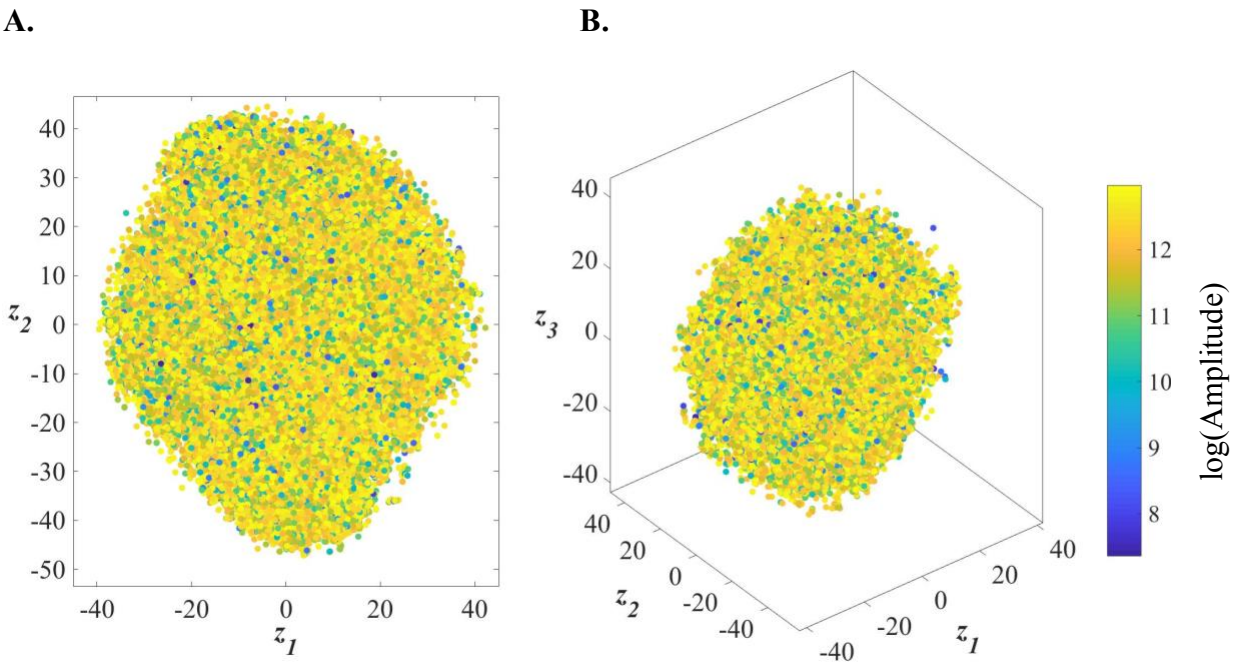



**Figure 5.** Comparison between **A.** two-dimensional embedding and **B.** three-dimensional embedding via *t*-SNE. All parameters remained constant for both embedding processes, except the embedding dimension. Coloring is proportional to the logarithm of the training set amplitudes, $\sum_{k,f} S(k,f;t)$. There is a 3% improvement in the cost of embedding as defined by the KL divergence (C 3) (6.4288 bits for two dimensions versus 5.8122 for three dimensions).

Moreover, an estimate of the probability density function (PDF), $b(\boldsymbol{z})$, was generated by convolving each point in the embedded space with a Gaussian of a small width ($\sigma = 1.3$) (Figure 4. B.). The space, $b(\boldsymbol{z})$ contains a number of resolved local maxima as indicated by regions of high PD. Potentially, the locations of these peaks correspond to stereotyped brain-states, either

behavior- or individual- specific. Throughout the duration of cohabitation, subjects transition through these mapped spaces, as they transition through brain-states.

As such, this low-dimensional embedded space of the LFP spectral features can be assessed in terms of the trajectory and velocity of transitions through the space, meaning the time-course and time it takes for animals to move through brain-states can be assessed. Specifically, the distribution of velocities within the embedded space is well represented by a two-component logarithmic Gaussian mixture model in which the two peaks of the Gaussian distributions are separated by an almost one-and-a-half-fold increase (Figure 6). The distribution of points in the low-velocity peak (approximately 33% of all time-points, $\mu = 2.0287$, $\sigma = 0.3561$) is separable from the distribution of the points in the high-velocity peak (approximately 33% of all time-points, $\mu = 3.0224$, $\sigma = 0.0612$). This suggests future analyses can be conducted to assess the temporal dynamics of brain-states during affiliative social behavior, namely, analyses of brain-state transition matrixes over different time scales and during different behaviors (Berman et al., 2016). Questions probing the temporal emergence of stereotyped neural dynamics over the course of pair bonding, and how the temporal structure of the neural dynamics bias organization of behaviors, can be further assessed.

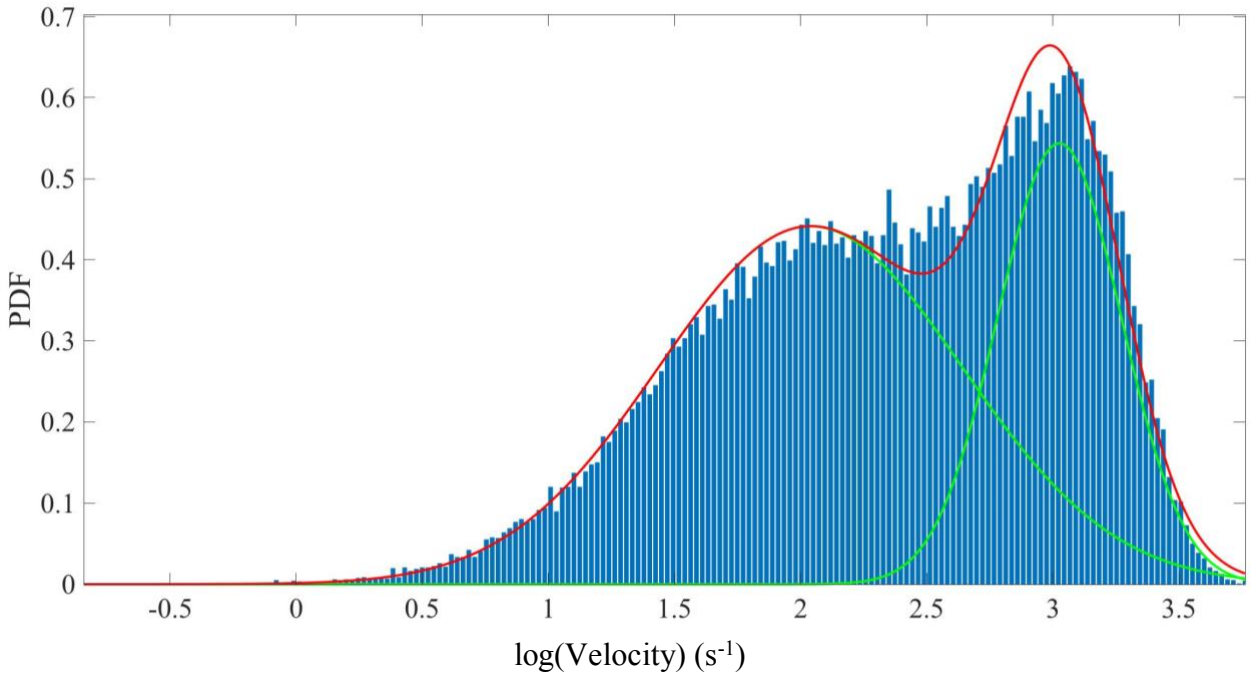



**Figure 6.** Histogram of velocities within the embedded behavioral space for the re-embedding of all data from all 9 hit subjects fitted to a two-component log-Gaussian mixture model. The blue bar chart represents the measured probability distribution, the red line is the fitted model, and the cyan and green lines are the mixture components of the fitted model. This represents two separable states of slow and fast velocity – resting and transition states through the embedded space.

**Intra-Animal and Intra-Behavior Comparisons**

All points for each hit and non-hit subjects during all behaviors, huddling, mating, self-grooming, and other (all behaviors not huddling, mating, or self-grooming), were embedded in a two-dimensional map to assess whether behaviors and corresponding brain-states are stereotyped across behaviors or across individuals. This is done to assess how variability between individual animals' brain-states across all behaviors compares to variability between behavior-specific brain-states across all animals. PDFs for each behavior, huddling, mating, self-grooming, and other, were compared pairwise via the JSD (5) and averaged for each hit (Figure 7. A.) and non-hit subject (Figure 7. B.), (i.e. the average JSD between pairs of behavior-specific maps for each

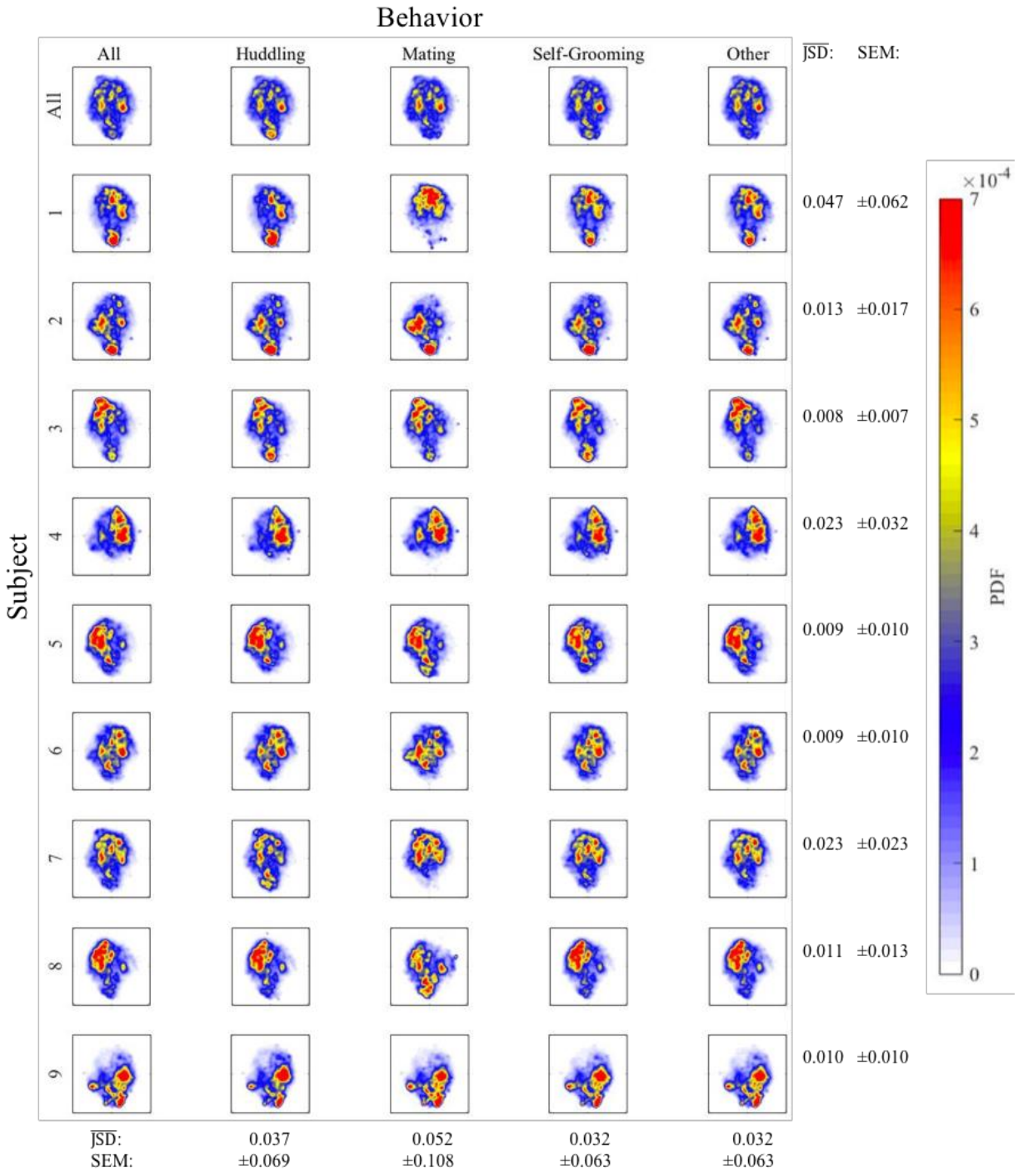
animal, across one row of maps, was computed). This is a measure of how dissimilar the behavior-specific PDFs are for each animal, reflecting intra-animal variability of PDFs. Furthermore, PDFs for one behavior were compared pairwise via the JSD (5) and averaged for all hit (Figure 7. A.) and non-hit subject (Figure 7. B.), (i.e. the average JSD between pairs of animal-specific maps for each behavior, down one column, was computed). This is a measure of how dissimilar the animal-specific PDFs are for each behavior, reflecting intra-behavior variability of PDFs.

For hit subjects, the median intra-behavior variability is significantly greater than the median intra-animal variability, suggesting mPFC and NAcc brain-states are more dissimilar between individuals than between behaviors for hit subjects ($p = .020$) (Figure 8. A.). Individual signatures may thus washout any effect seen by the behaviors. However, for non-hit subjects, the median intra-animal variability is significantly greater than the median intra-behavior variability, suggesting mPFC and within or bordering BNST brain-states are more dissimilar between behaviors than between individuals ($p = .001$) (Figure 8. B.).

All points for each hit and non-hit subject during all behaviors for the first hour of cohabitation after the first mating bout (Figure 9. A. and B.) and for the last hour of cohabitation (Figure 10. A. and B.) were embedded in a two-dimensional map to assess whether brain-states become stereotyped or distinct across behaviors or across individuals over time. The average JSD between PDFs for hit and non-hit subjects was computed as described above for the first hour after the first mating bout and for the last hour of cohabitation. For hit subjects, there are no significant differences between intra-animal and intra-behavior variability for the first hour after the first mating bout and for the last hour of cohabitation. Within both intra-animal and intra-behavior variability for hit subjects, there are no significant differences between the first hour

after the first mating bout and the last hour of cohabitation. (Figure 11. A.) For non-hit subjects, there are no significant differences between intra-animal and intra-behavior variability for the last hour of cohabitation (Figure 11. B.). Within both intra-animal and intra-behavior variability for non-hit subjects, there are no significant differences between the first hour after the first mating bout and the last hour of cohabitation. However, for the first hour after the first mating bout, the median intra-behavior variability is significantly greater than the median intra-animal variability ($p = 010$) (Figure 11. B.).

**A.**


Behavior
All
Huddling
Mating
Self-Grooming
Other
JSD:
SEM:
Subject
All
1
2
3
4
5
6
7
8
9
0.047 ±0.062
0.013 ±0.017
0.008 ±0.007
0.023 ±0.032
0.009 ±0.010
0.009 ±0.010
0.023 ±0.023
0.011 ±0.013
0.010 ±0.010
0.037 ±0.069
0.052 ±0.108
0.032 ±0.063
0.032 ±0.063
×10^-4
7
6
5
4
3
2
1
0
PDF

**B.**

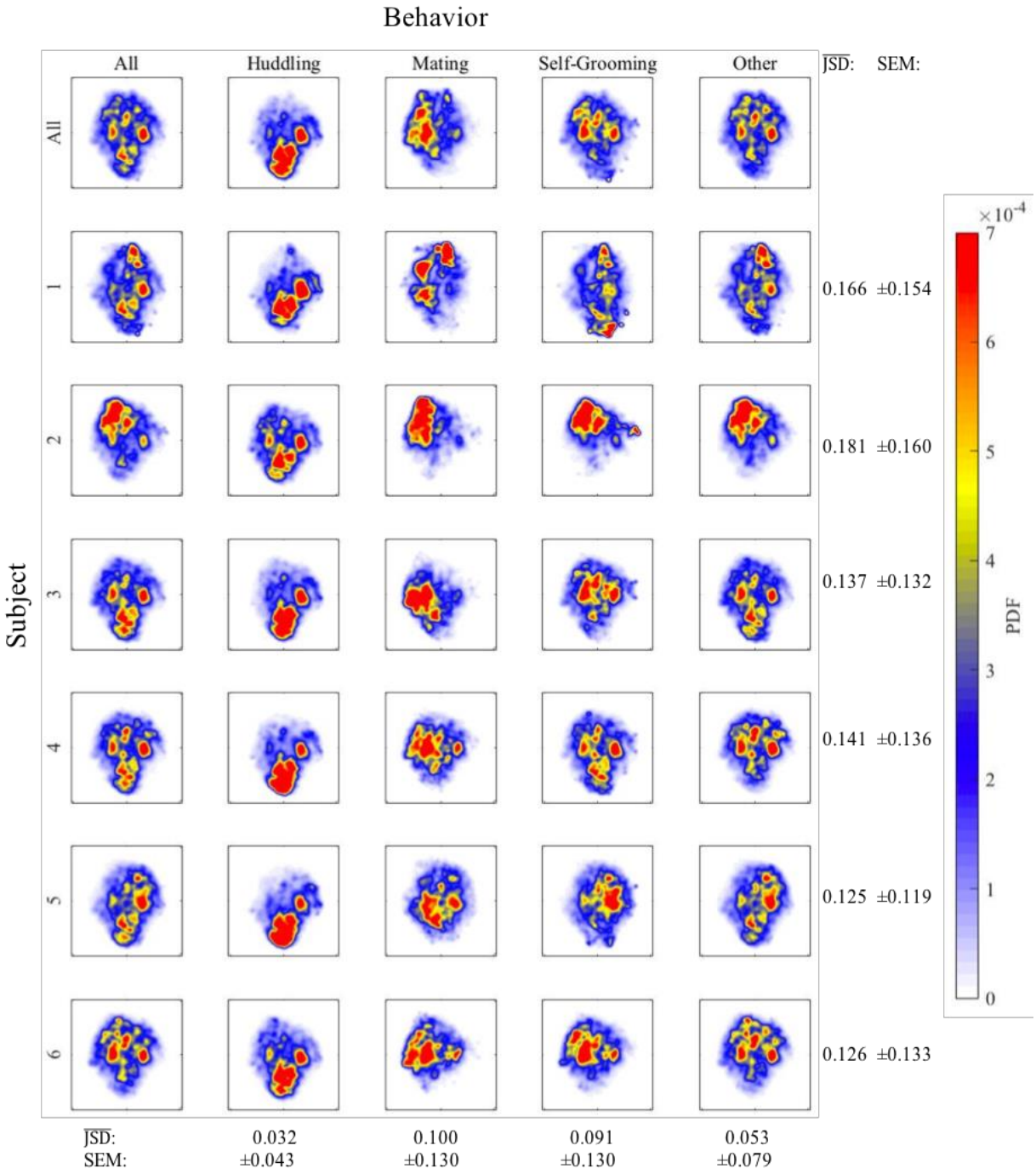
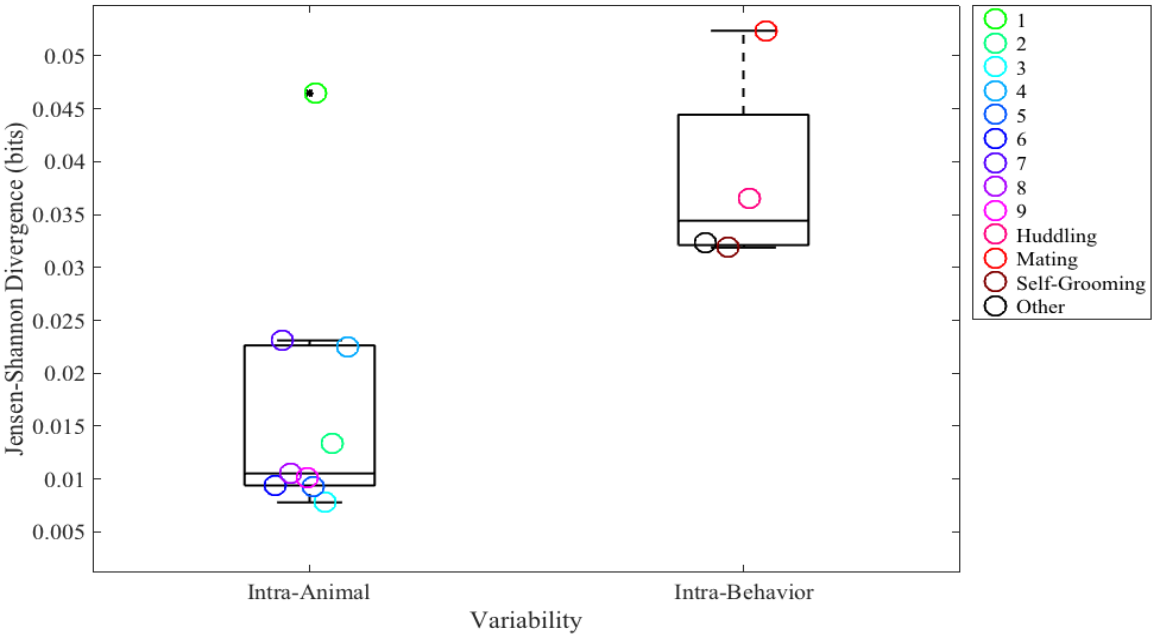



**Figure 7.** Jensen-Shannon divergence (bits) computed for all unique pairs of individual behaviors across all **A.** 9 hit subjects and **B.** 6 non-hit subjects ordered vertically from least to greatest latency to huddling for a total of five minutes, as defined by Amadei et al. (2017). This is a measure of dissimilarity between compared PDFs wherein values ranging from 0 to 1 indicate similar to dissimilar PDFs. The average intra-animal variability is located in the last column for each animal, and the average intra-behavior variability is located in the last row for each behavior.

