Interpretable Brain State Manifold for Characterizing Heterogeneity

JULIA HUIMING WANG

A thesis submitted in partial fulfillment of the requirements

for the degree of Doctor of Philosophy

School of Biological Sciences

Cold Spring Harbor Laboratory

June 13, 2023

"*Those who are clever, who have a Brain, never understand anything.*"

Winnie the Pooh

Acknowledgements

Firstly, I want to thank my advisor, Tatiana Engel, for first taking me on as an URP and then as her first graduate student. I want to thank her for her support and for constantly pushing me to become a better scientist.

I'd also like to thank everyone on my committee, Tony Zador, my chair, Saket Navlakha, and Jeremy Borniger. I have learned so much from all of you and have enjoyed learning your different approaches to science.

Thank you to the CSHL School of Biological Sciences staff, Alyson Kass-Eisler, Kim Graham, Kim Creteur, Monn Monn Myat and Alex Gann, for all their help throughout the years. Alyson, Monn, and Alex – thank you for calming me down and letting me bother you all the time.

It would be impossible to thank all of the friends that have supported me throughout the PhD, friends from Indiana, Stanford, and CSHL. I hope I have thanked you enough times in my life that you know who you are.

Thank you also to my previous mentors, Steve Baccus and Jennifer Kowalski for inspiring me to pursue neuroscience and giving me the confidence to do a PhD.

Lastly, I'd like to thank my family 妈妈，爸爸，咪咪，姥姥，姥爷. I am glad I can share this achievement with you.

Contents

List of Figures

List of Tables

List of Abbreviations

AE	**A**ctive **E**xploration
braivest	**Brai**n **St**ate **V**A**E**
CNN	**C**onvolutional **N**eural **N**etwork
DW	**D**rowsy **W**ake
EEG	**E**lectro**e**ncephalo**g**ram
ELBO	**E**vidence **L**ower **Bo**und
EMG	**E**lectro**m**yo**g**raphy
GMM	**G**aussian **M**ixture **M**odel
HMM	**H**idden **M**arkov **M**odel
KL	**K**ullback **L**eibler
LFADS	**L**atent **F**actor **A**nalysis via **D**ynamical **S**ystems
LFP	**L**ocal **F**ield **P**otential
NL	**N**eighbor **L**oss
NREM	**N**on **R**apid-**E**ye **M**ovement
PCA	**P**rinciple **C**omponents **A**nalysis
PSD	**P**ower **S**pectral **D**ensity
QW	**Q**uiet **W**ake
REM	**R**apid-**E**ye **M**ovement
RNN	**R**ecurrent **N**eural **N**etwork
SWS	**S**low **W**ave **S**leep
TN-VAE	**T**ime **N**eighbor **V**ariational **A**uto**e**ncoder
t-SNE	**t**-Distributed **S**tochastic **N**eighbor **E**mbedding
UDR	**U**nsupervised **D**isentanglement **R**anking
UMAP	**U**niform **M**anifold **A**pproximation and **P**rojection
VAE	**V**ariational **A**uto**e**ncoder

Dedicated to my biggest fan

爸爸，我达到我们的梦想～

Chapter 1

Introduction: Brain States, an Overview

What different states do we embody throughout the day? Most obviously, we spend part of the day asleep and another larger, part of the day, awake. Going deeper, different states of sleep may depend on if we are dreaming or not and different states of wake may depend on our arousal, attentiveness to surrounding stimuli, and task at hand. These behavioral states all have bases in the brain – they are controlled by several neuromodulators and expressed through different activity patterns in the brain, creating differing modes of operation known as *brain states* (Vyazovskiy et al., 2009; McNamara, 2019; Mircea M. Steriade and McCarley, 2013).

This chapter establishes a background of sleep and brain states. The chapter begins by defining brain states and (Chapter 1.1) and establishing their importance (Chapter 1.2). Then we discuss deviations from these classical brain states and limitations of existing methods of sleep classification (Chapter 1.3). This introduction contextualizes the work presented in this thesis, including the scientific questions explored and quantitative methods developed to answer those questions.

1.1 What is a brain state?

This thesis focuses mainly on the broad brain states of sleep and wake, which are usually split up into three basic states, Slow-Wave Sleep (SWS), Rapid-Eye Movement Sleep (REM), and wake, which are characterized by a combination of psychological, physiological and neural activity features (Fig. 1.1) (Scammell, Arrigoni, and Lipton, 2017). This neural activity is traditionally measured through Local Field Potential (LFP) or electroencephalogram (EEG), and is usually paired by a measure of intrinsic muscle activity (*muscle tone*) through electromyography (EMG) (McNamara, 2019; Hernan et al., 2017). SWS, which is also known as non-REM (NREM), is characterized by unconsciousness and slow-wave oscillations (1-4 Hz) in neural activity recordings, paired with low muscle tone. REM, on the other hand, is marked by the presence of rapid-eye movements, vivid dreams, and theta oscillations (4-8 Hz) in neural activity, along with muscle atonia (McNamara, 2019; Hernan et al., 2017). Wake is characterized by fast oscillations ($>$ 13 Hz), and an alert, attentive state, and an increase in muscle tone (Poulet and Crochet, 2019; McNamara, 2019; Hernan et al., 2017).

Several neuromodulators regulate the activation and inactivation of these different brain states. Norepinephrine, histamine, and serotonin have all been implicated in controlling wakefulness. Serotonin may also play a role in the ratio of REM to SWS sleep. Acetylcholine is responsible for desynchronization of neuron populations and an increase in acetylcholine corresponds to a decrease in slow wave rhythm (McNamara, 2019; Mircea M. Steriade and McCarley, 2013).

The daily cycling between these different states is known as the circadian rhythm. Humans are typically awake for the light period of the day, and then transition to sleep, aided by melatonin, for the dark period. During sleep, humans

progress through cycles of SWS and REM every ninety minutes, beginning with SWS. It is most common to wake after a REM period, although it is possible to wake from any sleep stage (McNamara, 2019).

There are several interesting deviations of sleep rhythms across species. In all primates, including humans, sleep is consolidated into one single, long bout during the dark phase of the day. Smaller non-primate mammals, such as the mice and rats, instead sleep periodically; this is likely due to their vulnerability as prey. One of the most interesting deviations in sleep cycles is the adaption of unilateral or uni-hemispheric sleep in avian animals and cetaceans. These animals are able to allow one side of the body to enter SWS while keeping the other side active. This adaption may be evolutionarily advantageous for acts such as swimming and flying (McNamara, 2019). Understanding the commonalities and differences in expression of sleep and brain states across species can further understanding of the usefulness of these states and their deviations.

1.2 Why are different brain states important?

Sleep deprivation can severely impair cognitive function, starting with memory loss, reduced clarity of thought, and emotional dysregulation. Long periods without sleep can even lead to visual hallucinations, delusions, and eventually death (McNamara, 2019). Knowing the dangerous effects of sleep deprivation, one might ask: Why is sleep so important and why can't humans live without it? Why do we have different brain states?

Sleep serves several critical roles in maintaining human health and brain function. First, during sleep, the body performs a variety of important metabolic, energy, and immune related functions (McNamara, 2019). Different stages of sleep

also play different roles in learning and memory. SWS is important for the consolidation of short-term memories into long-term storage; subsequently, REM plays a role in stabilizing these memories through a process called synaptic consolidation (Diekelmann and Born, 2010). Additionally, ongoing variations in brain state during wake can affect sensory responses and behavior (Harris and Thiele, 2011; McGinley et al., 2015; Hulsey et al., 2023; Engel et al., 2016). As a result, sleep disorders and impairment, such as insomnia, sleep apnea or narcolepsy, can have a profound and debilitating impact on one's life.

1.3 Classification of brain states

Brain states are traditionally classified from EEG or LFP recordings of brain activity, sometimes paired with EMG activity. First, raw signal is transformed to the frequency domain through Fourier transform or wavelet analysis. This frequency information is often sorted into bands, delta (1-4 Hz), theta (4-8 Hz), beta(13-29 Hz), and gamma (30-80 Hz) (Fig. 1.2A) (Hernan et al., 2017). Different states are then classified by the combination of power from different bands, often visualized through the power spectral density (PSD) (Fig. 1.2B)(Hernan et al., 2017; Watson et al., 2016). This basic method of classification involves human-expert visual labeling of each window of time (often 5 or 30 s). This requires expert knowledge of raw LFP/EEG data and PSD shapes (Alsolai et al., 2022). This can become intractable for large recordings over multiple days for many subjects, which motivates the need for automated methods of brain state labelling.

1.4 Deviations from traditional brain state definitions and limitations

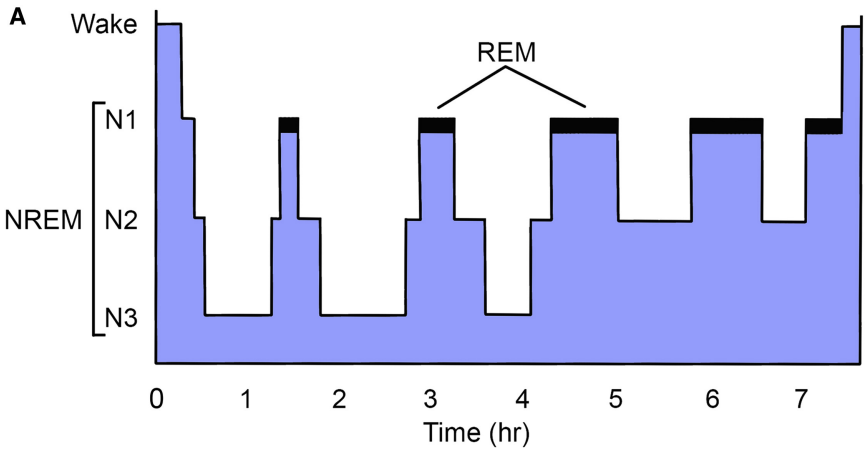
The three canonical brain states of SWS, REM, and wake are generally thought to be (1) discrete, (2) temporally persistent, and (3) spatially uniform across the cortex. However, each component of this conceptual model has been challenged by experimental evidence. First, within each canonical state, the patterns of neural activity vary continuously and exhibit signatures of distinct substates, such as quiet wake (QW) and active exploration (AE) substates of the wake state (Gervasoni et al., 2004; Poulet and Crochet, 2019) or light and deep substates of SWS (McNamara, 2019) (Fig. 1.3B, D) . These variations in neural activity within a state may be better captured not as discrete substates but as a continuum, which also describes transitions between states (Fig. 1.3C) (Gervasoni et al., 2004; Harris and Thiele, 2011). Second, the temporal persistence of states is often interrupted by microstates, which include short periods of wake-like activity in sleep (micro-arousal) and vice versa (micro-sleep), and can occur regularly as well as modulate behavior (Fig. 1.4) (Soltani et al., 2019; Watson et al., 2016). Third, these states may not be global and uniform across the cortex; instead some cortical areas may exhibit signatures of a different state of sleep or wake than others at the same time (Gervasoni et al., 2004; Funk et al., 2016; Vyazovskiy et al., 2011; Nir et al., 2011; Bernardi et al., 2019), suggesting the existence of local brain states or heterogeneity in the expression of brain states across areas (Fig. 1.5) . These observations altogether indicate that the canonical model of brain states is incomplete. However, no alternative framework exists to comprehensively and systematically quantify the spatiotemporal dynamics of global brain state. Developing such a conceptual framework requires large-scale recordings from multiple brain regions simultaneously during the natural variation of sleep and wake states, along with

analytical tools to quantify the expression and propagation of brain states across regions.

Extracting a quantitative description of brain-state dynamics from large-scale, multi-day neural recordings data is an open challenge. Traditionally, this classification of brain states into canonical REM, SWS, and wake from short segments of EMG and EEG/LFP data, known as sleep scoring, has been achieved through expert hand-labeling from the amplitude of neural oscillations of various frequencies. Hand-picked thresholds for combinations of these amplitudes are then used to classify states (Alsolai et al., 2022). These methods require expert intuition and are not robust to potential differences in state expression across subjects or brain regions. As a consequence, there is only an 63-90% consensus in expert labels of brain states (Alsolai et al., 2022). Moreover, manual adjustment of thresholds for different subjects and brain areas is laborious and does not scale to multi-day recordings from many subjects and multiple brain regions. To ease these issues, several supervised deep learning methods were developed to automatically classify brain states (Chambon et al., 2018; Gunnarsdottir et al., 2020; Prabhudesai, Collins, and Mainsah, 2019; Caldart et al., 2020; Allocca et al., 2019). However, these methods still treat expert-labels as ground truth for training, even though expert-labels are subject to inter-rater variability and constrained to three basic states (Fiorillo et al., 2019; Koch, Jennum, and Christensen, 2019). In addition, supervised learning models are trained to predict state labels from only a specific set of data, and are prone to fail when presented with data of a differing quality or with distributional shift (Barger et al., 2019). This can lead to a lack of consistency across models and problems in generalizing to recordings from different subjects, brain areas, or disease states (Fiorillo et al., 2019; Alsolai et al., 2022). Thus, supervised methods are unsuitable to elucidate deviations from the three canonical brain states.

In contrast, unsupervised learning methods do not rely on training labels but instead learn from inherent properties of the data such as statistical structure. Therefore, they are not constrained by existing expert-labels of three canonical brain states and can account for continuous variation in neural activity necessary to capture substates and transition states and have the flexibility to describe heterogeneity in expression of states across regions. In particular, unsupervised dimensionality reduction methods, such as variational autoencoders (VAEs) (Kingma and Welling, 2014) have been used extensively to analyze high-dimensional biological data (Luxem et al., 2022; Sussillo et al., 2016; Higgins et al., 2021; Wehmeyer and Noé, 2018) and can produce interpretable low-dimensional representations of high-dimensional data (Khemakhem et al., 2020b). Thus the VAE can be used to characterize a continuous low-dimensional representation—a manifold—of brain states.

In Chapter 2, we will introduce the VAE as a dimensionality technique and in Chapter 3, we will discuss the challenges produced in validation of VAEs and the solutions we developed. Finally in Chapter 4, we will use our developed method to investigate the questions we have posed around brain state characterization.



B

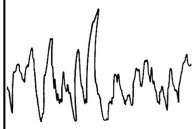
	Wake	NREM sleep	REM sleep
Psychological features	Varying amounts of alertness and attentiveness	Unconscious, or bland thoughts	Vivid, story-like dreams
Physiological features	Sympathetic tone variable	Sympathetic tone low; roving eye movements in light NREM sleep	Sympathetic tone variable; bursts of fast saccadic eye movements
EEG pattern (5 sec)			
Developmental changes	Short wake bouts in infants and young children	Deep NREM sleep abundant in children, but gradually decreases across adulthood	Abundant in infants, steady levels across adulthood; NREM-REM cycle short in infants

FIGURE 1.1: **a**, Over the night, a typical young adult rapidly enters deep NREM sleep (N3) and then cycles between NREM and REM sleep about every 90 min. As homeostatic sleep pressure dissipates across the night, NREM sleep become lighter and REM sleep episodes become longer. **b**, Features of wake, NREM sleep, and REM sleep. Figure from Scammell, Arrigoni, and Lipton, 2017.

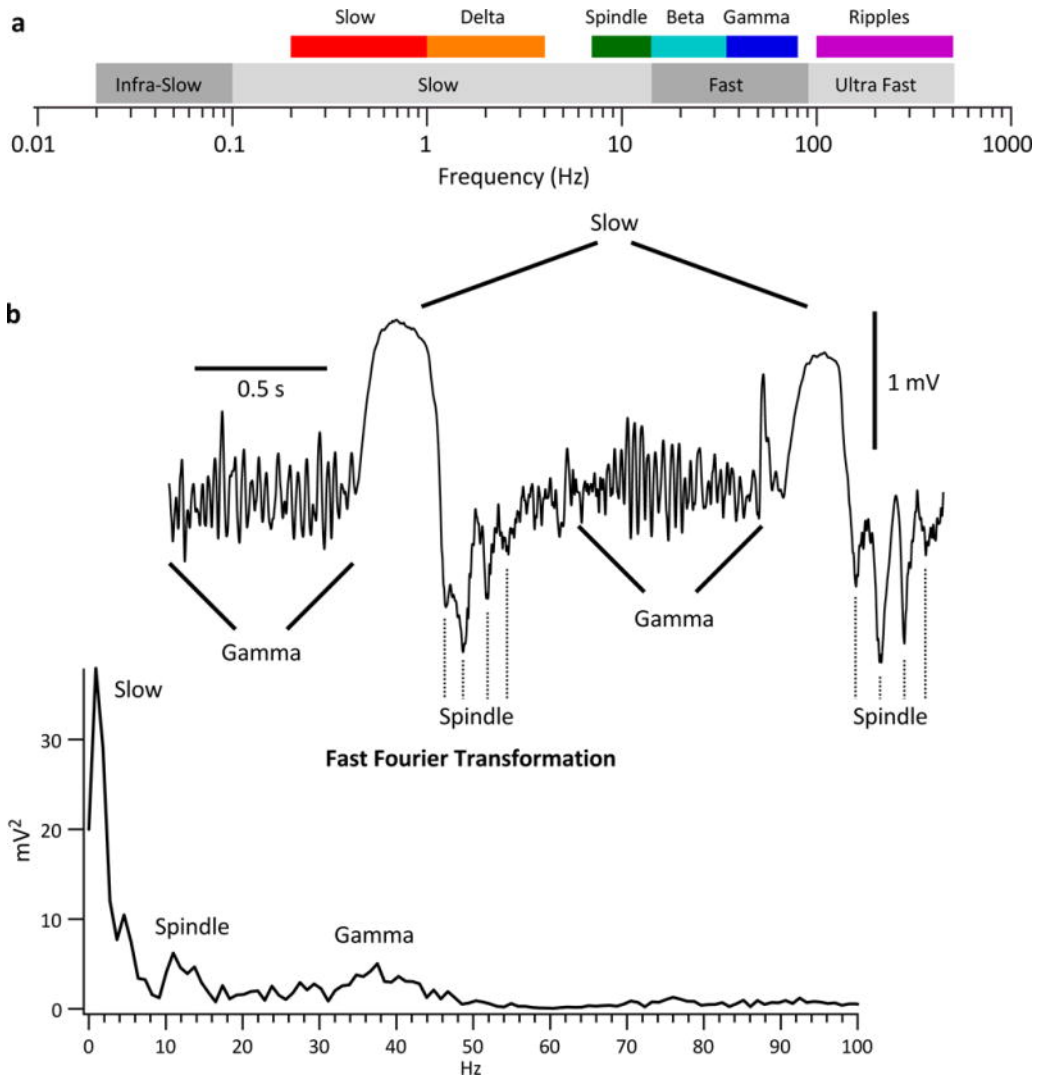


FIGURE 1.2: **Oscillations in the thalamocortical system. a**, Frequency band of oscillations recorded in the thalamocortical system. From: Bazhenov and Timofeev **b**, Above, a segment of local field potential recorded during slow-wave sleep in a cat's associative cortex; below, fast Fourier transformation of the signal shown above. Figure from Hernan et al., 2017.

