[2]$\Big(\Big)1 \parallel 2$

Distribution Agreement

Signature:

Feng Zhu Date

Overcoming Spatiotemporal Trade-offs in Calcium Imaging Using Deep Learning-based Dynamics Modeling

By

Feng Zhu
Doctor of Philosophy

Neuroscience

Chethan Pandarinath, Ph.D.
Advisor

Gordon Berman, Ph.D.
Committee Member

Eva Dyer, Ph.D.
Committee Member

Annabelle Singer, Ph.D.
Committee Member

Samuel Sober, Ph.D.
Committee Member

Accepted:

Kimberly J. Arriola, Ph.D., MPH
Dean of the James T. Laney School of Graduate Studies

Date

Overcoming Spatiotemporal Trade-offs in Calcium Imaging Using Deep Learning-based Dynamics Modeling

By

Feng Zhu
B.A., University of Illinois at Urbana-Champaign, IL, 2015

Advisor: Chethan Pandarinath, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Neuroscience
2022

Abstract

Overcoming Spatiotemporal Trade-offs in Calcium Imaging Using Deep Learning-based Dynamics Modeling

By Feng Zhu

Recent advances in neural interfaces have enabled monitoring of the activity of increasingly large neuronal populations. Among these techniques, two-photon (2p) calcium imaging is a powerful tool to probe how population-level computations relate to biological structure, as it can identify layers and cell types of interest. However, extracting fast patterns of neural activity from 2p data has proven challenging because of limitations on temporal resolution imposed by the spatiotemporal trade-offs inherent to 2p laser scanning. Noises and nonlinearities introduce additional challenges to analyzing 2p data. This dissertation bridges this gap by taking a dynamical systems approach to denoise and improve temporal resolution of 2p data. In Chapter 1, we provide a broad introduction to the challenges with 2p data and methods for modeling neural dynamics, and discuss the benefits of modeling dynamics from 2p data. In Chapter 2, we present a novel neural network training strategy that offers a principled solution to spatiotemporal trade-offs created by bandwidth limits in neural interfaces. This strategy enables inference of latent dynamics with spatiotemporal super resolution and is applicable for a wide range of neural interfaces. In Chapter 3, we detail the results of extending a state-of-the-art deep learning method that models neural dynamics in spiking activity for application to 2p imaging. We demonstrate that our new method outperforms standard methods in recovering high-frequency components in synthetic tests and predicting single-trial behaviors in 2p recordings from sensorimotor areas in mice performing a forelimb reach task. In Chapter 4, we present machine learning innovations that eliminate the need of deconvolution as a preprocessing step for our approach. This opens the door to modeling fast and complex dynamics from 2p data in settings where massive populations of neurons are imaged with extremely slow sampling rates as a trade-off. In sum, our work provides an avenue to overcome the limits of spatiotemporal trade-offs in 2p calcium imaging, enabling accurate inference of population dynamics across a wide range of sampling rates in vast populations with identified neurons.

Overcoming Spatiotemporal Trade-offs in Calcium Imaging Using Deep Learning-based Dynamics Modeling

By

Feng Zhu
B.A., University of Illinois at Urbana-Champaign, IL, 2015

Advisor: Chethan Pandarinath, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Neuroscience
2022

Acknowledgments

My PhD is a mixture of many things: pain and happiness, despair and hope, fear and bravery...This is an unbelievable journey that I could not complete without love and support from my advisor Chethan Pandarinath, my dissertation committee, my lab members, my friends, and my family (my parents, wife, and son). I am very grateful to these people.

First, I would like to thank my PhD advisor Chethan Pandarinath. Your enthusiasm for science, your ability to come up with novel ideas and turn those into actions, your extraordinary communication skills, your endless support to lab members, your kindness - you have defined what an amazing advisor and team leader should look like. Personally, your influence on me is way beyond my PhD. You opened the door for me to enter the field of computational neuroscience and machine learning. You see opportunities through a mess and you turn impossible things into success. You nail down actions from the most complicated inputs and critiques from others. You as a leader run at the front - how could I not follow your steps? I learned so much from you, and you will always be the role model of my life. I appreciate the wisdom and courage that you gave to me and I feel I am empowered to overcome any difficulties of my life.

I would also like to thank the members of my dissertation committee, Gordon Berman, Eva Dyer, Annabelle Singer, and Sam Sober. You have all provided insightful feedback to my projects and helpful career advice. I would like to thank my collaborators Matt Kaufman, Andrea Giovannucci, Hank Grier, Changjia Cai and Raghav Tandon on the multidisciplinary calcium imaging projects. You have provided extremely helpful feedback that shaped the path of the projects. I have learned a lot from you through our weekly meetings through the years, including designing appropriate analyses, interpreting results, and writing. I would like to thank Ellen Hess and Jeff Markowitz for guidance in the collaborative work on one-photon calcium imaging projects.

As one of the first graduate students that joined the Pandarinath lab, I have witnessed the growth of our lab. We have built an open and collaborative lab culture; our lab members have very strong connections and we are highly supportive to each other. I would like to thank my lab members for discussing research ideas, providing insightful feedback on projects and presentations, helping me get up to speed on a lot of the computational concepts and technical details and chatting about life. I would like to thank the students I have mentored. Seeing you learning new skills and making progress on the projects has been a pleasure of my PhD. I would also like to thank my cohort (Emory Neuroscience and BME Neuro students) for being supportive and providing inspiration.

Finally I would like to thank my friends and family. Zhenyu Gao has been my closest friend since we were undergraduates at University of Illinois at Urbana-Champaign. Thank you for being supportive to me and chatting about research, career and life. I would like to thank my parents. I always feel warm and hopeful because of the support from my parents. Thank you for giving me all the freedom I needed when I made a decision to study abroad and start a career in science to chase my dream. A special thank to my grandfather. My grandfather led me into the world of biology and science in general. He took me to the nature to observe cicadas and many other insects since I was in my childhood. Since then, I become very interested about living creatures, and then biology, and then the brain and neuroscience.

Most of all, I would like to thank my wife Xuan. You are the one that knows me the most, gets my joke every time, being with me for all of the difficult or happy times since I was 19 years old. I would not be where I am today without my wife's love and support. Thank you for being the audience of the show of my life, for being the tree trunk for me to stay like a cicada, for completing my soul. I would also like to thank my son Luca, for bringing happiness and light to our family.

Contents

List of Figures

Chapter 1

Introduction

1.1 Benefits and challenges with two-photon (2p) calcium imaging

Over the past decade, the ability to record from large populations of neurons has increased exponentially (Stevenson and Kording, 2011; Steinmetz et al., 2021; Demas et al., 2021). Such advances in neural interfaces are enabling new insights into how neural populations implement the computations necessary for motor, sensory, and cognitive processes (Vyas et al., 2020a). Among these techniques, two photon (2p) calcium imaging offers the revolutionary ability to monitor in vivo vast populations of neurons - rapidly increasing from tens of thousands to millions - in 3-D, with cell types of interest and layers identified (Demas et al., 2021; Pachitariu et al., 2017; Peron et al., 2015b). Therefore, 2p imaging is a powerful technique that has been widely used by neuroscientists to provide insights on how neural circuitry gives rise to function.

1.1.1 Challenges with analyzing 2p data

A key trade-off in 2p imaging, however, is that greater spatial sampling (i.e., more neurons) is typically associated with lower temporal resolution of sampling (Siegle et al., 2021). With 2p imaging, neurons are serially scanned by a focused laser beam that traverses the field of view (FOV), resulting in different neurons being sampled at different times within an imaging frame. As a consequence, a trade-off exists between the size of the FOV (and hence the number of neurons monitored), the sampling frequency, and the pixel size (and hence the signal-to-noise with which each neuron is sampled). Further, 2p calcium imaging provides a noisier and indirect measurement of neurons' spiking activity (Wei et al., 2020b; Peron et al., 2015a). The measured fluorescence transients are a low-passed transformation of the underlying spiking activity, with the time course of the transient dictated by the calcium indicator's rise and fall times, and also limited by the kinetics of calcium buffering (Wei et al., 2020b). Moreover, the transformation is nonlinear: the magnitude of fluorescence change produced by spiking is not linearly proportional to the number of spikes (Tian et al., 2009; Chen et al., 2013b; Scheuss et al., 2006). For example, a single spike might have little effect on the measured fluorescence, while two spikes might produce a substantial increase. These challenges together limit the fidelity with which the activity of large neuronal populations can be monitored and extracted via 2p, and thus limit our ability to link 2p activity to neural computation and behavior on fine timescales.

1.1.2 Computational methods for analyzing 2p data

In recent years, a variety of computational methods have been employed to combat the challenges with 2p data and improve inference of neuronal activity from 2p imaging data. Some previous work has made steps in overcoming the limits of modest 2p frame rates in attempts to infer the fast changes in neural firing rates that relate to fast behaviors. Efforts to chip away at this barrier have relied on regularities imposed by repeated stimuli or highly stereotyped behavior (Picardo et al., 2016; Mano et al., 2019), or jittered inferred events on sub-frame timescales to minimize the reconstruction error of the associated fluorescence (Hoang et al., 2020). However, these methods show limited improvements and could not be generalized to more naturalistic or flexible behaviors due to the requirement of stereotypy in the behavior or neural response.

More effort has been dedicated to the inference of neurons' spike times from 2p imaging data (Pnevmatikakis, 2019a). Established pipelines for the analysis of calcium imaging data (Freeman et al., 2014; Giovannucci et al., 2019; Kaifosh et al., 2014; Lu et al., 2018; Pachitariu et al., 2017; Pnevmatikakis et al., 2016; Romano et al., 2017; Zhou et al., 2018) usually feature infrastructure to handle large datasets (Freeman et al., 2014; Giovannucci et al., 2019), algorithms to correct for motion artifacts (Friedrich et al., 2017), routines to localize neurons from summary statistics (Apthorpe et al., 2016; Kaifosh et al., 2014; Kirschbaum et al., 2020; Mishne et al., 2018; Pachitariu et al., 2013; Reynolds et al., 2017; Soltanian-Zadeh et al., 2019) or methods based on matrix factorization to simultaneously tackle the problem of separating signals from overlapping fluorescence sources in both the spatial and temporal domains (Giovannucci et al., 2017, 2019; Inan et al., 2017; Keemink et al., 2018; Maruyama et al., 2014; Mishne and Charles, 2019; Mukamel et al., 2009; Pachitariu et al., 2017; Pnevmatikakis et al., 2016; Saxena et al., 2020; Zhou et al., 2018). The resulting fluorescence signals can then be processed to infer spike times by a variety of spike detection algorithms. A popular family of such algorithms exploits biophysical models of calcium spike generation to estimate a correlate of firing rate (Friedrich et al., 2017; Pachitariu et al., 2017; Pnevmatikakis et al., 2016; Vogelstein et al., 2010; Wei et al., 2020a; Yaksi and Friedrich, 2006) or single spikes (Deneux et al., 2016; Ganmor et al., 2016; Jewell et al., 2020; Oñativia et al., 2013; Pnevmatikakis et al., 2013; Vogelstein et al., 2009) to solve the spike extraction problem (Berens et al., 2018; Hoang et al., 2020; Rupprecht et al., 2021; Sebastian et al., 2021; Speiser et al., 2017; Theis et al., 2016) with promising results. Finally, some algorithms are trying to exploit the spatio-temporal relationship among neurons or across trials to build a better representation of spiking patterns directly from raw movies or from de-noised fluorescence traces (Picardo et al., 2016).