**A.**




**B.**

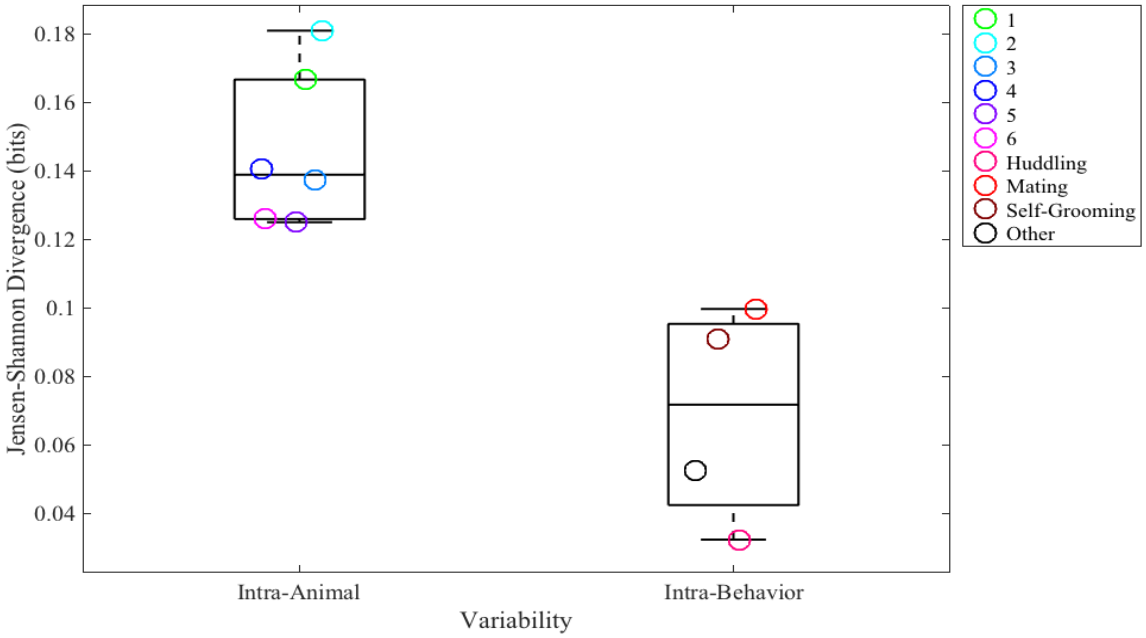
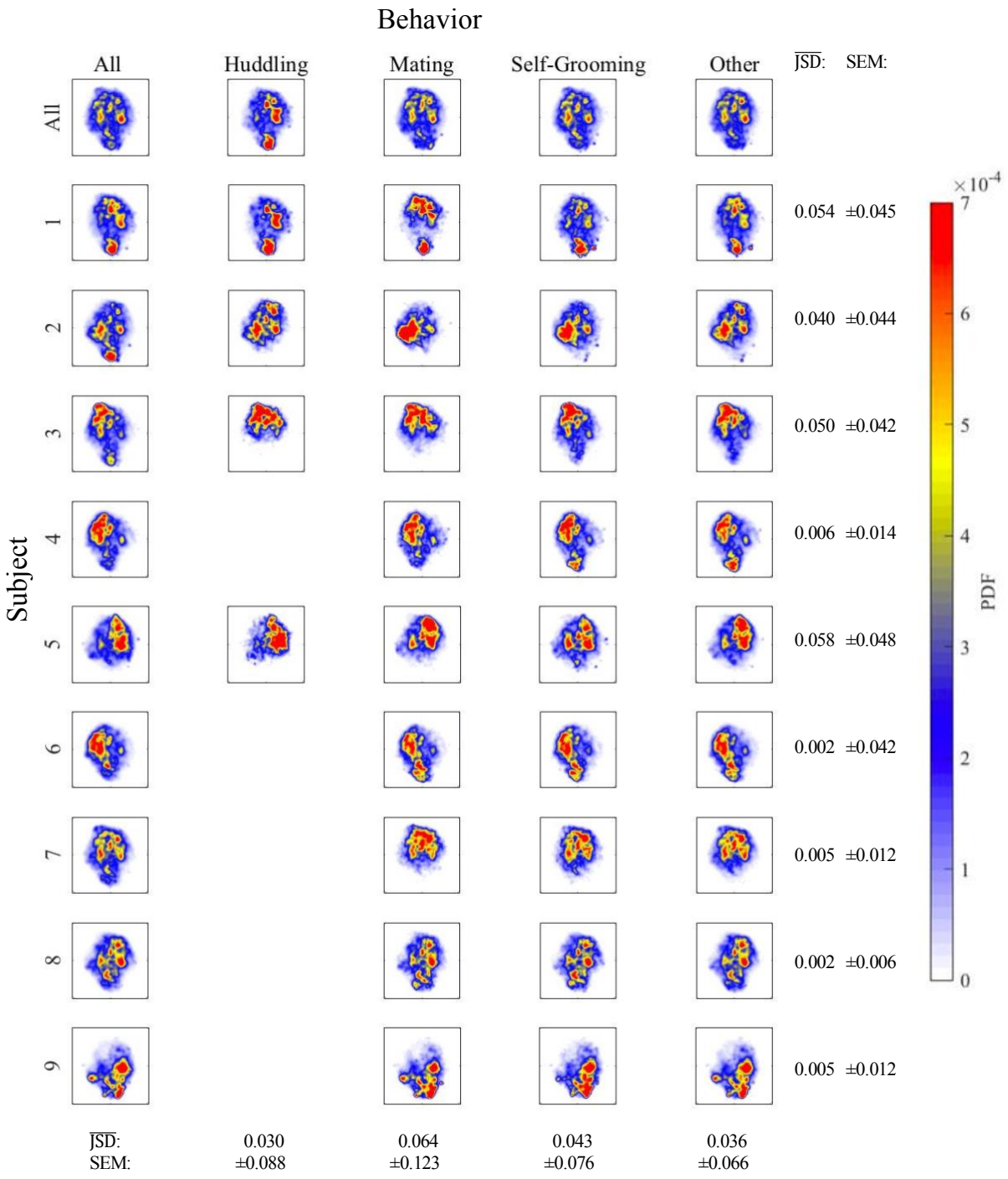



**Figure 8.** Jensen-Shannon divergence (bits) computed for all unique pairs of individual behaviors across all **A.** 9 hit subjects and **B.** 6 non-hit subjects. This is a measure of dissimilarity between compared PDFs wherein values ranging from 0 to 1 indicate similar to dissimilar PDFs. **A.** For hit subjects, the median intra-behavior variability (Median = 0.034) is significantly greater than the intra-animal variability (Median = 0.011) as determined by the Wilcoxon rank-sum test, p = .020. **B.** For non-hit subjects, the intra-animal variability (Median = 0.139) is significantly greater than the intra-behavior variability (Median = 0.072) as determined by the Wilcoxon rank-sum test, p = .001.

**A.**


Behavior
All
Huddling
Mating
Self-Grooming
Other
JSD: SEM:
Subject
All
1
2
3
4
5
6
7
8
9
0.054 ±0.045
0.040 ±0.044
0.050 ±0.042
0.006 ±0.014
0.058 ±0.048
0.002 ±0.042
0.005 ±0.012
0.002 ±0.006
0.005 ±0.012
JSD: 0.030 0.064 0.043 0.036
SEM: ±0.088 ±0.123 ±0.076 ±0.066
×10^-4
7
6
5
4
3
2
1
0
PDF

**B.**

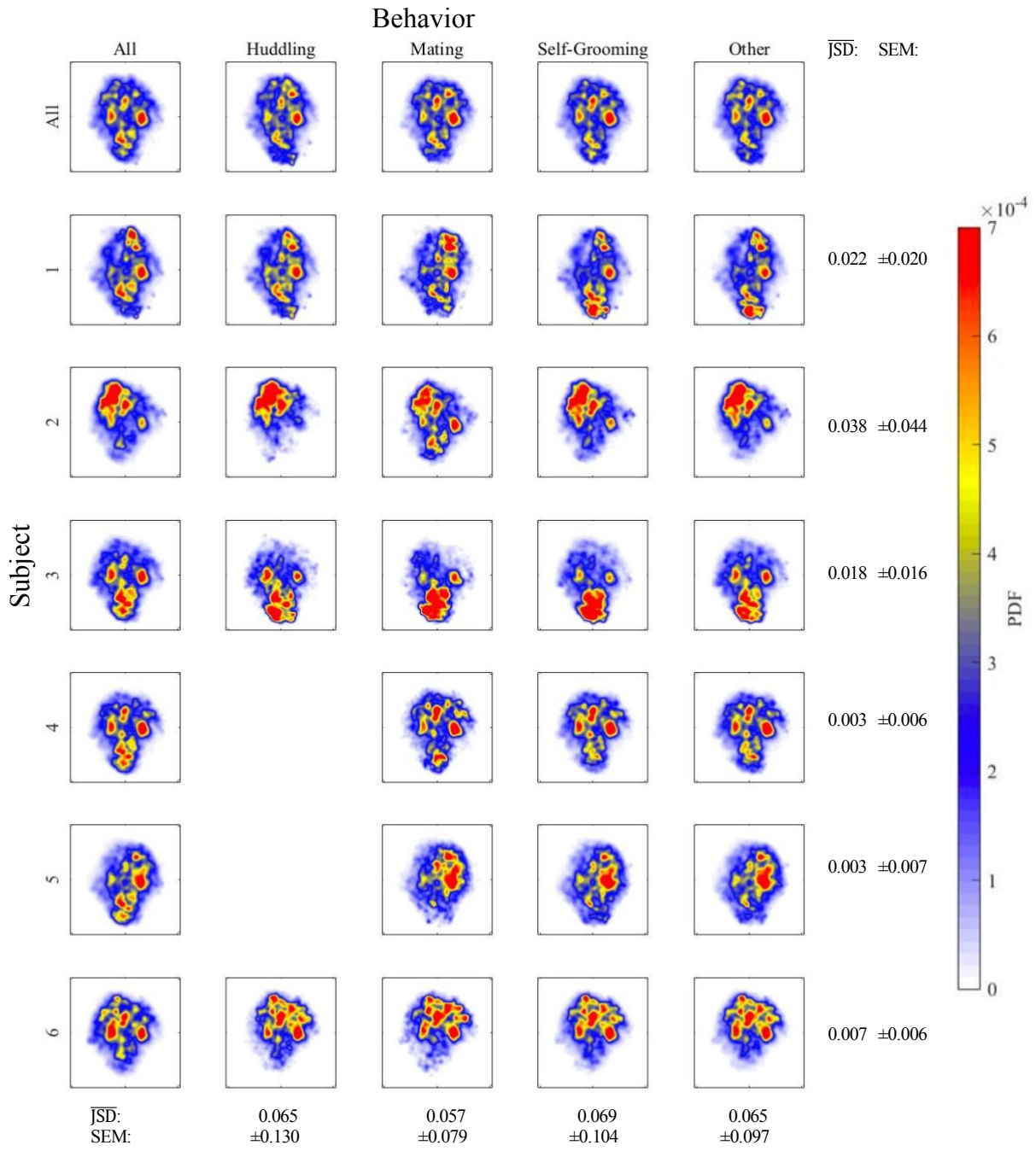
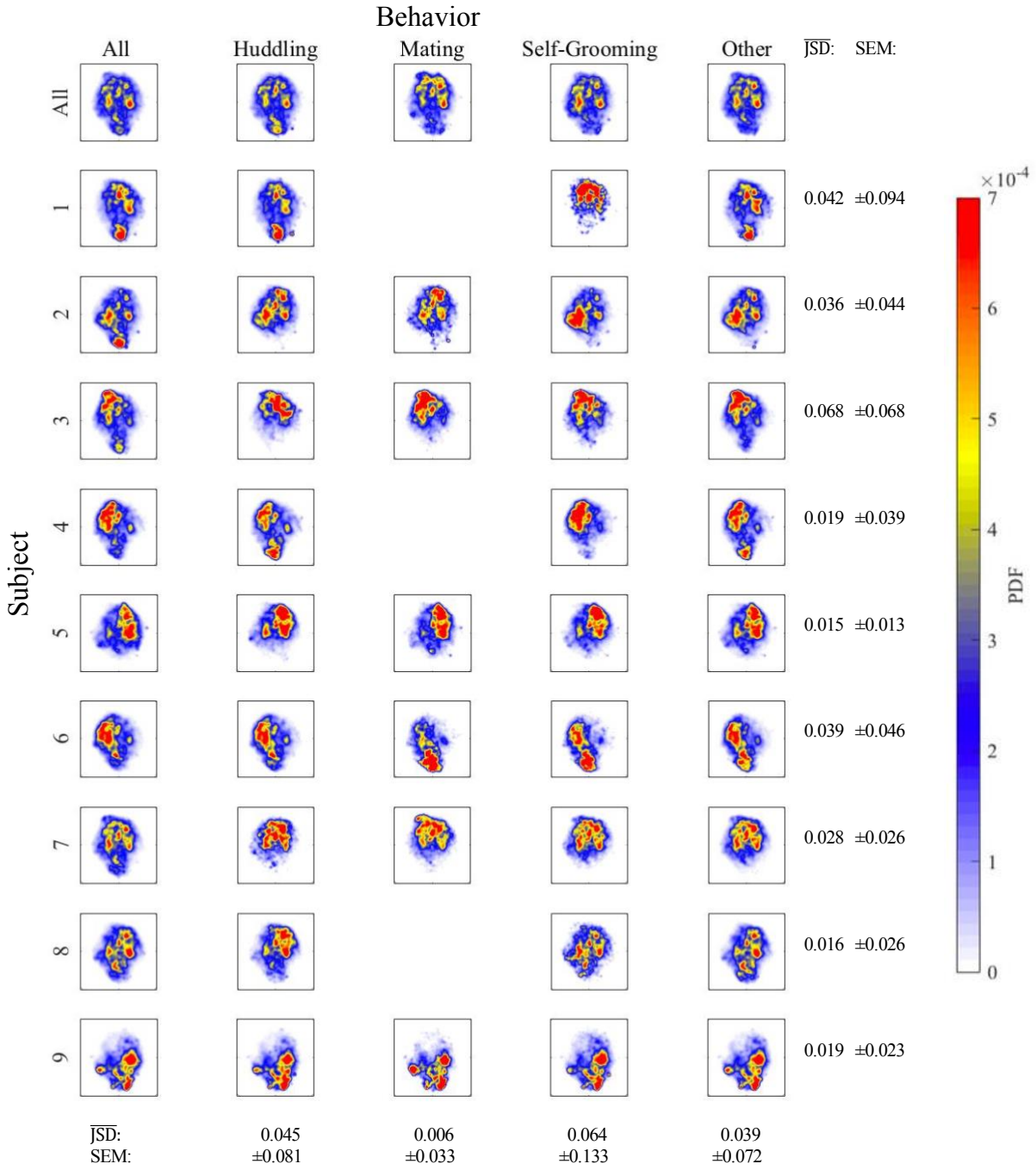



**Figure 9.** Jensen-Shannon divergence (bits) computed for all unique pairs of individual behaviors across all **A.** 9 hit subjects and **B.** 6 non-hit subjects for the first hour of cohabitation after the first bout of mating. Subjects are ordered vertically from least to greatest latency to huddling for a total of five minutes, as defined by Amadei et al. (2017). This is a measure of dissimilarity between compared PDFs wherein values ranging from 0 to 1 indicate similar to dissimilar PDFs. The average intra-animal variability is located in the last column for each animal, and the average intra-behavior variability is located in the last row for each behavior.

**A.**


Behavior
All
Huddling
Mating
Self-Grooming
Other
JSD:
SEM:
Subject
All
1
2
3
4
5
6
7
8
9
0.042 ±0.094
0.036 ±0.044
0.068 ±0.068
0.019 ±0.039
0.015 ±0.013
0.039 ±0.046
0.028 ±0.026
0.016 ±0.026
0.019 ±0.023
JSD: 0.045 0.006 0.064 0.039
SEM: ±0.081 ±0.033 ±0.133 ±0.072
×10^-4
7
6
5
4
3
2
1
0
PDF

**B.**

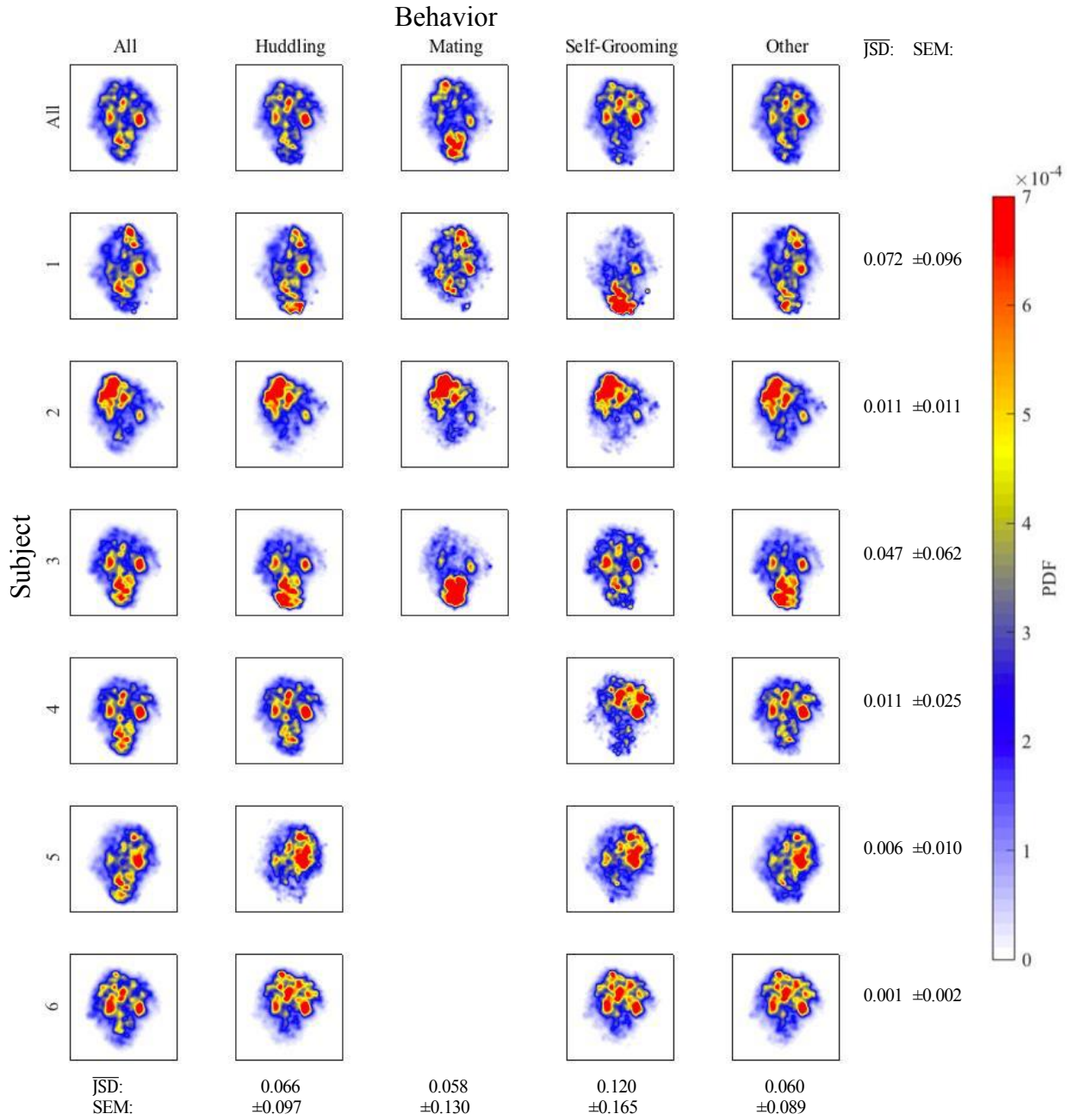
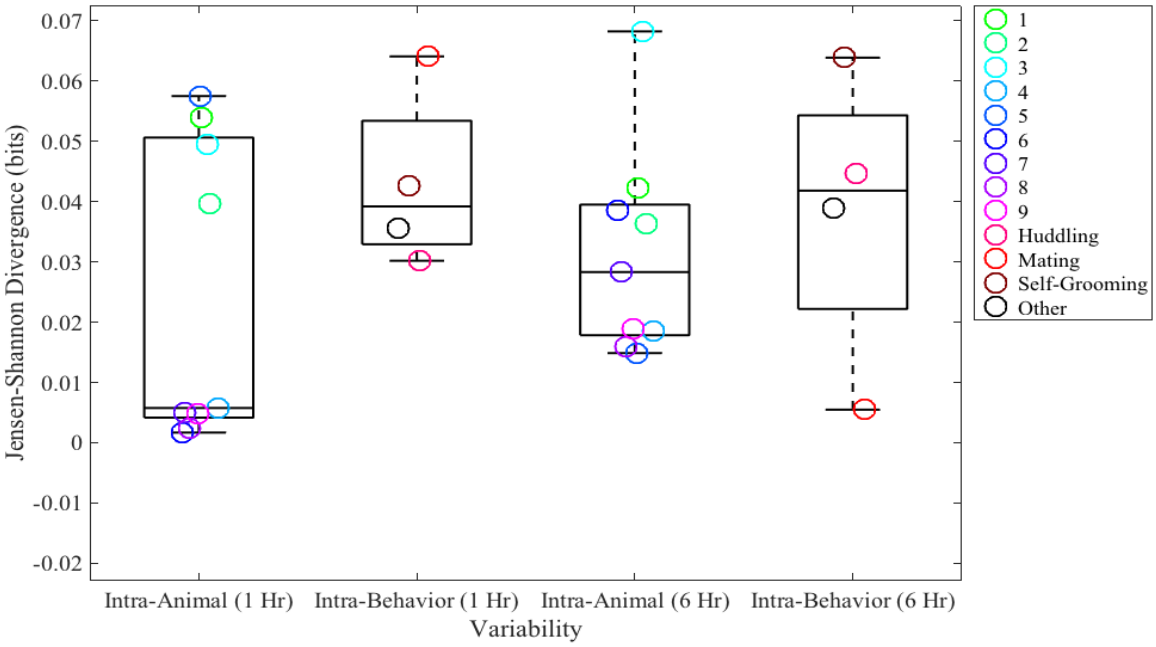



**Figure 10.** Jensen-Shannon divergence (bits) computed for all unique pairs of individual behaviors across all **A.** 9 hit subjects and **B.** 6 non-hit subjects for the last hour of cohabitation. Subjects are ordered vertically from least to greatest latency to huddling for a total of five minutes, as defined by Amadei et al. (2017). This is a measure of dissimilarity between compared PDFs wherein values ranging from 0 to 1 indicate similar to dissimilar PDFs. The average intra-animal variability is located in the last column for each animal, and the average intra-behavior variability is located in the last row for each behavior.