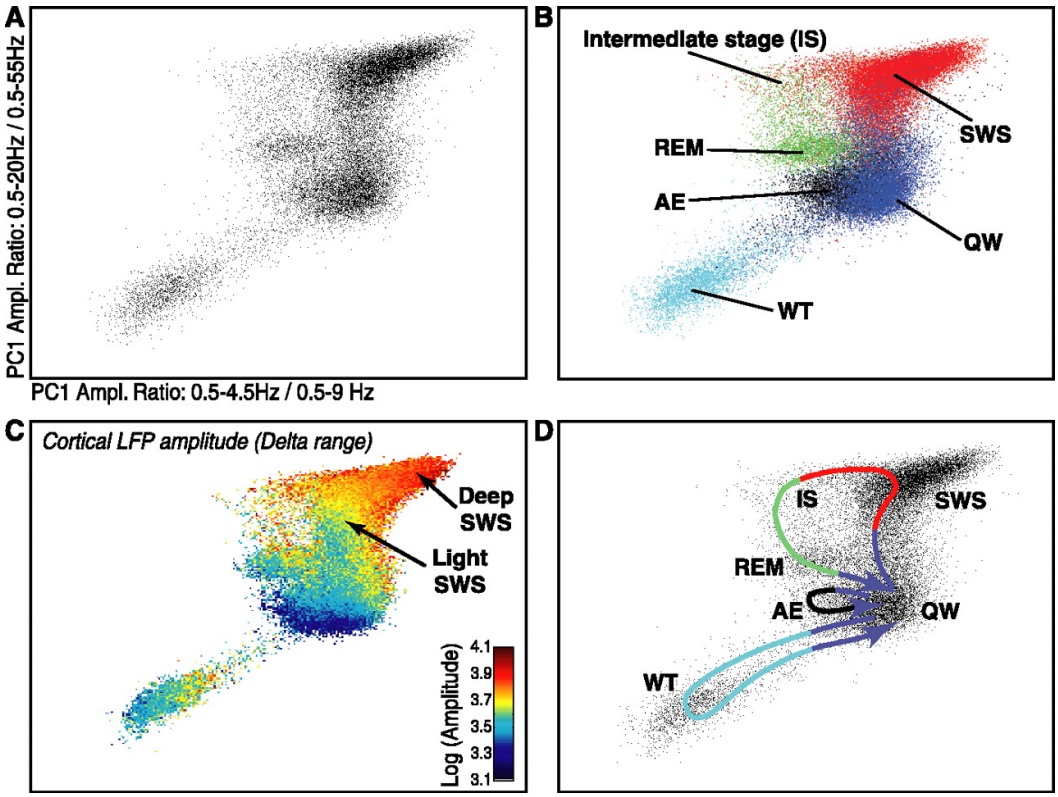


FIGURE 1.3: **Global brain states and two-dimensional state space.** **a**, Scatter plot of the two chosen LFP spectral amplitude ratios, in which four distinct clusters are clearly visible. Each dot corresponds to a 1 sec window for which the amplitude ratios were calculated (48 hr recording, rat 1; for clarity, only one-third of data points, evenly sampled, were plotted). **b**, When color coded according to the behavioral states visually identified, each cluster in the plot corresponds to a distinct state. **c**, The amplitude of cortical LFPs in the delta frequency range (1-4 Hz) is color coded. A fine distinction can be made between light SWS (high spindle density) and deep SWS (mostly composed of delta waves). **d** Transitions between states can be defined as specific trajectories connecting different clusters, with characteristic duration and speed. Typical trajectories are illustrated. Transitions from SWS to REM always course through the IS region. Trajectories also define the polarity of the different clusters. Entrance to and exit from the SWS cluster always occur on one end of the elongated SWS cluster. Figure from Gervasoni et al., 2004.

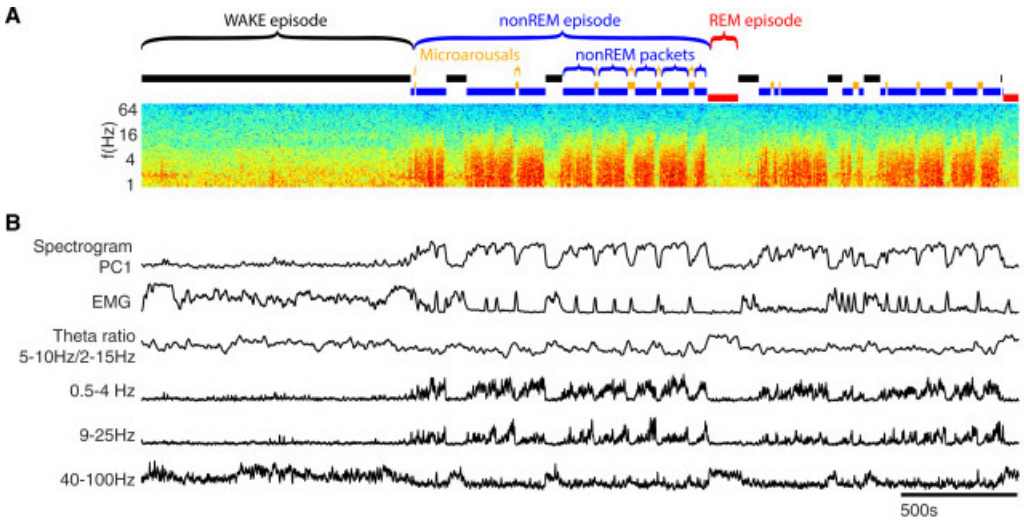


FIGURE 1.4: **Microarousal example a**, Time-power analysis of cortical local field potentials (LFPs). Time-resolved fast Fourier transform-based power spectrum of the LFP recorded from one site of a 64-site silicon probe in layer 5 of the orbitofrontal cortex. Epochs generated by manually approved automatic brain state segregation are shown above the spectrum. **b**, Metrics extracted for state classification. The first principal component (PC1) of the LFP spectrogram segregated nonREM packets from "other" epochs. Non-nonREM epochs with high theta power and low electromyogram (EMG) activity were designated as REM. Remaining epochs were termed either as WAKE (>40 s) or microarousal. Alternating epochs of nonREM packets and MAs comprise nonREM episodes. Integrated power in the delta (0.5–4 Hz), sigma (9–25 Hz) and gamma (40–100 Hz) bands over time is also shown. Figure from Watson et al., 2016.

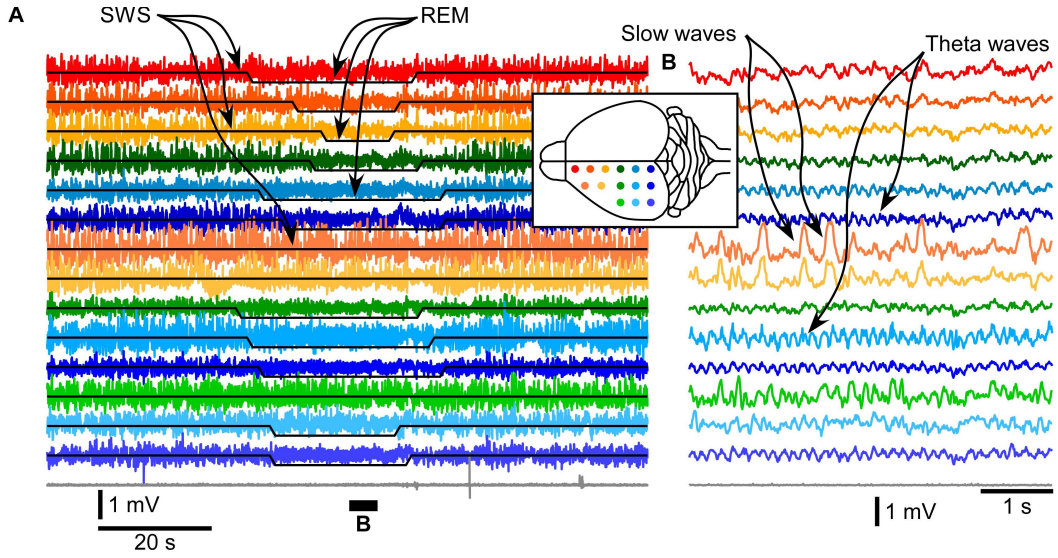


FIGURE 1.5: **Area-specific distribution of electrographic brain states. a**, A segment of 14 LFP channels and muscle (gray) recordings during SWS, REM sleep, and again SWS. Signals from electrodes are color coded and the location of electrodes is indicated in the inserted drawing. **b**, A short segment that was overall qualified as REM sleep, but two fronto-laterally located electrodes show clear slow-wave activity. Figure from Soltani et al., 2019

Chapter 2

Unsupervised Learning, Variational Autoencoders and Variations

As noted in Chapter 1, we turn to unsupervised learning methods, methods that do not rely on ground-truth labels, to analyze brain state data. In this chapter, we discuss the use of unsupervised learning techniques in characterizing biological data before defining the models we build on. This chapter serves as a technical introduction.

2.1 Unsupervised characterization of biological data

Supervised learning requires training a model to predict provided ground-truth labels. However, these methods are unsuitable when ground-truth labels are unknown, difficult to obtain, or unreliable. In these instances, unsupervised learning is useful as it does not rely on training labels but instead learns from inherent properties of the data such as statistical structure. In this section, we will discuss two types of unsupervised learning: dimensionality reduction and clustering.

2.1.1 Dimensionality reduction

Dimensionality reduction is a technique used to transform high-dimensional data to a low-dimensional representation. Intuitively, a large number (D) of variables is measured, but some of these variables may be correlated, so the data itself can be summarized by a smaller number of variables (K) (Cunningham and Yu, 2014). This would be useful in instances when some biological process exists on a low-dimensional manifold (a mathematical surface) but is measured by scientists in a high-dimensional manner. Dimensionality reduction allows one to recover and visualize the shape of that manifold. One example of high-dimensional data in neuroscience is the combination of activity of neurons in the brain, where each dimension corresponds to the activity of one neuron and each data point corresponds to a window of time. However, the activity of these neurons may be generated by some low-dimensional latent process that controls firing rate(Cunningham and Yu, 2014; Musall et al., 2019; Mante et al., 2013; Churchland et al., 2012; Gervasoni et al., 2004). In addition, visualizing the data in a lower-dimensional space may allow for easier correlation to some behavioral output, such as performance in a task (Cunningham and Yu, 2014; Musall et al., 2019; Churchland et al., 2012.

The lower-dimensional representation importantly must preserve relationships between data-points. The types of relationships that are preserved depends on the specific dimensionality reduction technique that is employed. The simplest method is Principal Components Analysis (PCA), which finds a linear combination of the high-dimensional features to define the low-dimensional representations. Other more sophisticated techniques, such as Uniform Manifold Approximation and Projection (UMAP) (McInnes, Healy, and Melville, 2020) or t-Distributed Stochastic Neighbor Embedding (t-SNE) Maaten and Hinton, 2008), allow for

non-linear combinations of variables and seek to preserve various types of distance between points. However, each of these methods has its own pitfalls. Linear methods may not be powerful enough to separate points. Nonlinear methods can be highly sensitive to noise, and the results may vary heavily depending on hyperparameters or initial conditions. If such a technique artificially separates points without a true basis in features of the data, it would be difficult to trust the resulting low-dimensional representation for future interpretation and analysis (Cunningham and Yu, 2014). This issue we will be discussed further in Chapter 3.

2.1.2 Clustering Techniques

While the goal of dimensionality reduction is often to simply visualize high-dimensional data in a lower-dimensional setting, one might also want to actually assign labels to each data-point. This can also be done in an unsupervised manner through clustering. The process of clustering generally involves defining clusters of points depending on their similarity to each other, although there are many different types of clustering techniques. While it is possible to apply clustering techniques to high-dimensional data, they are generally more effective on low-dimensional data. In this section, we will discuss two clustering techniques, Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs).

GMMs seek to define the distribution of data points provided as a mixture of Gaussian distributions. These distributions can then be treated as clusters, and if a data point is most probable to fall into a particular distribution, it is labeled as part of that cluster. These distributions are very flexible, and can be constrained or unconstrained to have different covariance structures. However, this great flexibility can make it difficult to train the GMM; thus, initialization conditions can be very important.

Hidden Markov Models (HMMs) are useful for defining clusters from time-series data. Markov chains are defined by transition probabilities between a set of discrete states, and these probabilities are only influenced by the current state, not any previous state. HMMs allow each of these states in the Markov chain to be measured through noisy emissions, thus making these states "hidden". Thus, a HMM is defined by a set of states, transition probabilities between states, and emission distributions for each state. Training an HMM on time-series data involves finding the set of states, transition probabilities, and emission distributions that most probably generated the observed data. Then each data-point can also be assigned to its most probable state, known as "decoding" states.

Validating a clustering algorithm and choosing the appropriate number of clusters can be a difficult task. This is often done through cross validation, where one holds out a proportion of the data as a validation set, and trains several models from different training and validation set splits. Then, certain metrics are calculated on the validation set and the model with best performance based on the desired metric is chosen. The choice of a suitable metric can be difficult. In the case of GMMs and HMMs, which are both probabilistic models, one can use the log likelihood of the data under the model, which can be thought of as a measure of the probability the data presented was generated by the model. Often, the log likelihood increases monotonically with the number of clusters, so it is impossible to choose the number of clusters based on highest log likelihood. Instead, scientists often use the "elbow method", a heuristic method that chooses the number of clusters based on the percent increase of log likelihood instead. When the log likelihood is plotted against the number of clusters there is often a sharp increase at first before the likelihood begins to plateau, and the number of clusters corresponding to the "elbow" is selected.

2.2 Variational Autoencoder Definition

One powerful method of dimensionality reduction is the variational autoencoder (VAE). The VAE was introduced in 2014 (Kingma and Welling, 2014) as a solution to a problem in Bayesian statistics of inference of continuous latent variables. Since, the VAE and its variants, has been used in a plethora of applications including dimensionality reduction.

The VAE can be thought of as parameterizing a generator of high-dimensional variables from continuous latent variables of a lower dimension. In particular, we assume that high-dimensional variables x are generated by some random process from a latent variable z drawn from prior distribution $p(z)$ (Fig. 2.1). The goal is to understand the parameters that generate z and x (θ), but doing so directly is intractable; so we estimate the true posterior ($p(z|x)$) with $q(z|x)$ and aim to maximize the Evidence Lower Bound (ELBO):

$$L(\theta, \psi; x^{(i)}) = E_{q_\psi(z|x^{(i)})}[\log p_\theta(x^{(i)}|z)] - D_{KL}(q_\psi(z|x^{(i)})||p_\theta(z)) \tag{2.1}$$

The full derivation for the variational lower bound from Kingma and Welling, 2014; Kingma and Welling, 2019 is presented in Appendix B.

The above definition can be thought of as the Bayesian perspective of the VAE. However, another, perhaps more intuitive, perspective is the neural network perspective (Fig. 2.2). In this perspective, we can relate the terms from the above VLB to an encoder ($q(z|x)$) and a decoder ($p(x|z)$). Thus the first term of the VLB can be thought of the reconstruction loss of x and $\hat{x}$ produced by the decoder.

Then, the bottleneck of the autoencoder (z), is the the output of the dimensionality reduction achieved by training the VAE, also known as the latent representation. However, the encoder produces a distribution of z defined by a mean (μ) and variance(σ), and this distribution is sampled before being passed to the decoder, thus making the autoencoder variational. The second term of the VLB is the Kullback-Leibler (KL) divergence between that distribution and a $p(z)$ (which is often a normal distribution), and this acts as a regularizer that prevents overfitting.

2.3 β-VAE

The β-VAE (Higgins et al., 2017) is a popular variation of the original VAE. This variation simply introduces a single hyperparameter beta that weights the KL divergence of the VLB term. Thus, the VLB becomes:

$$L(\theta, \psi; x^{(i)}) = E_{q_\psi(z|x^{(i)})}[\log p_\theta(x^{(i)}|z)] - \beta D_{KL}(q_\psi(z|x^{(i)})||p_\theta(z)) \tag{2.2}$$

In Higgins et al., 2017 they motivate that the flexibility afforded by the inclusion of the β term encourages disentanglement. Disentanglement is the idea that latent variables inferred from the VAE should be disentangled to single, independent generative factors that represent some true components of the data (Higgins et al., 2017; Khemakhem et al., 2020a). VAEs are good candidates for disentanglement, and Higgins et al., 2017 argue that setting $\beta > 1$ encourages disentanglement because the model is pushed to learn a more efficient latent representation. However, others have posited that true disentanglement can only be achieved with some level of supervision or strong inductive biases on the models and data (Khemakhem et al., 2020a; Mita, Filippone, and Michiardi, 2021). The focus on disentanglement has interesting and important implications for achieving latent representations of the model that are interpretable. In chapter 3, we will discuss our results exploring

how VAEs and β-VAE are insufficient in producing interpretable latent representations for neural data and our solution.

2.4 VAEs in neuroscience

One central goal in neuroscience is to discover underlying latent dynamics that govern observable outputs, such as spikes, behavior, or brain states (Cunningham and Yu, 2014). This latent to observable can be modeled as a VAE generative process, and many have leveraged VAEs to uncover these underlying dynamics. One such example, latent factor analysis via dynamical systems (LFADS) posits that neural spikes arise from a number of low-dimensional latent factors, and uses RNN-based sequential autoencoder to uncover these factors (Fig. 2.3) (Pandarinath et al., 2018). They and others argue that identifying these underlying latent factors will allow us to make sense of large amounts of neural activity data and relate these interpretable factors to some behavioral outputs. Since then, several others have used autoencoder-like architectures to disentangle underlying factors governing neural activity, some directly incorporating the relationship between these factors and behavioral output in their models.

These methods have been powerful and useful in furthering our understanding of principles governing neural activity. However, they are also not immune to the issues of all dimensionality reduction methods. In Chapter 3, we will discuss some of these issues in more detail and how we propose to solve them for our use of VAEs.

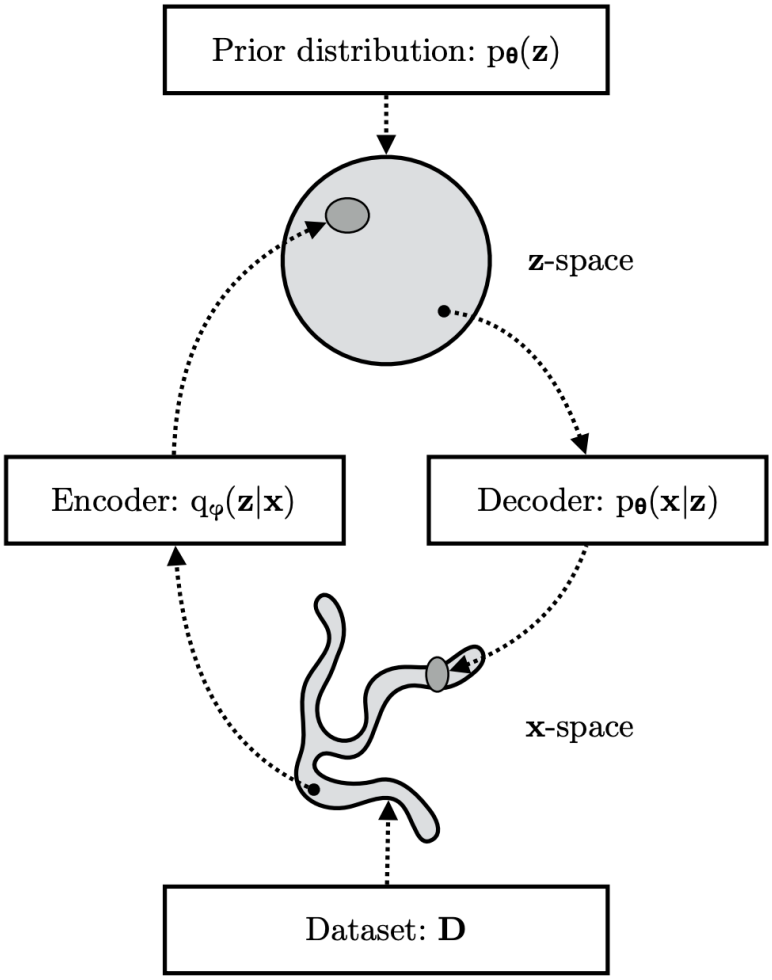


FIGURE 2.1: **Probabilistic view of the VAE** A VAE learns stochastic mappings between an observed x-space, whose empirical distribution $q_D(x)$ is typically complicated, and a latent z-space, whose distribution can be relatively simple (such as spherical, as in this figure). The generative model learns a joint distribution $p_\theta(x, z)$ that is often (but not always) factorized as $p_\theta(x, z) = p_\theta(z)p_\theta(x|z)$, with a prior distribution over latent space $p_\theta(x, z)$, and a stochastic decoder $p_\theta(x|z)$. The stochastic encoder $q_\psi(z|x)$, also called inference model, approximates the true but intractable posterior $p_\theta(z|x)$ of the generative model. Figure from Kingma and Welling, 2019.

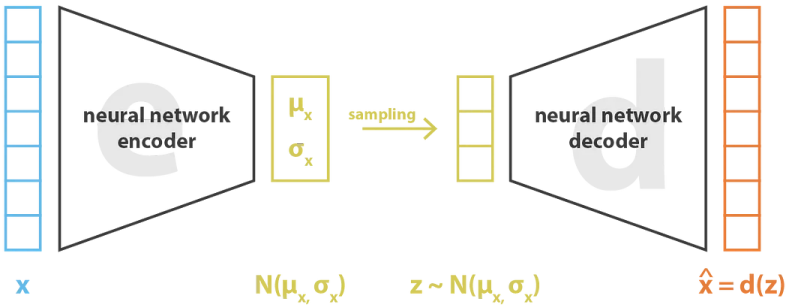


FIGURE 2.2: **Neural Network view of the VAE** In variational autoencoders, the loss function is composed of a reconstruction term (that makes the encoding-decoding scheme efficient) and a regularization term (that makes the latent space regular). Figure from Rocca, 2021.