Ideally the spikes-to-fluorescence transformation would be invertible, such that analyzing calcium events would be equivalent to analyzing spiking activity (Wei et al., 2020b). However, recent benchmarks illustrate that the spike inference algorithms described above all achieve limited correspondence to ground truth spiking activity obtained with electrophysiology, particularly on fine timescales (Berens et al., 2018; Pachitariu et al., 2018). Rather than focusing on the responses of individual neurons, an alternative approach is to leverage the coordinated patterns of activity across the neural population. Empirical studies have repeatedly demonstrated strong shared structure underlying the activity of large populations of neurons (Montijn et al., 2016), as might be expected from a highly interconnected network. Due to this coordination, population activity might serve as a rich, complementary source of information to improve inference from calcium signals.

There are multiple lines of work that have attempted to capture the population structure underlying calcium imaging signals, such as methods to identify repeated sequences (Mackevicius et al., 2019) or temporal factors that slowly change across trials (Williams et al., 2018) from neural population activity monitored with calcium imaging, or methods to extract recurring firing motifs directly from calcium imaging videos using a variational autoencoder framework (Kirschbaum et al., 2019). Dimensionality reduction methods have been developed to decouple evoked and spontaneous activities (Triplett et al., 2020) or to identify an input manifold of odor representations (Wu et al., 2018) from calcium signals. However, to our best knowledge, no study has demonstrated a precise inference of the population state from calcium imaging that corresponds closely with single-trial behavior on fine timescales (i.e., 10 millisecond).

1.1.3 A dynamical systems approach for 2p data

Among the broad family of population-level analysis, there is one class of methods that takes a dynamical systems approach. These dynamics models characterize patterns of covariation across a neuronal population to reveal the multi-dimensional internal state of the network (i.e., network state) as a whole. They find the consistent rules (i.e., dynamics) that govern how the network state evolve over time, and describes the time-varying activity of each neuron as a reflection of the underlying network state (Linderman et al., 2017; Zoltowski et al., 2020; Pandarinath et al., 2018b). For example, when applied to electrophysiological data, dynamics models assume that an individual neuron's spiking is a noisy observation of a latent "firing rate", which fluctuates in a coordinated way with the firing rates of other neurons in the population. Leveraging the dynamics inherent in the neural populations, dynamics models offer great promise in improving inference: the trajectory of network state inferred by dynamics models can reveal key insights into the computations being performed by the brain areas of interest (Vyas et al., 2020a), and the inferred network state can also enhance our ability to relate neural activity to behavior. For example, one state-of-the-art dynamics model is Latent Factor Analysis via Dynamical Systems (LFADS), which is a deep learning-based method that estimates network state from electrophysiological spiking data (Sussillo et al., 2016; Pandarinath et al., 2018b). When applied to data from motor, sensory, and cognitive regions, LFADS-inferred network state reveals close correspondence with single-trial behavior on a 5-10 millisecond timescale (Pandarinath et al., 2018b; Keshtkaran et al., 2021).

Given the success of dynamics models in uncovering network state from electrophysiological data, this dissertation develops new approaches based on dynamics to tackle challenges and achieve accurate inference of network state from activity monitored through 2p calcium imaging. We build upon the state-of-the-art dynamics model, LFADS, and develop innovations to adapt the model to fit calcium signals and utilize staggered sample times of the neurons to achieve sub-frame temporal resolution. We next evaluate the new method using both simulated 2p data in which activity reflects known, nonlinear dynamical systems, and with real 2p data from mice performing a water reaching task. Last, we identify a limitation of the new method at slower imaging rates due to the dependency of deconvolution, and develop additional innovations to eliminate the need of deconvolution. Ultimately, work from this dissertation provides an avenue to link network-level computation to biological structure (i.e., cell types and layers) in unprecedented ways.

In Chapter 1.2, we will detail the concepts of neural dynamics and methods to model dynamics, and discuss why modeling dynamics could mitigate the challenges with calcium imaging, such as low temporal resolution and distortions of the signals.

1.2 Model neural populations as dynamical systems

1.2.1 Definition and evidence of neural population dynamics

Early work to understand how neural activity has focused on analyzing individual neurons' activities and how they "represent" externally-measurable or controllable parameters (Georgopoulos et al., 1982; Hubel and Wiesel, 1959). However, the increased ability to get access to the activity of many neurons simultaneously has revealed many features at the level of neural population that are difficult to explain using the representational viewpoint (Churchland and Shenoy, 2007; Fetz, 1992). This has motivated a shift of the field towards understanding how neurons within a network coordinate their activity to perform computations.

A large body of work suggests that the activity of individual neurons within a large population is not independent, but instead is coordinated through a lower-dimensional, internal state that evolves following lawful dynamics in time (Yuste, 2015; Vyas et al., 2020a). Here we formulate such neural population dynamics as follows: At a given time point t, we can represent the observed activity of a population of N neurons as a vector $\boldsymbol{y}_t \in \mathbb{R}^N$. The latent, internal state underlying these observed neural activity ($\boldsymbol{y}_t$) can be represented as a vector $\boldsymbol{x}_t \in \mathbb{R}^D$, where $\boldsymbol{y}_t \approx h(\boldsymbol{x}_t)$ for some function h. In many brain areas, the dimension D of $\boldsymbol{x}_t$ to be far smaller than the number of possible observations N (Cunningham and Yu, 2014). The evolution of the latent state over time is largely driven by dynamics captured by a function f such that $\boldsymbol{x}_{t+1} \approx f(\boldsymbol{x}_t)$. Note that, in systems that receive inputs (i.e., a non-autonomous dynamical system), the governing rule of the evolution of the latent state is a function of both the current state of the system and the inputs.

Motor cortex, in particular, has emerged as a key area for understanding neural dynamics, as its activity is strongly governed by internal dynamics, yet also well related to observable behavior (Churchland and Shenoy, 2007; Churchland et al., 2012; Elsayed et al., 2016; Pandarinath et al., 2018b,a). A piece of groundbreaking work has revealed that activity in motor cortex during movement generation can be well explained by rotational dynamics, with the initial condition seeded by preparatory activity and consistent rotational rules governing the evolution of the population activity during movement execution (Churchland et al., 2012). The dynamical systems framework has then provided a new avenue to probe many functions of the motor cortex, including movement preparation and execution (Ames et al., 2014; Kaufman et al., 2014; Elsayed et al., 2016), motor learning (Sadtler et al., 2014; Golub et al., 2015), and generation of muscle activity (Russo et al., 2018). In addition to motor areas, computation through dynamics is increasingly recognized as a widespread phenomenon across many other brain areas (Vyas et al., 2020a). Modeling the activity of neural populations as latent dynamical systems has provided new insights into the computations involved in cognitive processes such as decision-making (Mante et al., 2013; Raposo et al., 2014; Carnevale et al., 2015), interval timing (Remington et al., 2018), and navigation (Harvey et al., 2012; Morcos and Harvey, 2016).

1.2.2 Methods for inferring neural dynamics

Recent empirical evidence of neural dynamics has highlighted the need to further develop tools capable of inferring latent structure and dynamics more efficiently and accurately. A common approach to estimate latent structure is principal component analysis (PCA). However, methods based on PCA usually require averaging across repeated trials as a denoising step to yield reasonable performance (Yu et al., 2009; Pandarinath et al., 2018a). As an alternative, factor analysis (FA) is used for single-trial analysis, by treating activity shared across neurons as ”signal” and independent activity as ”noisy” (Sadtler et al., 2014; Golub et al., 2015, 2018), or by introducing a set of ”trial factors” to account for variability across trials (Williams et al., 2018). However, a key limitation of PCA or FA is that they do not model the temporal dependency between time points, and thus cannot leverage dynamics to provide more precise estimation of the latent state.

A variety of methods have been developed to improve inference of latent states by modeling the interdependencies between time points. One class of such methods are based on Gaussian process (GP) (Yu et al., 2009; Lakshmanan et al., 2015; Zhao and Park, 2017). These methods assume that the latent factors evolve independently in time, with the smoothness of each factor dictated by its own characteristic time constant (Yu et al., 2009; Pandarinath et al., 2018a). Another class of methods are based on linear dynamical systems (LDS) (Macke et al., 2012; Gao et al., 2016; Kao et al., 2015). LDS-based methods assume that the latent state at a given time point is a linear transformation of its state at the previous time point, allowing interactions between latent factors (Macke et al., 2012; Pandarinath et al., 2018a). Because activity at different behavioral phases might be governed by different dynamics, switching LDS (SLDS) was developed to allow a set of linear dynamics to be learned to capture latent state at different phases (Petreska et al., 2011; Linderman et al., 2017; Glaser et al., 2020). Despite the improvements provided by GP- and LDS-based methods, they all rely on strong simplifying assumptions about the underlying dynamics. As a result, their performance of estimating latent dynamics breaks down when tested on nonlinear systems with higher complexity, as revealed by recent benchmarks on real neural population recordings (Pei et al., 2021; Sussillo et al., 2016; Pandarinath et al., 2018b).

Instead of simplifying assumptions about the underlying dynamics, a recently developed method uses recurrent neural networks (RNNs) which are powerful nonlinear function approximators, capable of modeling complex, highly nonlinear dynamics (Sussillo et al., 2016; Pandarinath et al., 2018b). Using a sequential autoencoder configuration, this method, known as Latent factor analysis via dynamical systems (LFADS), models neural population activity as an input-driven dynamical system and achieves a breakthrough in the ability to infer latent state and dynamics on single trials. The latent state inferred by LFADS can be linked to behavior with unprecedented accuracy on a moment-by-moment basis (Pandarinath et al., 2018b). Recent work further developed a powerful, large-scale hyperparameter optimization framework on top of LFADS, known as AutoLFADS (Keshtkaran and Pandarinath, 2019; Keshtkaran et al., 2021), to ensure the hyperparameters to be optimized properly and efficiently.

More recent effort on modeling neural dynamics has focused on utilizing new advances in computer science and deep learning. Besides RNN-based models (i.e., LFADS / AutoLFADS), a transformer-based model was developed to reduce inference time for potential applications in brain-machine interfaces (Ye and Pandarinath, 2021). A method based on neural ODE was developed to improve interpretability of learned dynamics (Kim et al., 2021). A control-based method was developed to tackle the challenge of inferring ongoing external inputs to the dynamical system (Schimel et al., 2022). Given the bloom of methods for modeling neural dynamics, a benchmark has been developed to systematically compare methods and coordinate efforts across various developers (Pei et al., 2021).

In this dissertation, we develop and test dynamics models for application to 2p calcium imaging data. We build our models on top of AutoLFADS, as it gives state-of-the-art performance on the benchmark (Pei et al., 2021) and has demonstrated superb ability to link neural activity to behavior across datasets from different brain areas (Pandarinath et al., 2018b) and recording modalities (Flint et al., 2020; Wimalasena et al., 2022).

1.2.3 How could modeling dynamics mitigate challenges with calcium imaging

As detailed in 1.1.1, analyzing 2p imaging data has proven challenging because of the distortions of neuronal activity (noises and nonlinearities) and limitations on temporal resolution. Here we provide reasonings on why modeling neural dynamics could tackle these challenges.

As shown in prior work (Pandarinath et al., 2018b), modeling dynamics is an effective way to denoise the neural population activity because of two key principles it relies on. First, simultaneously recorded neurons are not independent, but rather exhibit coordinated spatial patterns of activation that reflect the state of the network (Cunningham and Yu, 2014; Pandarinath et al., 2018a). Due to this coordination, a substantial fraction of a given neuron's activity can be inferred by knowing the activity of other neurons in the same network (Montijn et al., 2016; Yu et al., 2009). As a consequence, network states might also be reliably estimated even if the measurement of individual neurons' activity is unreliable (e.g., distorted due to noises or nonlinearities in 2p imaging). Second, the activation of these coordinated spatial patterns evolves over time in ways that are largely predictable based on consistent rules (dynamics) (Shenoy et al., 2013; Vyas et al., 2020a). Thus, while it may be challenging to accurately estimate the network's state based solely on activity observed at a single time point, knowledge of the network's dynamics provides further information to help constrain network state estimates using data from multiple time points. These principles of neural dynamics make it possible to substantially denoise 2p data and extract the underlying latent state with higher precision.