**A.**




**B.**

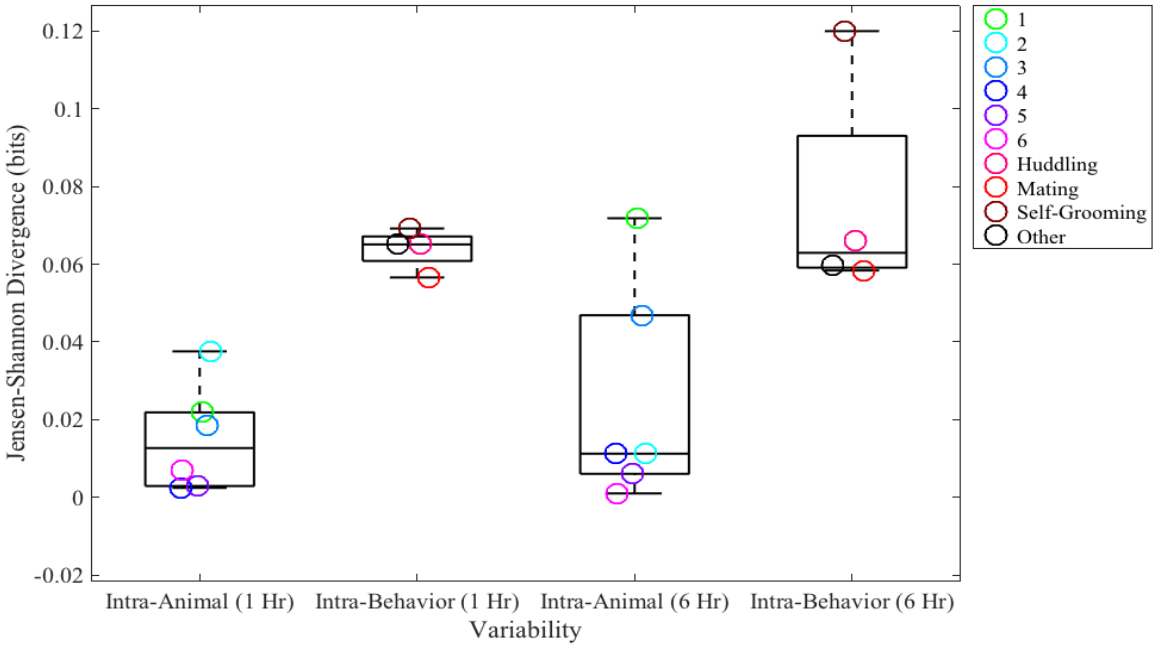



**Figure 11.** Jensen-Shannon divergence (bits) computed for all unique pairs of individual behaviors across all **A.** 9 hit subjects and **B.** 6 non-hit subjects for the first hour after the first mating bout (1 Hr) and last hour of cohabitation (6 Hr). This is a measure of dissimilarity between compared PDFs wherein values ranging from 0 to 1 indicate similar to dissimilar PDFs. **A.** For hit subjects, there are no significant differences between the median intra-animal or intra-behavior variability for 1 Hr or 6 Hr. **B.** For non-hit subjects, the median intra-behavior variability (Median = 0.065) is significantly greater than the median intra-animal variability (Median = 0.013) as determined by the Wilcoxon rank-sum test, p = .010**.**

**Identified Brain-States Corresponding to Regions of the Embedded Space**

The embedded space of the LFP data is comprised of peaks surrounded by valleys – local maxima and surrounding local minima – as indicated by the coloring of the PDFs (Figure 12. A.). By finding connected areas in the $(z_1, z_2)$ plane such that climbing up the gradient of PD consistently leads to the same local maximum, (i.e. conducting a watershed transform) (Meyer, 1994), 36 separable regions were identified (Figure 12. B.). Each of the identified regions contains a single local maximum of PD. Unique peak frequencies of the PSD for each brain region characterize the 36 unique regions where there is an identified local maximum of PD (Figure 12). Mean power of the mPFC, mean power of the NAcc, net mean power (mean mPFC power for all frequencies – mean NAcc power for all frequencies), peak frequency at which mPFC power is maximum, and peak frequency at which NAcc power is maximum were identified for each region (Table 1). Unique segmented regions of local maximum PDs are clustered by PSD similarity, in that the shape of the PSD between brain regions is the most similar feature (Figure 14, Figure D. 1 - 36). There is no particular pattern of mean wavelet coefficients and the corresponding standard errors, mean PSD of the mPFC, mean PSD of the NAcc, or net mean PSD; however, the combinations of the peak frequencies at which mPFC and NAcc power are maximum is unique for each identified region, (Table 1, Figure D. 1 - 36). Thus, the frequency at which the LFP signal is strongest, namely, power is maximum, is the most salient and separable feature for each identified region in the low-dimensional map.

Specifically, all regions correspond to similar power spectral densities (PSDs). This reflects how we defined our parameters, in that, regions should be clustered by similar power spectral density – this confirms *t*-SNE reliably distinguishes regions with separable spectral features. 12 of the 36 regions correspond to exact pairings of peak frequencies, meaning, for 12

of 36 regions, both mPFC and NAcc exhibited the maximum power at the same frequency. Furthermore, 19 of the 36 regions exhibited nearly-paired peak frequencies wherein the paired frequencies differed by up to 12 Hz; these regions will hereafter be referred to as nearly-paired. 7 of the 36 regions exhibited frequencies differing by greater than 10 Hz: 5, 10, 11, 12, 17, 18, 27, 31, and 32. First, the regions corresponding to strongest, paired and nearly-paired delta power (1 to 3 Hz) are 8, 9, 16, 19, 20, 29. Regions corresponding to strongest, paired and nearly-paired theta power (4 to 10 Hz) are 1, 2, 3, 4, 13, 14, 15, 22, 23, 24, 25, 26, 28, 30, 33, 34, 35, and 36. Additionally, regions corresponding to strongest, paired and nearly-paired gamma power (40 to 100 Hz) are 6, 7, 10, 12, and 21. Interestingly, region 5 contains points exhibiting highest power in the high-gamma band (79.248 Hz) in the mPFC, and in the theta band in NAcc. Furthermore, regions 17, 18, and 27 contain points exhibiting highest power in the theta band (approximately 4 Hz to 9 Hz) in the mPFC and in the high-gamma band (approximately 59 Hz to 79 Hz) in the NAcc. Regions 31 and 32 contain points exhibiting highest power in the delta band (approximately 1 Hz) in the mPFC and in the low-gamma band (approximately 27 to 36 Hz) in the NAcc. The only region corresponding to highest power in the high-gamma band (approximately 76 Hz) in the mPFC and in the low-gamma band (approximately 21 Hz) in the NAcc is region 11. Overall, pairwise strongest frequency band power determines clustering in the low-dimensional embedding. Next, the extent to which pairwise strongest frequency band power may be used to classify individuals or behaviors must be assessed. Given LFP data and mapping of the LFP data, it may be possible to predict what brain-state one individual is in or what behaviors that individual is expressing.

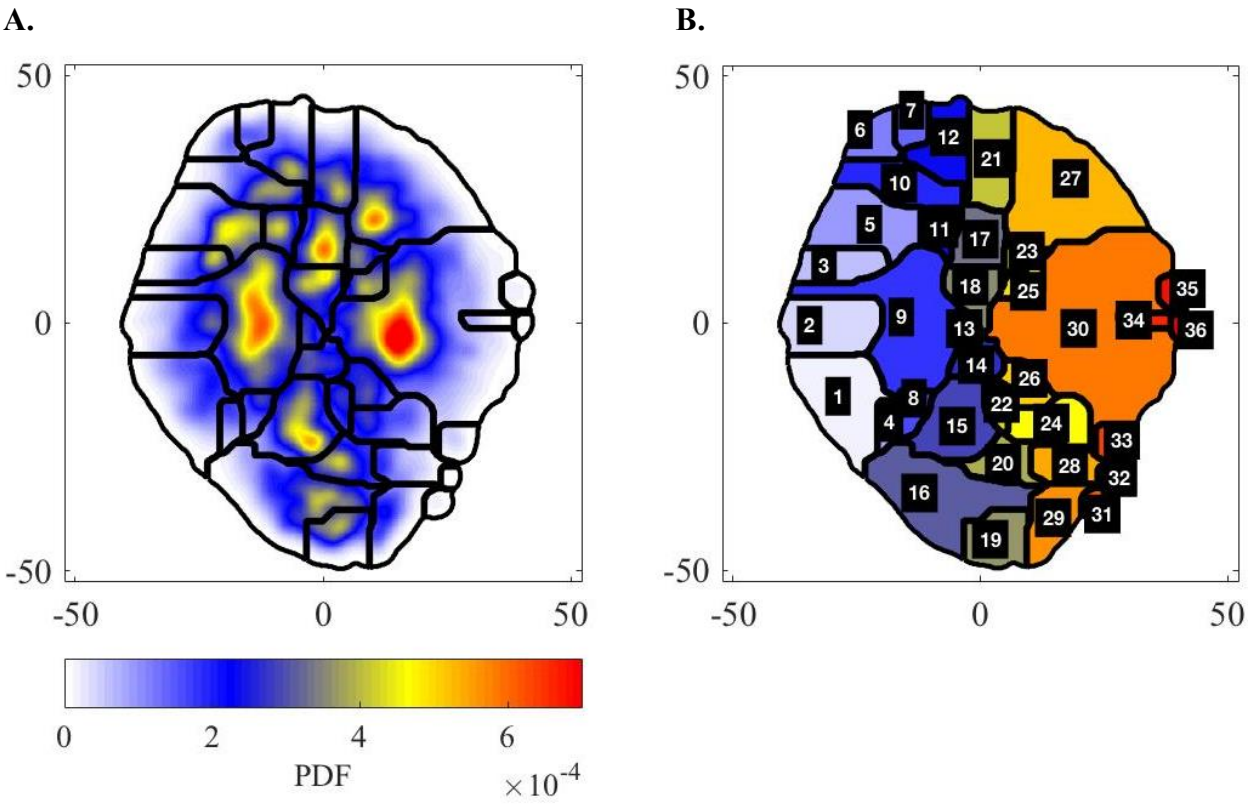



**Figure 12.** Segmentation into regions via a watershed transform. **A.** Boundary lines were obtained from performing a watershed transform on the PDF of the re-embedding of all 9 hit subjects data. The number of regions, unique local maxima, identified were 36. **B.** Labeled watershed map identifying 36 regions of unique local maximum PD. Here, coloring does not mean anything.

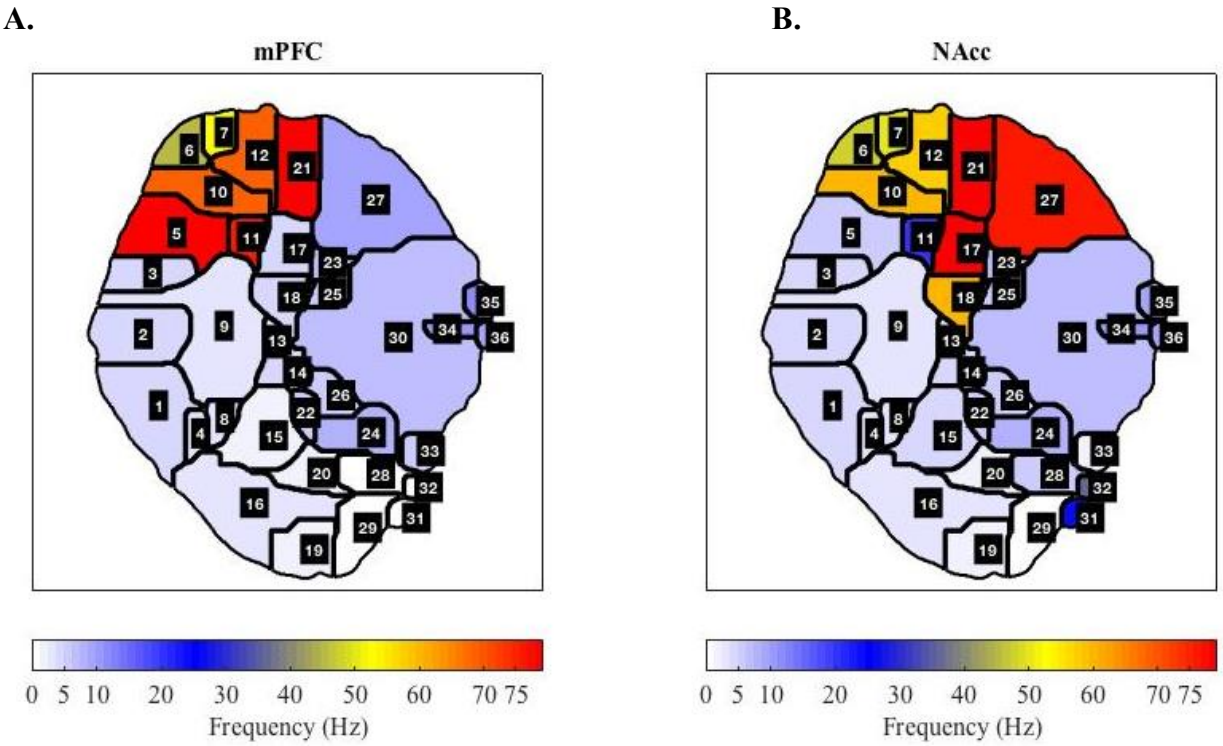



**Figure 13.** Labelled segmentation of the PDF for the re-embedding of 9 hit subjects into 36 regions via a watershed transform. Coloring corresponds to the peak frequency of the power spectrum for that region. **A.** The spectral features of LFPs recorded from the mPFC per region and **B.** The spectral features of LFPs recorded from the NAcc.

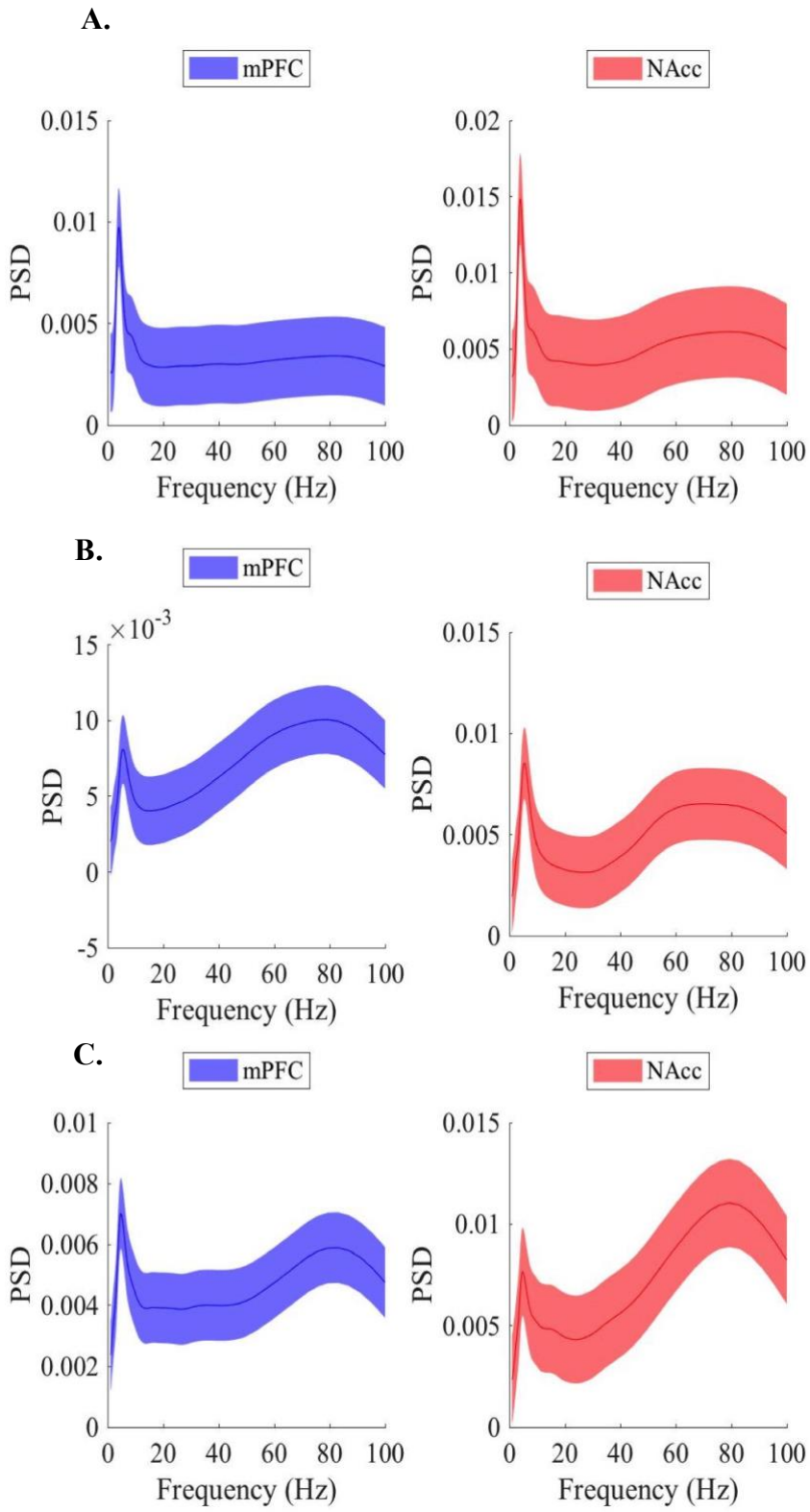



**Figure 14.** Examples of PSDs for signal from each brain region, mPFC and NAcc, within specific segmented regions as follows: **A.** Region 1, **B.** Region 5, and **C.** Region 17 in **Figure 13**. Coloring is blue for the mPFC and red for the NAcc. All PSDs for each identified region can be found in Appendix D.