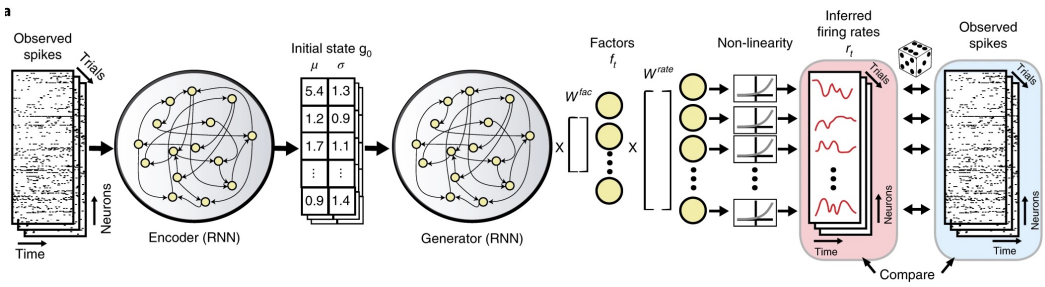


FIGURE 2.3: **LFADS schematic** Schematic overview of the LFADS architecture. Figure from Pandarinath et al., 2018.

Chapter 3

Overcoming overfitting for Variational Autoencoder representations of time series data

3.1 Abstract

Variational autoencoders (VAEs) have been used extensively to discover low-dimensional latent factors governing neural activity and animal behavior. However, without careful model selection, the uncovered latent factors may reflect noise in the data rather than true underlying features, rendering such representations unsuitable for scientific interpretation. Existing solutions to this problem involve introducing additional measured variables or data augmentations specific to a particular data type. We find that for time-series data, predicting the next point in time in VAEs mitigates learning of spurious features. In addition, we introduce a model selection metric based on smoothness over time in the latent space. We show that together these two constraints on VAEs to be smooth over time produce robust latent representations and faithfully recover latent factors on synthetic datasets.

3.2 Introduction

How do we know whether the latent representation obtained from our VAE is useful? In particular, if we want to draw scientific conclusions from this latent representation, we would want to make sure that it is *interpretable* and the features it learns are real. For example, if different instances of a model produced different representations from the same data, it would be impossible to meaningfully interpret these representations. Thus, robustness and reproducibility are prerequisites for interpretability.

However, standard VAEs do not produce robust representations (Locatello et al., 2019; Keshtkaran and Pandarinath, 2019), indicating that these models can learn spurious features unrelated to underlying structure of the data. There are two possible causes of how VAEs may learn spurious features. First, the VAE might learn a functional mapping that is irrelevant to the low-dimensional structure but sufficient for reconstructing the data (Keshtkaran and Pandarinath, 2019). The reconstruction goal of VAEs only requires the model to separate dissimilar points but does not incentivize the model to keep similar points close in the latent space. Thus, there may be several local optima in the loss landscape at which the VAE can map points in the latent space to points in the original space without learning the correct low-dimensional structure.

Second, the VAE can overfit to noise in the data by trying to separate data points based on the noise features. Traditional practice to avoid overfitting is regularization, which involves selection of hyperparameters by optimizing the model's performance on a held-out validation dataset. However, flexible machine learning models with many parameters can generalize well on unseen data despite learning to interpolate perfectly through noise in the training data (Belkin et al., 2018;

Bartlett et al., 2020). In these instances, many models can generalize equally well but have different learned features and therefore interpretations (Genkin and Engel, 2020). This type of overfitting is not an issue if generalization performance is the only goal, but it undermines interpretation of the latent representations learned by the model.

Solutions proposed to avoid the learning of spurious features involve 1) adding inductive bias to the model architecture or through regularization (Locatello et al., 2019; Khemakhem et al., 2020b) and 2) using appropriate model selection metrics to choose robust features that are reproducible across different model instances (Duan et al., 2016; Genkin and Engel, 2020). Here we extend these solutions to the VAEs applied to time series data. We first show that the standard VAEs are prone to learning spurious features. We then introduce an inductive bias in the VAE architecture and a model selection metric which both promote smoothness of latent factors over time. We show that with these changes, VAEs learn robust representations that correctly recover latent factors on synthetic datasets.

3.3 Related Work

3.3.1 Benign Overfitting

Keshtkaran and Pandarinath, 2019 showed that autoencoders can learn spurious features when recovering latents used to create a synthetic spiking dataset. They posit that such overfitting occurs because the model can learn to perform an identity transformation without extracting relevant low-dimensional information from the data. Another explanation might be that while autoencoders are incentivized to separate different points in the latent space, they have no incentive to keep similar points together, which is a problem shared by many dimensionality

reduction methods (Chari, Banerjee, and Pachter, 2021).

3.3.2 Identifiable VAEs and self-supervision

Efforts to use VAEs to uncover true and interpretable latents overlaps with efforts in creating identifiable VAEs. *Identifiability* of a certain model is the constraint that a model's parameterization of a certain set of observations is unique. Consequently, the goal of identifiable VAEs in *representation learning* is to recover low-dimensional factors that correspond to meaningful concepts of the original high-dimensional data, which is known as *disentanglement* (Higgins et al., 2017; Khemakhem et al., 2020b). However, it is nearly impossible to fully disentangle without inductive biases on models and data (Khemakhem et al., 2020b) As a consequence, there have been several efforts, in neuroscience and otherwise, to achieve disentangled representation using VAEs with inductive biases. These biases usually come in one of two forms: leveraging multiple data modalities for regularization or data augmentation for self-supervision. For instance, both pi-VAE (Zhou and Wei, 2020) and ID-VAE (Mita, Filippone, and Michiardi, 2021) introduce an auxillary variable, such as behavior or another neuron, and use the relationship between the generative latents and auxillary variables to regularize the model. Similarly, Liu et al., 2021 introduce SwapVAE, which uses both neural activity and hand dynamics during a monkey performing a reaching task. While introducing these auxillary variables does allow for better disentanglement, not all datasets have associated auxillary variables, making this method unsuitable.

In contrast, self-supervised learning in the form of data augmentation has also been introduced recently to achieve identifiable representations. By asking the model to predict the augmented data point, it is ensured that the model learns to preserve features that are shared between original and augmented data. Sinha and

Dieng, 2021 use a semantic-preserving transformation of an image (such as a rotation) as a self-supervised label, and Liu et al., 2021 drop or swap spikes in a spiking neural dataset as a form of data augmentation. These types of data augmentation are specific to the dataset presented and a semantic-preserving augmentation is not obvious for every type of data. For instance, removing time segments of local field potential (LFP) data may remove more important information than simply dropping individual spikes in a spiking dataset. Therefore, we were incentivized to find another type of inductive bias that could be applied broadly to time-series data.

3.3.3 Lack of Model Selection Metrics

Even with inductive biases added that lead to disentanglement, there is still a need for model selection metrics. In supervised settings, a metric such as validation classification accuracy can be used for model selection, especially if there is no need to measure the interpretability of the model. However, it is still common to use model performance as a model selection metric in unsupervised settings, despite the fact that model performance and interpretability are not necessarily correlated. Some recent work has proposed evaluation metrics that ask for model performance on predicting to an external variable (Higgins et al., 2017; Pei et al., 2021); however, these metrics still rely on some external label rather than evaluating the quality of the latent representations themselves. Thus, there is a need for metrics that can be used for VAE model selection. Duan et al., 2020 propose that such a metric should be based on similarity of latent representations between trained models, as disentangled representations should be similar whereas non-disentangled or overfit representations can have degeneracy. They introduce Unsupervised Disentanglement Ranking (UDR) which uses pairwise comparison between representation from models with same hyperparameter settings as a model

selection metric. However, their proposed metric is computationally complex in order to account for rotations of these representations, so we again leverage time to find another metric which can be easily computed despite rotations and show that this metric correlates with representation invariance across models.

3.4 Methods

3.4.1 Datasets

To study how standard VAEs learn spurious features and to test how our approach overcomes this problem, we use synthetic data with known low-dimensional structure as well as biological neural recordings data.

Gaussian clusters: We generated synthetic data points from 3 clusters mimicking the wake, rapid eye movement (REM) and slow wave sleep (SWS) brain states. Each cluster was modeled as a Gaussian distribution in 33-dimensional space, where dimensions represent local field potential (LFP) power in 30 frequency bands, total power of the electromyography (EMG) signal, body temperature, and accelerometer data with mean and variance of the Gaussian distributions matched to the biological data.

Spiral: We sampled data points uniformly along a two-dimensional spiral, with time index increasing monotonically along the spiral starting from the innermost point. We then nonlinearly embedded these points into 30-dimensional space and added Gaussian white noise.

Hidden Markov Model (HMM): We generated data points from a three-state HMM with emissions sampled from 31-dimensional Gaussian distributions with mean and variance and transition dynamics matched to wake, REM and SWS states in biological data.

Visual cortex LFP data: We used combined EMG and LFP data recorded from the visual cortex of a mouse during continuous 24-hr recordings over 12 days. Each data point represents a 2 second time window. For this 2-second time window, we extracted the LFP power in 30 frequency bands and total EMG power to form a 31-dimensional feature vector representing the signals. A subset of data points had expert-provided labels of wake, REM and SWS brain states.

3.4.2 Model Architecture

We envision that underlying neural dynamics evolve on a low-dimensional manifold via a Gaussian random walk. Then, these underlying neural dynamics are measured as spikes or potential, both of which exist in a higher dimensional space (combination of spikes from many neurons or combination of oscillation power from many frequency bands).

We add an inductive bias to the VAE architecture that promotes learning the underlying low-dimensional structure in time-series data. We assume that latent factors evolve smoothly in time over the latent space. Thus, we constrain the VAE architecture to be autoregressive by modifying the objective from reconstructing the original data point to predicting the next point in time. In this way, the function the model parameters learn encompasses both the generative function that maps low-dimensional latent factors to observables and the transition probability over time between points in the latent space. Thus, we minimize the following loss function:

$$L = \text{MSE}(x_{t+1}, d(e(x_t))) + \beta \text{KL}(q(z)||p(z)). \tag{3.1}$$

Here x_{t+1} is the data point at time $t+1$, $d(e(x_t))$ is the result of passing x_t through the encoder (e) and decoder (d), that is prediction of x_{t+1} from x_t. $q(z)$ is the distribution of points z_t in the latent space, which is the bottleneck layer of the VAE, and

$p(z)$ is the prior distribution. We allow for a weight on the KL divergence term β, which was proposed as the β-VAE Higgins et al., 2017. We call our autoregressive VAE the Time-Neighbor VAE (TN-VAE).

3.4.3 Model selection metric

To prevent learning spurious features due to overfitting to noise in the training and validation data, we define a model selection metric based on our assumption that latent factors evolve smoothly over time in the latent space. For each model instance, we calculate the distance in the latent space between the representations of each data point and the next point in time (normalized by the overall size of the latent manifold). We call this metric the Neighbor Loss (NL) defined as:

$$\text{NL} = \sum_{t=0}^{N-1} \frac{|z_{t+1} - z_t|}{\bar{z}}, \tag{3.2}$$

where $\bar{z}$ is the average distance from the origin across latent representations of all N points in the dataset. Theoretically, if transitions between points in the latent space follow a Gaussian random walk, then minimizing the absolute distance between latent representations of neighboring points in time is equivalent to maximizing their log-likelihood

3.4.4 Training Details

For each dataset, we trained a vanilla variational autoencoder and a TN-VAE. We use 2 layers of 250 units each for the encoder and decoder. We vary 3 hyperparameters: batch size (1000,5000,10000,50000), learning rate (1^{-2}, 1^{-3}, 1^{-4}, 1^{-5}), and β (1^{-2}, 1^{-3}, 1^{-4}, 1^{-5}). For each hyperparameter combination, we train 4 models with different training/validation splits and initializations. We additionally have a held-out test set that we use to evaluate each model. We use the

procrustes alignment distance between test set encodings as a proxy for similarity of trained models allowing for rotation.

3.5 Results

3.5.1 VAE learns spurious features

To illustrate that standard VAEs are prone to learning spurious features, we train a VAE on a synthetic dataset of points drawn from 3 high-dimensional Gaussian clusters (Fig. 3.1A). Both the training and validation loss decrease monotonically throughout 300 training epochs (Fig. 4.1B). Thus, the standard approach would be to select the model at the last training epoch that has the lowest validation loss. However, despite good reconstruction performance on validation data, the VAE not only fails to correctly separate the 3 clusters in the learned latent representation, but also learns spurious features. Over training epochs, the latent representation gradually loses the smooth Gaussian shape and feathers into streak-like patterns. (Fig. 3.1C). Moreover, separate model instances trained on different splits of the data achieve similar validation loss but uncover different latent representations (Fig. 3.1D), indicating that the VAE learns spurious features that are not robust.

3.5.2 Priors on smoothness over time promote learning true and not spurious features

We introduce a modified model architecture TN-VAE and a NL model selection metric which promote smoothness of latent factors over time (Methods). We use two synthetic datasets with known low-dimensional structure (Fig. 3.3) to test how these modifications affect the quality of learned latent representations in

comparison to the standard VAE and standard model selection based on the validation loss. For both datasets, the standard VAE with validation-loss model selection fails to uncover meaningful structure from the data and learns spurious features (Fig. 3.3A,B). TN-VAE introduces an inductive bias that pushes the model to learn meaningful features, but can still overfit to noise in the data (Fig. 3.3D,E). Using TN-VAE in combination with NL model selection metric results in most faithful reconstruction of the original low-dimensional structure in the data without learning spurious features (Fig. 3.3G,H).

We then apply each approach to biological LFP data recorded in the visual cortex of a mouse over normal day and night activity. While the ground truth structure in these data is unknown, the domain knowledge suggests the existence of 3 major clusters in these data corresponding to wake, REM, and SWS brain states. The standard VAE fails to separate the clusters at all and instead learns spurious features (Fig. 3.3C). The TN-VAE with validation-loss model selection separates the clusters but shows signs of overfitting to noise (Fig. 3.3F). Finally, TN-VAE with the NL model selection metric produces a smooth latent representation with 3 clusters that match the human-expert labeling of wake, REM, and SWS brain states (Fig. 3.3I).

3.5.3 NL metric selects for robust representations

Finally, we tested how our NL model selection metric corresponds with robustness of learned representations. For the spiral and LFP datasets, we chose a variety of different hyperparameter configurations and for each configuration, we trained 4 separate model instances with unique training/validation splits. Each model instance produces a latent representation of the same held out test dataset. We then measured the average procrustes distance Gower, 1975 between latent representations of the test set for each pair of model instances. Lower procrustes

distance indicates more consistency in representations across model instances and thus, can be used to assess the robustness of learned latent representations. We find that NL correlates well with procrustes distance, whereas validation loss is not correlated with how consistent the latent representations are across model instances (Fig. 3.4). Intuitively, if the model is underfit, there would be no correlation between the structure of the latent representation and time information. If the model is overfit, the model separates neighboring points in time based on noise features that are random across training instances. As a result, minimizing the NL metric as a criterion for model selection leads to selection of robust latent representations.

In this chapter we established that VAEs are prone to benign overfitting and adding inductive bias using smoothness over time can help avoid this. In addition, we saw that using the distance between neighboring points in time in the latent space can be used as an additional model selection metric, and when we do so, we find latent encodings that are robust over many training epochs. This allows us to be confident in our use of VAEs in analyzing brain state data.

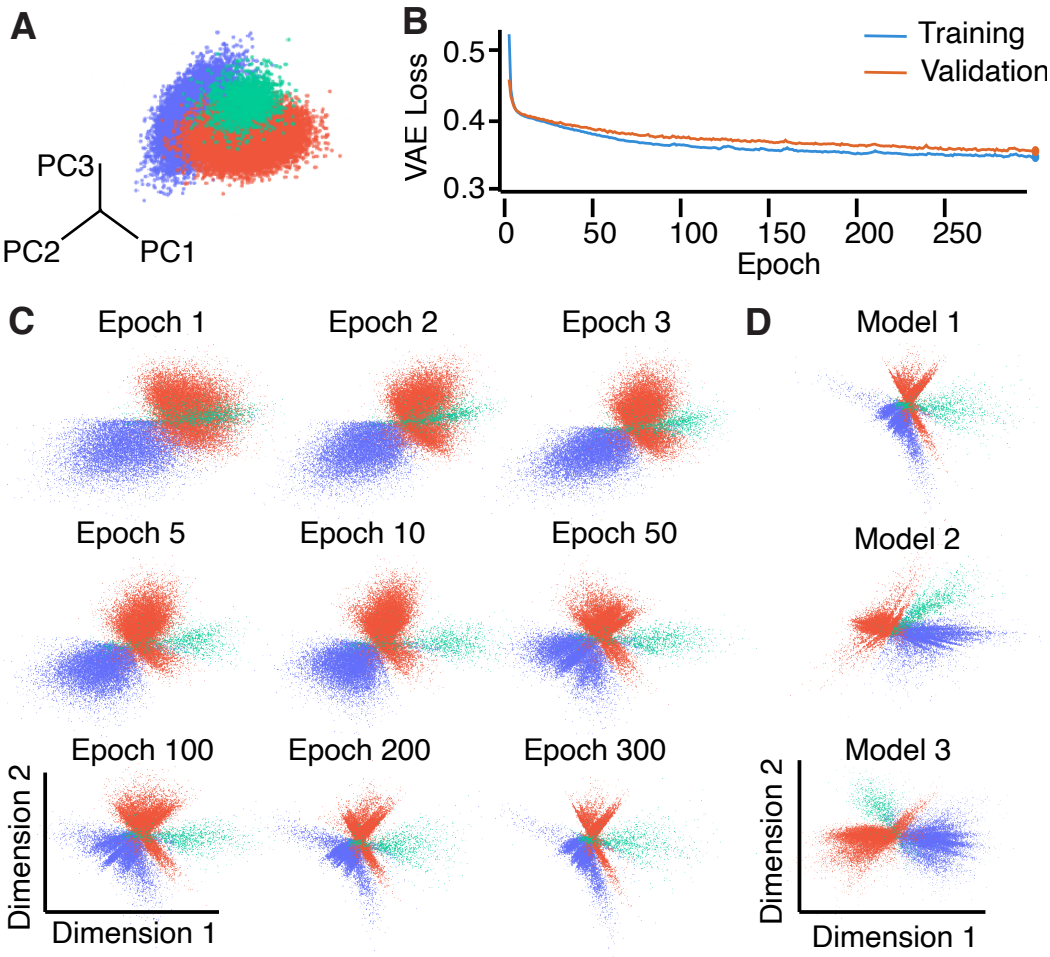


FIGURE 3.1: **A**, Synthetic dataset with 3 Gaussian clusters in 33-dimensional space, visualized by projecting on the first 3 principle components. Colors indicate the cluster from which each point was sampled. **B**, Training and validation loss for a VAE trained on the Gaussian clusters dataset. **C**, Latent representations learned by the VAE at different training epochs. **D**, Latent representations after 300 epochs in 3 separately trained VAEs with similar validation loss show lack of robustness.