Modeling dynamics also provides a solution to improve temporal resolution. As described in 1.1.1, in 2p experiments, neurons are serially scanned by a laser that traverses the field of view (FOV), resulting in different neurons being sampled at different times within an imaging frame. For each individual neuron, the sampling rate is relatively low, equal to the frame rate of imaging. Indeed, current population-level analysis methods treat 2p data as if all neurons within a FOV were sampled at the same time at the imaging frame rate. However, the fact that each neuron is sampled at staggered, known times within the frame could be employed to increase the time resolution at the level of population. The dynamical systems approach allows us to leverage such sub-frame timing information to increase temporal resolution. One way of doing so is to rebin the data into finer, sub-frame bins, with activity from each neuron assigned to the appropriate bins according to the sampled times. In this way, we create a sparse matrix with more precise timing and recast the underlying interpolation problem as a missing data problem that can be naturally handled by the dynamical systems approach. Due both to the fact that the dynamics imposes a significant amount of structure on the trajectory of the latent state and the fact that the dimensionality of latent state is typically far smaller than the number of observed neurons, it is possible to estimate the latent state without observing every neuron at every time bin. This approach opens the door for breaking the limit of space-time trade-offs inherent to 2p imaging (and many other neural interfaces) and achieving sub-frame temporal resolution.

In sum, the dynamical systems approach provides an avenue to precisely infer latent state from 2p imaging data with substantially improved temporal solution.

1.3 Dissertation overview

My research and dissertation focus on developing methods to improve inference of latent dynamics from 2p imaging data, with the goal of providing a better analytical tool of 2p data for the neuroscience community. This dissertation is divided into three main parts, each having a different focus (chapters 2, 3 and 4).

In Chapter 2, we formulate a fundamental challenge in neuroscience, the space-time trade-off, due to bandwidth limits that are inherent to a variety of modern neural interfaces. We develop a novel neural network training strategy, namely selective backpropagation through time (SBTT), to achieve spatiotemporal super resolution. We validate the applicability of SBTT across different recording modalities (i.e., electrophysiology and 2p calcium imaging), neural network architectures (i.e., RNNs and transformers), and training settings (i.e., transfer learning for practical usage in brain-machine interfaces).

In Chapter 3, we design and incorporate innovations tailored specifically for 2p data, and validate and provide a framework, namely Recurrent Autoencoder for Discovering Imaged Calcium Latents (RADICaL), to precisely infer single-trial latent dynamics from 2p data that is convenient to use by the neuroscience community. With synthetic data, we find that RADICaL outperforms other state-of-the-art methods in recovering higher-frequency features in the latent state, and provide evidence on when RADICaL will succeed or fail across a variety of experimental settings (i.e., sampling rate, frequency of the underlying features, noise level, nonstereotyped trial structure, way to deconvolve). With real 2p data from mice performing a forelimb reach task, we test RADICaL across 2 mice and 2 brain areas (M1 and S1). We find that RADICaL precisely infers latent state that shows a close correspondence with the animals' behavior, substantially outperforming alternate methods. We also demonstrate that RADICaL maintains high-quality inference even when neuronal populations are greatly reduced.

In Chapter 4, we identify a limitation of RADICaL described in Chapter 3. RADICaL relies on deconvolution as a pre-processing step, but deconvolution breaks down when sampling rate is low (e.g., < 8 Hz). To enable RADICaL to be generalized to slower sampling regimes which are common in 2p experiments, we eliminate the need of deconvolution by integrating a generative rate-to-fluorescence autoregressive model into RADICaL. We next develop a regularization strategy, per-neuron coordinated dropout, to better separate the inference of population-level, latent dynamics from inference of the neuron-level, calcium dynamics. Our new method, namely deconvolution-free RADICaL (DfRAD), demonstrates precise inference of latent state when the sampling rate is substantially reduced (e.g., 2 Hz) and enables the application of dynamics modeling of some of the most exciting, massive datasets in systems neuroscience.

Chapter 5 discusses future directions and summarizes the dissertation.

A comprehensive list of references is provided at the end of the dissertation.

Chapter 2

Deep inference of latent dynamics with spatio-temporal super-resolution using selective backpropagation through time

2.1 abstract

Modern neural interfaces allow access to the activity of up to a million neurons within brain circuits. However, bandwidth limits often create a trade-off between greater spatial sampling (more channels or pixels) and the temporal frequency of sampling. Here we demonstrate that it is possible to obtain spatio-temporal super-resolution in neuronal time series by exploiting relationships among neurons, embedded in latent low-dimensional population dynamics. Our novel neural network training strategy, selective backpropagation through time (SBTT), enables learning of deep generative models of latent dynamics from data in which the set of observed variables changes at each time step. The resulting models are able to infer activity for missing samples by

combining observations with learned latent dynamics. We test SBTT applied to sequential autoencoders and demonstrate more efficient and higher-fidelity characterization of neural population dynamics in electrophysiological and calcium imaging data. In electrophysiology, SBTT enables accurate inference of neuronal population dynamics with lower interface bandwidths, providing an avenue to significant power savings for implanted neuroelectronic interfaces. In applications to two-photon calcium imaging, SBTT accurately uncovers high-frequency temporal structure underlying neural population activity, substantially outperforming the current state-of-the-art. Finally, we demonstrate that performance could be further improved by using limited, high-bandwidth sampling to pretrain dynamics models, and then using SBTT to adapt these models for sparsely-sampled data.

2.2 Introduction

Modern systems neuroscientists have access to the activity of many thousands to potentially millions of neurons via multi-photon calcium imaging and high-density silicon probes (Stringer et al., 2019a; Demas et al., 2021; Jun et al., 2017; Steinmetz et al., 2021). Such interfaces provide a qualitatively different picture of brain activity than was achievable even a decade ago.

However, neural interfaces increasingly face a trade-off – the number of neurons that can be accessed (capacity) is often far greater than the number that is simultaneously monitored (bandwidth). For example, with 2-photon calcium imaging (2p; **Fig. 2.1a**, *top*), hundreds to thousands of neurons are serially scanned by a laser that traverses the field of view, resulting in different neurons being sampled at different times within an imaging frame. As a consequence, a trade-off exists between the size of the field-of-view (and hence the number of neurons monitored), the sampling frequency, and the signal-to-noise with which each neuron is sampled. Whereas current analysis methods treat 2p data as if all neurons within a field-of-view were sampled at the same time at the imaging frame rate, the fact that each neuron is sampled at staggered, known times within the frame could be employed to increase the time resolution.

Electrophysiological interfaces face similar trade-offs (**Fig. 2.1a**, *bottom*). With groundbreaking high-density probes such as Neuropixels and Neuroseeker (Jun et al., 2017; Steinmetz et al., 2021; Raducanu et al., 2017), simultaneous monitoring of all recording sites is either not currently possible or limits the signal-to-noise ratio, so users typically monitor a selected subset of sites within a given recording session. For example, Neuropixels 2.0 probes contain up to 5120 electrodes, 384 of which can be recorded simultaneously (Steinmetz et al., 2021). In other situations, power constraints might make it preferable to restrict the number of channels that are simultaneously monitored, such as in wireless or fully-implanted applications where battery life and heat dissipation are key challenges (Miranda et al., 2010; Borton et al., 2013;

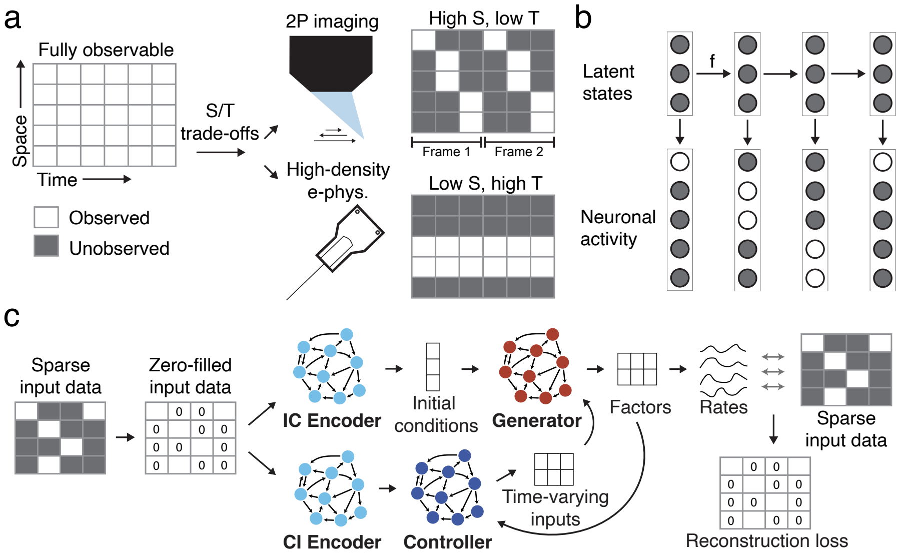


Figure 2.1: Exploiting space-time trade-offs in neural interfaces using SBTT. (a) In 2-photon calcium imaging (top), individual neurons are serially scanned at a low frame rate, resulting in staggered sample times. In modern electrophysiological recordings (bottom), bandwidth or power constraints prevent simultaneous monitoring of all recording sites. (b) Observed neuronal activity reflects latent, low-dimensional dynamics (captured by the function f). (c) SBTT applied to a sequential autoencoder for inferring latent dynamics from neural population activity.

Simeral et al., 2021). As newer interfacing strategies provide a pathway to hundreds of thousands of channels for revolutionary brain-machine interfaces (Sahasrabuddhe et al., 2021; Musk et al., 2019), neural data processing strategies that can leverage dynamic deployment of recording bandwidth might allow substantial power savings.

Solutions to these space-time trade-offs may come from the structure of neural activity itself. A large body of work suggests that the activity of individual neurons within a large population is not independent, but instead is coordinated through a lower-dimensional, latent state that evolves with stereotyped temporal structure (**Fig. 2.1b**). We can represent the state at time t as a vector $\boldsymbol{x}_t \in \mathbb{R}^D$ that evolves according to dynamics captured by a function f such that $\boldsymbol{x}_{t+1} \approx f(\boldsymbol{x}_t)$. Rather than directly observing the latent state $\boldsymbol{x}_t$, we observe neural activity that we represent as $\boldsymbol{y}_t \in \mathbb{R}^N$, where $\boldsymbol{y}_t \approx h(\boldsymbol{x}_t)$ for some function h. Due both to the fact that f imposes a significant amount of structure on the trajectory of the $\boldsymbol{x}_t$'s and the fact that we typically expect the dimension D of $\boldsymbol{x}_t$ to be far smaller than the number of possible observations N, one might expect that it should be possible to estimate the $\boldsymbol{x}_t$'s without observing every neuron at every time step (i.e., measuring only some of the elements of each $\boldsymbol{y}_t$), just as we generally infer latent states from only a fraction of the neurons in a given area. If so, principled exploitation of the space-time trade-off of neural interfaces might achieve higher-fidelity or more bandwidth-efficient characterization of neural population activity.

To our knowledge, no methods have demonstrated inference of dynamics from data in which the set of neurons being monitored changes dynamically at short intervals. To address this challenge, we introduce *selective backpropagation through time* (SBTT; **Fig. 2.1c**), a method to train deep generative models of latent dynamics from data where the identity of observed variables varies from sample to sample. Here we explore applications of SBTT to state space modeling of neural population activity that obeys low-dimensional dynamics.