**Table 1.** Summary of identified region features – maximum power for mPFC signal, maximum power for NAcc signal, peak frequency for mPFC signal, peak frequency for NAcc signal, and net mean power (mean power for mPFC signal – mean power for NAcc signal). Values within the bolded box represent the peak frequencies for each brain region, which are the most distinct features between segmented regions.

| Region | mPFC Max PSD | NAcc Max PSD | mPFC Peak F (Hz) | NAcc Peak F (Hz) | Net Mean PSD |
|---|---|---|---|---|---|
| 1 | 0.010 | 0.015 | 4.037 | 3.854 | -0.192 |
| 2 | 0.010 | 0.012 | 3.854 | 3.854 | -0.020 |
| 3 | 0.009 | 0.010 | 4.229 | 4.229 | 0.063 |
| 4 | 0.008 | 0.013 | 3.678 | 3.511 | -0.202 |
| 5 | 0.010 | 0.009 | 79.248 | 5.337 | 0.084 |
| 6 | 0.008 | 0.017 | 45.349 | 47.508 | -0.155 |
| 7 | 0.009 | 0.018 | 52.140 | 49.770 | -0.119 |
| 8 | 0.008 | 0.011 | 2.535 | 2.535 | -0.110 |
| 9 | 0.007 | 0.008 | 2.915 | 2.915 | -0.009 |
| 10 | 0.011 | 0.008 | 68.926 | 59.948 | 0.083 |
| 11 | 0.008 | 0.007 | 75.646 | 20.565 | 0.028 |
| 12 | 0.010 | 0.012 | 68.926 | 57.224 | -0.030 |
| 13 | 0.007 | 0.008 | 3.511 | 3.511 | -0.162 |
| 14 | 0.007 | 0.008 | 4.037 | 3.854 | -0.092 |
| 15 | 0.008 | 0.007 | 2.310 | 4.863 | 0.013 |
| 16 | 0.011 | 0.008 | 2.783 | 2.915 | 0.075 |
| 17 | 0.007 | 0.011 | 4.641 | 79.248 | -0.118 |
| 18 | 0.006 | 0.007 | 4.863 | 59.948 | -0.117 |
| 19 | 0.016 | 0.009 | 1.748 | 1.748 | 0.133 |
| 20 | 0.010 | 0.006 | 1.668 | 1.668 | 0.081 |
| 21 | 0.008 | 0.015 | 79.248 | 79.248 | -0.122 |
| 22 | 0.007 | 0.008 | 6.136 | 6.136 | -0.020 |
| 23 | 0.009 | 0.010 | 7.055 | 7.055 | -0.063 |
| 24 | 0.008 | 0.009 | 7.743 | 8.498 | -0.062 |
| 25 | 0.009 | 0.010 | 5.857 | 6.136 | -0.107 |
| 26 | 0.009 | 0.009 | 5.857 | 6.136 | 0.027 |

| | | | | | |
|---|---|---|---|---|---|
| 27 | 0.007 | 0.010 | 9.326 | 75.646 | -0.151 |
| 28 | 0.009 | 0.010 | 1.048 | 5.857 | -0.159 |
| 29 | 0.012 | 0.013 | 1.150 | 1.205 | -0.125 |
| 30 | 0.007 | 0.010 | 6.428 | 6.428 | -0.160 |
| 31 | 0.015 | 0.012 | 1.000 | 27.186 | -0.540 |
| 32 | 0.012 | 0.009 | 1.098 | 35.938 | -0.047 |
| 33 | 0.005 | 0.014 | 5.857 | 1.668 | -0.413 |
| 34 | 0.011 | 0.013 | 10.235 | 10.235 | -0.001 |
| 35 | 0.013 | 0.005 | 11.233 | 9.326 | 0.341 |
| 36 | 0.010 | 0.014 | 9.426 | 9.326 | -0.048 |

**DISCUSSION**

Here, it has been shown that the neural dynamics, namely corticostriatal signal recorded over the course of prairie vole pair bonding can be mapped in a low-dimensional space (Figure 4. B.). That mapping can further be assessed by segmenting regions that contain local maxima (Figure 12) and by assessing the speed at which animals transition between these regions (Figure 6). With this mapping, we can assess subtle dynamics and structure of the signal that differ between and within individuals during specific behaviors. We can further assess subtle dynamics and structure of the signal that emerge over time, by mapping the neural dynamics at different time-scales, such as one hour after the first mating bout compared to the last hour of cohabitation to assess the temporal emergence of behavior-specific brain-states over the course of pair bonding. Given LFP data and this mapping of the LFP data, it may be possible to predict what brain-state one individual is in or what behaviors that individual is expressing, conducting neural decoding of the LFP signal to understand and identify the neural dynamics contributing to pair bond formation, and social behavior at large.

For hit subjects, intra-behavior variability supersedes that of intra-animal variability (Figure 8. A.), meaning individual subject differences in the neural signal from mPFC and NAcc

during behaviors were greater than the difference between neural signal exhibited during specific behaviors. This may be explained by variable electrode placement, which must be examined by sorting individuals by most anterior to posterior electrode placement. It is possible that the more similar the electrode placement is between animals, the more similar the corresponding maps may be. Although Amadei et al. (2017) did not observe a significant effect of electrode placement on mean non-huddling net modulation, this does not necessarily mean electrode placement cannot account for subtle differences in signal because Amadei et al. (2017) assayed a specific feature of the signal – the amount signal is modulated by low-frequency phase-high-frequency amplitude coupling from the mPFC to NAcc and NAcc to mPFC. Analyses utilized here more generally considered the spectral features of the signal throughout all points in time of the recording.

However, for non-hits subjects, intra-animal variability supersedes that of intra-behavior variability (Figure 8. B.), meaning the difference between neural signal from mPFC and within or bordering BNST exhibited during specific behaviors was less than individual subject differences in the neural signal. This is particularly interesting because the BNST is also a brain-region considered part of the "social brain network", in that it works to mediate social anxiety and support recognition of conspecifics (Greenberg et al., 2010; Lee et al., 2008). Perhaps there are signature neural features of signal from the BNST to be identified for future use to understand and decode social behaviors, given neural signal. Future work to assess this include identifying the regions present in non-hits animals by using the established mappings of PDFs via *t*-SNE and segmented regions via watershed transform. Overall, by using the mapping and segmentation of the mapping (Figure 12) we can now assess spectral features present in identified regions for mappings of neural signal during specific behaviors and at specific points

in time over the course of pair bonding for both hit and non-hit subjects. Besides assessing how mapping of the neural dynamics changes over course intervals of time (Figure 9 -11), we can further assess more local changes of the neural dynamics over shorter intervals of time.

Here, only LFP data was used to construct mappings. It is theorized that LFP signal is a measure of inhibitory activity (Buzsáki, Anastassiou, & Koch, 2012; Herreras, 2016), and as such, these constructed mappings may only be reflective of inhibitory neural signal. This may be appropriate for analyzing recordings of neural activity in the NAcc, as neuronal inhibition in the NAcc facilitates encoding of reward-directed behavior (Taha and Fields, 2006), however, it may not be appropriate for assessing the activity in the mPFC or BNST. With neuronal spike data, the underlying neural dynamics of pair bonding could be more thoroughly analyzed to assess for behavior-specific features relevant for neural decoding (Holdgraf et al., 2017; Munuera, Rigotti, & Salzman, 2018).

**Future Directions**

As we identified the low-dimensional embedded space of the LFP spectral features can be assessed in terms of the trajectory and velocity of transitions through the space, meaning the time-course and time it takes for animals to move through brain-states, we can now assess the transitions in more depth (Figure 6). The data suggest there are two separable states of rest and activity in which animals transition through brain-states. Now, the temporal and hierarchical pattern of "visiting" brain-states can be assessed, in that the probability an animal will go from one brain-state to the next after a certain time can be analyzed as in Berman et al., 2016. In turn, this can facilitate elucidation of how transitions between brain-states are reflective of transitions between behavior. Thus, a specific question that can be posed is as follows: do transitions between brain-states predict transitions between behaviors or vice-versa? Naturally, that allows

us to assess if there is hierarchical control of brain-states on behavioral-states and vice versa, and if so, what that hierarchy entails.

Moreover, with the identified spectral features identified for specific behaviors exhibited over the course of pair bonding, we may be able to classify a support vector machine (SVM) to determine when a specific social behavior is occurring given the neural signal. SVM classification and neural nets are used to decode behavior from spectrogram representations of LFP signal (Mehring et al., 2003; Niketeghad et al., 2014). As such, if we can identify spectral features present during specific behaviors or at specific points in time over the course of pair bonding, we may be able to classify those specific behaviors given LFP signal. To test this, we must first identify specific spectral features present in the identified segmented regions of the low-dimensional map (Figure 9), then identify what features correspond to what behaviors (i.e. which regions are represented for which behaviors), and then use the identified features to train an SVM. After training an SVM, we can analyze data from a different population of subjects to assess if the identified spectral features of the LFP signal are sufficient for classifying behaviors during pair bonding.

Then, to assess the conclusions of Amadei et al., using *t*-SNE, that is, to map the phase-amplitude coupling dynamics, we can use the phase-amplitude information represented as the Kullback-Liebler modulation index (KL MI) as input. This would require changing the distance metric used to separate points, namely, using a different cost function which is not the $D_{KL}$. However, using this input for the mapping will allow us to assess how the net modulation changes over time and with respect to specific behaviors exhibited during pair bonding. First, the LFP signal must be filtered and transformed in accordance with Amadei et al.'s method (2017) by applying a low- and high-pass filter, applying a Hilbert transform to extract phase and

amplitude signal at low and high frequencies, then computing the KL MI according to Tort et al., 2008. The distance metric, (i.e. the cost function minimized during *t*-SNE implementation), must be changed to assess the dissimilarity and similarity of points over time, represented by the KL MI over time – we will no longer be assessing PDFs.

Overall, as Amadei et al. (2017) were the first to elucidate the electrophysiological underpinnings of pair bonding, we are now working more towards decoding the neural basis of pair bonding and affiliative social behavior at large. With the analysis of subtler dynamics present in the neural signal, we may be able to deconstruct the complex mechanism of pair bonding and affiliative social behavior as a whole. If we identify unique behavior-specific brain-states present during pair bonding, and thus social bonding, we may be able to better assess and decompose this highly complex phenomenon.

**REFERENCES**

Ahern, TH, Modi, ME., Burkett, JP & Young, LJ. Evaluation of two automated metrics for analyzing partner preference tests. *J. Neurosci. Methods* 182, 180–188 (2009).

Aragona BJ et al. Nucleus accumbens dopamine differentially mediates the formation and maintenance of monogamous pair bonds. *Nat. Neurosci.* 9, 133–139 (2006).

Billings JCW, Medda A, Shakil S, Shen X, Kashyap A, Chen S, Abbas A, Zhang X, Nezafati M, Pan W-J, Berman GJ, & Keilholz SD. Instantaneous brain dynamics mapped to a continuous state space. *NeuroImage* 162, 344-352 (2017).

Block A, Dhanji H, Thompson-Tardif SF, & Floresco S. B. Thalamic–prefrontal cortical–ventral striatal circuitry mediates dissociable components of strategy set shifting. *Cereb. Cortex* 17, 1625–1636 (2007).

Buzsáki G, Anastassiou CA, & Koch C. The origin of extracellular fields and currents--EEG, ECoG, LFP and spikes. *Nat. Rev. Neurosci.* 13(6), 407-420 (2012).

Cover TM & Thomas JA. Elements of information theory, 2nd ed. Hoboken, NJ: Wiley-Interscience (2006).

Cowen AS & Keltner D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *PNAS* 114(38), E7900-E7909 (2017).

Daubechies I. Ten lectures on wavelets. Philadelphia, PA: SIAM. (1992).

De Vries, GJ & Miller MA. Anatomy and function of extrahypothalamic vasopressin systems in the brain. *Prog. Brain Res.* 119, 3–20 (1998).

De Vries, GJ & Panzica, GC. Sexual differentiation of central vasopressin and vasotocin systems in vertebrates: different mechanisms, similar endpoints. *Neurosci.* 138, 947–955 (2006).

Dimitriadis G, Neto J., and Kampff A. *t*-SNE visualization of large-scale neural recordings. *bioRxiv* 1–22 (2016).

Donaldson ZR, Spiegel L & Young LJ. Central vasopressin V1a receptor activation is independently necessary for both partner preference formation and expression in socially monogamous male prairie voles. *Behav. Neurosci.* 124, 159–163 (2010).

Floresco SB. The nucleus accumbens: an interface between cognition, emotion, and action. *Annu. Rev. Psychol.* 66, 25–52 (2015).

Fuglede B & Topsøe F. Jensen-Shannon divergence and Hilbert space embedding. *IEEE Conference Publication* (2004).

Getz LL, Carter CS, & Gavish L. The mating system of the prairie vole, *Microtus ochrogaster*: Field and laboratory evidence for pair-bonding. *Behav. Ecol. and Sociobiol*. 8(3), 189–194 (1981).

Goodson JL, Kabelik D. Dynamic limbic networks and social diversity in vertebrates: from neural context to neuromodulatory patterning. *Front. Neuroendocrinol.* 30(4), 429-441 (2009).

Greenberg GD, Laman-Maharg A, Campi KL,Voigt H, Orr VN, Schaal L, & Trainor BC. Sex differences in stress-induced social withdrawal: role of brain derived neurotrophic factor in the bed nucleus of the stria terminalis. *Front. Behav. Neurosci.* 7(223) (2014).

Herreras O. Local Field Potentials: Myths and Misunderstandings. *Front. Neural Circuits*. 10(101) (2016).

Holdgraf CR, Rieger JW, Micheli C, Martin C, Knight RT, & Theunissen FE. Encoding and Decoding Models in Cognitive Electrophysiology. *Front. Syst. Neurosci.* 11(61) (2017).

Insel TR, Shapiro LE. Oxytocin receptor distribution reflects social organization in monogamous and polygamous voles. *PNAS* 89, 5981–5985. (1992).

Jongen HT, Meer K, Triesch E. Optimization theory. Boston, MA: Kluwer Academic Publishers. (2004).

Kleiman, D. G. Monogamy in mammals. *Q. Rev. Biol.* 52, 39–69 (1977).

Lagarias JC, Reeds JA, Wright MH, Wright PE. Convergence properties of the Nelder Mead simplex method in low dimensions. *SIAM J. Optim*. 9, 112–147 (1998).

Lebow MA & Chen A. Overshadowed by the amygdala: the bed nucleus of the stria terminalis emerges as key to psychiatric disorders. *Molecular Psychiatry.* 21, 450–463 (2016).

Lee Y, Fitz S, Johnson PJ, & Shekhar A. Repeated Stimulation of CRF Receptors in the BNST of Rats Selectively Induces Social but not Panic-Like Anxiety. *Neuropsychopharm.* 33, 2586–2594 (2008)

Lim, M. M. et al. Enhanced partner preference in a promiscuous species by manipulating the expression of a single gene. *Nature* 429, 754–757 (2004).

Lin, J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory* 37 (1), 145–151 (1991).

Liu Y & Wang ZX. Nucleus accumbens oxytocin and dopamine interact to regulate pair bond formation in female prairie voles. *Neuroscience* 121 (3), 537-44 (2003).

Lobo MK, Covington HE III, Chaudhury D, Friedman AK, Sun H, Damez-Werno D, Dietz DM, Zaman S, Koo JW, Kennedy PJ, et al. Cell type-specific loss of BDNF signaling mimics optogenetic control of cocaine reward. *Science* 330, 385–390 (2010).

McGraw LA & Young L. The prairie vole: an emerging model organism for understanding the social brain. *Trends Neurosci*. 33(2), 103 (2010).

Mehring C, Rickert J, Vaadia E, de Oliveira SC, Aertsen A, & Rotter S. Inference of hand movements from local field potentials in monkey motor cortex. *Nat Neurosci*. 6(12), 1253-4 (2003).

Meyer F. Topographic distance and watershed lines. *Signal Process*. 38, 113–125 (1994).

Munuera J, Rigotti M, & Salzman CD. Shared neural coding for social hierarchy and reward value in primate amygdala. *Nature Neurosci.* 21, 415–423 (2018)

Nicola SM. The nucleus accumbens as part of a basal ganglia action selection circuit. *Psychopharm. (Berl.)* 191, 521–550 (2007).

Niketeghad S, Hebb AO, Nedrud J, Hanrahan SJ, & Mahoor MH, Single trial behavioral task classification using subthalamic nucleus local field potential signals. *Engineering in Medicine and Biology Society (EMBC) 14th IEEE International Conference*, 3793-3796, (2014)

Ross, H. E. et al. Characterization of the oxytocin system regulating affiliative behavior in female prairie voles. *Neuroscience* 162, 892–903 (2009).

Ross, H. E. et al. Variation in oxytocin receptor density in the nucleus accumbens has differential effects on affiliative behaviors in monogamous and polygamous voles. *J. Neurosci*. 29, 1312-1318 (2009).

Sofroniew MV. Morphology of vasopressin and oxytocin neurones and their central and vascular projections. *Prog Brain Res*. 60, 101-14 (1983).

Stanley DA & Adolphs R. Toward a neural basis for social behavior. *Neuron* 30;80(3), 816-26 (2013).

Stephens GJ, Osborne LC, & Bialek W. Searching for simplicity in the analysis of neurons and behavior. *PNAS* 108, 15565-15571 (2011).

Stuber GD. et al. Excitatory transmission from the amygdala to nucleus accumbens facilitates reward seeking. *Nature* 475, 377–380 (2011).

Svendsend E. Pair formation, duration of pair-bonds, and mate replacement in a population of beavers (*Castor canadensis*). *Can. J. Zool*. 67, 336-340 (1989).

Taha SA & Fields HL. Inhibitions of Nucleus Accumbens Neurons Encode a Gating Signal for Reward-Directed Behavior. *J. Neurosci.* 26(1), 217-222 (2006).

Tinbergen, N. The Study of Instinct. New York, NY: Clarendon Press/Oxford University Press. (1951).

Tort, A. B. et al. Dynamic cross-frequency couplings of local field potential oscillations in rat striatum and hippocampus during performance of a T-maze task. *PNAS* 105, 20517–20522 (2008).

van der Maaten L & Hinton G. Visualizing data using *t*-SNE. *J. Mach. Learn. Res.* 9: 85 (2008).

Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bull*. 1, 80–83 (1945).

Williams, JR, Catania, KC, & Carter, CS. Development of partner preferences in female prairie voles (*Microtus ochrogaster*): the role of social and sexual experience. *Horm. Behav.* 26, 339–349 (1992).

Young, LJ & Wang Z. The neurobiology of pair bonding. *Nat. Neurosci.* 7: 1048–1054 (2004).

Young LJ, Lim MM, Gingrich B, & Insel TR. Cellular mechanisms of social attachment. *Horm. Behav.* 40, 133–138 (2001).

## Appendix A. Experiments

All procedures were approved by the Emory University Institutional Animal Care and Use Committee. Experimental subjects were fifteen adult, sexually-naïve female prairie voles (*M. ochrogaster*) 76-154 days of age at the start of experimentation. All animals were socially-housed in same-sex duos or trios until electrode implantation surgery – after surgery, they were separated and housed individually. Food (Laboratory Rabbit Diet HF 5326, LabDiet) and water were provided ad libitum during a 14:10 hour light:dark cycle.

Partner animals used in behavioral experiments were adult, sexually experienced males under 1.5 years of age. They were matched by age (within 61 days) and weight (within approximately 5 g) for each female.

## Appendix B. Morlet Wavelet Decomposition

Specifically, the transform, $W_{s,\tau}[y(t)]$ was calculated via

$$W_{s,\tau}[y(t)] = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} y(t) \psi^*\left(\frac{t-\tau}{s}\right) dt \quad \text{(B 1)}$$

where

$$\psi(\eta) = \pi^{-1/4} e^{i\omega_0 \eta} e^{-1/2\eta^2} \quad \text{(B 2)}$$

is the Morlet wavelet kernel or mother wavelet, $y_i(t)$ is the continuous LFP time-series data, $s$ is the time scale of interest, $\tau$ is a point in time, and $\omega_0$ is a non-dimensional parameter which determined time-frequency resolution, set to 5 here.

Notably, the Morlet wavelet has the additional property that the time scale, $s$, is related to the Fourier frequency, $f$, by

$$s(f) = \frac{\omega_0 + \sqrt{2+\omega^2{}_0}}{4\pi f}, \quad \text{(B 3)}$$

This can be derived by maximizing the response to a pure sine wave, $A(s,f) = \left|W_{s,\tau}\left[e^{2\pi i f t}\right]\right|$with respect to $s$. However, $A(s,\omega)$is disproportionally large when responding to pure sine waves of lower frequencies. To correct for this, we find a scalar function $C(s)$ such that $C(s)A(s,\omega^*) = 1$ for all $s$, where $\omega^*$ is $2\pi$ times the Fourier frequency, ($B$ 3). For a Morlet wavelet, this function is

$$C(s) = \frac{\pi^{-1/4}}{\sqrt{2s}} e^{-1/4(\omega_0 + \sqrt{2+\omega^2{}_0})} \quad \text{(B 4)}$$

where again, $\omega_0$ is from the relation of the time-scale to frequency.

Accordingly, we can define our power spectrum as,

$$S(k,f;\tau) = \frac{1}{C(s(f))}\left|W_{s(f),\tau}[y_k(t)]\right|, \quad \text{(B 5)}$$

determined from the transform taken from $s(f)$to $\tau$ for a time-series $y_i(t)$. This is important, as the power spectral density is the input for the non-linear embedding. Last, we use a dyadically spaced set of frequencies between $f_{min} = 1$ Hz and the Nyquist frequency, $N_f(f_{max} = 100$ Hz), via

$$f_i = f_{max} 2^{\frac{i-1}{N_f - 1} log_2 \frac{f_{max}}{f_{min}}} \quad \text{(B 6)}$$

for $i = 1, 2, \ldots N_f$ and their corresponding scales via equation (A 3). This creates a wavelet spectrogram that is resolved at multiple time scales for each of the first 100 frequencies.