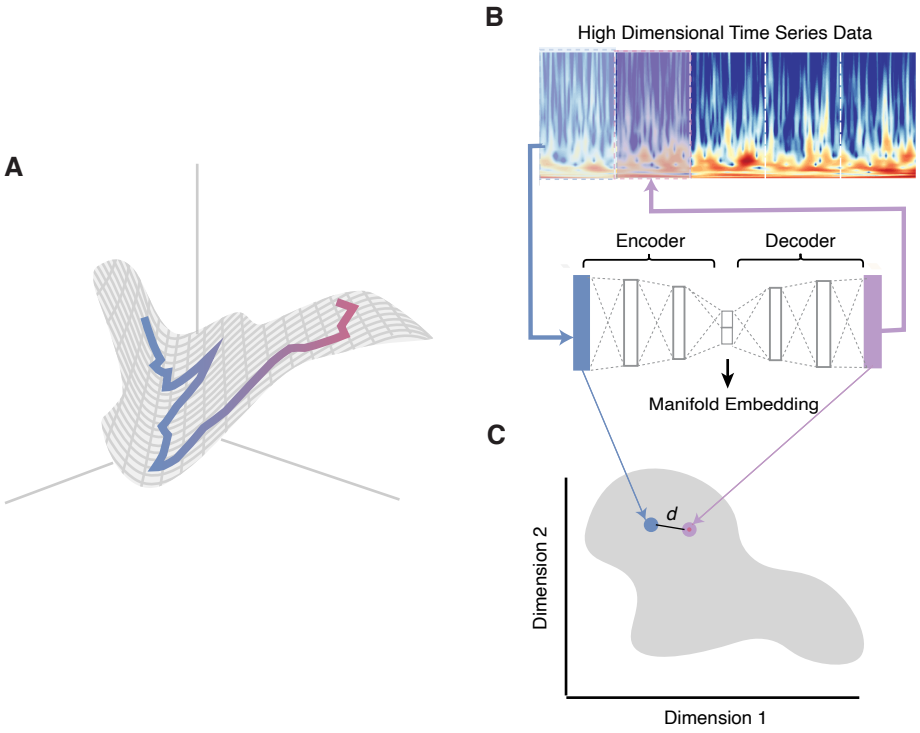


FIGURE 3.2: **Adding time as an inductive bias to VAE A**, We envision neural data as a random walk on a low-dimensional surface in a high-dimensional space. **B**, Neighbor VAE Architecture. High-dimensional time series data is binned to desired time windows and represented as a high-dimensional vector. This is passed into a VAE that predicts the vector for the next point in time. **C**, The distance between neighboring points in time on the manifold embedding discovered by the VAE is used as a validation metric.

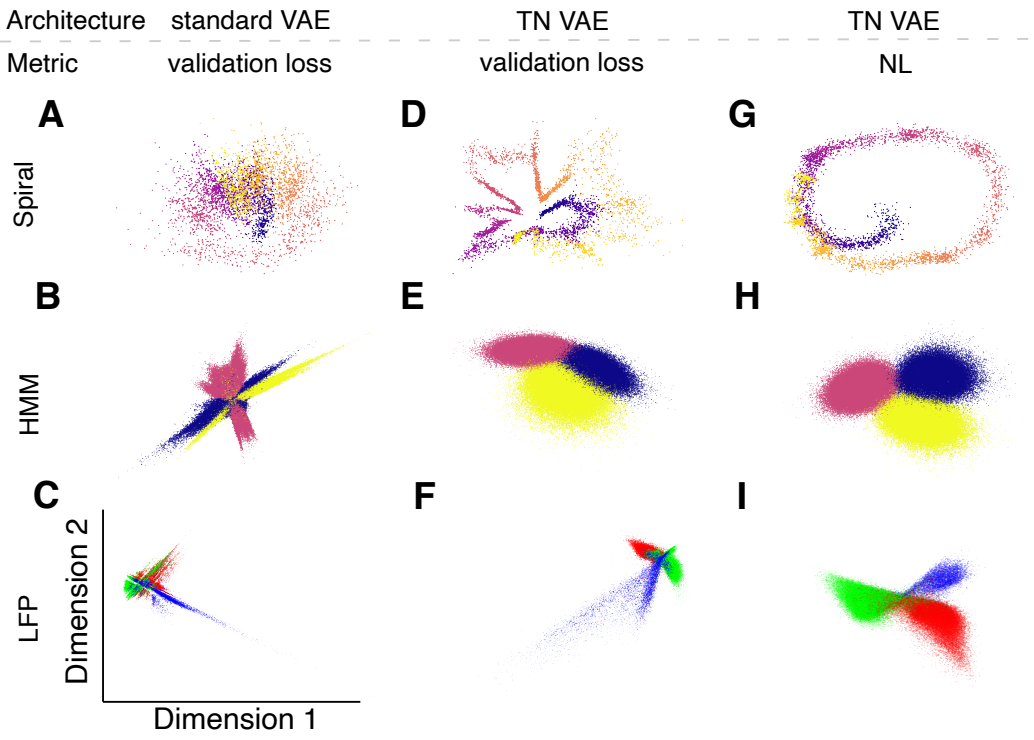


FIGURE 3.3: Latent representations on a held-out test set learned by the standard VAE with validation-loss model selection (**A-C**), TN-VAE with validation-loss model selection (**D-F**), and TN-VAE with NL model selection (**G-I**) for three datasets: synthetic spiral (**A, D, G**, points colored by the position on the original spiral), synthetic HMM (**B, E, H**, points colored by the 3 HMM states), and biological LFP data (**C, F, I**, points colored by expert-provided labels of wake, REM, and SWS brain states).

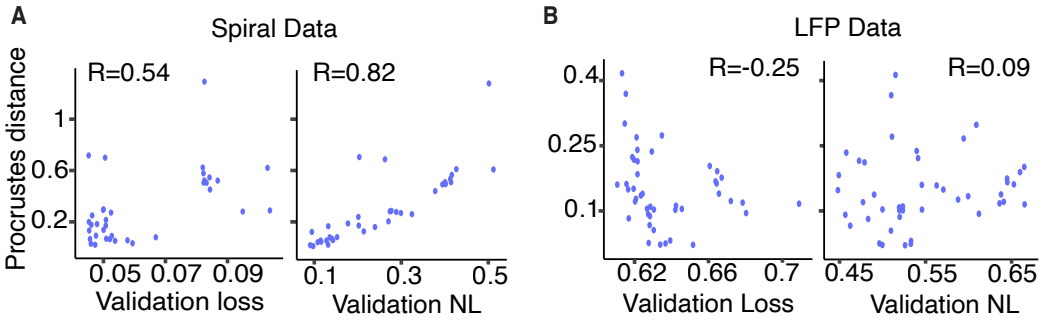


FIGURE 3.4: For each hyperparameter combination, we train 4 model instances on different training/validation splits. We measure the procrustes distances between latent representations learned by each pair of model instances on a test set. We also calculate the average validation loss and average validation NL over the 4 model instances for each hyperparameter combination. NL correlates more with procrustes distance than validation loss for synthetic spiral (**A**) and LFP data (**B**).

Chapter 4

A manifold of heterogeneous vigilance states across the cortex

4.1 Abstract

Brain states are conventionally divided into wake, slow wave sleep (SWS) and rapid eye movement (REM) sleep based on distinct patterns of neural activity and muscle tone. These brain states are conventionally thought to be discrete, temporally sustained, and spatially global. Recent evidence indicates that this conventional definition of brain states may be insufficient, but such analyses have not been done systematically with large-scale neural recordings. Here we show this insufficiency using simultaneously recorded multi-day electromyogram (EMG) and local field potentials (LFP) across the cortex. We developed a computational approach to place these recordings on a low-dimensional manifold visualization. With this manifold, we characterized 9 substates of sleep and wake and their differences in expression and dynamics throughout the cortex. Particularly, we found a lack of REM-like activity in the lateral somatosensory cortex and an increase in theta

rhythm in frontal cortex during wake. Our work provides a comprehensive quantification of deviations from canonical brain-state definition with a novel computational framework for analyzing brain states.

4.2 Introduction

In this study, we visualize brain states on a continuous manifold and systematically quantify a new conceptual model of brain states across the cortex. We recorded extracellular LFP recorded continuously across multiple sites in the mouse cortex for several days during normal wake and sleep cycles. We then developed a model based on a VAE to discover a continuous manifold from these recordings. The model not only accurately separated the three canonical brain states on the manifold, but also captured microstates and transition states. It did so by uncovering a continuum of states that could be quantified as substates with reproducible transition dynamics. When applied to LFP activity recorded simultaneously on multiple electrodes across the cortex, our model uncovered heterogeneity of sleep state expression across the cortex, specifically lack of of REM-like sleep in frontal-lateral regions and increase in AE-like wake in frontal regions of the cortex. Lastly, we found that further coexistence of differing states across regions occurred during global state transitions or as spatially local microstates. Through characterizing brain states on a low-dimensional manifold, we show that brain states are not always discrete, temporally persistent, and spatially uniform, thus providing a comprehensive description of spatiotemporal dynamics of brain states across the cortex.

4.3 Methods

4.3.1 Data Collection

Surgery, recordings, and canonical state detection are performed as described in Soltani et al., 2019. Briefly, mice are implanted with custom-made electrodes 600 um from the cortical surface in either 14 or 3 different cortical regions. Recordings were collected continuously for 24 hours per day for at least 3 weeks. Data is recorded at 1000 Hz. Custom-written routine was used to automatically detect states of the brain, with 5 second windows with a sliding time window of 1s (Bukhtiyarova et al., 2016; Soltani et al., 2019).

4.3.2 Artifact Removal

Artifacts are common in electrophysiological recordings and need to be removed before data analysis. Artifacts can arise from mouse movement affecting electrode recording. To remove artifacts, we highpass filtered the raw LFP data at 0.8 Hz and removed. We also remove any time windows where at any point LFP amplitude is greater than 1.5 mV.

4.3.3 Data preprocessing

First, we normalized raw LFP and EMG recordings through z-score for each session, where mean and standard deviation was calculated separately for each session. Then we performed continuous wavelet transformation using Complex Morelet Wavelets. The formulation of the complex Morelet wavelets is as follows:

$$\psi(t) = \frac{1}{\sqrt{\pi B}} \exp^{-\frac{t^2}{B}} \exp^{\mathrm{j}2\pi C t} \tag{4.1}$$

where B is the normalized bandwidth and C is the normalized center frequency. We used 30 wavelet bands with center frequencies evenly spaced on a logarithmic scale between 1 Hz and 50 Hz. For each time step, a wavelet coefficient is obtained for each frequency as a result of the wavelet transformation. The wavelet coefficients are at the same sample frequency as the original data, 1000 Hz. We then calculate power spectral density (PSD) through $\log_2(|c_t|^2)$, where c_t is the wavelet coefficient at time t. We downsample the PSD to 100 Hz and then bin to 2-second windows. We average the PSD over time within each window. The 30 PSD coefficients of EMG are integrated to obtain the average EMG power. The 30 frequency bands of LFP are combined with the average EMG power to create a 31-dimensional vector input to our model. We z-score each feature of our input vector separately, with mean and standard deviation calculated per session. We denote the feature vector for the 2-second time window t as x_t.

4.3.4 Variational Autoencoder (VAE)

We pass our 31-dimensional vector consisting of 30 frequency bands of LFP and the average EMG power into a type of variational autoencoder(VAE) known as a β-VAE (Higgins et al., 2017). A variational autoencoder consists of an encoder, which projects a high dimensional point onto a distribution low dimensional representation and a decoder which decodes a point on the low dimensional representation back to the high dimensional space. By training the autoencoder to successfully reconstruct the original high-dimensional data, we ensure that the low-dimensional representation, also known as the bottleneck layer, contains the most informative representation to separate dissimilar points. The variational autoencoder also adds regularization by enforcing a prior on the low-dimensional distribution. We modify the β-VAE loss function such that it seeks to predict the next window in time in order to minimize benign overfitting (citation) and to ensure

the model learns robust features without creating spurious clusters. We minimize the following loss function:

$$L = MSE(x_{t+1}, \hat{x}_{t+1}) + \beta KL(q(z)||p(z)). \tag{4.2}$$

Here x_{t+1} is the feature vector at time $t+1$, and $\hat{x}_{t+1}$ is its prediction from the feature vector x_t at the preceding time t. $q(z)$ is the distribution of points z_t in the latent space, which is the bottleneck layer of the β-VAE, and $p(z)$ is the prior distribution. The space on which the low-dimensional points lie represents the low-dimensional manifold that brain states evolve on.

Our encoder consisted of two 250-dimensional fully connected layers and a 2-dimensional latent representation. The decoder architecture also consists of two 250-dimensional fully connected layers, and ReLU activations were used throughout. We specifically constrain one dimension of this 2-d latent representation to be the average EMG power and learn the other dimension from the β-VAE. We did so because average EMG power is the most important factor in determining the wakefulness of the animal. Thus, we have one dimension that represents wakefulness and another that represents cortical LFP activity. Both dimensions are passed to the decoder during training. We used a β value of 0.001. For VAEs trained on data from a single electrode, we used learning rate 10^{-4} and batch size $10,000$. These models are coded using Tensorflow and Tensorflow-Probability.

4.3.5 Model Validation.

For each subject, we compiled 10–12 sessions (subject 0– 10 sessions, subject 1 – 12 sessions) of approximately 24-hour recordings (some recordings are more or less depending on when the experimentalist decided to start or stop the recording). This dataset was then split randomly into 4 training and validation sets,

where 80% of datapoints are in the training set and 20% are in the validation set. One additional session was reserved for testing.

Unsupervised deep learning techniques lack standardized metrics and procedures for model validation. In particular, a type of overfitting known as benign overfittingBartlett et al., 2020, where the model learns spurious features while validation loss still decreases, is common. We hypothesized that multiple models trained on different subsets of data from the same mouse with the same hyperparameter configuration should achieve the same latent manifold. If these models achieve diverging manifolds, this is an indication of benign overfitting. Thus, in order to choose the ideal hyperparameter combination for our model, we trained multiple models for each hyperparameter configuration on different training/validation splits. We show that our selected hyperparameter combination finds a robust manifold across different models (Figure supp.). We trained separate models on datasets for each subject, but we also found that a model trained on data from one subject could be applied to data from another subject with the same result as a model trained on data from that subject.

4.3.6 Frequency Band Calculation

To calculate the power for frequency bands delta (1 to 4 Hz), theta (4 to 12 Hz), alpha (8 to 10 Hz), sigma (10 to 16 Hz), beta (16-29 Hz), and gamma($> 30Hz$) (Hernan et al., 2017), we calculated the spectral power of each 2-second window using Fourier transform. We then integrated the spectral power within the corresponding range and divided by the total spectral power.

4.3.7 HMM Fitting

A Hidden Markov Model (HMM) is useful in analyzing time-series data, such as the data on our latent manifold. A Hidden Markov Model learns the initial probabilities, transition probabilities, and emissions for any number of states to define a state sequence corresponding to the data. We trained our model using Gaussian emissions to fit the sequence of points on the latent manifold. In order to determine the number of states, we performed 3-fold cross validation for 2-14 states and calculated the average negative log likelihood per number of states. We chose the the number of states when the negative log likelihood decreased by less than 5%. For this HMM, we found 8 states by this procedure.

4.3.8 Microstate Identification

We obtain labels fitting a HMM to the encodings, finding eight states. Then, we obtain a smoothed state sequence by taking the mode of 30 second sliding windows, in which the mode accounts for at least 20 seconds of that window. We define microstates as disagreements between the local 2-second window label and the smoothed label.

4.3.9 Analysis of 14 Electrodes

Next, we aggregated data from all 14 electrodes to obtain a common manifold for each electrode. We trained the model used a learning rate of 10^{-5} and batch size of $50,000$. Each model is trained for 300 epochs, where the model sees the entire training dataset per epoch. We performed the same model validation as with the single electrode model.

4.3.10 Multi-HMM

For HMM on the 14-electrode manifold, we trained a modified HMM, known as a Multi-HMM, that would require the same emissions for different electrodes but allow different initial probability and transition probability matrices for each electrode. Using the same cross-validation procedure, we find 10 states.

4.4 Results

4.4.1 A VAE for describing a manifold of brain states

We recorded LFP data across the cortex of mice during normal sleep-wake cycles. To achieve our goal of characterizing non-canonical brain states, we developed a dimensionality-reduction method based on the VAE. The VAE is more powerful and robust than other dimensionality-reduction methods, but lacks the bias and constraints of supervised classification methods. We envision brain states as traversing a latent manifold slowly over time. To discover this latent manifold, we first extracted power of different frequency bands for two second time windows of LFP and EMG recordings(Fig. 4.1B). For each time window, we combine the power from 30 frequency bands of LFP with average EMG power, resulting in a high-dimensional representation. In order to represent each data point on a two-dimensional manifold, we presented the high-dimensional representation to a β-VAE (Higgins et al., 2017) that seeks to predict the next window in time. The β-VAE learned a two-dimensional embedding, where one dimension is explicitly set as the average EMG power, which is essential in separating wake states from sleep states, and the second latent dimension is learned by the model. Thus, our manifold defines brain states by two axes – 1) EMG, which defines the "wakefulness", and 2) cortical activity.

We first used our method to characterize brain states on a single electrode in the visual cortex. When we did so, we found that the resulting manifold separates three basic states, corresponding with expert-provided canonical labels (Fig. 4.1C). In order to further validate that our model is able to separate these states, we used a Gaussian Mixture Model (GMM) clustering algorithm to find 3 clusters. We found an average of 82% accuracy of these cluster labels compared to canonical labels by human experts, echoing inter-expert accuracy (Fig. 4.1D). This was reproducible across mice (Fig. C.2).

It is important that our models produce interpretable and robust representations of the brain state space for any further analysis. Because unsupervised learning techniques generally lack validation metrics, these models can be prone to benign overfitting, or learning features that are not present in the data due to noise (Bartlett et al., 2020; Pei et al., 2021). We found that models that learn these spurious features are often degenerate–these features are not reproducible across different training instances. Thus, we use the latent representation reproducibility across different initializations and training/validation set splits to demonstrate that our model is not overfitting (Fig. C.1B). We found that variability between trained models is less than the variability between different sessions and subjects, which are subject to natural variation in the mouse's daily environment or experiences (Fig. C.1C). In addition, we found that a model trained on data from one mouse when applied to a session from another mouse achieves the same representation as if we had trained on the data from second mouse, further underscoring that our representation learns real features(Fig. C.1D). Thus, our use of these validation methods ensures we can continue to interpret results from our manifold.

One way to interpret the brain state manifold is by understanding the features that the VAE finds important in defining the manifold. We can do so by analyzing how different frequency bands tile the latent space (Fig. 4.1E). We found that our model extracts features that correspond with previous knowledge of states: Delta power is relatively lower in REM sleep than in SWS, and theta power is relatively higher in REM sleep and active wake than in SWS or in transition regions (Hernan et al., 2017). Beta and gamma powers are also relatively higher in wake than in sleep states, with beta power being especially high during sleep to wake transition regions (Hernan et al., 2017). In addition, theta rhythm is lower in wake and quiet wake than in active wake or REM sleep (Hernan et al., 2017). Because the model nonlinearly combines frequency bands to create the low-dimensional manifold, it is able to capture more information than simply comparing EMG to any single frequency band.

4.4.2 HMM reveals dynamics of brain states and microstates

Next, we sought to quantify the temporal patterns of brain states for this electrode. Although our manifold indicates a continuum of states, segmenting the manifold into substates can be useful for this quantification. In order to do so, we fit a Hidden Markov Model (HMM) to the encodings. HMMs are a unique type of clustering algorithm that also learns transitions between the defined clusters. Using the HMM, we found 8 states corresponding to three substates of SWS, Active Exploration (AE), Wake, Quiet Wake (QW), Drowsy Wake (DW) and REM (Fig. 4.2A, D, Fig. C.3). The number of states was determined by finding the elbow in the negative log likelihood by number of states, and state labels were determined corresponding by the frequency bands that are dominant in that state with known substates (Gervasoni et al., 2004; Hernan et al., 2017). By analyzing the dynamics of transitions between these 8 states, we found certain transitions do

not occur (Fig. 4.2B). Consistent with previous literature (Gervasoni et al., 2004), wake states cannot transition to REM state without passing through SWS. In addition, transitions from sleep to wake and vice versa consistently pass through an intermediary quiet or drowsy wake period. Lighter SWS states transition to deeper SWS states in succession, but deeper SWS only transition to REM or SWS1. These states are also reproducible in the second mouse (Fig. C.4a,b).

When looking at the duration lengths of each of these 8 states, we found that most states have a unimodal distribution peaking around 20 seconds(Fig. 4.2C). However, Drowsy Wake, Quiet Wake, and Wake have a bimodal distribution, with an additional peak at less than 10 seconds. These short duration states are likely microstates or short periods of one state within another state (Fig. 4.2E). In order to verify this, we identify microstate instances throughout the sleep wake cycle. Microstates are defined as a mismatch between a 2-second local state with a surrounding 30-second global state(Fig. 4.2F, (see Methods)) We found that these microstates, while rare, do occur consistently, with some types of microstates consistently more present than others(Fig. 4.2G). The majority of microstates occur within the two transition regions, reflecting the state duration graphs. Other microstate types reflect the transition graph of the HMM, suggesting that microstates may be a probing of a possible transition.