This paper is organized as follows. Section 2.3 provides an overview of related work. Section 2.4 details SBTT and its integration with sequential autoencoders for modeling neural population dynamics. Section 2.5 demonstrates the effectiveness of this solution in achieving more efficient and higher-fidelity inference of latent dynamics in applications to electrophysiological and calcium imaging data.

2.3 Related work

There is a long and rich literature on methods for system identification, particularly in the case of *linear* dynamical systems. The last several years have witnessed a burst of activity in establishing a more robust theoretical understanding of when and how well these methods work. Particularly relevant to our approach,(Hardt et al., 2018) shows that under suitable conditions on the dynamical system, performing gradient descent on the reconstruction loss of observed data can provably recover the parameters of the system despite the nonconvexity of the problem. Additional guarantees are provided in (Simchowitz et al., 2018; Hazan et al., 2018; Oymak and Ozay, 2019; Tsiamis and Pappas, 2019; Lee and Zhang, 2020) which make varying assumptions on the underlying dynamics and the observation function, the existence of an observable control input, and the stochasticity of the dynamical system. Adversarial noise models are further considered in (Simchowitz et al., 2019, 2020). We emphasize, however, that all of the above works limit their focus to *linear* dynamical systems where the observations are *fully sampled*, i.e., where all of $\boldsymbol{y}_t = \boldsymbol{H}\boldsymbol{x}_t$ is measured for all t.

In the case of a linear observation model ($\boldsymbol{y}_t = \boldsymbol{H}\boldsymbol{x}_t$) but where we observe only a subset of the elements of each $\boldsymbol{y}_t$, the problem is reminiscent of the *low-rank matrix completion* problem (Davenport and Romberg, 2016). Specifically, by letting $\boldsymbol{Y}$ and $\boldsymbol{X}$ denote the matrices whose columns are given by the $\boldsymbol{y}_t$ and $\boldsymbol{x}_t$ respectively, we can write $\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{X}$. If $D \ll N$, this is a low-rank matrix, and hence could be recovered from a random sampling of $O(D \log N)$ elements of each column of $\boldsymbol{Y}$ (Davenport and Romberg, 2016). However, this strategy essentially assumes that there is no relationship between the $\boldsymbol{x}_t$ – one would expect to obtain significant improvements by exploiting the dynamical structure among the $\boldsymbol{x}_t$ imposed by f. Indeed, in (Xu and Davenport, 2016, 2017) the authors show that if the dynamics f are *known*, then it is possible to significantly reduce the sampling requirements. However, the question of *learning* such an f from undersampled observations has again not been addressed in this literature.

In some application domains, there have been hints in this direction. In particular, in the related contexts of recommendation systems (Hidasi et al., 2015; Wu et al., 2017b) and student knowledge tracking (Piech et al., 2015; Xu and Davenport, 2020) there have been successful empirical efforts aimed at learning dynamical systems for modeling how user preferences/knowledge change over time. While such approaches have also had to confront the issue of missing observations (items that are not rated or questions that are not answered), they are aided by the existence of rich sources of additional metadata (e.g., tags) that lead to fundamentally different approaches than what we take here.

Within our application domain, a variety of methods have been developed to infer latent dynamical structure from neural population activity on individual trials, including those based on Gaussian processes (Yu et al., 2009; Duncker et al., 2019; Zhao and Park, 2017; Wu et al., 2017a), linear (Macke et al., 2012; Gao et al., 2016; Kao et al., 2015) and switching linear dynamical systems (Petreska et al., 2011; Linderman et al., 2017; Glaser et al., 2020), and nonlinear dynamical systems such as recurrent neural networks (Pandarinath et al., 2018b; Keshtkaran and Pandarinath, 2019; She and Wu, 2020; Keshtkaran et al., 2021), hidden Markov models (Hernandez et al., 2018), neural ODEs (Kim et al., 2021), and transformers (Ye and Pandarinath, 2021). Variants of these methods accommodate cases where the particular observed neurons change over long time periods (e.g., over the course of days) (Pandarinath et al., 2018b; Nonnenmacher et al., 2017; Kao et al., 2017), but these are not appropriate for cases where neurons are intermittently sampled on short timescales. As described below, several of these methods would be amenable to using SBTT to adapt to intermittent sampling, as SBTT should be applicable to any neural network architecture that learns weights via backpropagation through time.

2.4 Selective backpropagation through time

2.4.1 Overview

SBTT is a learning rule for updating the weights of a neural network that allows backpropagation of loss for the portions of data that are present while preventing missing data from corrupting the gradient signal. The technique optimizes the model to reconstruct observed data while extrapolating to the unobserved data. The implementation of SBTT is related to other approaches that augment network inputs and cost functions to reflect different subsets of the data matrix across samples, in particular coordinated dropout (Keshtkaran and Pandarinath, 2019), masked language modeling (Devlin et al., 2018), and DeepInterpolation (Lecoq et al., 2020). Though not designed for missing data, these previous approaches split fully-observed data into two portions - a portion that is provided at the input to the network, and a portion that is used to compute loss at the output. SBTT uses a similar strategy to accommodate missing data, by zero-filling missing input points and aggregating only losses for observed data points at the output. Though prior work has adopted a similar strategy to handle missing data (Hurwitz et al., 2019), the contribution of SBTT is integrating the strategy with models with a temporal component to learn a dynamical system. To demonstrate SBTT, we provide code for a basic experiment using a sequential autoencoder and Lorenz dataset (`https://github.com/snel-repo/sbtt-demo`).

2.4.2 Illustration with a simple linear dynamical system

We begin by describing our approach in the context of a simple linear dynamical system. In the case where we have no (observable) inputs, we can model a linear dynamical system as

$$\begin{aligned} \boldsymbol{x}_{t+1} &= \boldsymbol{A}\boldsymbol{x}_t + \boldsymbol{w}_t \\ \boldsymbol{y}_t &= \boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{z}_t. \end{aligned}$$

Here, $\boldsymbol{x} \in \mathbb{R}^D$ represents a hidden state, $\boldsymbol{y} \in \mathbb{R}^N$ represents our observations, and $\boldsymbol{w}_t$ and $\boldsymbol{z}_t$ represent noise. The matrix $\boldsymbol{A}$ models the dynamics of the hidden state, and $\boldsymbol{H}$ models the observation function of our system. In this setting, our task is to learn the parameters $\boldsymbol{A}$ and $\boldsymbol{H}$ given the observations $\boldsymbol{y}_0, \ldots, \boldsymbol{y}_{T-1}$ as well as the initial system state $\boldsymbol{x}_0$.

SBTT is a variation of standard back-propagation where loss terms attributed to missing observations are ignored when computing back-propagation updates. Concretely speaking, consider a linear recurrent network that can learn this linear model using a least squares loss

$$\mathcal{L} = \frac{1}{T}\sum_{t=0}^{T-1} \frac{1}{2}\|\boldsymbol{y}_t - \boldsymbol{H}\boldsymbol{x}_t\|_2^2.$$

If the observation vector $\boldsymbol{y}_t$ contains a missing entry at index i, the least squares loss would not contain the $(y_t^i - (Hx_t)^i)^2$ term, where the superscript i represents the ith index of a vector. If $\boldsymbol{o}_t = \boldsymbol{H}\boldsymbol{x}_t$ is taken to be the output of the recurrent network at time step t, then the loss with respect to the outputs of the network is

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{o}_t} = \frac{1}{T}(\boldsymbol{o}_t - \boldsymbol{y}_t). \tag{2.1}$$

SBTT requires that loss terms, and subsequently loss gradients, related to missing observations are ignored. This means that elements in the gradient vector (2.1) are ignored and set to 0 at indices i where the corresponding observations, y_t^i, are missing. This gradient is then back-propagated through time to obtain gradients with respect to model parameters $\boldsymbol{A}$ and $\boldsymbol{H}$ as shown below

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{H}} = \sum_{t=0}^{T-1} \frac{\partial \mathcal{L}}{\partial \boldsymbol{o}_t} (\boldsymbol{x}_t)^\intercal, \frac{\partial \mathcal{L}}{\partial \boldsymbol{A}} = \sum_{t=1}^{T-1} \frac{\partial \mathcal{L}}{\partial \boldsymbol{x}_t} \boldsymbol{x}_{t-1}^\intercal,$$

where $\frac{\partial L}{\partial \boldsymbol{x}_t}$ is recursively computed using back-propagation through time:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{x}_t} = \boldsymbol{A}^\intercal \frac{\partial \mathcal{L}}{\partial \boldsymbol{x}_{t+1}} + \boldsymbol{H}^\intercal \frac{\partial \mathcal{L}}{\partial \boldsymbol{o}_t}.$$

These parameters can then be updated using gradient descent.

2.4.3 Integration with a deep generative model of neural population dynamics

Here we will demonstrate the use of SBTT with a recently developed framework for inferring nonlinear latent dynamics from neural population recordings. This framework, Latent Factor Analysis via Dynamical Systems (LFADS), is a sequential variational auto-encoder (SVAE), detailed in (Pandarinath et al., 2018b). LFADS models single-trial latent dynamics by learning the initial state of the dynamical system, the dynamical rules that govern state evolution, and any time-varying inputs that cannot be explained by the dynamics (i.e., in the case of a non-autonomous dynamical system). Briefly, a bidirectional RNN encoder operates on the neural spiking sequence $\mathbf{y}(t)$ and produces a conditional distribution over initial condition $\mathbf{z}$, $Q(\mathbf{z}|\mathbf{y}(t))$. A Kullback-Leibler (KL) divergence penalty is applied as a regularizer for divergence between the uninformative prior $P(\mathbf{z})$ and $Q(\mathbf{z}|\mathbf{y}(t))$. The initial condition is then drawn from $Q(\mathbf{z}|\mathbf{y}(t))$ and mapped to an initial state for a generator RNN, which learns to approximate the dynamical rules underlying the neural data. A controller RNN takes as input the state of the generator at each time step, along with a time-varying encoding of $\mathbf{y}(t)$ (produced by a second bidirectional RNN encoder), and injects a time-varying input $\mathbf{u}(t)$ into the generator. Similar to $\mathbf{z}$, $\mathbf{u}(t)$ is drawn from a parameterized time-varying distribution of $Q(\mathbf{u}(t)|\mathbf{y}(t))$ produced by the controller. A second KL penalty is applied between $P(\mathbf{u}(t))$ and $Q(\mathbf{u}(t)|\mathbf{y}(t))$. At each time step, the generator state evolves with input from the controller and the controller receives delayed feedback from the generator. The generator states are linearly mapped to factors, which are in turn mapped to the firing rates of the neurons using a linear mapping followed by an exponential nonlinearity. LFADS assumes a Possion emission model for the observed spiking activity. The optimization objective combines the reconstruction cost of the observed spiking activity (i.e., the Poisson likelihood of the observed spiking activity given the rates produced by the generator network), the KL penalties described above, and L2 regularization penalties on the weights of the recurrent networks. During training, network weights are optimized using stochastic gradient descent and backpropagation through time.

The first step in applying SBTT to LFADS is to zero-fill the missing data before feeding it into the initial condition (IC) and controller input (CI) encoders. After passing the data through the remaining hidden layers, we use the resulting rate estimates to compute a reconstruction loss (Poisson negative log-likelihood) for each observed neuron-timepoint and aggregate by taking the mean. The modified reconstruction loss is combined with other losses as in the standard LFADS model. The network only optimizes for reconstruction of observed data and is free to interpolate at unobserved points.