## Appendix C. *t*-distributed Stochastic Neighbor Embedding Implementation

For our initial embedding using *t*-SNE, we largely follow the method introduced by van der Maaten & Hinton (2008), minimizing the cost function

$$C = D_{KL}(P||Q) = \sum_{ij} p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad \text{(C 1)}$$

where $p_{ij} = {}^{1}/_{2}\left(p_{j|i} + p_{i|j}\right)$, and

$$q_{ij} = \frac{(1+\Delta_{ij}^2)^{-1}}{\sum_k \sum_{l \neq k} (1+\Delta_{k,l}^2)^{-1}} \quad \text{(C 2)}$$

and $\Delta_{ij}$ is the Euclidean distance between points $i$ and $j$ in the embedded space. The cost function is optimized through a gradient descent procedure that is preceded by an early exaggeration period, allowing for the system to more readily escape local minima. This is much like to pulling the points, exaggerating the extent to which they are separated, or maximally dissimilar.

The memory complexity of this algorithm prevents the practical number of points from exceeding 35,000. The solution here is to generate an embedding using a selection of roughly 4,375 data points from each of the 8 subjects out of the 9 total hit subjects observed (out of approximately 4,000,000 data points per individual). To ensure that these points create a representative sample, we perform *t*-SNE on 20,000 randomly selected data points from each individual. This embedding is then used to estimate a probability density by convolving each point with a two-dimensional Gaussian whose width is equal to the distance from the point to its $N_{embed}$ = 5 nearest neighbors. This space is segmented by applying a watershed transform to the inverse of the PDF, creating a set of regions (Meyer, 1994). Finally, points are grouped by the region to which they belong, and the number of points selected out of each region is proportional to the integral over the PDF in that region. This is performed for all datasets, yielding a total of approximately 35,000 data points in the training set.

Given the embedding resulting from applying *t*-SNE to our training set, we wish to embed additional points, i.e. re-embed new points, into our brain-state space by comparing each with the training set individually. Mathematically, let $X$ be the set of all feature vectors in the training set, $X'$ be their associated embeddings via *t*-SNE, $z$ be a new feature vector that we would like to embed according to the mapping between $X$ and $X'$, and $z$ be the embedding of $z$ that we would like to determine.

As with the *t*-SNE cost function, we embedded $z$ by enforcing that its transition probabilities in the two spaces were as similar as possible. Like before, the transitions in the full space, $p_{j|z}$, are given by

$$p_{j|z} = \frac{exp(-d(z,j)^2/2\sigma_z^2)}{\sum_{x \in X} exp(-d(z,k)^2/2\sigma_z^2)'} \quad \text{(C 3)}$$

where $d(z,j)$ is the Kullback–Leibler divergence ($D_{KL}$) between $z$ and $x \in X$, and $\sigma_z$ is once again found by constraining the entropy of the condition transition probability distribution, using the same parameters as for the *t*-SNE embedding. Similarly, the transition probabilities in the embedded space are given by

$$q_{i|\zeta} = \frac{(1+\Delta_{\zeta,i}^2)^{-1}}{\sum_{x' \in X'}(1+\Delta_{\zeta,x'}^2)^{-1}} \quad \text{(C 4)}$$

where $\Delta_{\zeta,x'}$ is the Euclidean distance between $\zeta$ and $y \in X'$.

For each $\zeta$, we then found the $\zeta^*$

$$\zeta^* = \arg\min_{\zeta} D_{KL}(p_{x|z} || \, q_{y|\zeta}) \quad \text{(C 5)}$$

$$= \arg\min_{\zeta} \sum_{x \in X} p_{x|z} \log \frac{p_{x|z}}{q_{y(x)|\zeta}} \quad \text{(C 6)}$$

that minimizes the $D_{KL}$ between the transition probability distributions in the two spaces.

As before, this is a non-convex function, leading to potential complexities in performing our desired optimization. However, if we start a local optimization, using the Nelder–Mead simplex algorithm (Jongen, Meer, & Triesch, 2004; Lagarias, et al.,1998) from a weighted average of points, $\zeta_0$ , where

$$\zeta_0 = \sum_{x \in X} p_{x|z} y(x), \tag{C 7}$$

this point is almost always within the basin of attraction of the global minimum. To ensure that this is true in all cases, however, we also performed the same minimization procedure, but starting from the point $y(x^*)$, where

$$x^* = \arg\max_x p_{x|z} \tag{C 8}$$

This returned a better solution. Because this embedding can be calculated independently for each value of z, the algorithm scales linearly with the number of points. We also made use of the fact that this algorithm is embarrassingly parallelizable, allowing the algorithm to run faster. Moreover, because we have set our transition entropy, $H$, to be equal to 5, there are rarely more than 50 points to which a given $z$ has a non-zero transition probability. Accordingly, we sped up our cost function evaluation considerably by only allowing $p_{x|z} > 0$ for the nearest 200 points to $z$ in the original space.

Lastly, we found the space of brain-states for the non-hits subjects by embedding these data into the space created with the hits training set. We find that the mean extent to which all points embedded well in the space is 93.7%, indicating all hit subjects' data embedded well. When re-embedding all non-hits animals, that value was 93.7%. If above 90%, we determined the data were embedded well.