These results underscore that brain states are in fact, not persistent but instead can be transient; specifically, certain substates in the transition from sleep to wake are most transient. One caveat is that when we apply this same analysis to other electrodes in the cortex, we find there can be differences in the states revealed by the HMM (Fig. C.9), leading us to investigate further how these states manifest across the cortex.

4.4.3 Heterogeneous expression of brain states across the cortex

Next, we wished to understand how brain states may be expressed differently across the cortex. To do so, we visualized the encodings of different cortical areas on a shared manifold. Thus, we next fit our model to data from all 14 electrodes across the cortex. We also sought to describe this shared latent space via a Hidden Markov Model; however the basic HMM is insufficient in this case. Instead, we wanted to leverage a HMM that constrains each electrode to have the same emissions but allows for different initial probability and transition matrices, which we term a MultiHMM (see methods). Again finding the elbow, we found the optimal number of states to be 9 for both animals with 14 electrodes. We found the states to be consistent between animals, further underscoring that any differences between animals when training on one electrode are due to small differences in electrode placement. These states are the same as in the single electrode case – SWS1, SWS2, SWS3, REM, DW, QW, Wake, AE – with an extra SWS state, SWS4 (Fig. 4.3A).

When we compare the points on the manifold for each of the 14 electrodes, we found stark differences (Fig. 4.3A). Specifically, the lateral somatosensory cortex completely lacks the REM cloud. Electrodes in the frontal medial area show more AE than Wake. These differences are highlighted in the percent of time spent in each state per electrode(Fig. 4.3B), and are reproducible across mice (Fig. C.5). In order to verify that the lack of REM in somatory sensory cortex was real, we identified time points of "global REM", where 11 or more electrodes were in REM as determined by expert labels. We then calculated the average spectrogram for medial and lateral somatosensory cortex for each of these time points. We saw clearly that the average spectrogram for lateral somatesensory cortex reflected the "global SWS" spectrogram (Fig. 4.3C).

One previously existing notion of local sleep is the presence of local slow-waves in awake miceVyazovskiy et al., 2011, specifically sleep-deprived mice. In our results, the presence co-existence of DW in QW or Wake in differing areas reflected these slow-waves in awake mice, with DW exhibiting more slow waves (Fig. C.6A). We additionally found that frontal areas spend more time in DW than other areas (Fig. 4.3B, Fig. C.6B). Lastly, we found that these time periods of slow waves in wake last no more than 30 seconds, which has not been described previously (Fig. C.6C).

4.4.4 Spatiotemporal dynamics of sleep and wake

Our results from the previous section indicate that heterogeneity exists in sleep and wake across the cortex. Some of that heterogeneity is accounted for by differences in state expression, such as lack of slow waves during REM in lateral somatosensory cortex or increase in theta power (AE) in frontal area during Wake. However, there may also be other instances of non-global states across the brain, instances when different parts of the brain are in different states at the same time. In fact, when we labeled each time point with 14 labels - one per electrode from the MultiHMM - we found that only around 60% of time points are global (all 14 electrodes are in the same state) (Fig. 4.4A, Fig. C.7A). In fact, at some points in time, only 4 electrodes are in the same state. We went on to define "global states" as 11 or more electrodes being in the same state. We found that non-global states are largely less than 30 seconds in duration (Fig. 4.4B, Fig. C.7B). We can further break this down by looking at the globality per state. Not surprisingly, we found that REM, AE, Wake, and Transition states are the most non-global states, while SWS states are largely global (Fig. 4.4C, Fig. C.7C). We discussed previously that REM was non-global because of the differential expression of REM in the lateral area of the cortex, with some electrodes completely missing REM expression.

In addition, Wake and AE are non-global separately, but together with QW and DW form a relatively global state (Fig 4.4D). However, the remaining non-global dynamics have yet to be characterized thus far.

We hypothesized that two contributors to non-global dynamics are transition states and microstates. It is known that individual slow waves propagate from one side of the cortex to the order (Tononi, 2009). However, it is unknown whether, on a larger time scale, brain states start to transition in one area before another. In order to illustrate this, we found all instances of transitions from any one state to another state, and visualized the average transitions by coloring each point by the average position on the manifold. Because it would be intractable to visualize all transitions between 9 states, we returned to three basic states. We founnd that specific spatiotemporal patterns arise for each of these transition types. Transition from SWS to REM sleep arises first from the medial posterior cortex. ransitions to SWS appear more global. Interestingly, transitions from REM to SWS passes through Wake for a few seconds. This has been described in previous literature (reference: need from Jeremy). In addition, we find that microstates occur spatially non-globally. Microstates of REM in SWS or SWS in REM only co-occur between any two electrodes a maximum of 30% of the time (Fig. C.8). Altogether, these results provide a complex and rich picture of varying spatiotemporal heterogeneity in sleep states across the cortex, which provides an introduction to future studies further describing the complexity of brain states.

4.5 Main Figures

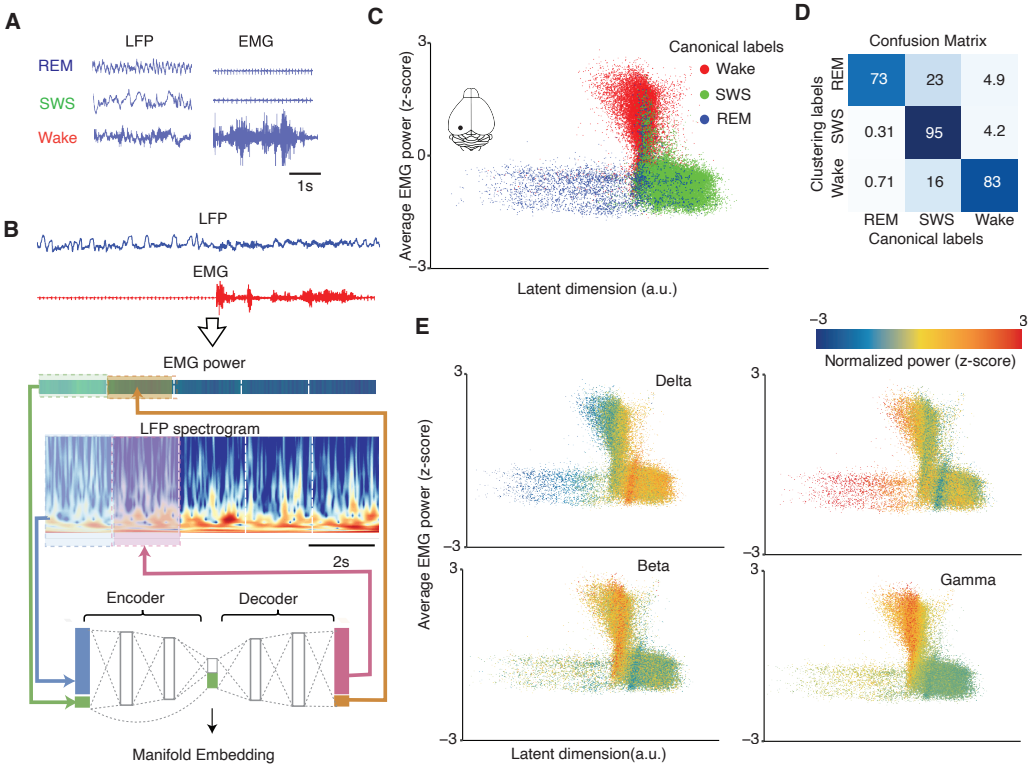


FIGURE 4.1: **VAE Framework for brain state manifold** **a**, Example LFP and EMG traces recorded during REM, SWS and Wake. **b**, Model architecture. For each 2-second window, we compute EMG power and wavelet spectrogram of LFPs, which become inputs to the variational auto-encoder (VAE). The VAE is trained to predict the next point in time. **c**, Two-dimensional latent manifold of the EMG and LFP activity from a single channel in the visual cortex for 72 hours of recording reveals clusters which largely agree with human- expert labeling of wake, SWS, and REM states (color code). **d**, Confusion matrix between labels generated by human of GMM clusters of latent manifold. GMM clustering achieves performance similar to inter-expert agreement. **e**, Latent manifold from (c) colored by average power from frequency bands (top to bottom) delta, theta, beta, and gamma. Each point representing a two second window in time is colored by normalized spectral amplitude ratio for each frequency band divided by total spectral power.

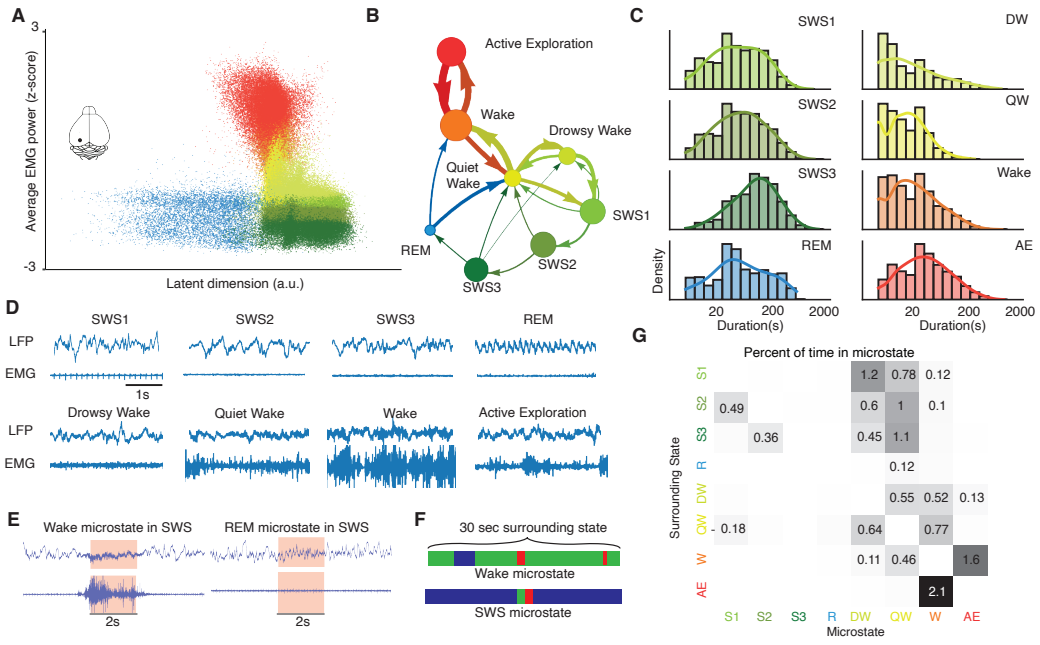


FIGURE 4.2: **HMM reveals dynamics of brain states and microstates a**, Encodings colored by each of 8 states fit by HMM. **b**, A diagram of most likely transitions. Circle size correlates with frequency of that state, and arrow width corresponds to likelihood of transition between states. Going from any wake state to REM is highly unlikely without first passing through SWS. SWS goes from lighter to deeper before transitioning to REM. **c**, State lengths of each of the 8 states fit by the HMM. **d**, example raw LFP and EMG traces from each of the 8 discovered states. **e**, example raw LFP and EMG traces frmo microstates. **f**, Microstates are identified by mismatches between local state labels for 2 second windows and the surrounding 30-sec state label (obtained from the mode, see Methods). **g**, Percent of time spent in microstate out of total time in surrounding state. Color of numbers corresponds to labels in panel b.

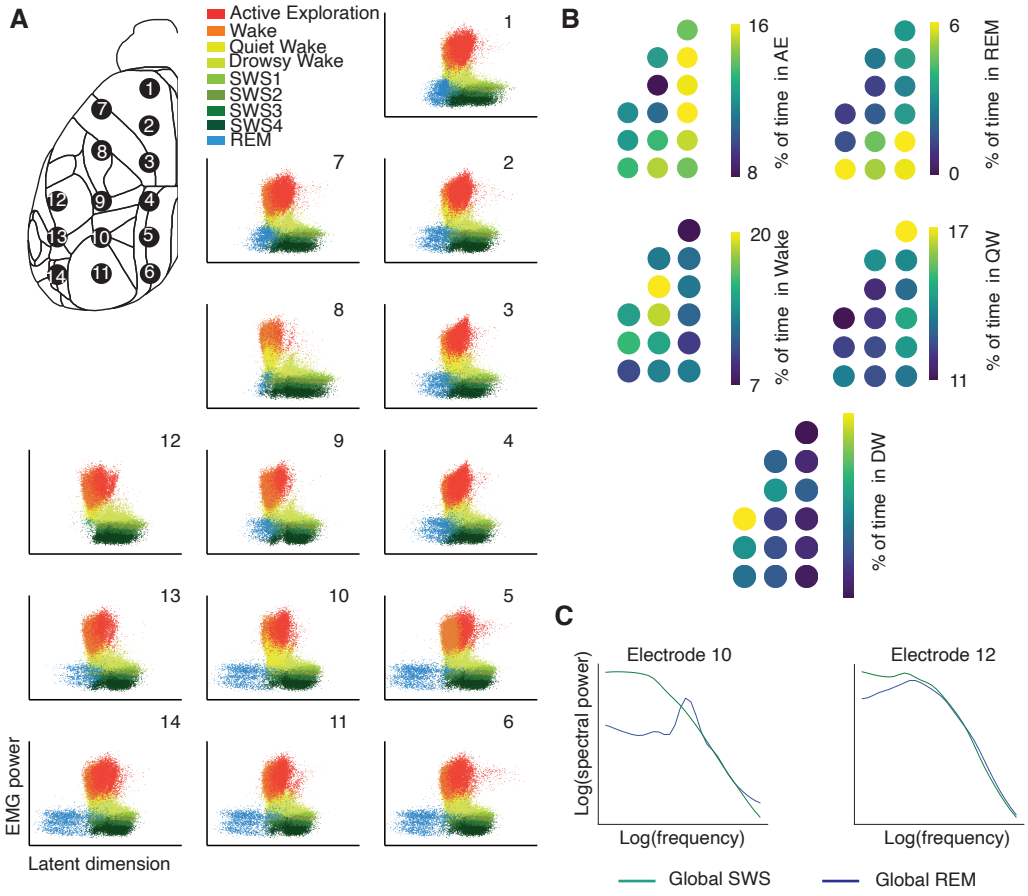


FIGURE 4.3: **Heteregeneous expression of REM sleep across the cortex a**, LFP is recorded from 14 probes simultaneously across the cortex. A model is fit to data compiled from all 14 probes. Encodings are shown for each separate probe, colored by one of 9 HMM states fit. REM sleep (blue) is missing or partial for some electrodes. **b**, Percent of time spent in Active Exploration(AE) (top left) and REM (top right), Wake (bottom left), and QW (bottom right) for each of the 14 electrodes. **c**, Periods of global REM and global SWS are defined as time points where greater than 10 electrodes are in that state. For somatosensory 12 (left) and somatosensory 10 (right) electrodes, the average spectrogram for global REM (blue) and global SWS (green) are shown.

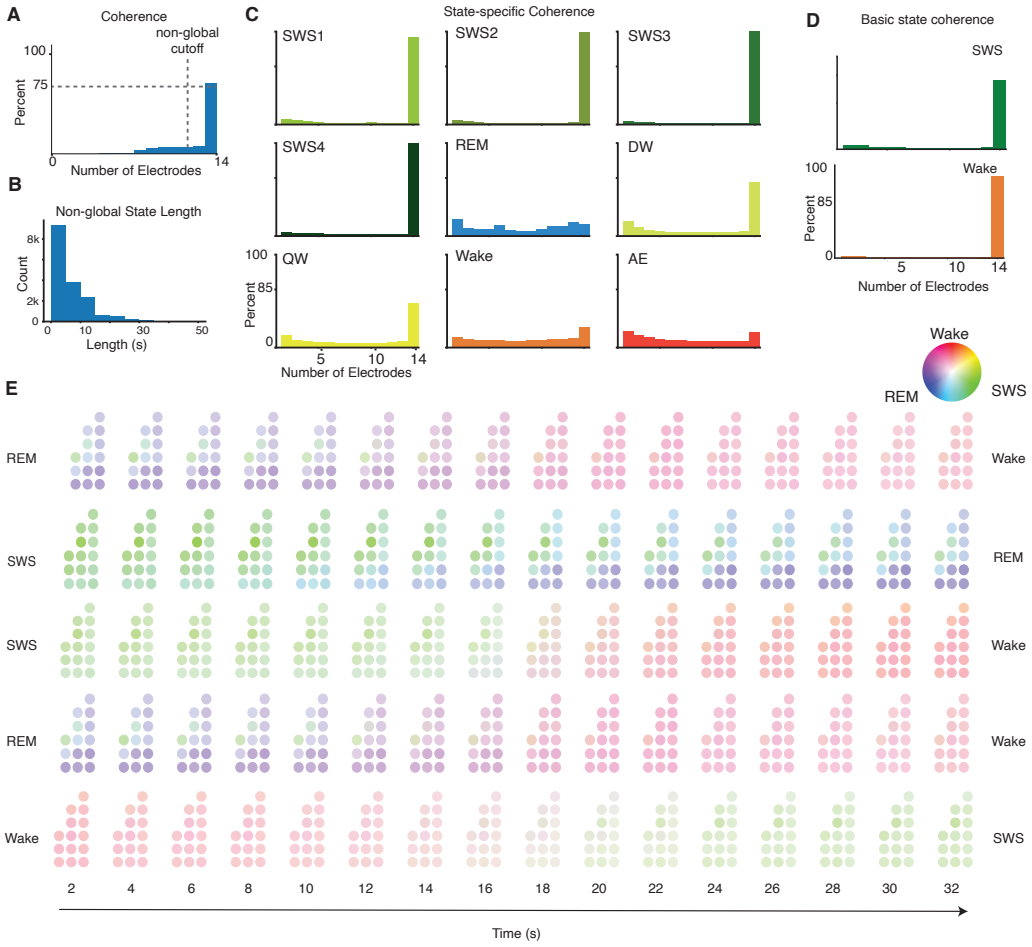


FIGURE 4.4: **Spatiotemporal dynamics of brain states. a**, Coherence histogram of states across 14 electrodes. Coherence is defined as the maximum number of electrodes in the same state (found by the HMM) at any point in time. While most time points are global (14 electrodes in the same state), there are a significant amount of time points with 13 or less electrodes in coherence. **b**, Histogram of the lengths of non-global states (states where less than 13 electrodes are in agreement.) Most are less than 30 seconds in length, likely corresponding to micro-states and transition states. **c**, Histogram of number of electrodes (1-14) in each particular state at any point in time divided by the total amount of time any electrode is in that state. REM and Wake states show the least amount of time in a global state. **d**, same as (c) but for basic states of SWS and Wake. While substates of wake are non-global, wake as a whole is global. **e**, Average dynamics for transitions from REM to Wake and SWS to REM show spatiotemporal heterogeneity. Colors denote the position on the manifold of that electrode. Each electrode graph represents 2 seconds in time.

Chapter 5

Discussion and perspectives

The work in this thesis was guided by the goal of finding an unbiased characterization of brain states. In Chapter 1, we provided a brief overview on brain states and their importance before looking at some deviations to canonical brain states. In Chapter 2, we overviewed some unsupervised methods that are useful for analyzing biological data including dimensionality reduction and clustering, and we discuss our decision to use the variational autoencoder (VAE) to describe our brain states. Motivated to ensure the results from the VAE were suitable for further scientific interpretation, we explored the relationship between benign overfitting and representation learning in VAEs and showed that using consistency over time prevents overfitting to spurious features (Chapter 3). Lastly, we applied our modified to brain state data and characterized a manifold of brain states. We comprehensive described a non-canonical view of brain states, including transitions, microstates and heterogeneity in brain states across cortical regions.