Throughout this paper we use population-based training along with coordinated dropout, together known as AutoLFADS, to optimize our models (Keshtkaran and Pandarinath, 2019; Keshtkaran et al., 2021; Jaderberg et al., 2017). This framework is essential for achieving reliably high-performing LFADS models, regardless of dataset statistics.

2.5 Experiments

2.5.1 High performance with limited bandwidth on primate electrophysiological recordings

A key target application of AutoLFADS with SBTT is to enable reduced sampling of electrodes: either to enable recording from larger populations of electrodes with limited bandwidth (such as with Neuropixels), or to reduce power consumption (such as for fully-implantable brain-machine interfaces). To investigate the performance of AutoLFADS models trained with SBTT, we started with a large and well-characterized dataset containing electrophysiological recordings from macaque primary motor and dorsal premotor cortex (M1/PMd) (Churchland et al., 2010, 2021). The data were collected during a delayed reaching task, in which the monkey made both straight and curved reaches from a center position, around virtual barriers (the maze), to one of 108 possible target positions. The dataset consisted of 2296 trials with 202 sorted units aligned to movement onset in a window from 250 ms before to 450 ms after this point. Spike counts were binned at 10 ms (70 bins). We held out 50 randomly selected units from modeling to use for evaluation of inferred latent factors. We simulated various missing data scenarios for the remaining 152 units by randomly masking a fraction of the observations at each time step for each trial (**Fig. 2.2a**, *top*).

For each of the masked datasets, we used AutoLFADS with SBTT to robustly train neural dynamics models. Latent factors and firing rates were inferred for all time steps, despite the missing (masked) observations. Even with 70% dropped samples, the inferred firing rates showed structure comparable to the model of fully observed data (**Fig. 2.2a**, *bottom*).

To determine whether the models were able to capture biologically relevant information from sparsely sampled data, we evaluated the inferred latent factors in terms of their ability to predict hand velocity (**Fig. 2.2b**) and the spiking activity of held-out units (**Fig. 2.2c**). As a recognizable baseline, we trained a Gaussian Process Factor Analysis (GPFA) model (40 latent dimensions, 20 ms bins) on the fully observed dataset (Yu et al., 2009; Denker et al., 2018). GPFA is a commonly-used and versatile method for extracting latent structure from neural population activity, and these parameters have been validated on this dataset in prior work (Pandarinath et al., 2018b). We trained simple linear decoders to predict hand velocity from the inferred latent factors with an 80 ms delay (50/50, trial-wise train-test split), and evaluated using the coefficient of determination, averaged over x- and y-dimensions. For AutoLFADS with SBTT, decoding performance showed a minimal decline until around 80% of the data had been dropped, with some models outperforming the GPFA baseline using as little as 15% of the original data (**Fig. 2.2b**). To measure how well the models captured the population structure, we trained generalized linear models (GLMs) (Paninski, 2004; Jas et al., 2020) to predict the spikes for the held out units and evaluated fit quality using pseudo-R^2 (pR^2). Similar to the decoding results, we found that AutoLFADS with SBTT captured population structure significantly better than fully observed GPFA, and that the information content of the factors declined slowly until about 80% missing samples (**Fig. 2.2c**).

To evaluate the importance of modeling latent dynamics for accurate inference with sparsely observed data, we also trained NDT with selective backpropagation on the same datasets (Ye and Pandarinath, 2021). We found that decoding performance from inferred firing rates declined faster than for AutoLFADS with SBTT, but NDT still outperformed GPFA with up to 40% missing data (**Fig. 2.2b**).

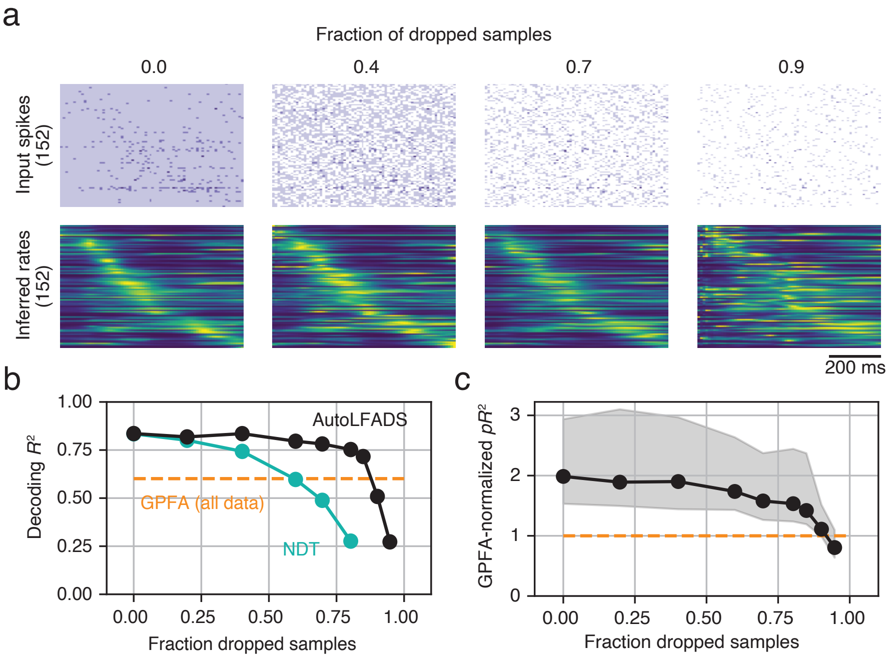


Figure 2.2: SBTT allows inference of latent dynamics from M1/PMd electrophysiology data with sparse observations. (a) Spike count input and inferred rate output of LFADS for the same example trial with increasingly sparse observations. Masked data are shown in white, observed zeroes are shown in light purple, and nonzero spike counts are shown in darker shades. Units are sorted by timing of firing rate peaks for the fully sampled model. (b) Accuracy of linear hand velocity decoding from inferred latent factors. (c) Quality of GLM fits from inferred latent factors to 50 held-out units. pR^2 values for each held-out unit are normalized to the corresponding values achieved by the GPFA baseline. Points denote the median across all units. Shaded areas depict the 25th and 75th quantiles.

2.5.2 Recovery of high frequency features in simulated 2P calcium imaging data

High-frequency features of neural responses are generally assumed to be lost in 2P imaging due to limited scanning speeds and indicator kinetics. We hypothesized that some of the loss is actually due to standard 2P data processing, which discards information regarding sub-frame sampling time of individual neurons, and that SBTT could recover some of this information. The inherently staggered sampling of neurons due to raster scanning can be treated as a time series with missing values and higher temporal resolution than the frame rate. We tested SBTT on both simulated and real calcium imaging data. In both cases, we adapted AutoLFADS to better account for the statistics of deconvolved calcium activity (AutoLFADS-ZIG) by substituting the underlying Poisson emission model with a Zero-Inflated Gamma distribution (Wei et al., 2020a). In our experiments we compared three methods: AutoLFADS-ZIG with SBTT (ALFADS-SBTT), a standard frame-resolution version of AutoLFADS-ZIG without SBTT (ALFADS), and Gaussian smoothing of deconvolved calcium activity.

We generated artificial 2P data from a population of simulated neurons (278 neurons) whose firing rates were linked to the state of an underlying Lorenz system (Zhao and Park, 2017; Sussillo et al., 2016). To assess the ability to reconstruct latent dynamics at different frequencies, we simulated Lorenz systems with different speeds. For each Lorenz system we report the Z dimension power spectrum peak, which contains the most concentrated and highest frequencies. Fluorescence traces were simulated from the spike trains using an order 1 autoregressive model followed by a non-linearity and injected with 4 sources of noise. Firing rates were simulated with a sampling frequency of 100Hz, and a ”location” was randomly chosen for each simulated neuron, such that sampling times for different neurons were staggered to simulate 2p laser scanning sampling times. This produced fluorescence traces with one of three possible associated phases (0,11,22ms) and overall sample rate 33 Hz. We

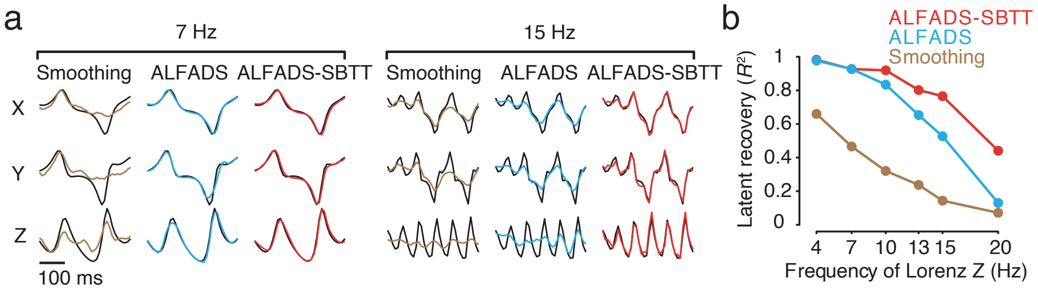


Figure 2.3: SBTT improves inference of high-frequency dynamics from simulated 2P data with known dynamical structure. (a) True and inferred Lorenz latent states (X/Y/Z dimensions) for a single example trial from Lorenz systems simulated at two different frequencies (7Hz and 15Hz). Black: ground truth. Colored: inferred. (c) Performance in estimating the Lorenz Z dimension as a function of Lorenz speed was quantified by variance explained (R^2) for all three methods. The speed of the Lorenz dynamics was quantified based on the peak location of the power spectra of Lorenz Z dimension, with a sampling frequency of 100Hz.

deconvolved neural activity from the fluorescence traces using the OASIS algorithm (Friedrich et al., 2017) as implemented in the CaImAn package (Giovannucci et al., 2019).

For ALFADS-SBTT we used the sub-frame phase information to generate intermittently-sampled data. In contrast, for both ALFADS and Gaussian smoothing, we discarded phase information and collapsed samples into a single time bin per frame, as is standard in 2p imaging data processing. To evaluate the performance in recovering the ground truth Lorenz states, we trained a mapping from the output of each method (i.e., the inferred event rates from ALFADS-SBTT and ALFADS, and smoothed deconvolved events by Gaussian smoothing; signals were interpolated to 100 Hz for the latter methods) to the ground truth Lorenz states using cross-validated ridge regression. We used R^2 between the true and inferred Lorenz states as a metric of performance.

The true and predicted Lorenz states for two example trials are illustrated in **Fig. 2.3a**. The performance of smoothing and ALFADS dropped substantially for higher Lorenz state frequencies, while ALFADS-SBTT maintained reasonable estimates ($R^2 \approx 0.8$) up to 15Hz (**Fig. 2.3a & b**) and never dropped below 0.4 in the range of tested frequencies.

2.5.3 Improved representation of hand kinematics in mouse 2P calcium imaging data

We next applied SBTT to real 2P calcium imaging data we collected from motor cortex in a mouse performing a forelimb water grab task. The dataset comprised 475 trials in which the mouse was cued by a tone to reach to a left or right spout and retrieve a droplet of water with its right forepaw. Pyramidal cells expressing the GCaMP6f calcium indicator were imaged with a two-photon microscope at a 31 Hz frame rate, and a subset of 439 modulated neurons within the field-of-view (FOV) were considered for analysis (FOV shown in **Fig. 2.4a**, *left*; example calcium traces in **Fig. 2.4a**, *right*). The mouse's forepaw position was tracked in 3D at 150 Hz with stereo cameras and DeepLabCut Mathis et al. (2018). Calcium events were deconvolved with OASIS (Friedrich et al., 2017; Giovannucci et al., 2019).