**Appendix D. Identified Spectral Features (PSD) For Each Region**

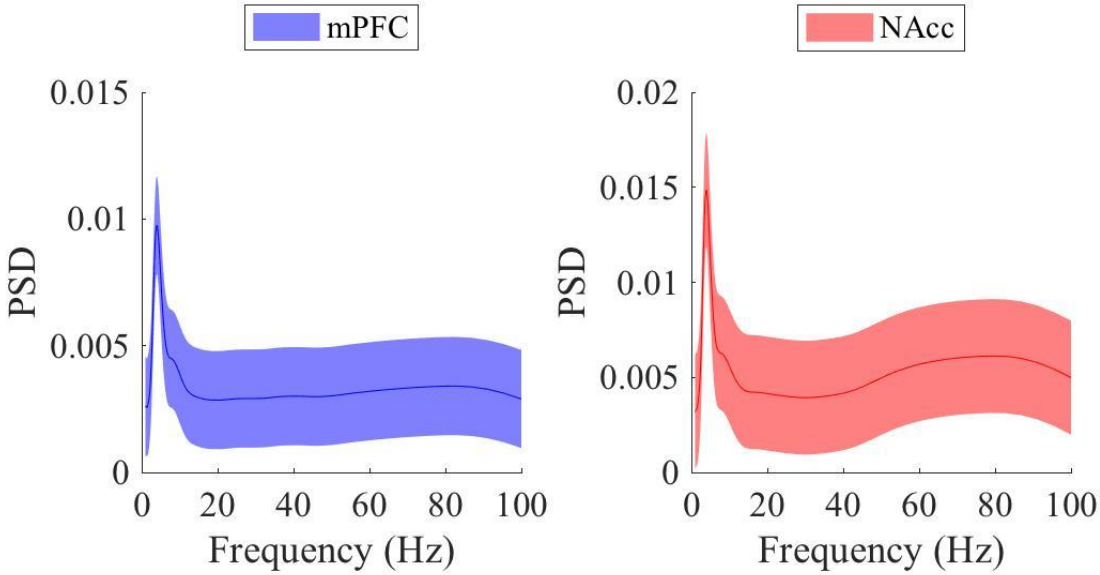



**Figure D. 1.** Region 1 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

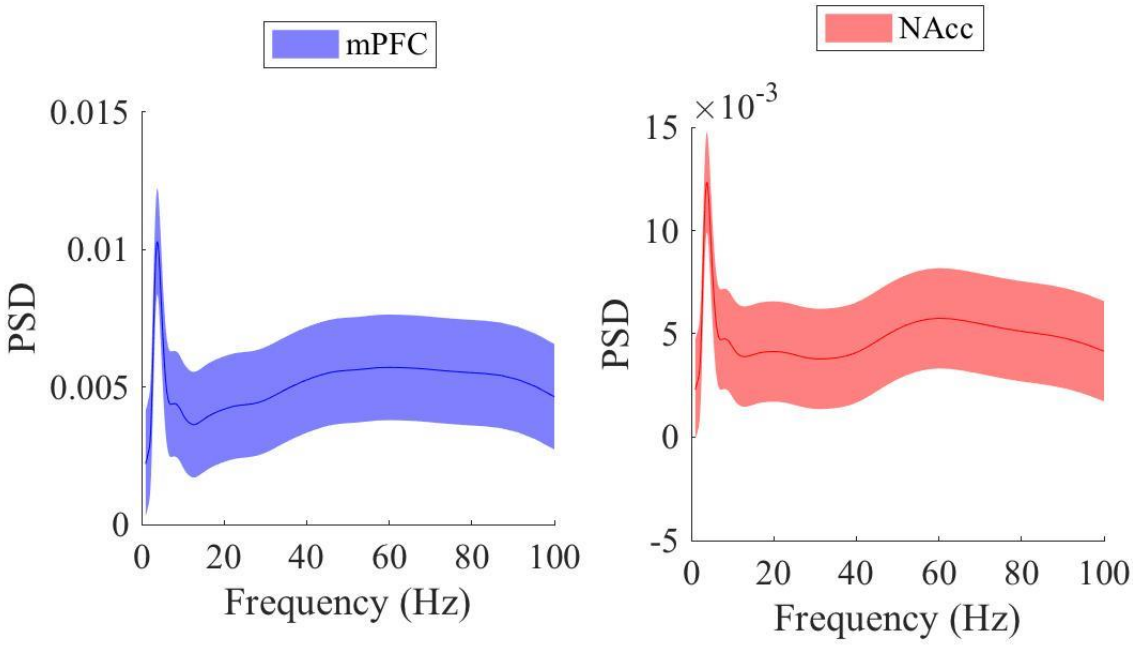



**Figure D. 2.** Region 2 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

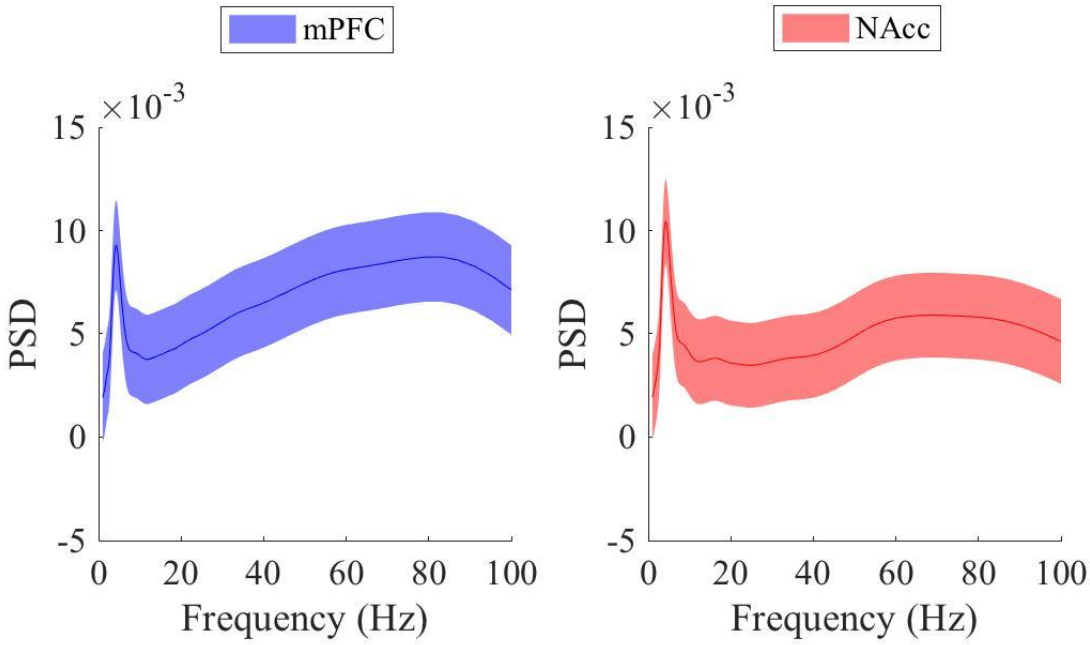



**Figure D. 3.** Region 3 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

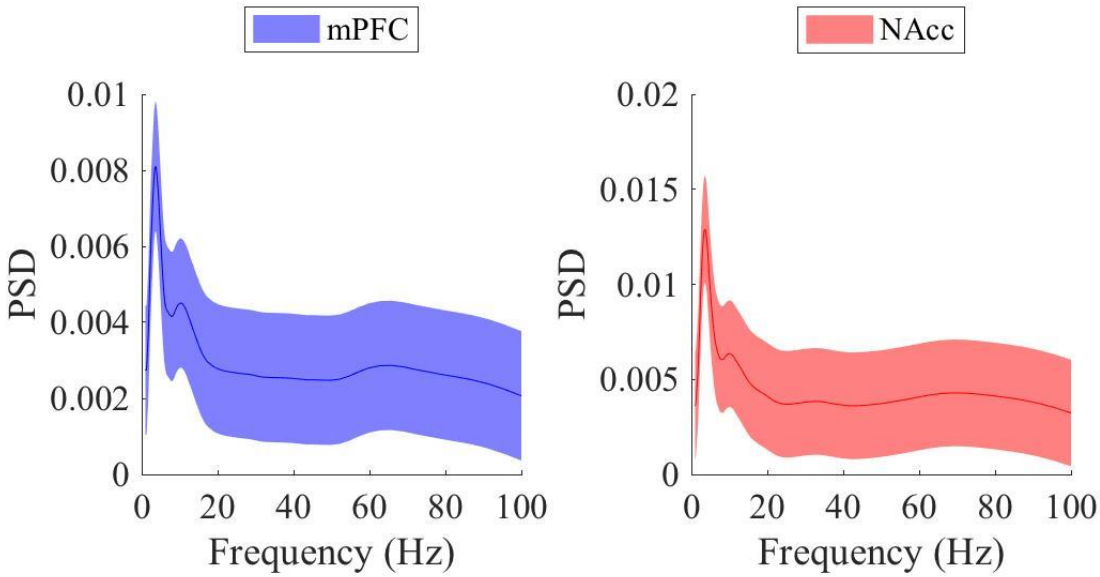



**Figure D. 4.** Region 4 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

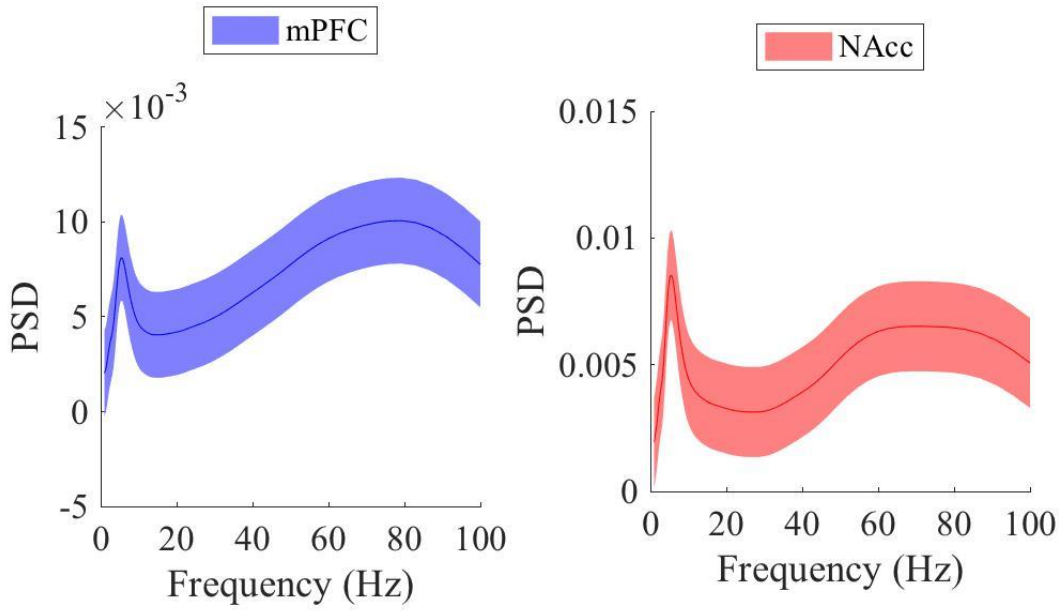



**Figure D. 5.** Region 5 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

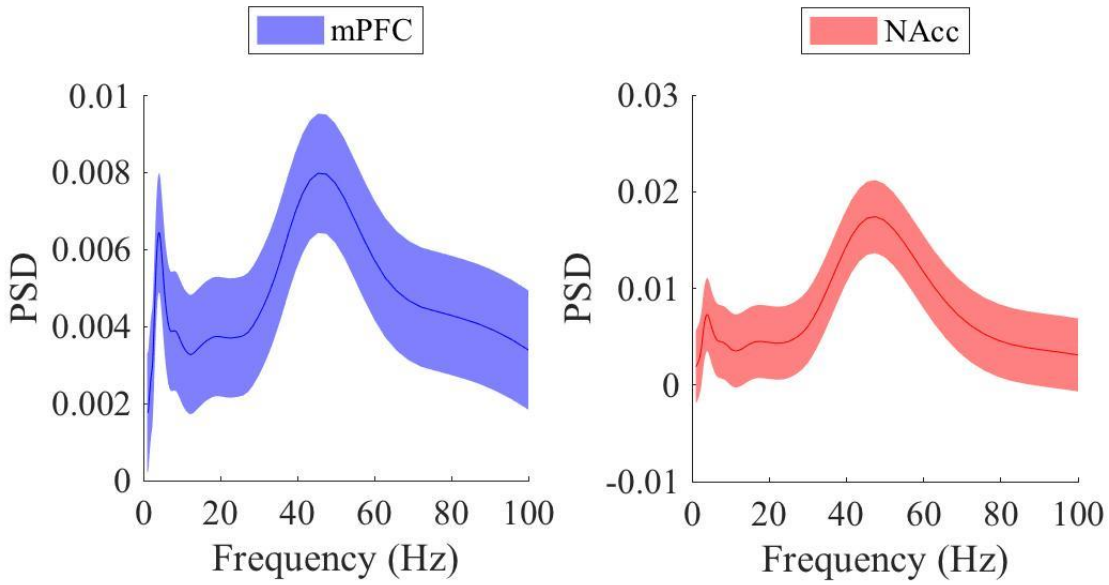



**Figure D. 6.** Region 6 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

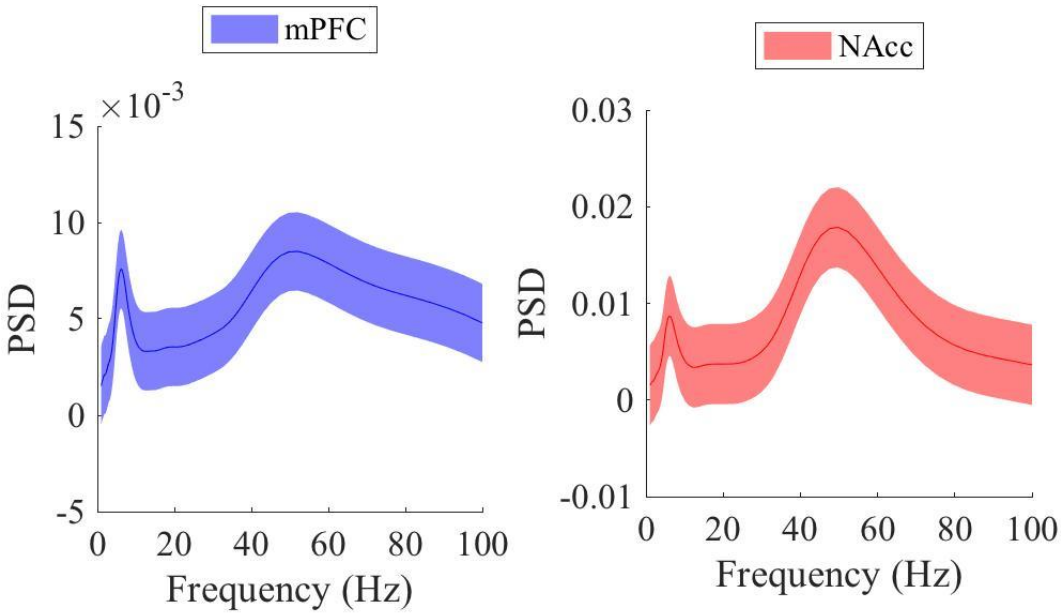



**Figure D. 7.** Region 7 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

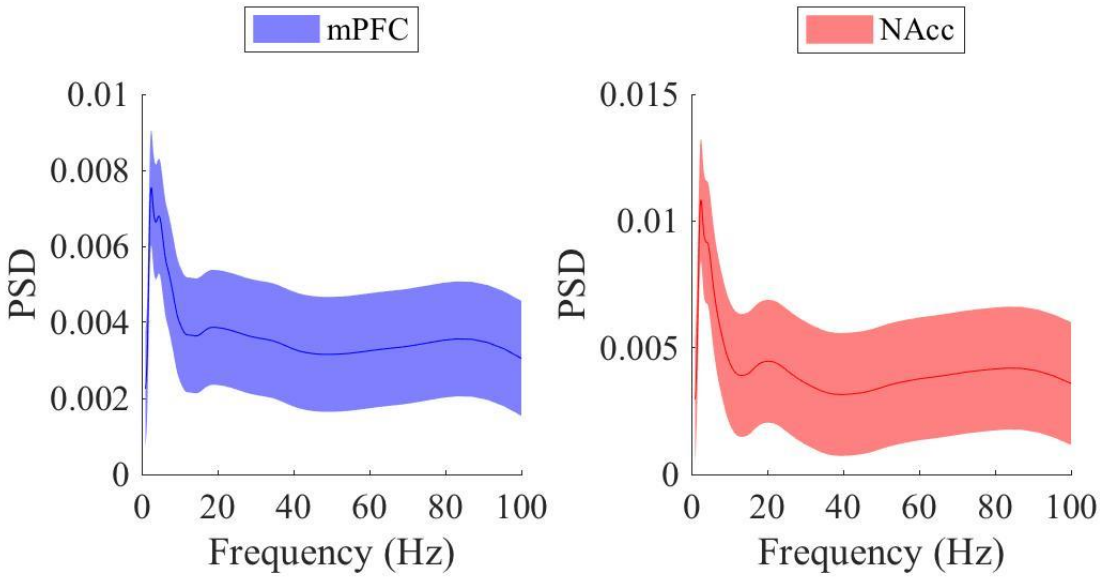



**Figure D. 8.** Region 8 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

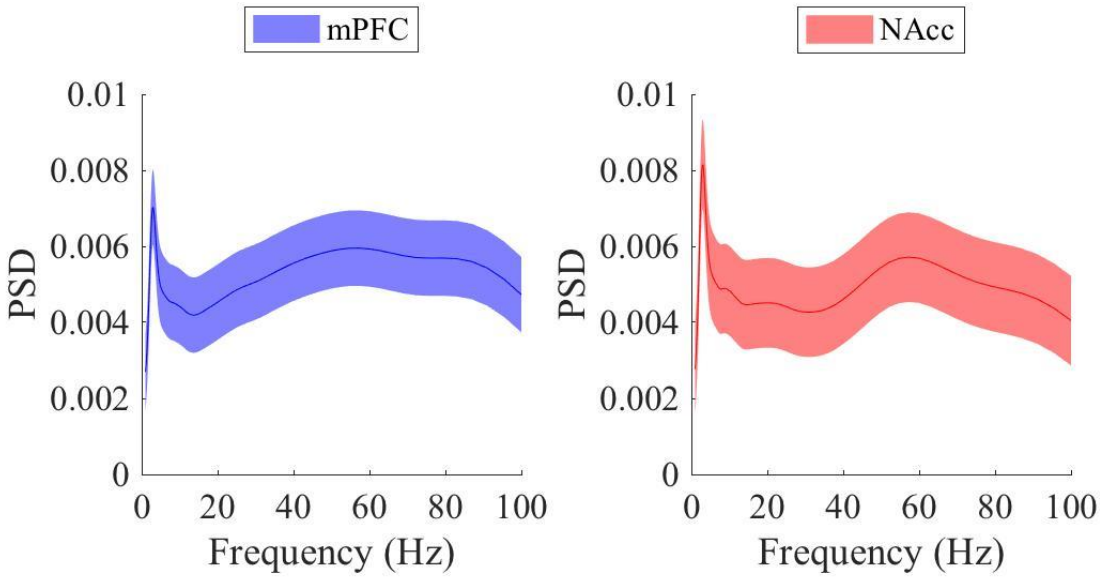



**Figure D. 9.** Region 9 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

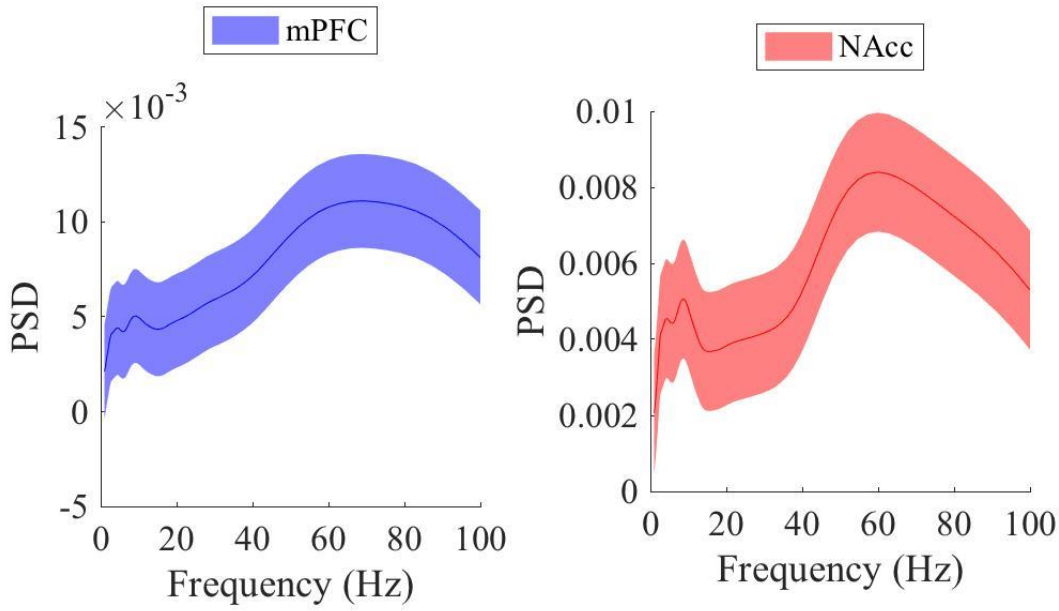



**Figure D. 10.** Region 10 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

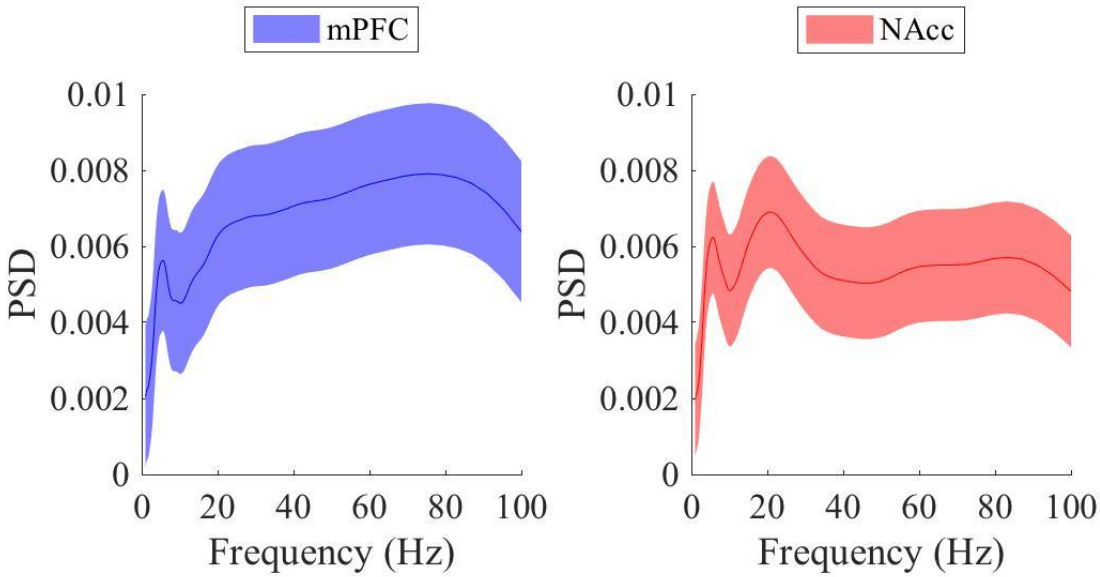



**Figure D. 11.** Region 11 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

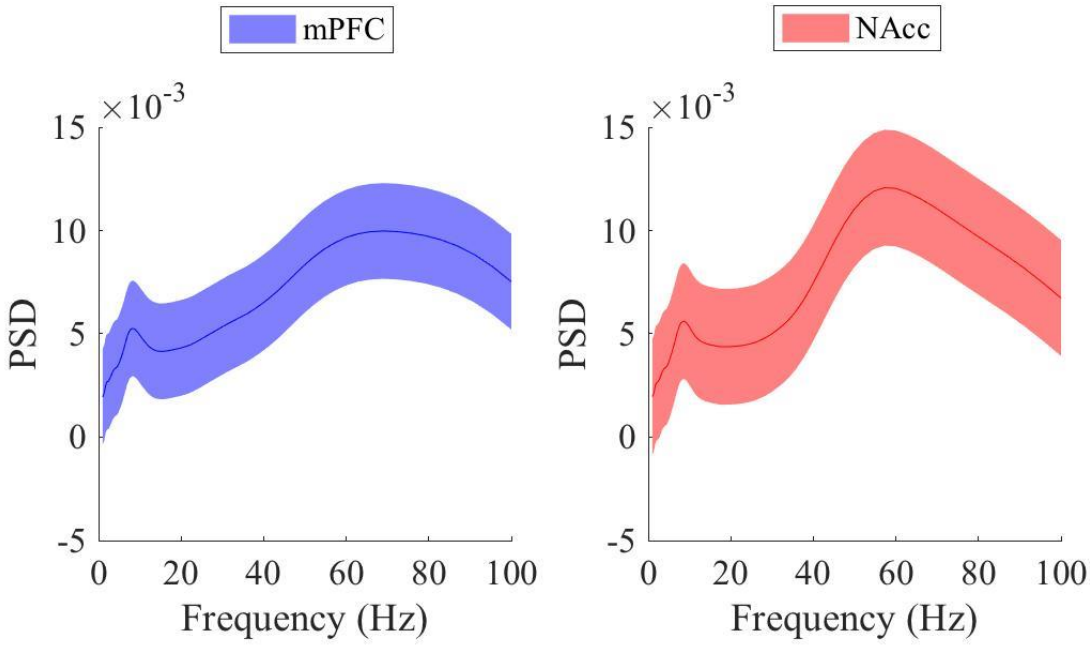