5.1 VAEs and representation learning in neuroscience

VAEs and other dimensionality reduction methods have been used to analyze high-dimensional biological data. Specifically in neuroscience it is thought

that some underlying latent representation generates the high dimensional output we see as neuron activity. We can recover these underlying latents with VAEs (Sussillo et al., 2016; Cunningham and Yu, 2014. However, many of the studies using this approach fail to discuss the model selection process for choosing an appropriate VAE. Without careful model selection, it is difficult to be confident that in any further scientific interpretation from the inferred latents, as VAEs are prone to overfitting to spurious noise despite low validation loss (Bartlett et al., 2020; Belkin et al., 2018). In this work, we showed that we can use consistency of latent representations over time as both an inductive bias on model architecture and as a model selection metric. The use of time in variational autoencoder architectures is not a novel concept (Schneider, Lee, and Mathis, 2022; Liu et al., 2021; Sedler, Versteeg, and Pandarinath, 2023), but it has not been explored in depth as a method to avoid benign overfitting. In addition, current model selection metrics for VAEs either rely on an external variable as a label (Higgins et al., 2017; Pei et al., 2021) or measure similarity between trained models (Duan et al., 2020) Thus, our contributions in this thesis are to 1) show that vanilla VAEs are prone to benign overfitting 2) show how using time in VAE architecture is an effective inductive bias and regularizer and 3) introduce a novel model selection metric based also in smoothness over time of latent embeddings.

One way to recover meaningful and interpretable representations of data by VAEs is through identifiable VAEs. It is important to note that true identifiability cannot be achieved without an external label (Khemakhem et al., 2020b), and we do not try to mathematically prove our methods achieve identifiability. However, given the idea that identifiability is in opposition to degeneracy (Duan et al., 2020; Genkin and Engel, 2020) (meaningful representations should be similar and representations that overfit to noise can be different), we show that our model selection

metric (NL) is a suitable metric when such an external label is not sufficient. Another caveat to note is that our method relies on the assumption that the data evolves on some low-dimensional manifold slowly over time. Thus, our method would be not be suitable where this assumption is does not hold.

5.2 Brain state manifold characterization

We effectively use a modified VAE to describe a manifold of brain states that corresponds with previously described states of SWS, REM, and Wake. Through analyzing this manifold, we not only confirm known phenomena about state transitions and the existence of micro-sleep and micro-wake periods, we also further characterize stereotyped transitions between substates and the frequency of different types of microstates. When we apply our model to data recorded simultaneously from 14 electrodes across the cortex, we find hetereogeneous expression of several states, especially the lack of REM-like activity in lateral somatosensory and frontal areas and increase in AE in frontal areas.. We also find that the non-globality of states goes beyond this lack of REM-like activity in certain regions, so we characterize the spatiotemporal patterns and find stereotyped non-global transition patterns as well as microstates that are local both in time and space.

Most classification methods, whether by experts or autonomous algorithms, treat brain states of wake and sleep as discrete categories. In contrast, we model brain states as continually traversing on a manifold, which allows us to capture the complexity and heterogeneity of neural activity patterns across sleep and waking, beyond the simple model of three discrete states. The first proposition of brain states appearing on a manifold was by Gervasoni et al., 2004; however the field has mainly continued to rely on the discrete state definition. Previous characterizations of the manifold relied on simple linear dimensionality reduction approaches

(Gervasoni et al., 2004); instead we use a novel approach to determine a robust nonlinear manifold that can be conserved across brain regions. This continuous manifold also allows us to analyze differences between brain regions not just in the amount of time spent in various states but in the expression of those states. We do, however, also recognize that segmenting the manifold into a collection of states provides a powerful framework for more precise quantification.

The existence of microstates, including micro-arousals and micro-sleeps has been documented previously (Watson et al., 2016; Soltani et al., 2019); but their exact definition and purpose remain ambiguous. While some define a conscious state change to be at least 500 ms (Libet et al., 1967; Edelman, 2003), it is possible that even longer states (up to 6 seconds) would be undetected by an individual and certainly shorter than canonical notions of states. In this study, we provide a novel systemic quantification of the frequency of different micro-states. We find, surprisingly, the most common types of microstates reflect the transition graph between states; perhaps micro-states represent a probing of a transition. Other theories indicate that the frequency of microstates may be related to experience or stress (Maness et al., 2022; Smith et al., 2022).

Differing types of "local sleep" have been introduced in the past (Rector et al., 2009; Krueger et al., 2019) - such as local slow waves in awake mice (Vyazovskiy et al., 2011) and the presence of slow waves in REM sleep (Soltani et al., 2019; Nazari et al., 2022). By applying our model to large scale recordings across the cortex, we provide results that unify and extend previous notions of local sleep. First, we find the absence of REM in the lateral area of the cortex. Our results are in accordance with (Nazari et al., 2022) which also found non-uniform presence of slow-waves during REM sleep in the frontal lateral somatosensory areas. There are several possible explanations for differential expression of theta rhythm during

REM sleep. One possibility is that the lateral area is farthest from the hippocampus, where theta rhythm is generated during sleep, and theta rhythm measured in the cortex is due to volume conductance rather than actual theta rhythm present in neurons in the cortex. Instead, Nazari et al. (Nazari et al., 2022) found that these areas were less projected to by cholinergic neurons from the basal forebrain, allowing for one possible explanation of the persistence of slow-waves in these areas during global REM. In the present study, we further characterize these differences in the presence of slow-waves during REM sleep as heterogeneity on a shared manifold across cortical areas. Second, we find heterogeneity in expression of differing types of wake throughout the cortex. The medial regions seem to express more theta power in wake. We also find the presence of slow waves (Quiet Wake) to be in accordance with previous studies of slow waves in wake (Vyazovskiy et al., 2011). In contrast to previous studies (Vyazovskiy et al., 2011) that studied these local slow waves in sleep-deprived mice, we show these waves are present in normal sleep-wake cycles and occur more in the frontal regions. It is unclear what the function of heterogeneity of brain states may be. Previous studies have shown that brain regions that are more used during the wake periods may require greater sleep intensity or duration (Krueger et al., 2019; Rector et al., 2009). Furthermore, we find that microstates occur heterogeneously, which has been heretofore undescribed. It would be interesting in future studies to further analyze the patterns and frequencies of these spatially heterogeneous microstates.

5.3 Looking forward

We have provide a flexible and unbiased method for characterizing brain states. This method, which exists as a publicly available github repository (https://github.com/engellab/braivest), can be applied to characterize other types

of brain state heterogeneity. One such type of heterogeneity might be over disease states. For example, disruption of sleep is an indicator of cancer progression. Specifically, tumor progression in cancer patients triggers changes in sleep patterns through the disruption of neuromodulators governing sleep and wake transitions (Borniger et al., 2018). However, exactly how the sleep expression in the brain changes throughout cancer progression is yet to be characterized. Additionally, another type of abnormal brain state is the seizure state. We show in preliminary results that our model is effective in separating out seizure instances from normal sleep and wake time periods (Fig. D.1). In the future it may also be possible to use this method to distinguish different types of seizures (Lüttjohann, Fabene, and Luijtelaar, 2009). Lastly, this tool can be powerful for understanding more fine-grained wake states, such as differences in arousal and attention states. We show that using our model on voltage imaging traces from the parietal cortex of mice running recovers a latent dimension that correlates with change in pupil size (Fig. D.2). We can apply the same method to characterize differences in arousal or attention that may be correlated with performance on some task.

Another avenue for future study is the exploration of different model architectures to explore brain states. For instance, we can explore a VAE that takes in data from all 14 electrodes simultaneously to obtain a manifold that represents the entire brain's state. We find even more of a continuous manifold rather than discrete states with this method, and find areas on the manifold corresponding to transitions between states (Fig. D.3). We could also explore other architectures that would take in a sequence of timepoints, such as CNNs with attention or RNNs. These architectures may be effective in defining substates that are defined not only on the activity in that moment but also on surrounding activity (i.e. microstates).

Appendix A

Supplemental Tables

TABLE A.1: Normal hippocampal, neocortical, and thalamocortical oscillations

Oscillation	Behavioral State	Known Origin
Infraslow (seconds to minutes)	All states, including natural states and anesthesia	Unknown, possibly metabolic
Slow (0.1–1 Hz) and delta (1–4 Hz)	SWS, several forms of anesthesia (e.g., ketamine-xylazine, urethane)	Intracortical, but thalamus actively contributes to synchronization
Theta rhythm (4–12 Hz)	Exploration, REM sleep	Septohippocampal network (unclear whether theta activity recorded in neocortex also originates in septohippocampal network); theta rhythm in rodents is more pronounced than in other species

TABLE A.1: Normal hippocampal, neocortical, and thalamocortical oscillations

Oscillation	Behavioral State	Known Origin
Alpha rhythm (8–10 Hz)	In primates and carnivores dominates occipital lobe during wakefulness with eyes closed	Unknown, but in vitro data suggest that thalamus has sufficient machinery to generate continuous alpha rhythm
Spindles (10–16 Hz), transient waxing and waning oscillations usually lasting 0.5–2 s; currently spindles are subdivided into slow (10–13 Hz) and fast (13–16 Hz)	Predominantly recorded during stage 2 sleep, can be recorded during deep sleep	Spindles have been described as being generated in thalamus; however, several properties of slow spindles raise questions about involvement of other brain regions
Sigma rhythm (10–15 Hz)	All states of vigilance, but stronger during SWS	Unknown; sometimes the terms sleep spindles and sigma rhythm are used to describe the same phenomenon
Mu rhythm (7–14 Hz) and associated beta rhythm (20 Hz)	Quiet wakefulness, over somatosensory cortex; stops when the movement is present	Unknown; some suggest that this rhythm is generated by somatosensory thalamic nuclei

TABLE A.1: Normal hippocampal, neocortical, and thalamocortical oscillations

Oscillation	**Behavioral State**	**Known Origin**
Beta rhythm (16–29 Hz)	Waking and drowsiness; may also be accentuated by GABAergic medications, e.g., benzodiazepines	Likely cortical
Gamma rhythm; currently subdivided into low gamma (30–80 Hz) and high gamma (90–120 Hz)	Waking or modulated by sleep slow waves	Likely local cortical networks, but can be synchronized with thalamic activities
Neocortical ripples (140–200 Hz)	All states of vigilance, but different power: wake < REM < SWS < anesthesia < seizure onset zone	Intracortical (for neocortical ripples), depend on gap junctions

Appendix B

Derivations

B.1 VAE ELBO Derivation

This section is adapted from Kingma & Welling's original paper and accompanying introduction text.

We seek to maximize the marginal likelihood given as $\log p_\theta(x)$.

$$\begin{aligned}
\log p_\theta(x) &= E_{q_\psi(z|x)}[\log p_\theta(x)] \\
&= E_{q_\psi(z|x)}\left[\log \frac{p_\theta(x,z)}{p_\theta(z|x)}]\right] \\
&= E_{q_\psi(z|x)}\left[\log \frac{p_\theta(x,z)q_\psi(z|x)}{q_\psi(z|x)p_\theta(z|x)}\right] \\
&= E_{q_\psi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\psi(z|x)}\right] + E_{q_\psi(z|x)}\left[\log \frac{q_\psi(z|x)}{p_\theta(z|x)}\right]
\end{aligned}$$

The first term is the ELBO ($L(\theta, \psi; x)$) and the second term is $D_{KL}(q_\psi(z|x)||p_\theta(z|x))$, which is the KL divergence of the approximate from the true posterior.
Since $D_{KL}(q_\psi(z|x)||p_\theta(z|x)) >= 0$,

$$\log p_\theta(x) >= L(\theta, \psi; x) = E_{q_\psi(z|x)}[-\log q_\psi(z|x) + \log p_\theta(x, z_=)] \qquad \text{(B.1)}$$

This can be rewritten as:

$$L(\theta, \psi; x^{(i)}) = E_{q_\psi(z|x^{(i)})}[\log p_\theta(x^{(i)}|z)] - D_{KL}(q_\psi(z|x^{(i)})||p_\theta(z)) \tag{B.2}$$

B.2 The likelihood calculation for Neighbor Loss

Assume datapoints traverse a 2-D space via random Gaussian walk, i.e. x_{t+1} is drawn from a circular Gaussian distribution centered at x_t with standard deviation σ. The log likelihood of this function is:

$$\begin{aligned}
\log(L) &= \log(\prod_{t=1}^{n} \frac{1}{2\pi\sigma^2} exp(-(\frac{(x_{t+1}-x_t)^2 + (y_{t+1}-y_t)^2}{2\sigma^2}))) \\
&= \sum_{t=1}^{n} \log(\frac{1}{2\pi\sigma^2} exp(-(\frac{(x_{t+1}-x_t)^2 + (y_{t+1}-y_t)^2}{2\sigma^2}))) \\
&= -\frac{n}{2}\log 2\pi - n\log\sigma - \frac{1}{2\sigma}\sum[(x_{t+1}-x_t)^2 + (y_{t+1}-y_t)^2]
\end{aligned}$$

Thus maximizing the log likelihood is equivalent to minimizing sum of the absolute distance between each point and its neighboring point in time.

Appendix C

Chapter 4 Supplemental Figures

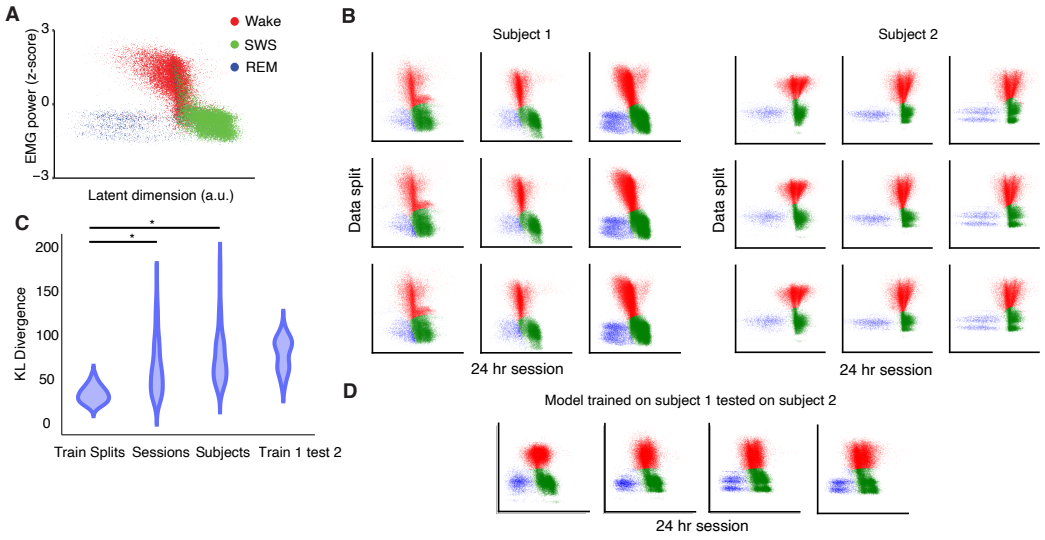


FIGURE C.1: **Robustness across training splits as validation metric** **a**, Two-dimensional latent manifold of the EMG and LFP activity from a single channel in the visual cortex for 72 hours of recording reveals clusters which largely agree with human- expert labeling of wake, SWS, and REM states (color code). (Same as Fig1C) **b**, Models trained on different training/validation splits of the dataset show robust latent representations. For mouse 1 (left) and mouse 2 (right), each row represents a model trained on a different data split and each column represents a different 24 hr session. Points are colored by GMM clustering labels. **c**, Violin plot showing KL divergence of encodings between data splits, sessions, subjects, and training on one subject and appylying to another. KL divergence between encodings obtained from models trained on different data splits have significantly less than between subjects and sessions. Subject and session variability is due to natural differences in electrode placement or mouse daily activity. **d**, Encodings from a model trained on data from mouse 1 and applied to mouse 2 still separate three clusters. Each encoding plot represents a different 24 hr session.

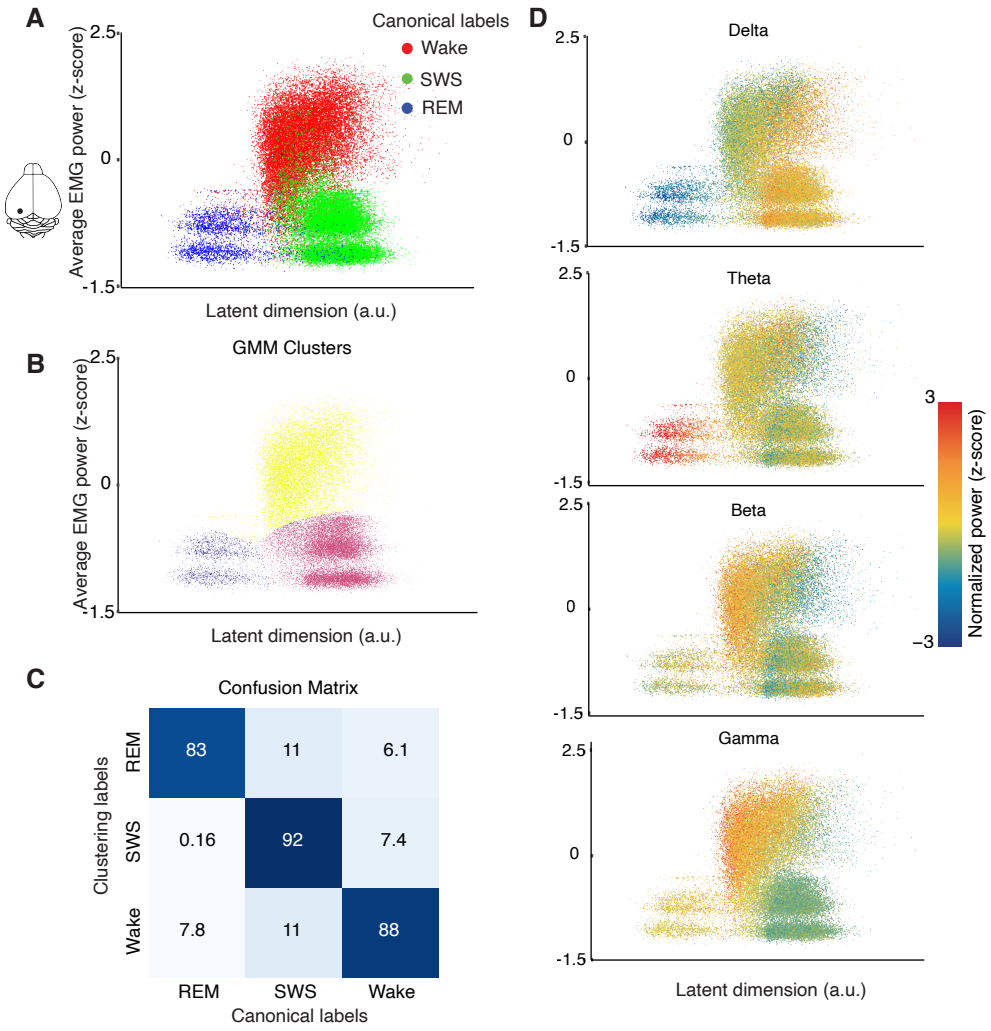


FIGURE C.2: **Encoding manifold for second mouse a**, Two-dimensional latent manifold of the EMG and LFP activity from a single channel in the visual cortex for 72 hours of recording reveals clusters which largely agree with human- expert labeling of wake, SWS, and REM states (color code). **b**, Same as (a) but colored by GMM clustering labels. **c**, Confusion matrix between labels generated by human of GMM clusters of latent manifold. GMM clustering achieves performance similar to inter-expert agreement. **d**, Latent manifold from (a) colored by average power from frequency bands (top to bottom) delta, theta, beta, and gamma. Each point representing a two second window in time is colored by normalized spectral amplitude ratio for each frequency band divided by total spectral power.

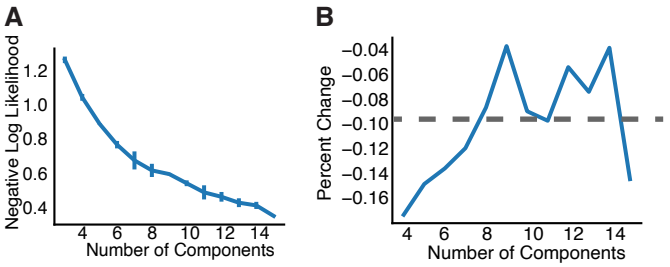


FIGURE C.3: **HMM state determination a**, Negative Log Likelihood on test examples from fitting a HMM to variable number of states on data from subject 1. **b**, Percent change of negative log likelihood from (a). A cutoff of 5% change was used to determine the number of states.