2P data for ALFADS-SBTT were processed analogously to the simulations, using neuron locations within the FOV to inform the intermittent sampling times. Trials represented a window spanning 200 ms before to 800 ms after the mouse's reach onset. This resulted in 100 time points per trial for ALFADS-SBTT, and 31 time points per trial for ALFADS and Gaussian smoothing. For both ALFADS-SBTT and ALFADS, trials were split into 80/20 train/validation.

To compare representations inferred by ALFADS-SBTT and ALFADS, we first evaluated how closely the single-trial event rates inferred for each neuron resembled that neuron's peri-stimulus time histogram (PSTH). PSTHs were calculated by taking the average of the Gaussian-smoothed deconvolved events across trials within each experimental condition. Because the mouse's reaches were not stereotyped to each spout (i.e., left or right), we subgrouped trials into 4 finer conditions based on forepaw Z position during the reach. ALFADS-SBTT single-trial event rates were more strongly correlated with neurons' PSTHs compared to those inferred by ALFADS (**Fig. 2.4b**).

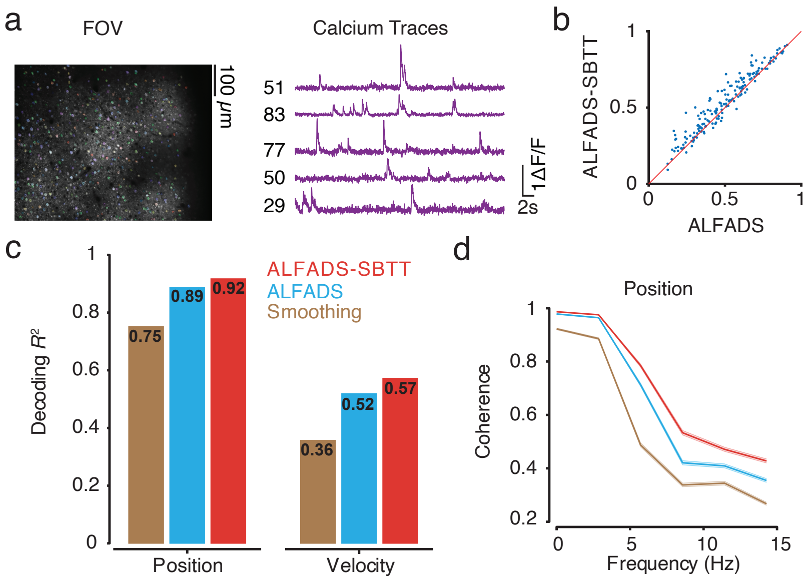


Figure 2.4: SBTT improves inference of latent dynamics from mouse 2P calcium imaging data. (a) Left: an example field-of-view (FOV), colored by neurons. Right: calcium traces (dF/F) from a single trial for 5 example neurons. (b) Performance of capturing empirical PSTHs was quantified by computing the correlation coefficient r between the inferred single-trial event rates and empirical PSTHs, comparing ALFADS vs ALFADS-SBTT. Each point represents an individual neuron. (c) Decoding performance was quantified by computing the R^2 between the true and decoded position (left) and velocity (right) across all trials. (d) Quality of reconstructing the kinematics across frequencies was quantified by measuring coherence between the true and decoded position for all three methods.

We next decoded the mouse's single-trial forepaw kinematics (position and velocity) based on each model's output. Decoding was performed using ridge regression with 5-fold cross validation. We used R^2 between the true and predicted hand positions and velocities as a metric of performance. R^2 was averaged across XYZ behavioral dimensions and all 5 folds of the test sets. Decoding using ALFADS-SBTT inferred rates outperformed results from smoothing deconvolved events, or from the ALFADS inferred rates (**Fig. 2.4c**). Because the improvement of decoding performance for position is modest, we further assessed how the improvement was distributed as a function of temporal frequency. We computed the coherence between the true and decoded positions for each method (**Fig. 2.4d**). Consistent with the simulations, ALFADS-SBTT predictions showed higher coherence with true position than predictions from other methods, with improvements more prominent at higher frequencies (5-15Hz).

2.5.4 Using high-bandwidth observations to improve performance in low-bandwidth conditions

In implantable or wireless applications, using the device's full interface bandwidth might incur significant power costs, which would burden users with frequent battery recharging. However, it may be possible to leverage high-bandwidth recordings from limited time periods to learn models of latent dynamics, and then switch to low-bandwidth modes for subsequent long-term operation, in order to minimize ongoing power use. Such an approach is enabled by the stability of latent dynamics over months to years (Pandarinath et al., 2018b; Kao et al., 2017; Gallego et al., 2020).

We tested these ideas on the same electrophysiological dataset described in section 2.5.1. After training AutoLFADS models on the fully sampled data, we retrained the initial condition and controller input encoders using SBTT on each of the sparsely sampled datasets. The weights for the rest of the network remained fixed. In this way, the dynamical rules learned from the fully sampled data are maintained, while the mappings from data to the initial conditions and controller inputs are adapted for sparse data. Retraining the encoding networks in this way (**Fig. 2.5**, "Retrained sparse") maintained performance to high levels of missing data, outperforming AutoLFADS trained on fully observed data but run with missing data (**Fig. 2.5**, "Trained full") or training directly on sparsely-sampled data (**Fig. 2.5**, "Trained sparse", same as in **Fig. 2.2b**). These results show that dynamics models are learned most accurately on fully observed data, but that the learned dynamics can be used to model sparsely sampled data if models are adapted to the sparser domain using SBTT.

2.6 Discussion

We introduced SBTT, a novel approach for learning latent dynamics from irregularly or sparsely sampled time series data. In experiments on real electrophysiology

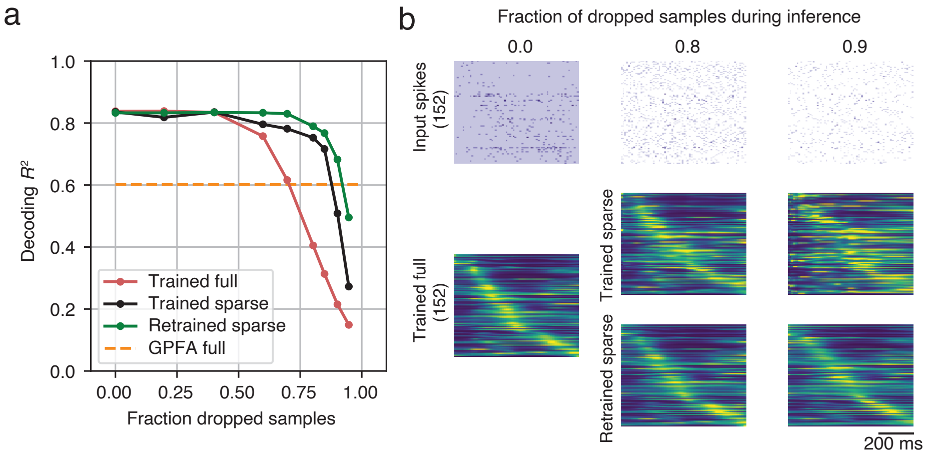


Figure 2.5: Retraining full-data LFADS encoders on sparse data improves decoding performance. (a) Hand velocity decoding performance in function of dropped samples (as in **Fig. 2.2b**). "Trained full" indicates training on fully observed data and inference on sparse data. "Trained sparse" indicates training and inference on sparse data. "Retrained sparse" indicates training on fully observed data, followed by encoder retraining and inference on sparse data. (b) Spike count input and inferred rate output of LFADS. Conventions are as in **Fig. 2.2a**.

data from macaque motor cortex, we show that models trained with SBTT learn biologically relevant neural dynamics with up to 80% masked training data. On data from a synthetic 2P calcium imaging simulation, we show that models trained with SBTT capture high frequency features of the latent dynamics that are not captured at frame resolution. We also showed improved behavioral decoding performance on real 2P imaging data from mouse M1. Finally, we demonstrate that retraining the early layers of a full-data model on sparse datasets using SBTT can substantially improve decoding performance at the most challenging sparsity levels, outperforming models trained on the sparse data alone. Taken together, these results clearly show that SBTT is a valuable technique for training models with irregularly or sparsely sampled time series data.

2.6.1 Limitations

Though we made an effort to characterize performance across multiple potential applications, it remains untested how this approach would generalize to other experimental settings (microscopes, calcium indicators, expression levels), model systems, and brain areas or tasks with more complex or higher-dimensional dynamics (Keshtkaran et al., 2021), but we are optimistic that these properties will extend to AutoLFADS models that use SBTT in other settings. Applications to brain-machine interfaces await incorporation of neural network-based dynamics models into closed-loop, real time systems. We also note that hardware implementations of intermittent sampling for electrophysiology are still largely unexplored, and might incur time or power costs when switching between channels. This might change the point at which intermittent sampling is beneficial from a power or performance perspective. We hope that this work indicates new directions for future generations of recording hardware that focus on high interface capacities and rapid switching between contacts.

2.6.2 Broader impact

Our results could pave the way to substantially decreased power consumption for fully-implantable brain-machine interfaces. Ultimately, this should result in more reliable and less burdensome assistive devices for people with disabilities. Further, expanding the information that can be gathered through a given recording bandwidth has scientific implications, and could enable neuroscientists to ask new questions via larger-scale studies of the brain.

Like any resource-intensive technology, this technique has the potential to increase inequity by only benefiting those who can afford the most advanced neural interfaces. Efforts to deploy such technologies should weigh input from ethicists to ensure that everyone benefits from these scientific innovations (Klein et al., 2015; Goering et al., 2021).

Chapter 3

A deep learning framework for inference of single-trial neural population dynamics from calcium imaging with sub-frame temporal resolution

3.1 abstract

In many brain areas, neural populations act as a coordinated network whose state is tied to behavior on a millisecond timescale. Two-photon (2p) calcium imaging is a powerful tool to probe such network-scale phenomena. However, estimating network state and dynamics from 2p measurements has proven challenging because of noise, inherent nonlinearities, and limitations on temporal resolution. Here we describe RADICaL, a deep learning method to overcome these limitations at the population level. RADICaL extends methods that exploit dynamics in spiking activity for ap-

plication to deconvolved calcium signals, whose statistics and temporal dynamics are quite distinct from electrophysiologically-recorded spikes. It incorporates a novel network training strategy that capitalizes on the timing of 2p sampling to recover network dynamics with high temporal precision. In synthetic tests, RADICaL infers network state more accurately than previous methods, particularly for high-frequency components. In 2p recordings from sensorimotor areas in mice performing a forelimb reach task, RADICaL infers network state with close correspondence to single-trial variations in behavior, and maintains high-quality inference even when neuronal populations are substantially reduced.

3.2 Introduction

In recent years, advances in neural recording technologies have enabled simultaneous monitoring of the activity of large neural populations (Stevenson and Kording, 2011; Steinmetz et al., 2021; Demas et al., 2021). These technologies are enabling new insights into how neural populations implement the computations necessary for motor, sensory, and cognitive processes (Vyas et al., 2020a). However, different recording technologies impose distinct tradeoffs in the types of questions that may be asked (Wei et al., 2020b; Siegle et al., 2021; Chen et al., 2013b). Modern electrophysiology enables access to hundreds to thousands of neurons within and across brain areas with high temporal fidelity (Steinmetz et al., 2021). Yet in any given area, electrophysiology is limited to a sparse sampling of relatively active, unidentified neurons (Siegle et al., 2021) (**Fig. 3.1a**). In contrast, two photon (2p) calcium imaging offers the ability to monitor the activity of vast populations of neurons - rapidly increasing from tens of thousands to millions (Demas et al., 2021; Pachitariu et al., 2017; Peron et al., 2015b) - in 3-D, often with identified layers and cell types of interest (Chen et al., 2013a, 2015). Thus 2p imaging is a powerful tool for understanding how neural circuitry gives rise to function.