**Figure D. 12.** Region 12 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

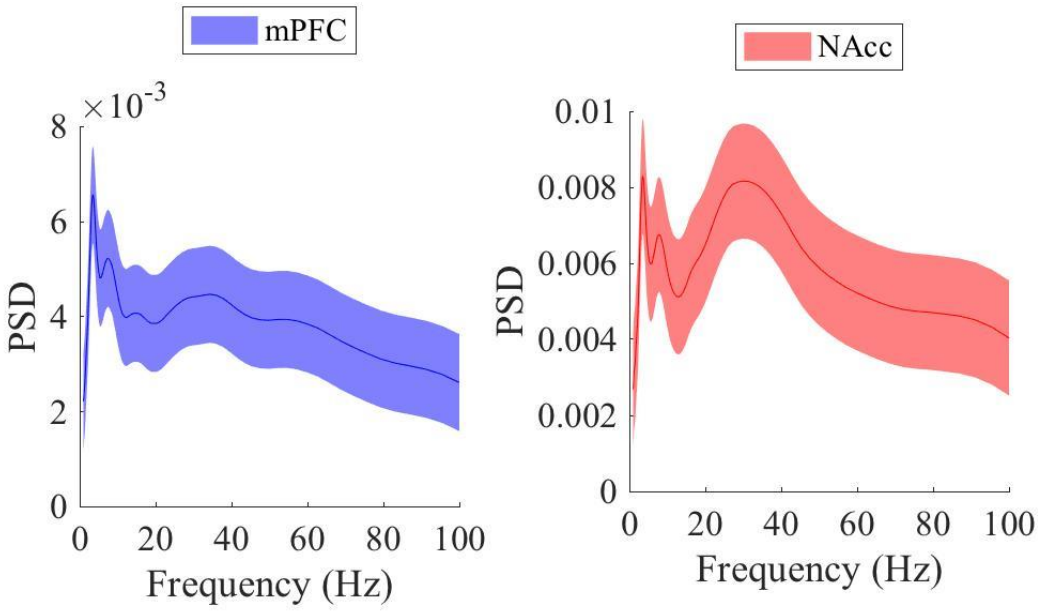



**Figure D. 13.** Region 13 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

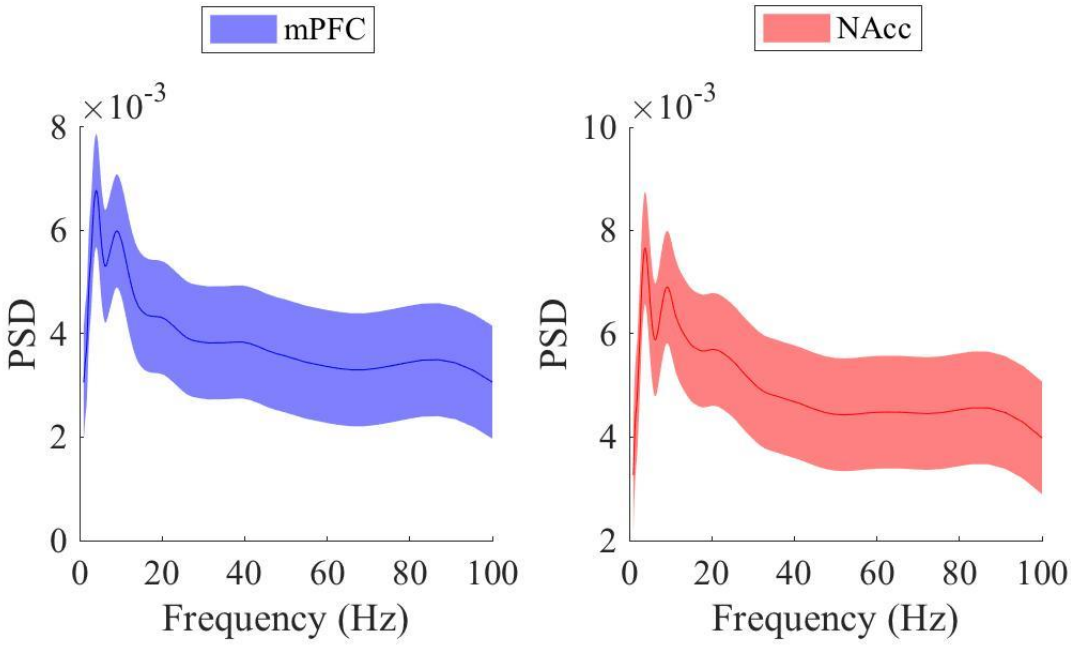



**Figure D. 14.** Region 14 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

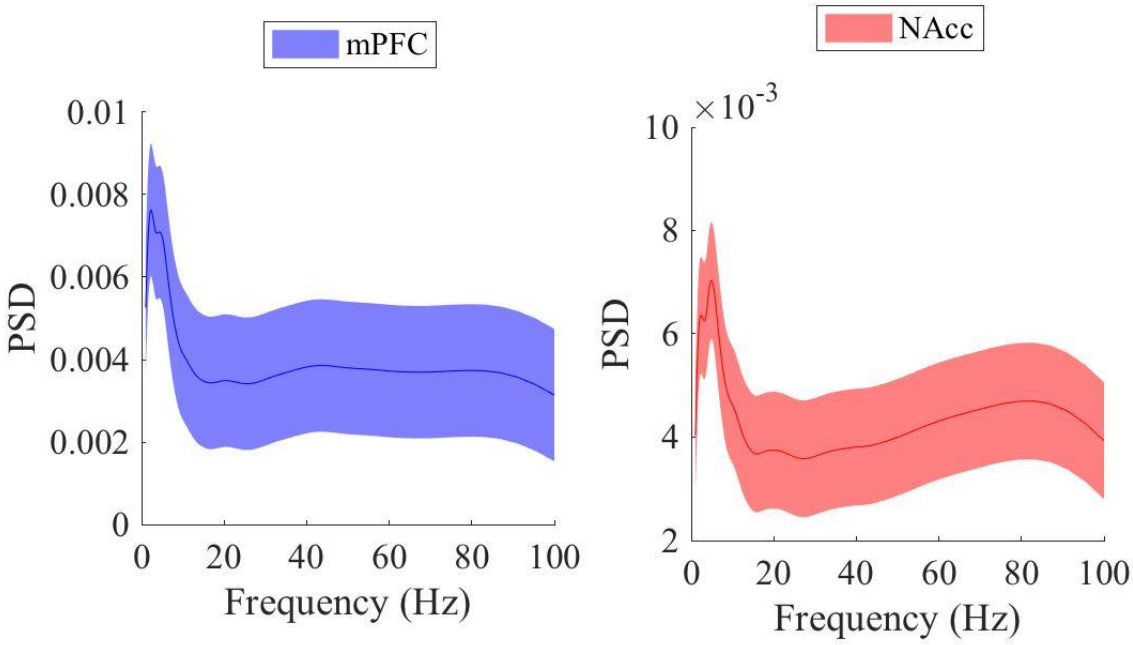



**Figure D. 15.** Region 15 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

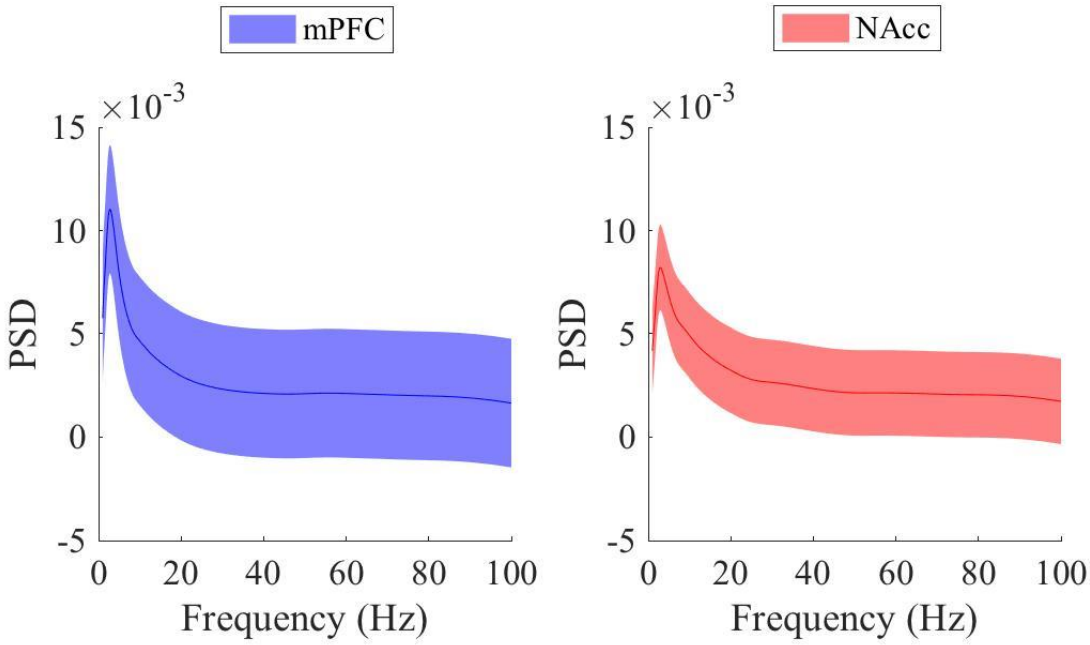



**Figure D. 16.** Region 16 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

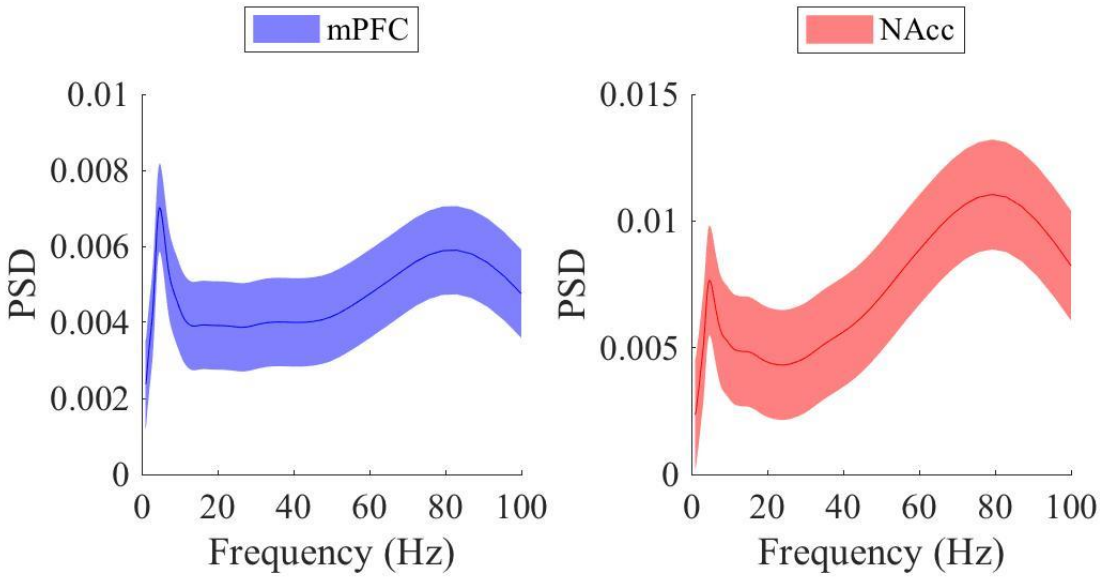



**Figure D. 17.** Region 17 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

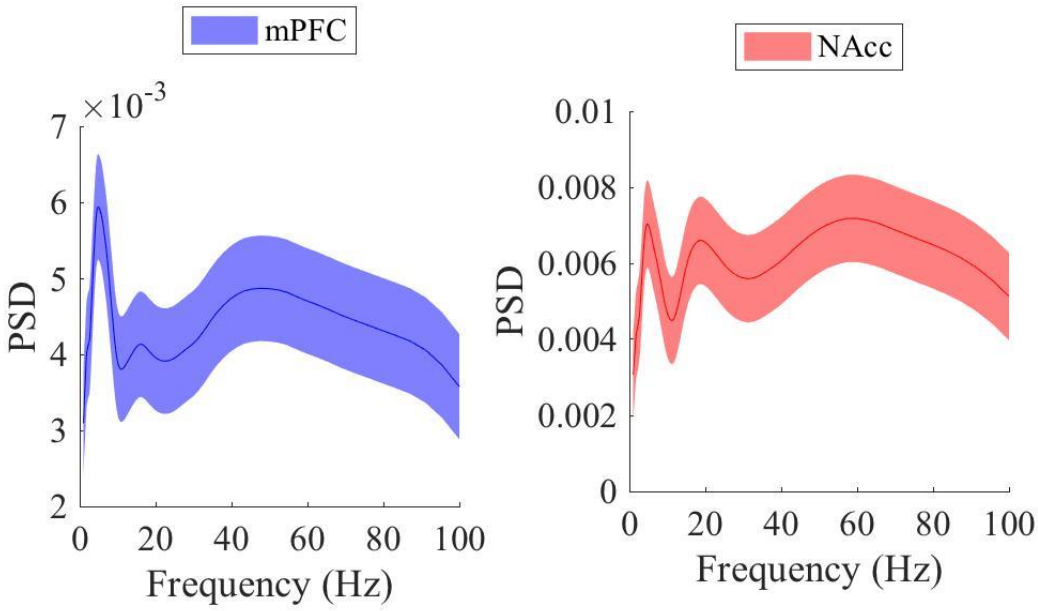



**Figure D. 18.** Region 18 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

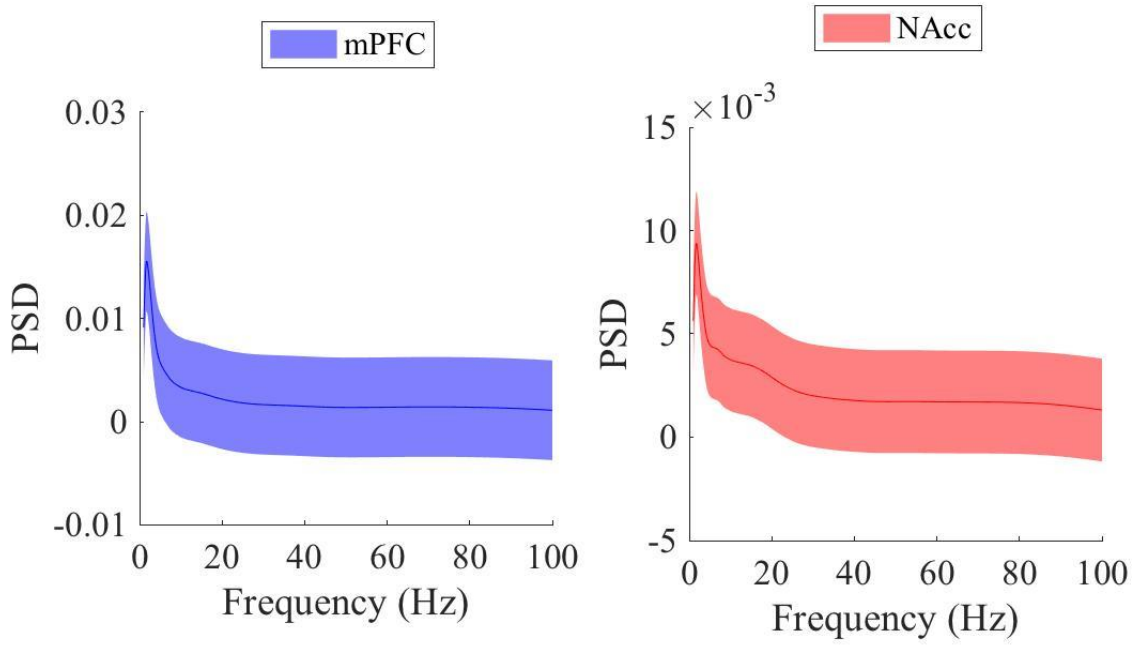



**Figure D. 19.** Region 19 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

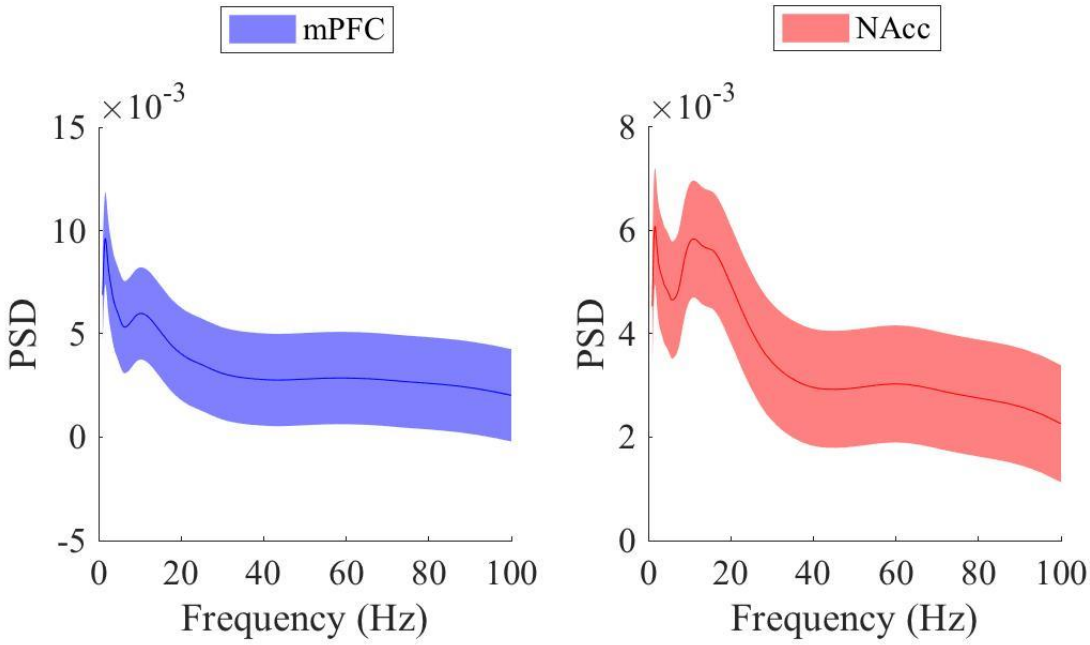



**Figure D. 20.** Region 20 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

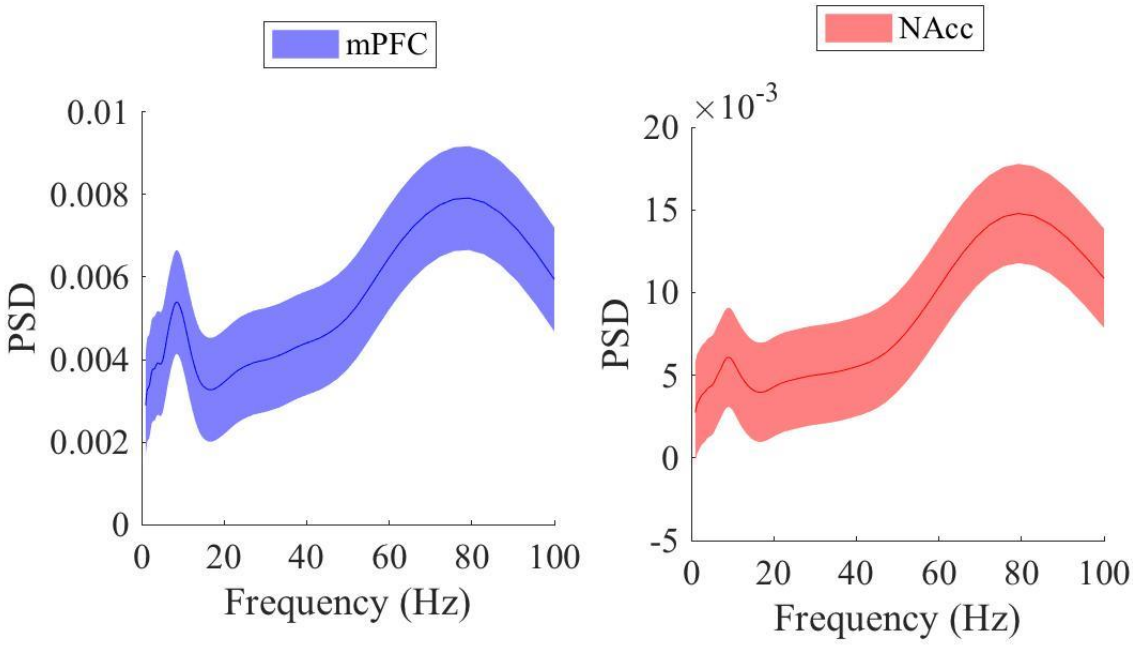



**Figure D. 21.** Region 21 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

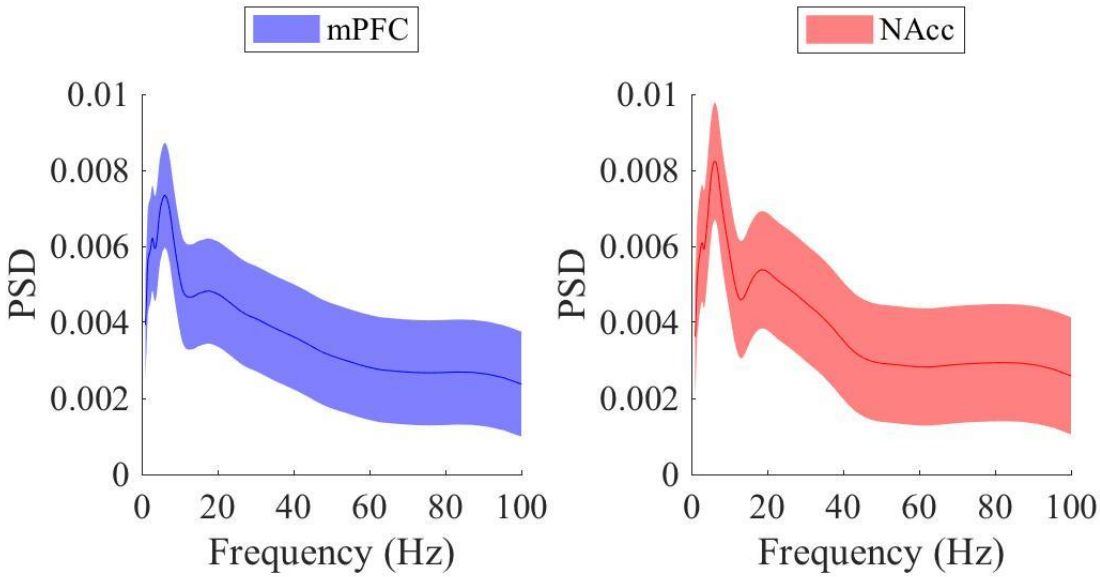



**Figure D. 22.** Region 22 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

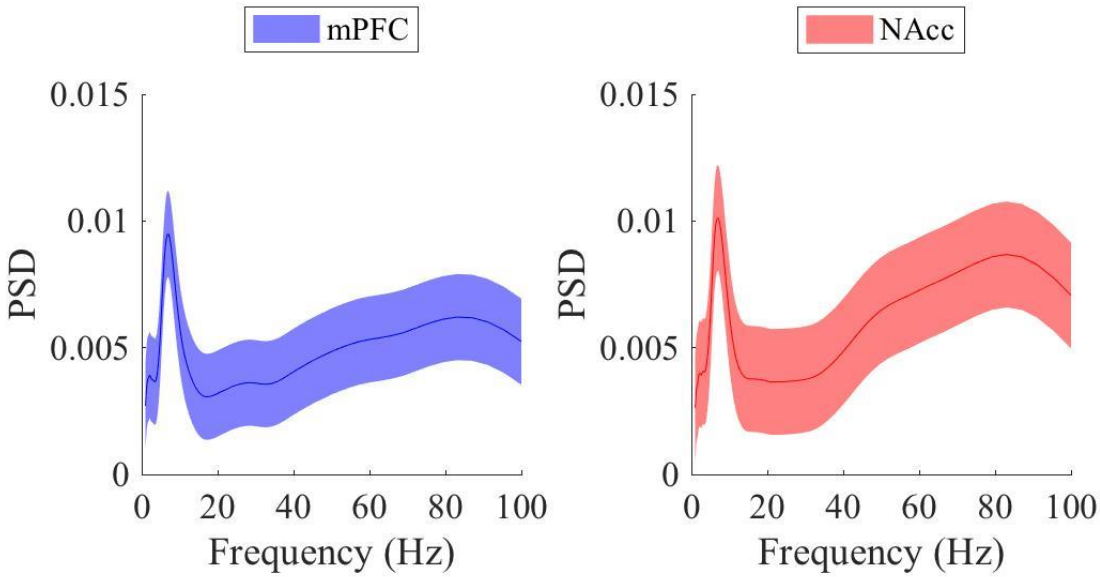



**Figure D. 23.** Region 23 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

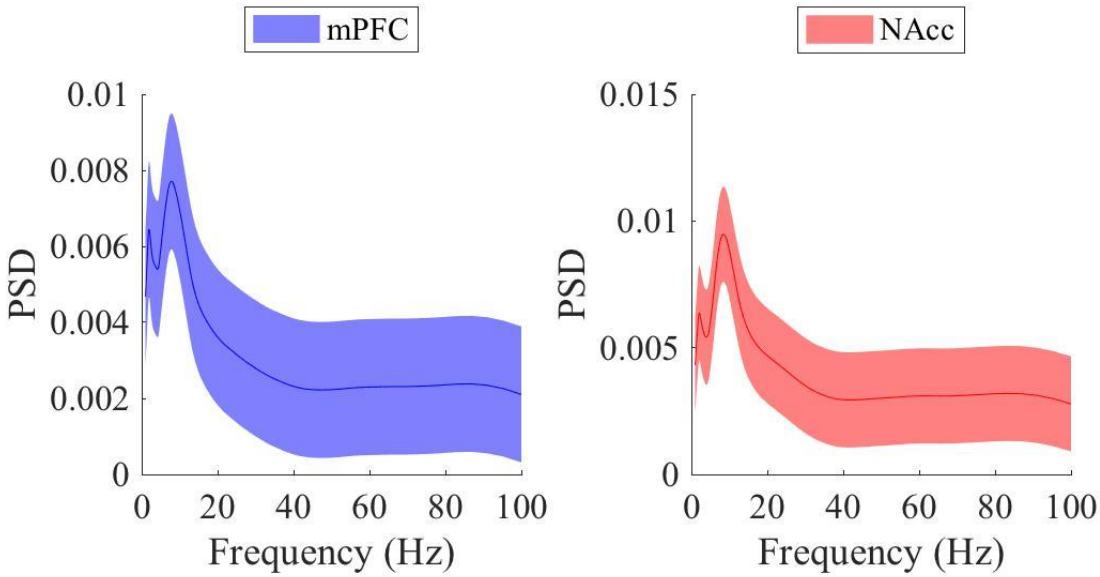



**Figure D. 24.** Region 24 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

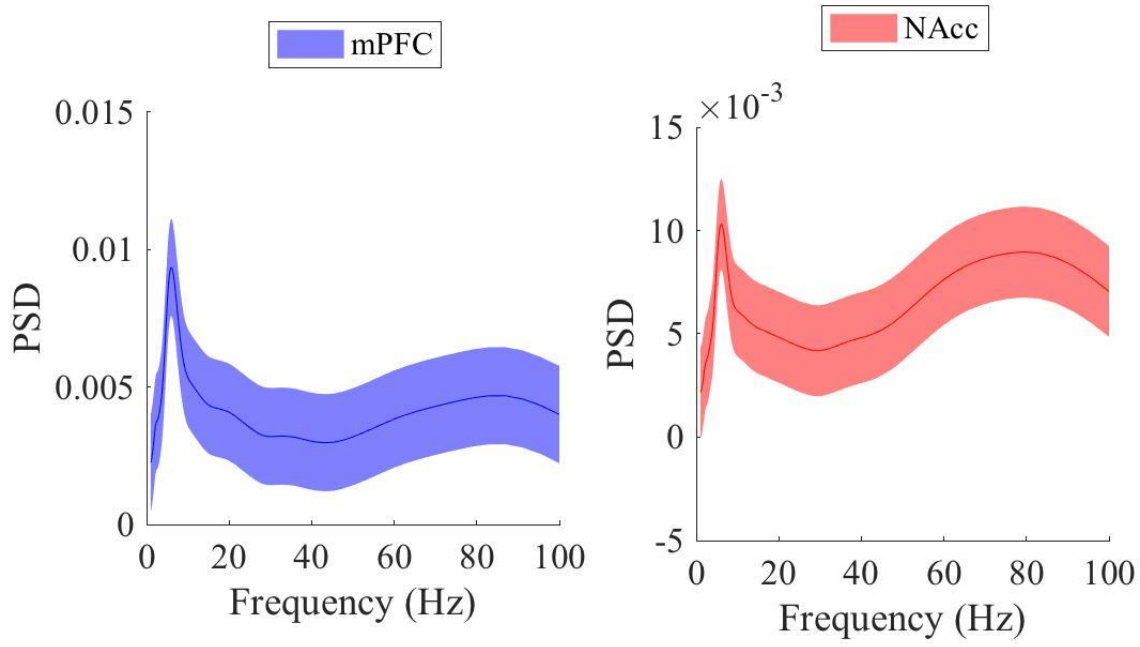