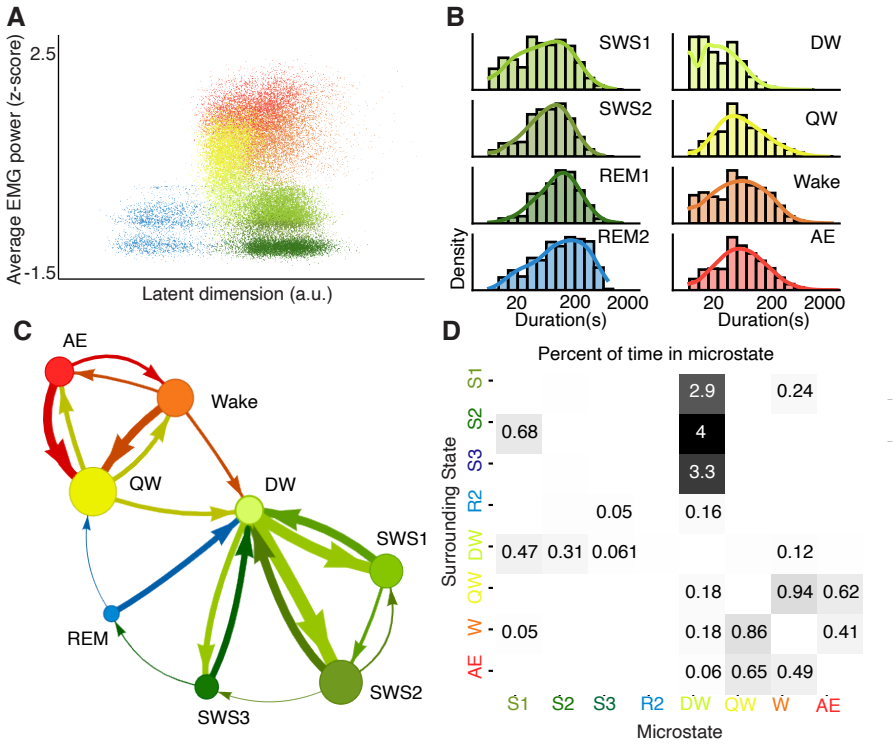


FIGURE C.4: **HMM and microstates on subject 2 a**, Encodings colored by each of 8 states fit by HMM. **b**, A diagram of most likely transitions. Circle size correlates with frequency of that state, and arrow width corresponds to likelihood of transition between states. **c**, State lengths of each of the 8 states fit by the HMM. **d**, Percent of time spent in microstate out of total time in surrounding state. Color of numbers corresponds to labels in panel b.

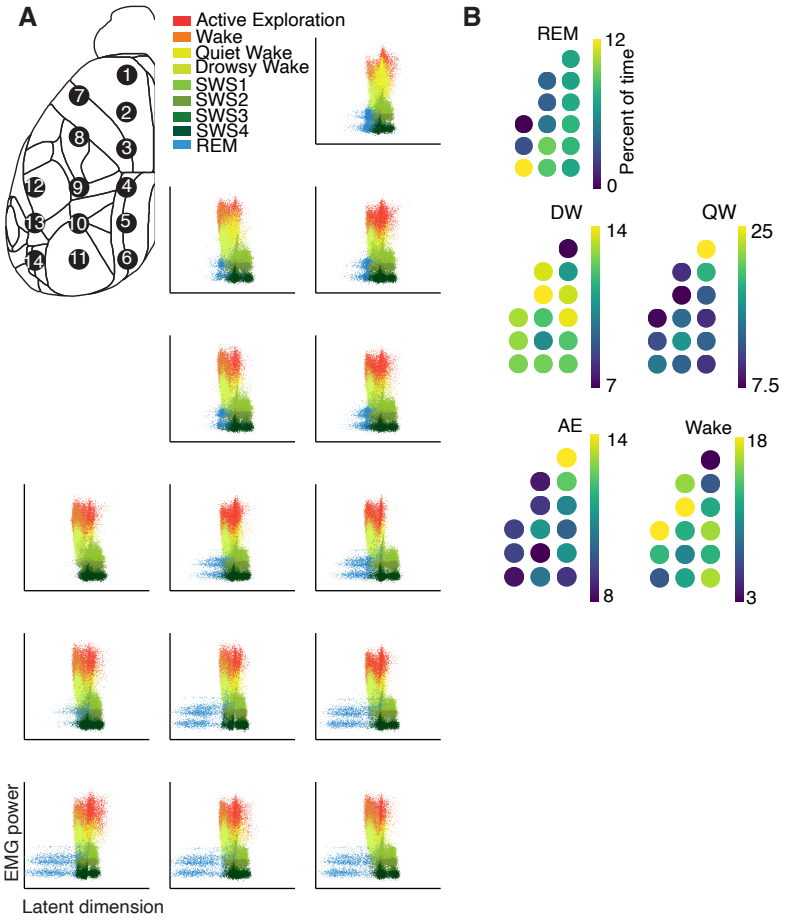


FIGURE C.5: **Heterogeneity of cortical areas for subject 2 a**, Encodings are shown for each separate probe, colored by one of 9 HMM states fit. REM sleep (blue) is missing or partial for some electrodes. **b**, Percent of time spent in REM, QW, DW, AE and Wake for each of the 14 electrodes.

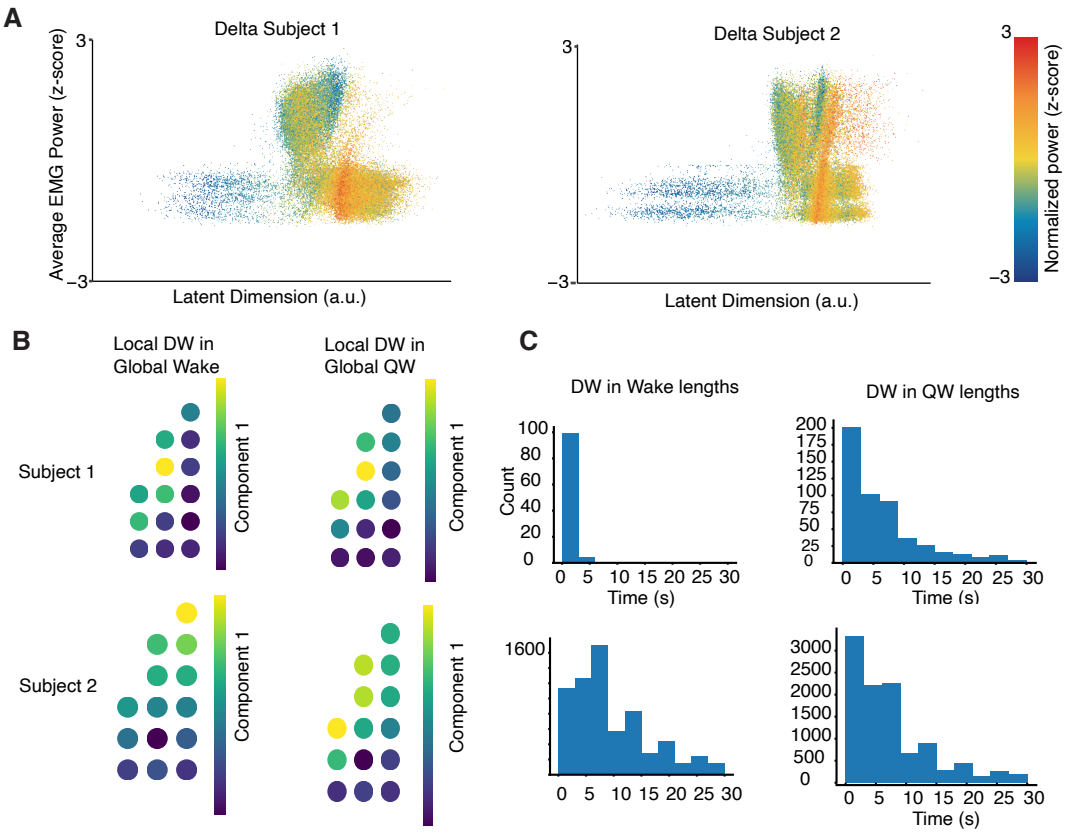


FIGURE C.6: **Local slow waves in wake** **a**, Encodings are shown from the model trained on data from all 14 probes, where each point, representing a 2-second window of time, is colored by normalized delta power. **b**, All instances of local drowsy wake (DW) in Wake (left) and Quiet Wake (right) are compiled and PCA is performed on the combined encodings from 14 electrodes. The contribution of each electrode to the top component can be visualized to understand how much that electrodes contributes the variation in local sleep. We see consistent results across subject 1 (top) and subject 2 (bottom). **c**, We calculated the lengths of each instance of DW in Wake (left) or QW (right). A histogram of all lengths is shown for subject 1 (top) and subject 2 (bottom).

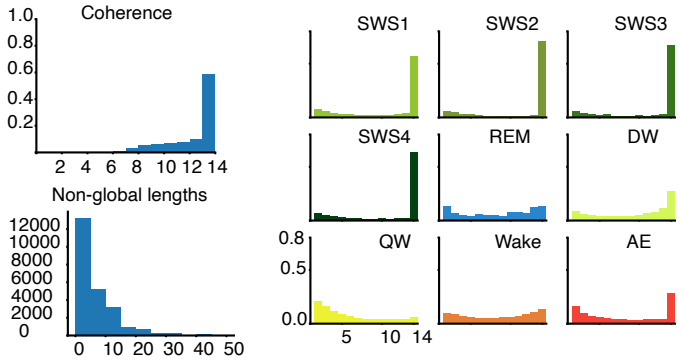


FIGURE C.7: **Spatiotemporal dynamics for subject 2 a**, Coherence histogram of states across 14 electrodes. Coherence is defined as the maximum number of electrodes in the same state (found by the HMM) at any point in time. **b**, Histogram of the lengths of non-global states (states where less than 13 electrodes are in agreement.) **c**, Histogram of number of electrodes (1-14) in each particular state at any point in time divided by the total amount of time any electrode is in that state.

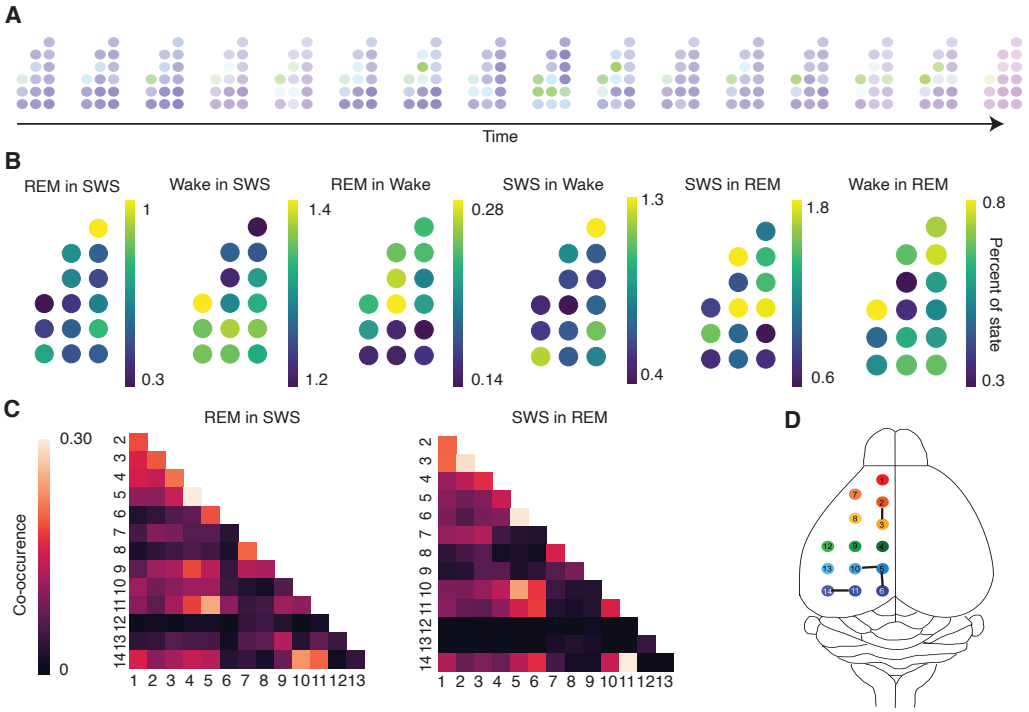


FIGURE C.8: **Spatially heterogenous microstates a**, Example microstate of SWS in REM, each point represents 2 seconds of time **b**, Percent of time spent in each of the labelled microstates for each probe. **c**, Heatmap of co-microstate occurence for REM in SWS and SWS in REM.

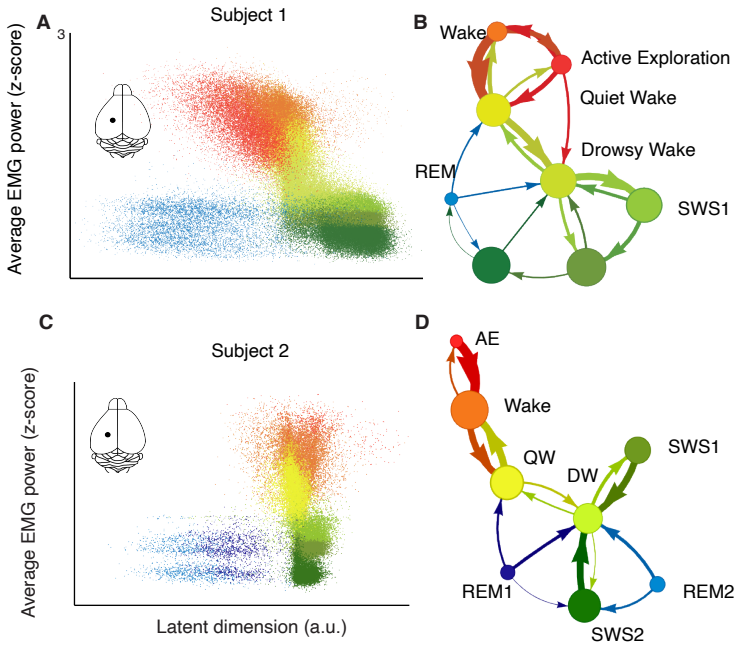


FIGURE C.9: **Differences in HMM states across electrodes a**, HMM states discovered from encodings obtained from fitting a model to data from medial somatosensory probe from subject 1. **b**, Transition diagram for HMM for medial somatosensory probe, subject 1. **c**, HMM states discovered from encodings obtained from fitting a model to data from medial somatosensory probe from subject 2. 2 SWS states and 2 REM states are discovered. **d**, Transition diagram for HMM for medial somatosensory probe, subject 2.

Appendix D

Additional Supplemental Figures

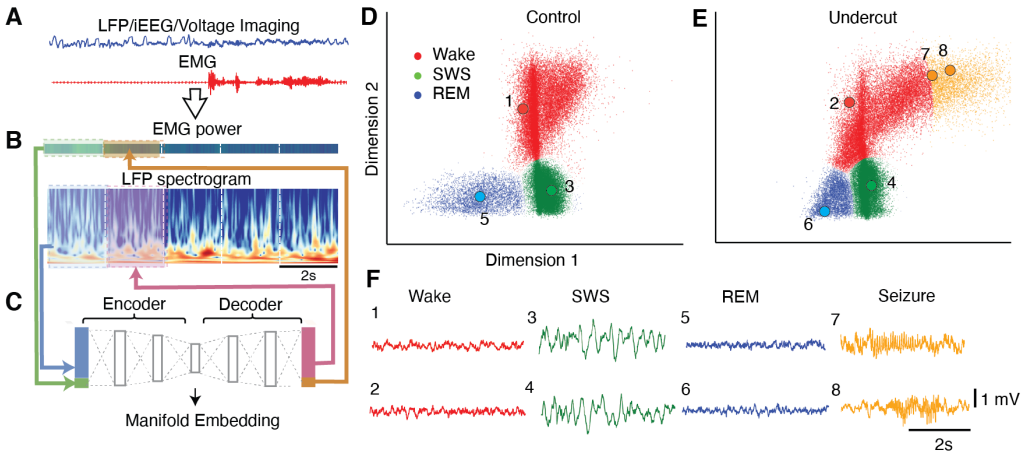


FIGURE D.1: **VAE separates seizure from normal brain state in latent representation a-c**, Model architecture. For each 2-second window, we compute EMG power and wavelet spectrogram of LFPs, which become inputs to the variational auto-encoder (VAE). The VAE is trained to predict the next point in time. **d-e**, Latent encodings for 24 hours of a normal (**d**) and undercut (**e**) mouse. The undercut mouse shows many time periods with high EMG power that correspond to epileptic events. **f**, Raw LFP traces from normal and undercut mice for Wake (1-2), SWS (3-4), and REM (5-6) data. Undercut mice show normal activity in these three states but have additional periods of seizures (7-8).

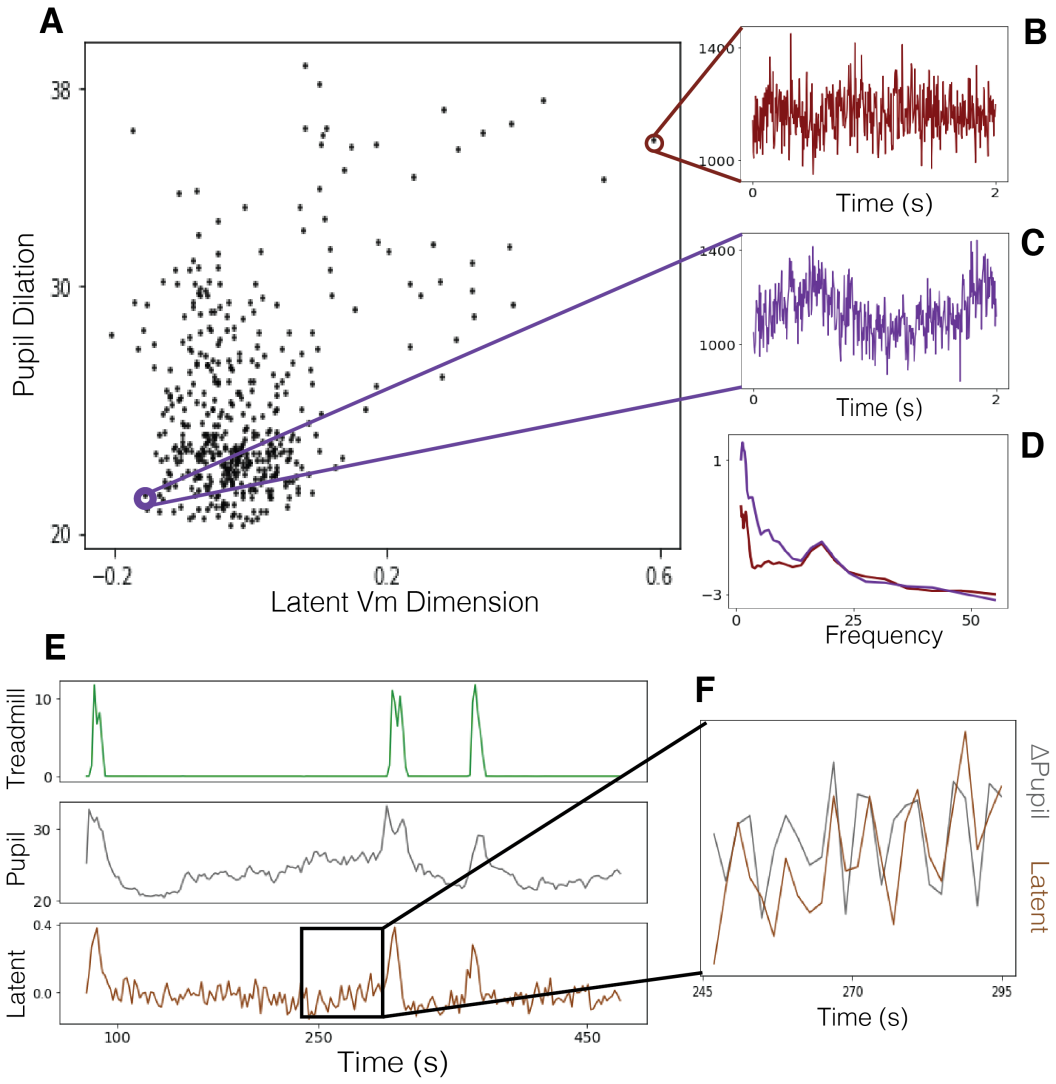


FIGURE D.2: **VAE uncovers latent dimension of arousal a**, We learn latent dimensions from spectrograms corresponding to two second windows of Vm. A singular latent dimension learned from the VAE correlates with pupil dilation ($r^2 = 0.4, p < 10e-10$). An example two second window with low pupil dilation shows low frequency fluctuations **(b)**, while an example two second window with high pupil dilation does not **(c)**. **d**, The spectrograms corresponding to **b** and **c**. **e**, The latent dimension correlates broadly with running and pupil dilation. **f**, Outside of running, an example suggests a relationship between the latent dimension and the derivative of pupil dilation, which is consistent with previous patch recordings.