A key tradeoff, however, is that the fluorescence transients measured via calcium imaging are a low-passed and nonlinearly-distorted transformation of the underlying spiking activity (**Fig. 3.1b**). Further, because neurons are serially scanned by a laser that traverses the field of view (FOV), a trade-off exists between the size of the FOV (and hence the number of neurons monitored), the sampling frequency, and the pixel size (and therefore the signal-to-noise with which each neuron is sampled). These factors together limit the fidelity with which the activity of large neuronal populations can be monitored and extracted via 2p, and thus limit our ability to link activity measured with 2p imaging to neural computation and behavior on fine timescales. Although a large amount of effort has been dedicated to improving the inference of spike trains from 2p calcium data (Pnevmatikakis, 2019b), recent benchmarks illustrate that a variety of algorithms to infer calcium events all achieve limited correspondence to ground truth spiking activity obtained with electrophysiology, particularly on fine timescales (Berens et al., 2018; Pachitariu et al., 2018).

Rather than focusing on the responses of individual neurons, an alternative approach is to characterize patterns of covariation across a neuronal population to reveal the multi-dimensional internal state of the network as a whole. These “latent variable models”, or simply “latent models”, describe each neuron’s activity as a reflection of the whole network’s state over time. For example, when applied to electrophysiological data, latent models assume that an individual neuron’s spiking is a noisy observation of a latent “firing rate”, which fluctuates in a coordinated way with the firing rates of other neurons in the population. Despite their abstract nature, the trajectory of network state inferred by latent models can reveal key insights into the computations being performed by the brain areas of interest (Vyas et al., 2020a). Inferred network state can also enhance our ability to relate neural activity to behavior. For example, one state-of-the-art deep learning method to estimate network state from electrophysiological spiking data is Latent Factor Analysis via Dynamical Systems (LFADS) (Sussillo et al., 2016; Pandarinath et al., 2018b). In applications to data from motor, sensory, and cognitive regions, LFADS uncovers network state that corresponds closely with single-trial behavior on a 5-10 millisecond timescale (Pandarinath et al., 2018b; Keshtkaran et al., 2021).

Building on the success of latent models for electrophysiological data, here we develop an approach to achieve accurate inference of network state from activity monitored through 2p calcium imaging. We first begin with LFADS and evaluate network state inference using simulated 2p data in which activity reflects known, nonlinear dynamical systems, and with real 2p data from mice performing a water reaching task. LFADS uncovers network state with substantially higher accuracy than standard approaches (e.g., deconvolution plus Gaussian smoothing). We then develop the Recurrent Autoencoder for Discovering Imaged Calcium Latents (RADICaL) to improve inference over LFADS through innovations tailored specifically for 2p data. In particular, we modify the network architecture to better account for the statistics of deconvolved calcium signals, and develop a novel network training strategy that exploits the staggered timing of 2p sampling of neuronal populations to achieve precise, sub-frame temporal resolution. Our new approach substantially improves inference from 2p data, shown in synthetic data through accurate recovery of high-frequency features (up to 20 Hz), and in real data through improved prediction of neuronal activity, as well as prediction of single-trial variability in hand kinematics during rapid reaches (lasting 200-300 ms). Ultimately, RADICaL provides an avenue to tie precise, population-level descriptions of neural computation with the anatomical and circuit details revealed via calcium imaging.

3.3 Results

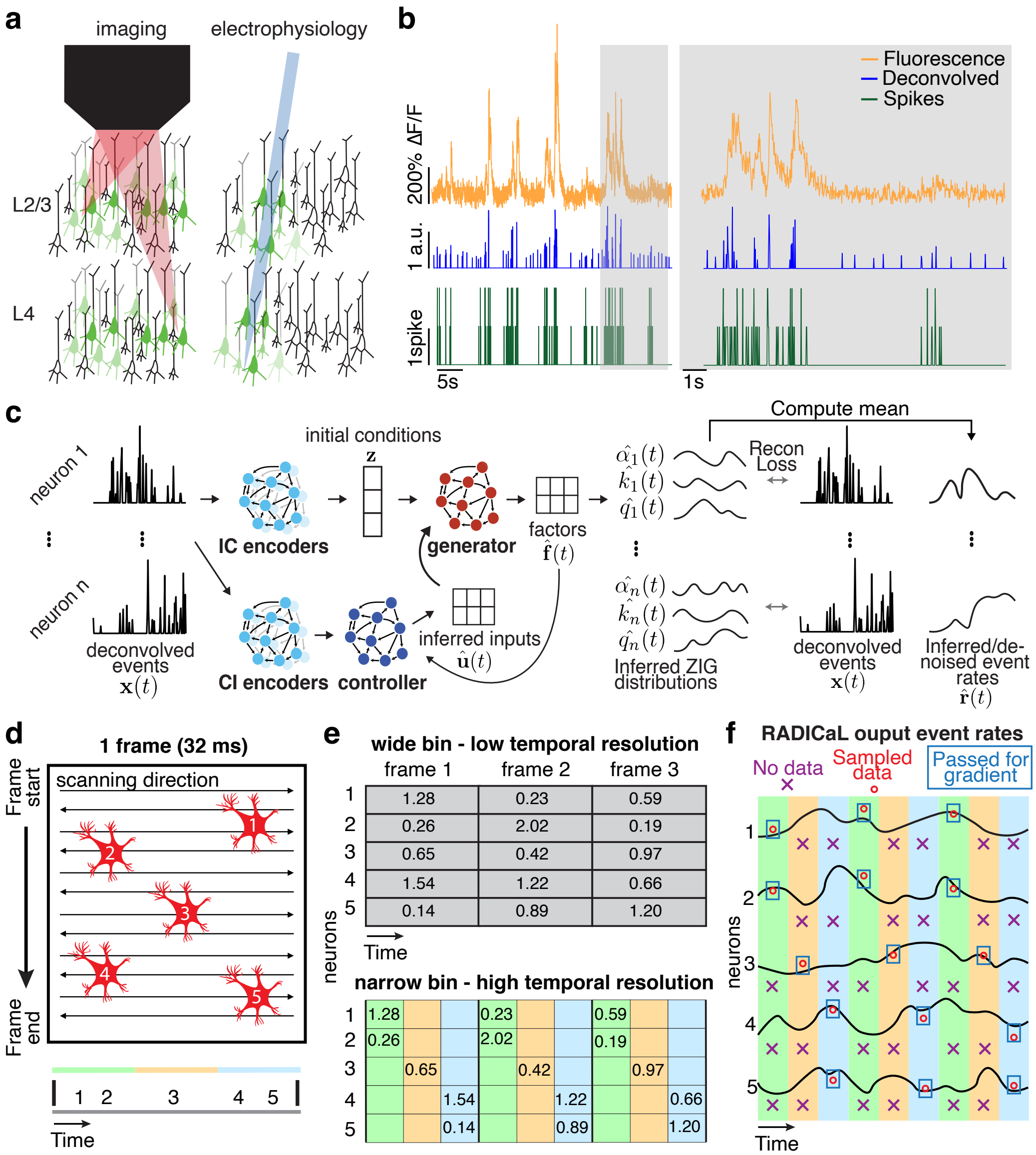
Inferring network state from 2p imaging data using dynamics Dynamical systems models such as LFADS rely on two key principles to infer network state from neural population activity. First, simultaneously recorded neurons exhibit coordinated patterns of activation that reflect the state of the network (Cunningham and Yu, 2014; Pandarinath et al., 2018a). Due to this coordination, network state might be reliably estimated even if the measurement of individual neurons' activity is unreliable. Second, these coordinated patterns evolve over time based on consistent rules (dynamics) (Vyas et al., 2020a; Shenoy et al., 2013). Thus, while it may be challenging to accurately estimate the network's state based on activity at a single time point, knowledge of the network's dynamics provides further information to help constrain network state estimates using data from multiple time points.

To apply these principles to improve inference from 2p data, we extended LFADS to produce RADICaL (**Fig. 3.1c**). Both LFADS and RADICaL model neural population dynamics using recurrent neural networks (RNNs) in a sequential autoencoder configuration (details in Methods, and in previous work (Sussillo et al., 2016; Pandarinath et al., 2018b)). This configuration is built on the assumption that the network state underlying neural population activity can be approximated by an input-driven dynamical system, and that observed activity is a noisy observation of the state of the dynamical system. The dynamical system itself is modeled by an RNN (the 'generator'). The states of the generator are linearly mapped onto a latent space to produce a 'factors' representation, which is then transformed to produce the time-varying output for each neuron (detailed below). The model has a variety of hyperparameters that control training and prevent overfitting, whose optimal settings are not known a priori. To ensure that these hyperparameters were optimized properly for each dataset, we built RADICaL on top of a powerful, large-scale hyperparameter optimization framework we recently developed known as AutoLFADS (Keshtkaran et al., 2021; Keshtkaran and Pandarinath, 2019).

3.3.1 Novel features of RADICaL

RADICaL incorporates two major innovations over LFADS and AutoLFADS. First, we modified RADICaL's observation model to better account for the statistics of deconvolved events. In LFADS, discrete spike count data are modeled as samples from an underlying time-varying Poisson process for each neuron. However, deconvolving 2p calcium signals results in a time series of continuous-valued events, with imperfect correspondence to the actual spike times and counts (Berens et al., 2018). These deconvolved events can be better approximated at each timepoint by a zero-inflated gamma (ZIG) distribution, which combines a gamma distribution to model the calcium event magnitudes and a point mass that represents the elevated probability of zero values (Wei et al., 2020a). In RADICaL, deconvolved events are therefore modeled as samples from a time-varying ZIG distribution whose parameters are taken from the output of the generator RNN (**Fig. 3.1c**; details in Methods). We define the network state at any given time point as a vector containing the inferred (i.e., de-noised) event rates of all neurons, where the de-noised event rate is taken as the mean of each neuron's inferred ZIG distribution at each time point (equation (3.3) in Methods). The de-noised event rates are latent variables that are tied to the underlying network state at each time point. Because of the complicated transformation from generator states to individual neurons' activity, we used the de-noised event rates as the model output for subsequent analyses to compare methods as directly as possible.