**Figure D. 25.** Region 25 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

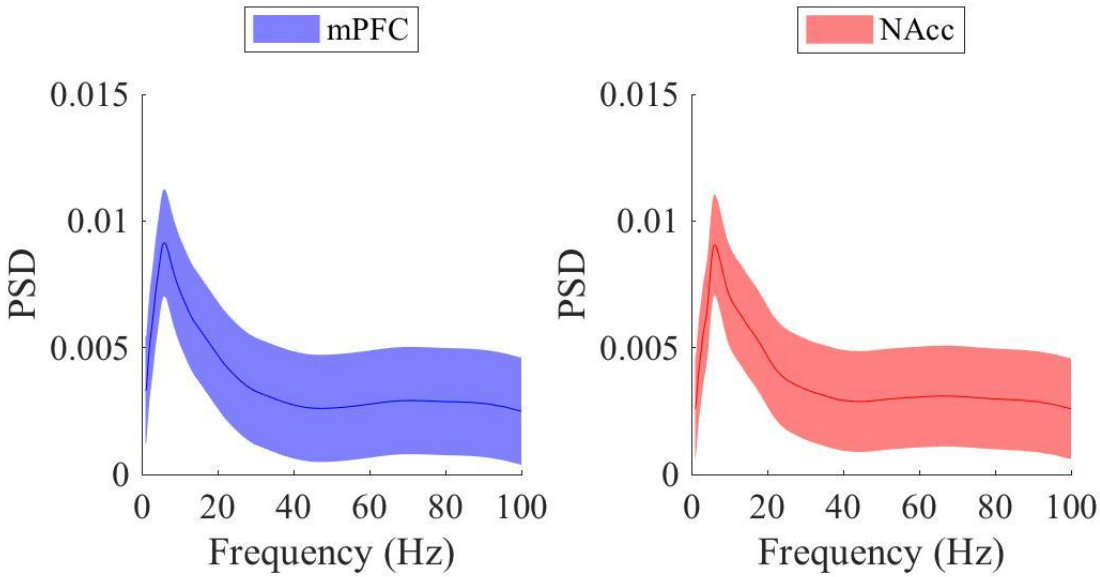



**Figure D. 26.** Region 26 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

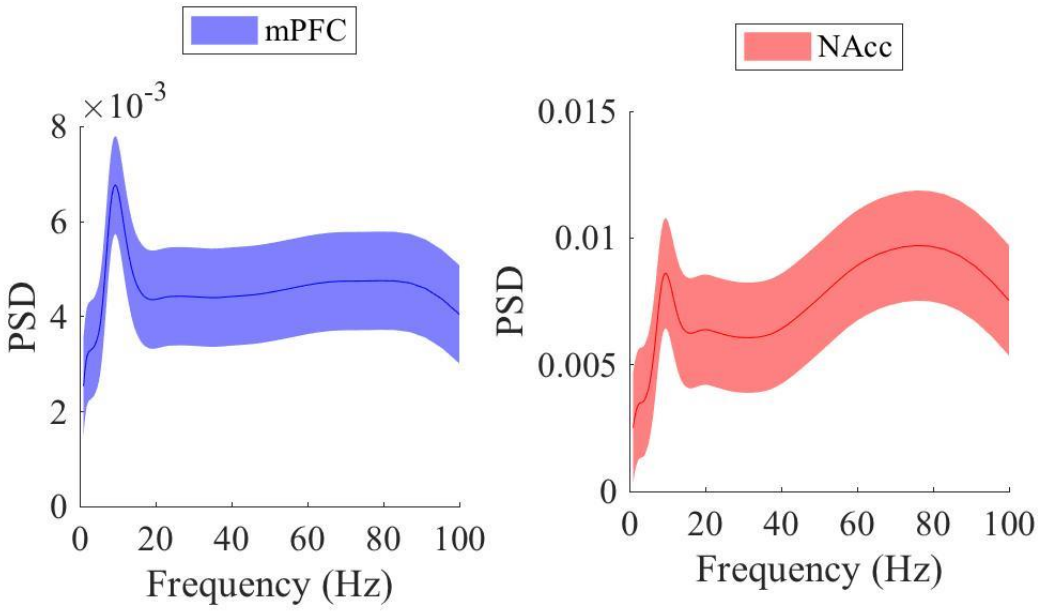



**Figure D. 27.** Region 27 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

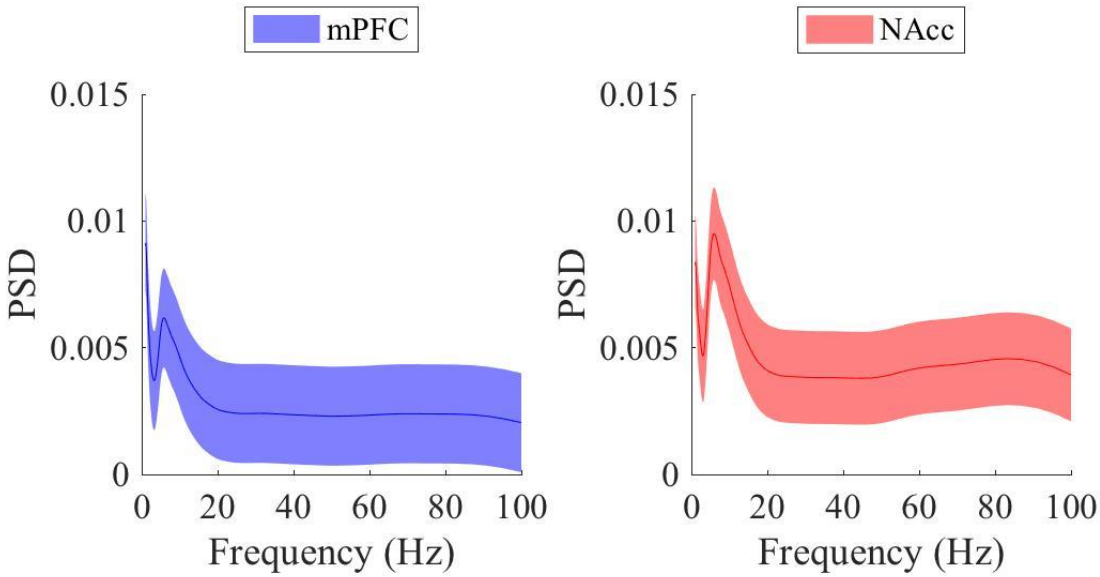



**Figure D. 28.** Region 28 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

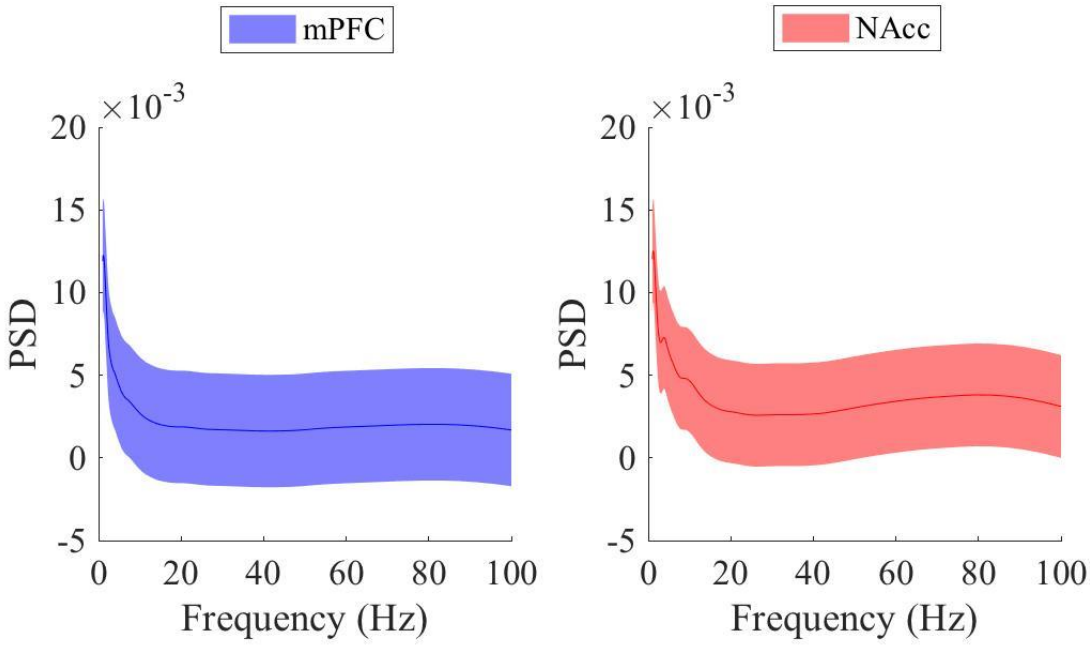



**Figure D. 29.** Region 29 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

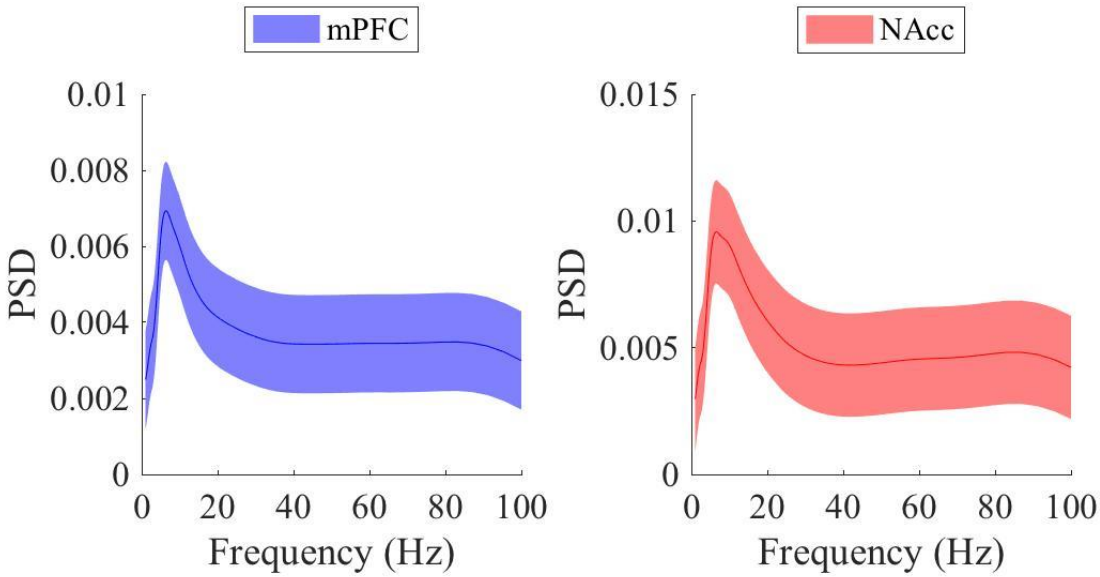



**Figure D. 30.** Region 30 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

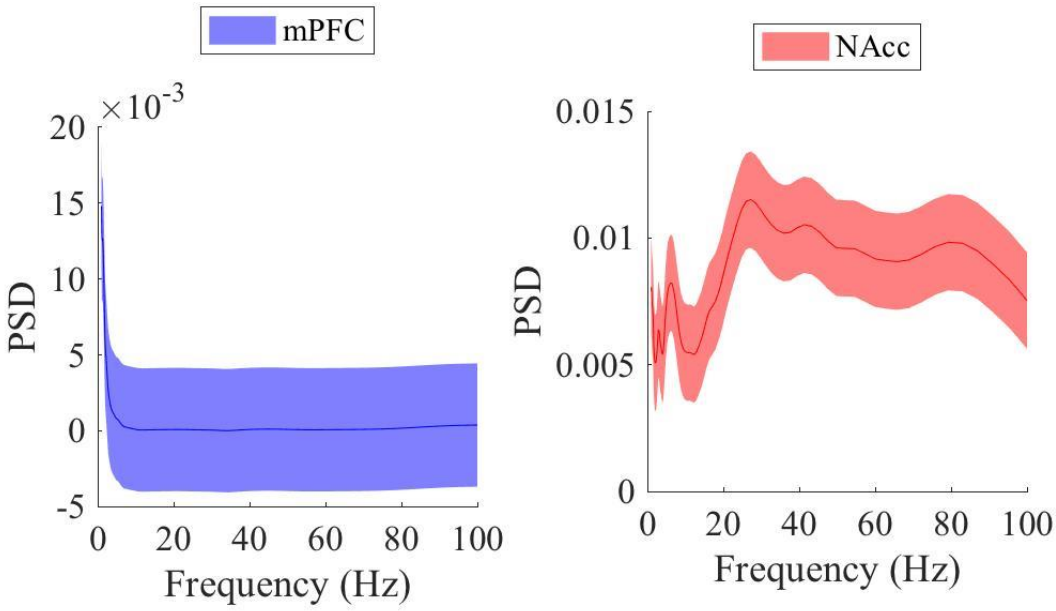



**Figure D. 31.** Region 21 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

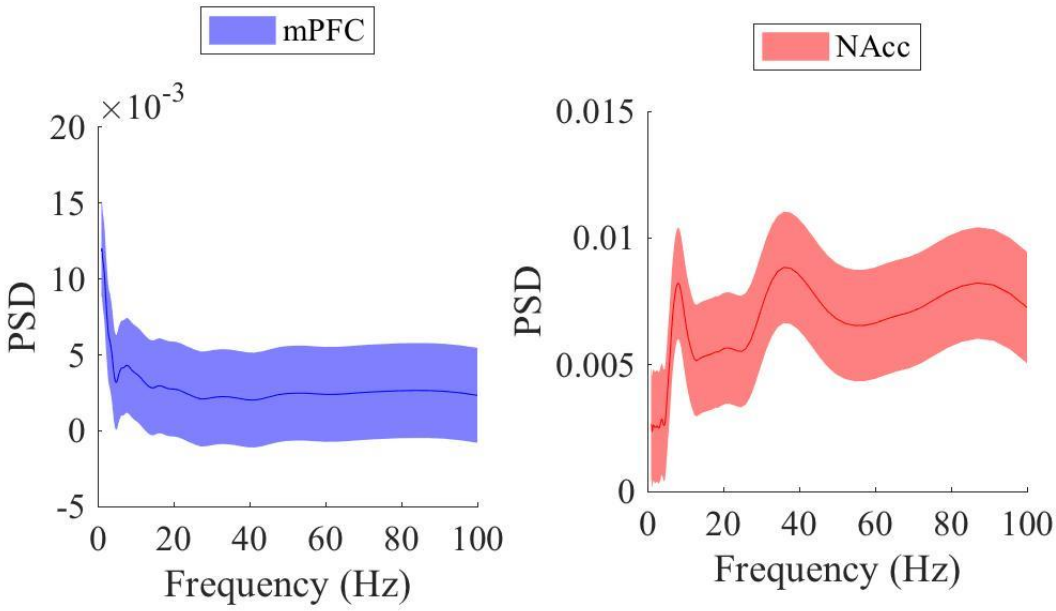



**Figure D. 32.** Region 32 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

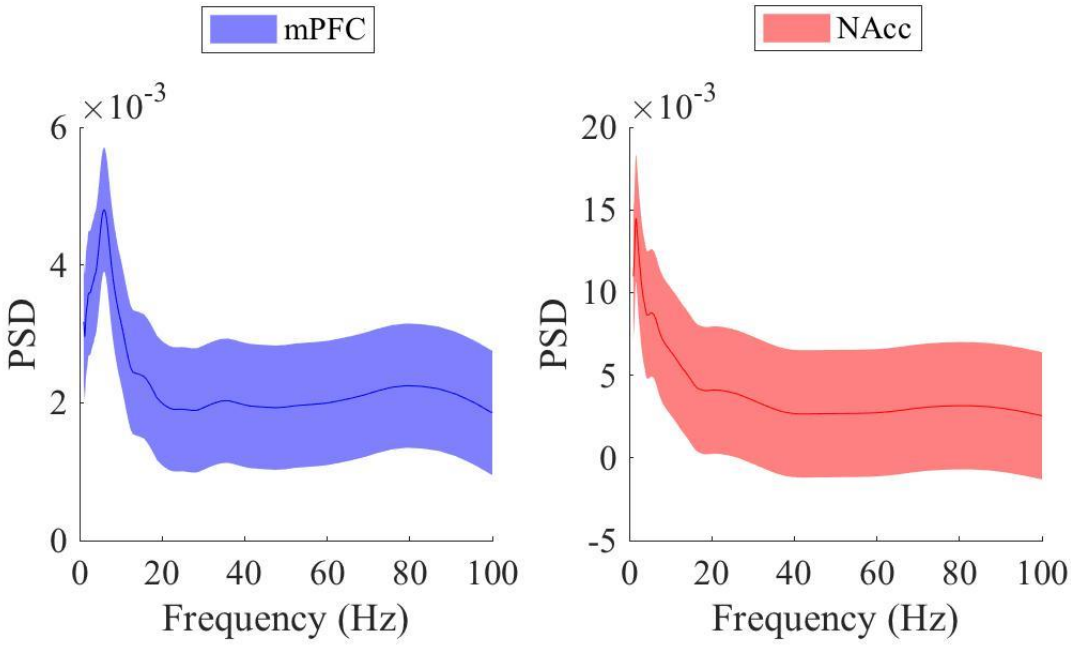



**Figure D. 33.** Region 33 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

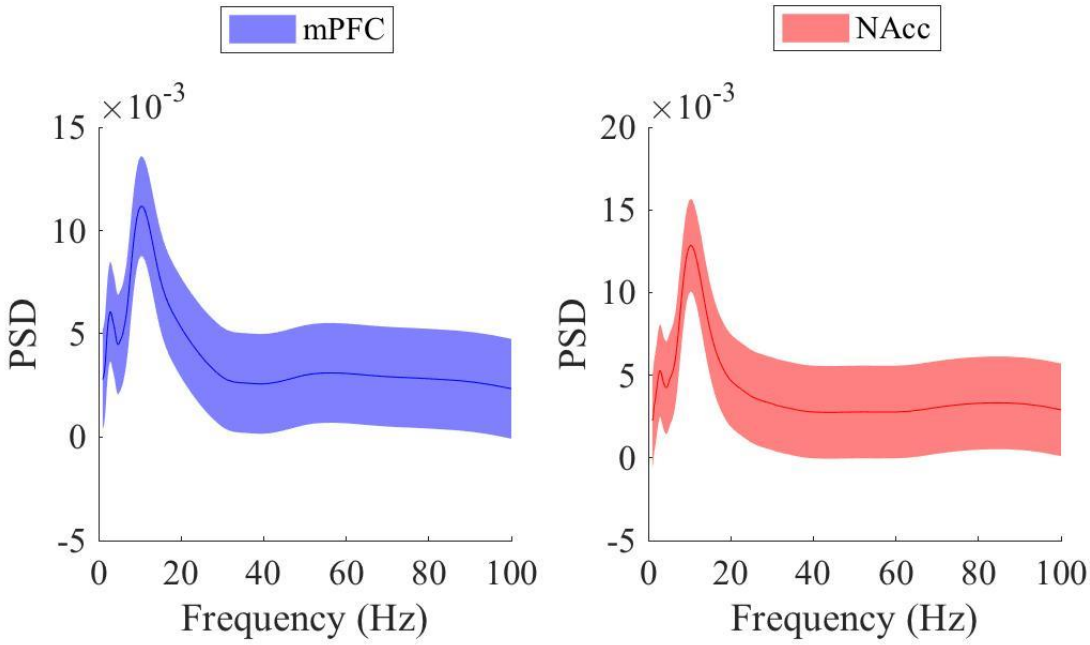



**Figure D. 34.** Region 34 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

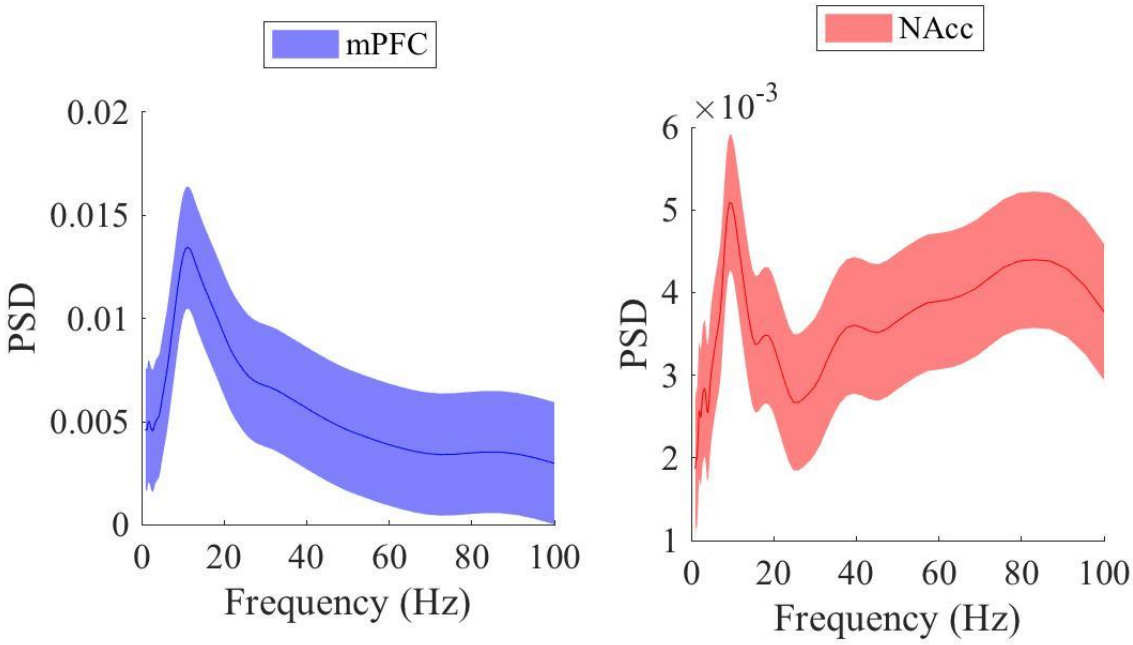



**Figure D. 35.** Region 35 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.

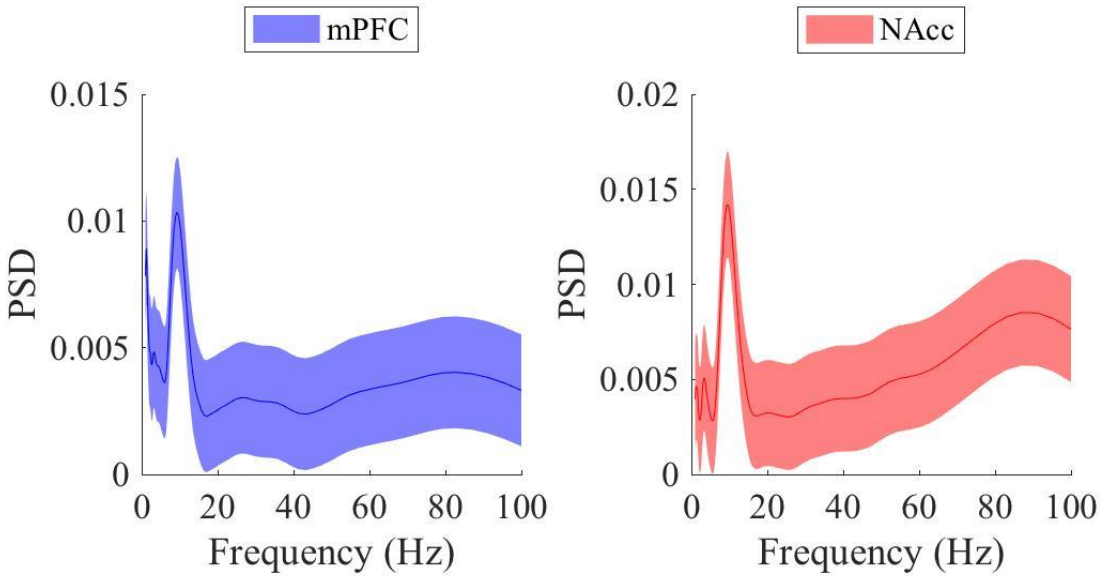



**Figure D. 36.** Region 36 PSD for Morlet wavelet decomposed LFP signal recorded from the mPFC and NAcc.