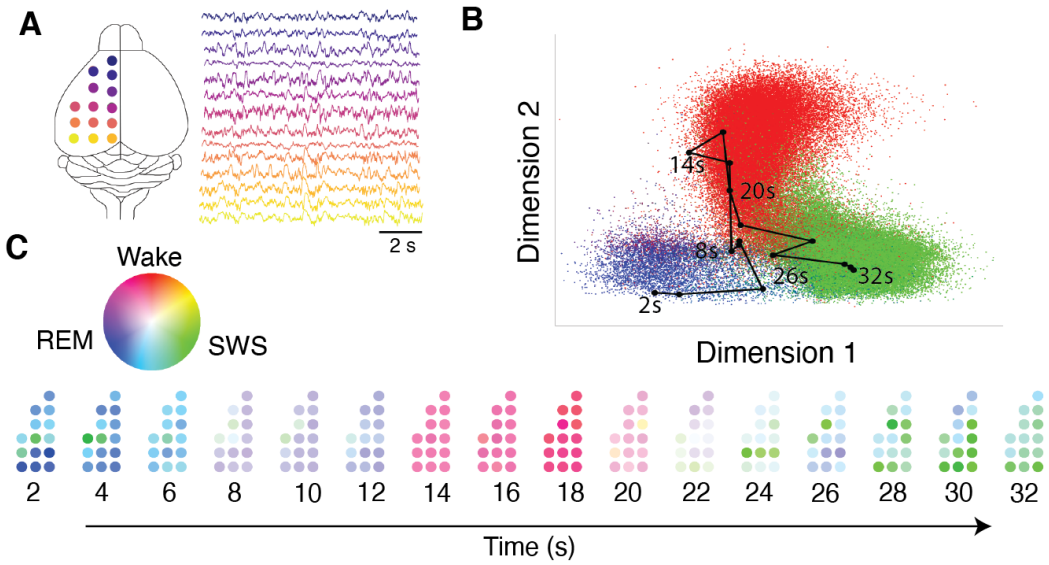


FIGURE D.3: **Manifold of spatial brain states a**, Data is recorded in 14 electrodes across the cortex. Each data point fed in the VAE corresponds to data from all 14 electrodes together for the 2 second window in time. **b**, Latent encodings show that 3 clusters are separated but with more smoothness corresponding to periods where there is heterogeneity in brain states across the cortex. Each point is colored from expert labels, with color corresponding to percent of electrodes in Wake (Red), SWS (green) and REM (blue). A trajectory of transition from SWS to REM is shown. **c**, The corresponding trajectory from **b** is shown as individual electrodes in the cortex, where each electrode is colored based on which state that electrode is in during that time window.

Bibliography

Allocca, Giancarlo et al. (2019). “Validation of ‘Somnivore’, a Machine Learning Algorithm for Automated Scoring and Analysis of Polysomnography Data”. English. In: *Frontiers in Neuroscience* 13. Publisher: Frontiers. (Visited on 08/18/2020).

Alsolai, Hadeel et al. (2022). “A Systematic Review of Literature on Automated Sleep Scoring”. In: *IEEE Access* 10. Conference Name: IEEE Access, pp. 79419–79443.

Barger, Zeke et al. (Dec. 2019). “Robust, automated sleep scoring by a compact neural network with distributional shift correction”. en. In: *PLOS ONE* 14.12. Publisher: Public Library of Science, e0224642. (Visited on 12/17/2020).

Bartlett, Peter L. et al. (Dec. 2020). “Benign overfitting in linear regression”. en. In: *Proceedings of the National Academy of Sciences* 117.48, pp. 30063–30070. (Visited on 01/10/2023).

Belkin, Mikhail et al. (2018). “Reconciling modern machine learning practice and the bias-variance trade-off”. en. In: p. 23.

Bernardi, Giulio et al. (Apr. 2019). “Regional Delta Waves In Human Rapid Eye Movement Sleep”. en. In: *Journal of Neuroscience* 39.14. Publisher: Society for Neuroscience Section: Research Articles, pp. 2686–2697. (Visited on 02/05/2023).

Borniger, Jeremy C. et al. (July 2018). "A Role for Hypocretin/Orexin in Metabolic and Sleep Abnormalities in a Mouse Model of Non-metastatic Breast Cancer". en. In: *Cell Metabolism* 28.1, 118–129.e5. (Visited on 12/17/2020).

Bukhtiyarova, Olga et al. (2016). "Supervised semi-automatic detection of slow waves in non-anaesthetized mice with the use of neural network approach". en. In: *Translational Brain Rhythmicity* 1.1. (Visited on 02/05/2023).

Caldart, Carlos S. et al. (July 2020). *Sleep Identification Enabled by Supervised Training Algorithms (SIESTA): An open-source platform for automatic sleep staging of rodent polysomnographic data*. en. preprint. Neuroscience. (Visited on 08/18/2020).

Chambon, Stanislas et al. (Sept. 2018). "A Deep Learning Architecture to Detect Events in EEG Signals During Sleep". In: *2018 IEEE 28th International Workshop on Machine Learning for Signal Processing (MLSP)*. Aalborg: IEEE, pp. 1–6. ISBN: 978-1-5386-5477-4. (Visited on 08/18/2020).

Chari, Tara, Joeyta Banerjee, and Lior Pachter (Aug. 2021). *The Specious Art of Single-Cell Genomics*. en. Pages: 2021.08.25.457696 Section: New Results. URL: https://www.biorxiv.org/content/10.1101/2021.08.25.457696v1 (visited on 03/22/2023).

Churchland, Mark M. et al. (July 2012). "Neural population dynamics during reaching". en. In: *Nature* 487.7405. Number: 7405 Publisher: Nature Publishing Group, pp. 51–56. (Visited on 03/20/2023).

Cunningham, John P. and Byron M. Yu (Nov. 2014). "Dimensionality reduction for large-scale neural recordings". en. In: *Nature Neuroscience* 17.11. Number: 11 Publisher: Nature Publishing Group, pp. 1500–1509. (Visited on 03/20/2023).

Diekelmann, Susanne and Jan Born (Feb. 2010). “The memory function of sleep”. en. In: *Nature Reviews Neuroscience* 11.2. Number: 2 Publisher: Nature Publishing Group, pp. 114–126. (Visited on 02/10/2023).

Duan, Sunny et al. (Feb. 2020). *Unsupervised Model Selection for Variational Disentangled Representation Learning*. arXiv:1905.12614 [cs, stat]. URL: http://arxiv.org/abs/1905.12614 (visited on 03/06/2023).

Duan, Yan et al. (Nov. 2016). “RL^2: Fast Reinforcement Learning via Slow Reinforcement Learning”. en. In: *arXiv:1611.02779 [cs, stat]*. arXiv: 1611.02779. (Visited on 04/25/2020).

Edelman, Gerald M. (Apr. 2003). “Naturalizing consciousness: A theoretical framework”. In: *Proceedings of the National Academy of Sciences* 100.9. Publisher: Proceedings of the National Academy of Sciences, pp. 5520–5524. (Visited on 02/05/2023).

Engel, Tatiana A. et al. (Dec. 2016). “Selective modulation of cortical state during spatial attention”. en. In: *Science* 354.6316, pp. 1140–1144. (Visited on 03/20/2023).

Fiorillo, Luigi et al. (Dec. 2019). “Automated sleep scoring: A review of the latest approaches”. en. In: *Sleep Medicine Reviews* 48, p. 101204. (Visited on 11/21/2022).

Funk, Chadd M. et al. (Feb. 2016). “Local Slow Waves in Superficial Layers of Primary Cortical Areas during REM Sleep”. eng. In: *Current biology: CB* 26.3, pp. 396–403.

Genkin, Mikhail and Tatiana A. Engel (Nov. 2020). “Moving beyond generalization to accurate interpretation of flexible models”. en. In: *Nature Machine Intelligence* 2.11. Number: 11 Publisher: Nature Publishing Group, pp. 674–683. (Visited on 03/25/2023).

Gervasoni, Damien et al. (Dec. 2004). "Global Forebrain Dynamics Predict Rat Behavioral States and Their Transitions". en. In: *Journal of Neuroscience* 24.49. Publisher: Society for Neuroscience Section: Behavioral/Systems/Cognitive, pp. 11137–11147. (Visited on 12/17/2020).

Gower, J. C. (Mar. 1975). "Generalized procrustes analysis". en. In: *Psychometrika* 40.1, pp. 33–51. (Visited on 03/19/2023).

Gunnarsdottir, Kristin M. et al. (Feb. 2020). "A novel sleep stage scoring system: Combining expert-based features with the generalized linear model". en. In: *Journal of Sleep Research*. (Visited on 08/27/2020).

Harris, Kenneth D. and Alexander Thiele (Sept. 2011). "Cortical state and attention". en. In: *Nature Reviews Neuroscience* 12.9. Number: 9 Publisher: Nature Publishing Group, pp. 509–523. (Visited on 02/24/2023).

Hernan, Amanda E. et al. (Nov. 2017). "Methodological standards and functional correlates of depth in vivo electrophysiological recordings in control rodents. A TASK1-WG3 report of the AES/ILAE Translational Task Force of the ILAE". eng. In: *Epilepsia* 58 Suppl 4.Suppl 4, pp. 28–39.

Higgins, Irina et al. (2017). "-VAE: LEARNING BASIC VISUAL CONCEPTS WITH A CONSTRAINED VARIATIONAL FRAMEWORK". en. In: *International Conference of Learning Representations*.

Higgins, Irina et al. (Dec. 2021). "Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons". en. In: *Nature Communications* 12.1, p. 6456. (Visited on 11/16/2021).

Hulsey, Daniel et al. (Mar. 2023). *Transitions between discrete performance states in auditory and visual tasks are predicted by arousal and uninstructed movements*. en. Pages: 2023.03.02.530651 Section: New Results. URL: https://www.biorxiv.org/content/10.1101/2023.03.02.530651v1 (visited on 03/17/2023).

Keshtkaran, Mohammad Reza and Chethan Pandarinath (2019). "Enabling hyperparameter optimization in sequential autoencoders for spiking neural data". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. (Visited on 03/06/2023).

Khemakhem, Ilyes et al. (Oct. 2020a). "ICE-BeeM: Identifiable Conditional Energy-Based Deep Models Based on Nonlinear ICA". In: *arXiv:2002.11537 [cs, stat]*. arXiv: 2002.11537. (Visited on 11/03/2021).

Khemakhem, Ilyes et al. (June 2020b). "Variational Autoencoders and Nonlinear ICA: A Unifying Framework". en. In: *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*. ISSN: 2640-3498. PMLR, pp. 2207–2217. (Visited on 09/27/2021).

Kingma, Diederik P. and Max Welling (May 2014). *Auto-Encoding Variational Bayes*. arXiv:1312.6114 [cs, stat]. URL: http://arxiv.org/abs/1312.6114 (visited on 11/18/2022).

— (2019). "An Introduction to Variational Autoencoders". In: *Foundations and Trends® in Machine Learning* 12.4. arXiv:1906.02691 [cs, stat], pp. 307–392. (Visited on 03/18/2023).

Koch, Henriette, Poul Jennum, and Julie A. E. Christensen (2019). "Automatic sleep classification using adaptive segmentation reveals an increased number of rapid eye movement sleep transitions". en. In: *Journal of Sleep Research* 28.2. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/jsr.12780, e12780. (Visited on 05/18/2022).

Krueger, James M. et al. (Feb. 2019). "Local sleep". en. In: *Sleep Medicine Reviews* 43, pp. 14–21. (Visited on 02/05/2023).

Libet, B. et al. (Dec. 1967). "Responses of Human Somatosensory Cortex to Stimuli below Threshold for Conscious Sensation". In: *Science* 158.3808. Publisher: American Association for the Advancement of Science, pp. 1597–1600. (Visited on 02/05/2023).

Liu, Ran et al. (Nov. 2021). "Drop, Swap, and Generate: A Self-Supervised Approach for Generating Neural Activity". In: *arXiv:2111.02338 [cs]*. arXiv: 2111.02338. (Visited on 01/20/2022).

Locatello, Francesco et al. (June 2019). "Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations". In: *arXiv:1811.12359 [cs, stat]*. arXiv: 1811.12359. (Visited on 11/12/2021).

Luxem, Kevin et al. (Nov. 2022). "Identifying behavioral structure from deep variational embeddings of animal motion". en. In: *Communications Biology* 5.1. Number: 1 Publisher: Nature Publishing Group, pp. 1–15. (Visited on 02/05/2023).

Lüttjohann, Annika, Paolo F. Fabene, and Gilles van Luijtelaar (Dec. 2009). "A revised Racine's scale for PTZ-induced seizures in rats". en. In: *Physiology & Behavior* 98.5, pp. 579–586. (Visited on 03/25/2023).

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing Data using t-SNE". In: *Journal of Machine Learning Research* 9.86, pp. 2579–2605.

Maness, Eden B. et al. (Oct. 2022). "Role of the locus coeruleus and basal forebrain in arousal and attention". en. In: *Brain Research Bulletin* 188, pp. 47–58. (Visited on 02/05/2023).

Mante, Valerio et al. (Nov. 2013). "Context-dependent computation by recurrent dynamics in prefrontal cortex". en. In: *Nature* 503.7474. Number: 7474 Publisher: Nature Publishing Group, pp. 78–84. (Visited on 03/20/2023).

McGinley, Matthew J. et al. (Sept. 2015). "Waking State: Rapid Variations Modulate Neural and Behavioral Responses". en. In: *Neuron* 87.6, pp. 1143–1161. (Visited on 03/17/2023).

McInnes, Leland, John Healy, and James Melville (Sept. 2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. arXiv:1802.03426 [cs, stat]. URL: http://arxiv.org/abs/1802.03426 (visited on 03/20/2023).

McNamara, Patrick (2019). *The neuroscience of sleep and dreams*. The neuroscience of sleep and dreams. Pages: xiv, 263. New York, NY, US: Cambridge University Press. ISBN: 978-1-107-17110-7 978-1-316-62974-1.

Mircea M. Steriade and Robert W. McCarley (2013). *Brainstem Control of Wakefulness and Sleep*. Springer International Publishing. ISBN: 978-1-4757-4669-3.

Mita, Graziano, Maurizio Filippone, and Pietro Michiardi (Feb. 2021). *An Identifiable Double VAE For Disentangled Representations*. arXiv:2010.09360 [cs, stat]. URL: http://arxiv.org/abs/2010.09360 (visited on 02/19/2023).

Musall, Simon et al. (Oct. 2019). "Single-trial neural dynamics are dominated by richly varied movements". en. In: *Nature Neuroscience* 22.10. Number: 10 Publisher: Nature Publishing Group, pp. 1677–1686. (Visited on 03/20/2023).

Nazari, Mojtaba et al. (Mar. 2022). *Regional variation in cholinergic terminal activity determines the non-uniform occurrence of cortical slow-wave activity during REM sleep*. en. preprint. Neuroscience. (Visited on 04/07/2022).

Nir, Yuval et al. (Apr. 2011). "Regional Slow Waves and Spindles in Human Sleep". English. In: *Neuron* 70.1. Publisher: Elsevier, pp. 153–169. (Visited on 02/05/2023).

Pandarinath, Chethan et al. (Oct. 2018). "Inferring single-trial neural population dynamics using sequential auto-encoders". en. In: *Nature Methods* 15.10, pp. 805–815. (Visited on 08/31/2020).

Pei, Felix et al. (Sept. 2021). "Neural Latents Benchmark '21: Evaluating latent variable models of neural population activity". en. In: *arXiv*. (Visited on 01/10/2023).

Poulet, James F. A. and Sylvain Crochet (2019). "The Cortical States of Wakefulness". In: *Frontiers in Systems Neuroscience* 12. (Visited on 04/07/2022).

Prabhudesai, Kedar S., Leslie M. Collins, and Boyla O. Mainsah (Mar. 2019). "Automated feature learning using deep convolutional auto-encoder neural network for clustering electroencephalograms into sleep stages". en. In: *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*. San Francisco, CA, USA: IEEE, pp. 937–940. ISBN: 978-1-5386-7921-0. (Visited on 08/27/2020).

Rector, David M. et al. (May 2009). "Physiological markers of local sleep". en. In: *European Journal of Neuroscience* 29.9, pp. 1771–1778. (Visited on 02/05/2023).

Rocca, Joseph (Mar. 2021). *Understanding Variational Autoencoders (VAEs)*. en. URL: https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73 (visited on 03/20/2023).

Scammell, Thomas E., Elda Arrigoni, and Jonathan O. Lipton (Feb. 2017). "Neural Circuitry of Wakefulness and Sleep". English. In: *Neuron* 93.4. Publisher: Elsevier, pp. 747–765. (Visited on 02/10/2023).

Schneider, Steffen, Jin Hwa Lee, and Mackenzie Weygandt Mathis (Oct. 2022). *Learnable latent embeddings for joint behavioral and neural analysis*. arXiv:2204.00673

[cs, q-bio]. URL: http://arxiv.org/abs/2204.00673 (visited on 03/20/2023).

Sedler, Andrew R., Christopher Versteeg, and Chethan Pandarinath (Feb. 2023). *Expressive architectures enhance interpretability of dynamics-based neural population models*. arXiv:2212.03771 [cs, q-bio]. URL: http://arxiv.org/abs/2212.03771 (visited on 03/06/2023).

Sinha, Samarth and Adji B. Dieng (May 2021). "Consistency Regularization for Variational Auto-Encoders". In: *arXiv:2105.14859 [cs]*. arXiv: 2105.14859. (Visited on 09/27/2021).

Smith, Jennifer et al. (Dec. 2022). *Regulation of stress-induced sleep fragmentation by preoptic glutamatergic neurons*. en. Pages: 2022.11.30.518589 Section: New Results. URL: https://www.biorxiv.org/content/10.1101/2022.11.30.518589v1 (visited on 02/05/2023).

Soltani, Sara et al. (2019). "Sleep–Wake Cycle in Young and Older Mice". In: *Frontiers in Systems Neuroscience* 13. (Visited on 04/07/2022).

Sussillo, David et al. (Aug. 2016). *LFADS - Latent Factor Analysis via Dynamical Systems*. arXiv:1608.06315 [cs, q-bio, stat]. URL: http://arxiv.org/abs/1608.06315 (visited on 01/09/2023).

Tononi, Giulio (Apr. 2009). "Slow Wave Homeostasis and Synaptic Plasticity". In: *Journal of Clinical Sleep Medicine* 5.2 suppl. Publisher: American Academy of Sleep Medicine, S16–S19. (Visited on 02/05/2023).

Vyazovskiy, Vladyslav V. et al. (Sept. 2009). "Cortical Firing and Sleep Homeostasis". en. In: *Neuron* 63.6, pp. 865–878. (Visited on 12/17/2020).

Vyazovskiy, Vladyslav V. et al. (Apr. 2011). "Local sleep in awake rats". en. In: *Nature* 472.7344. Number: 7344 Publisher: Nature Publishing Group, pp. 443–447. (Visited on 04/07/2022).

Watson, Brendon O. et al. (May 2016). “Network Homeostasis and State Dynamics of Neocortical Sleep”. en. In: *Neuron* 90.4, pp. 839–852. (Visited on 04/07/2022).

Wehmeyer, Christoph and Frank Noé (June 2018). “Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics”. In: *The Journal of Chemical Physics* 148.24. Publisher: American Institute of Physics, p. 241703. (Visited on 02/05/2023).

Zhou, Ding and Xue-Xin Wei (2020). “Learning identifiable and interpretable latent models of high-dimensional neural activity using pi-VAE”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 7234–7247. (Visited on 03/22/2023).