Second, we developed a novel neural network training strategy, selective backpropagation through time (Zhu et al., 2021b) (SBTT), that leverages the precise sampling times of individual neurons to enable recovery of high-frequency network dynamics. Since standard 2p microscopes rely on point-by-point raster scanning of a laser beam to acquire frames, it is possible to determine the sample times for each neuron with high precision within the frame (**Fig. 3.1d**). To leverage this information to improve inference of high-frequency network dynamics on single trials, we recast the underlying interpolation problem as a missing data problem: we treat imaging a whole frame as sequentially imaging multiple, smaller bands containing different neurons. In this framing, each neuron is effectively sampled sparsely in time, i.e., the majority of time points for each neuron do not contain valid data (**Fig. 3.1e**). Such sparsely sampled data creates a challenge when training the underlying neural network: briefly, neural networks are trained by adjusting their parameters (weights), and performing this adjustment requires evaluating the gradient of a cost function with respect to weights. SBTT allows us to compute this gradient using only the valid data, and ignore the missing samples (**Fig. 3.1f**; see Methods). Because SBTT only affects how we compute the gradient and update the weights, the network still infers event rates for every neuron at every time point, regardless of whether samples exist at that time point or not. This allows the trained network to accept sparsely-sampled observations as input, and produce high-temporal resolution event rate estimates as its output.



wide bin - low temporal resolution

	frame 1	frame 2	frame 3
1	1.28	0.23	0.59
2	0.26	2.02	0.19
3	0.65	0.42	0.97
4	1.54	1.22	0.66
5	0.14	0.89	1.20

narrow bin - high temporal resolution

1	1.28			0.23			0.59		
2	0.26			2.02			0.19		
3		0.65			0.42			0.97	
4			1.54			1.22			0.66
5			0.14			0.89			1.20

Figure 3.1: Improving inference of network state from 2p imaging. (a) Calcium imaging offers the ability to monitor the activity of many neurons simultaneously, in 3-D, often with cell types of interest and layers identified. In contrast, electrophysiology sparsely samples the neurons in the vicinity of a recording electrode, and may be biased toward neurons with high firing rates. (b) Calcium fluorescence transients are a low-passed and lossy transformation of the underlying spiking activity. Spike inference methods may provide a reasonable estimate of neurons' activity on coarse timescales (left), but yield poor estimates on fine timescales (right; data from (Chen et al., 2013b)). (c) RADICaL uses a recurrent neural network-based generative model to infer network state - i.e., de-noised event rates for the population of neurons - and assumes a time-varying ZIG observation model. For any given trial, the time-varying network state can be captured by three pieces of information: the initial state (i.e., "initial condition") of the dynamical system (trial-specific), the dynamical rules that govern state evolution (shared across trials), and any time-varying external inputs (i.e., "inferred inputs") that may affect the dynamics (trial-specific). (d) Top: in 2p imaging, the laser's serial scanning results in different neurons being sampled at different times within the frame. Bottom: individual neurons' sampling times are known with sub-frame precision (colors) but are typically analyzed with whole-frame precision (gray). (e) Sub-frame binning precisely captures individual neurons' sampling times but results in neuron-time points without data. The numbers in the table indicate the deconvolved event in each frame. (f) SBTT is a novel network training method for sparsely sampled data that prevents unsampled time-neuron data points from affecting the gradient computation.

3.3.2 RADICaL uncovers high-frequency features from simulated data

We first tested RADICaL using simulated 2p data where the underlying network state is known and parameterizable. We hypothesized that the new features of RADICaL would allow it to infer higher-frequency features with greater accuracy than standard approaches, such as Gaussian-smoothing the deconvolved events ("smth-dec"), smoothing the simulated fluorescence traces themselves ("smth-sim-fluor"), or state-of-the-art tools for electrophysiology analysis, such as AutoLFADS. We generated synthetic spike trains by simulating a population of neurons whose firing rates were linked to the state of a Lorenz system (Sussillo et al., 2016; Zhao and Park, 2017) (detailed in Methods and **Fig. A.1a**). We ran the Lorenz system at various speeds, allowing us to investigate the effects of temporal frequency on the quality of network state recovery achieved by different methods. In the 3-dimensional Lorenz system, the Z dimension contains the highest-frequency content (**Fig. A.1b**). Here we denote the frequency of each Lorenz simulation by the peak frequency of the power spectrum of its Z dimension (**Fig. A.1c**).

We used the synthetic spike trains to generate realistic noisy fluorescence signals consistent with GCAMP6f (detailed in Methods and **Fig. A.2**). To recreate the variability in sampling times due to 2p laser scanning, fluorescence traces were simulated at 100 Hz and then sub-sampled at 33.3 Hz, with offsets in each neuron's sampling times consistent with spatial distributions across a simulated FOV. We then deconvolved the generated fluorescence signals to extract events (Friedrich et al., 2017; Giovannucci et al., 2019). Because RADICaL uses SBTT, it could be applied directly to the deconvolved events with offset sampling times. In contrast, for both AutoLFADS and smth-dec, deconvolved events for all neurons were treated as all having the same sampling times (i.e., consistent with the frame times), as is standard in 2p imaging (detailed in Methods).

Despite the distortions introduced by the fluorescence simulation and deconvolution process, RADICaL was able to infer event rates that closely resembled the true underlying rates (**Fig. 3.2a**). To assess whether each method accurately inferred the time-varying state of the Lorenz system, we mapped the representations from the different approaches - i.e., the event rates inferred by RADICaL or AutoLFADS, the smoothed deconvolved events, and the smoothed simulated fluorescence traces - onto the true underlying Lorenz states using cross-validated ridge regression. We then quantified performance using the coefficient of determination (R^2), which quantifies the fraction of the variance of the true latent variables captured by the estimates. **Fig. 3.2b** shows the Lorenz Z dimension for example trials from three Lorenz speeds, as well as the recovered values for three of the methods. RADICaL inferred latent states with high fidelity (R^2 >0.8) up to 15 Hz, and significantly outperformed other methods across a range of frequencies (**Fig. 3.2c**; performance for the X and Y dimensions is shown in **Fig. B.1**; p<0.05 for all frequencies and dimensions, paired, one-sided t-Test, detailed in Methods). Notably, performance in estimating latent states was improved due to both of the innovations in RADICaL, with SBTT contributing more (**Fig. B.2**). To test RADICaL's ability in estimating single-trial dynamics for a task that lacks a repetitive trial-structure, we varied the simulation so that each trial had a unique initial condition for the Lorenz system. RADICaL accurately inferred the latent states on single trials (**Fig. A.3a**) and outperformed AutoLFADS and smth-dec at high Lorenz oscillation frequencies (**Fig. A.3b**).

To better understand the regimes in which RADICaL recovers the underlying latent variables well or poorly, we performed variants of the simulation experiments along 4 additional axes: imaging speed (**Fig. A.4**), high frequency structure in the latent variables (**Fig. B.3**), noise levels (**Fig. B.4**), and whether RADICaL could be effective when used with algorithms that infer spike times instead of event rates, such as MLspike (Deneux et al., 2016) (**Fig. B.5**). In all cases we found that

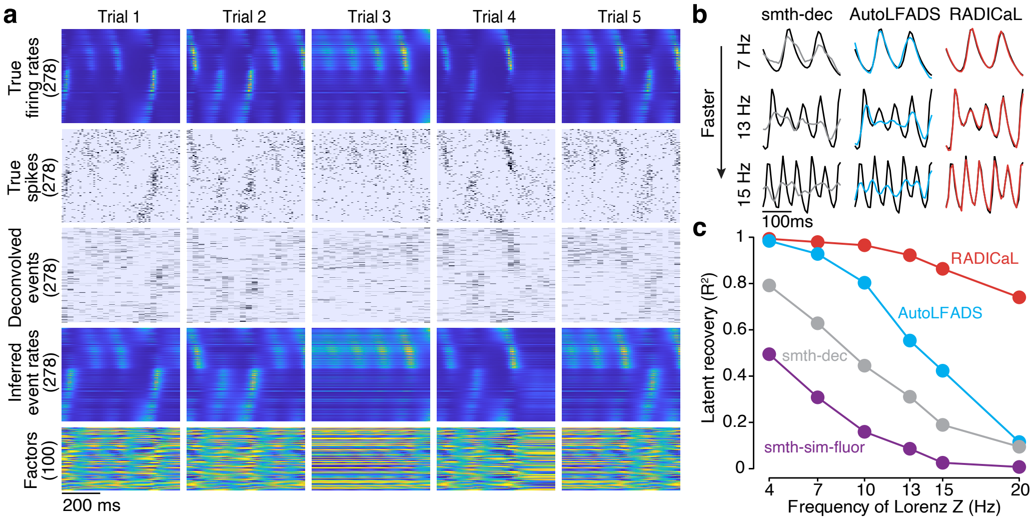


Figure 3.2: Application of RADICaL to synthetic data. (a) Example firing rates and spiking activity from a Lorenz system simulated at 7 Hz, deconvolved calcium events (inputs to RADICaL), and the corresponding rates and factors inferred by RADICaL. Simulation parameters were tuned so that the performance in inferring spikes using OASIS matched previous benchmarks (Berens et al., 2018) (see Methods). (b) True and inferred Lorenz latent states (Z dimension) for a single example trial from Lorenz systems simulated at three different Lorenz oscillation frequencies. Black: true. Colored: inferred. (c) Performance in estimating the Lorenz Z dimension as a function of simulation frequency was quantified by variance explained (R^2) for all 4 methods.

RADICaL substantially outperformed alternate approaches. However, as expected, our analysis showed that deconvolution itself performs poorly at very slow sampling rates (e.g., 2Hz and below), and for very high frequency content (e.g., >20 Hz), and thus RADICaL's performance in those regimes is limited by the use of deconvolution as a preprocessing step.

These simulations demonstrate RADICaL's performance in various circumstances, but the parameter space of possible experiments is very large (calcium indicators, expression patterns, imaging settings, etc.) and an exhaustive search of this parameter space is infeasible. Thus, we next benchmarked performance on real data to demonstrate RADICaL's utility in the real world.

3.3.3 RADICaL improves inference in a mouse “water grab” task

We next tested RADICaL on 2p recordings from mice performing a forelimb water grab task (**Fig. 3.3a**, top). We analyzed data from four experiments: two mice with two sessions from each mouse, in which different brain areas were imaged (M1, S1). Our task was a variant of the water-reaching task of (Galiñanes et al., 2018). In each trial, the mouse was cued by the pitch of an auditory tone to reach to a left or right spout and retrieve a droplet of water with its right forepaw (**Fig. 3.3a**, bottom; see Methods). The forepaw position was tracked at 150 frames per second with DeepLabCut (Mathis et al., 2018) for 420-560 trials per experiment. To test whether each method could reveal structure in the neural activity at finer resolution than left vs. right reaches, we divided trials from each condition into subgroups based on forepaw height during the reach (**Fig. 3.3a**, top right; see Methods). Two-photon calcium imaging from GCaMP6f transgenic mice was performed at 31 Hz, with 430-543 neurons within the FOV in each experiment (**Fig. 3.3b**).

With real datasets, a key challenge when benchmarking latent variable inference is the lack of ground truth data for comparison. A useful first-order assessment is whether the event rates inferred for individual trials match the empirical peri-stimulus time histograms (PSTHs), i.e., the rates computed by averaging noisy single-trial data across trials with similar behavioral characteristics (Pandarinath et al., 2018b; Keshtkaran et al., 2021). While this approach obscures meaningful across-trial variability, it provides a ‘de-noised’ estimate that is useful for coarse performance quantification and comparisons. To compute empirical PSTHs, we averaged the smoothed deconvolved events (smth-dec rates) across trials within each subgroup.

We found that RADICaL-inferred event rates recapitulated features of individual neurons’ activity that were apparent in the empirical PSTHs, both when averaging across trials, but also on individual trials (**Fig. 3.3c**). Importantly, RADICaL is an unsupervised method, meaning that it was not provided any behavioral information, such as whether the mouse reached to the left or right on a given trial, or which subgroup a trial fell into. Yet the single-trial event rates inferred by RADICaL showed clear separation not only between left and right reach conditions, but also between subgroups of trials within each condition. This separation was not clear with the single-trial smth-dec rates. We quantified the correspondence between the single-trial inferred event rates and the empirical PSTHs via Pearson's correlation coefficient (r; see Methods). RADICaL single-trial event rates showed substantially higher correlation with the empirical PSTHs than smth-dec rates (**Fig. 3.3d**) or those inferred by AutoLFADS (**Fig. A.5**). Importantly, these improvements were not limited to a handful of neurons, but instead were broadly distributed across the population. Within the trials modeled by RADICaL, we found there was a subset of right reaches from Mouse1/S1 that were "loopy" and atypical, showing multiple large peaks in hand speed (**Fig. 3.3e**, top). The RADICaL single-trial event rates exhibited distinct patterns of neural responses for these atypical trials (**Fig. 3.3e**, bottom), demonstrating RADICaL's ability to automatically capture idiosyncrasies of single-trial activity that are common in experiments that constrain behavior less tightly.

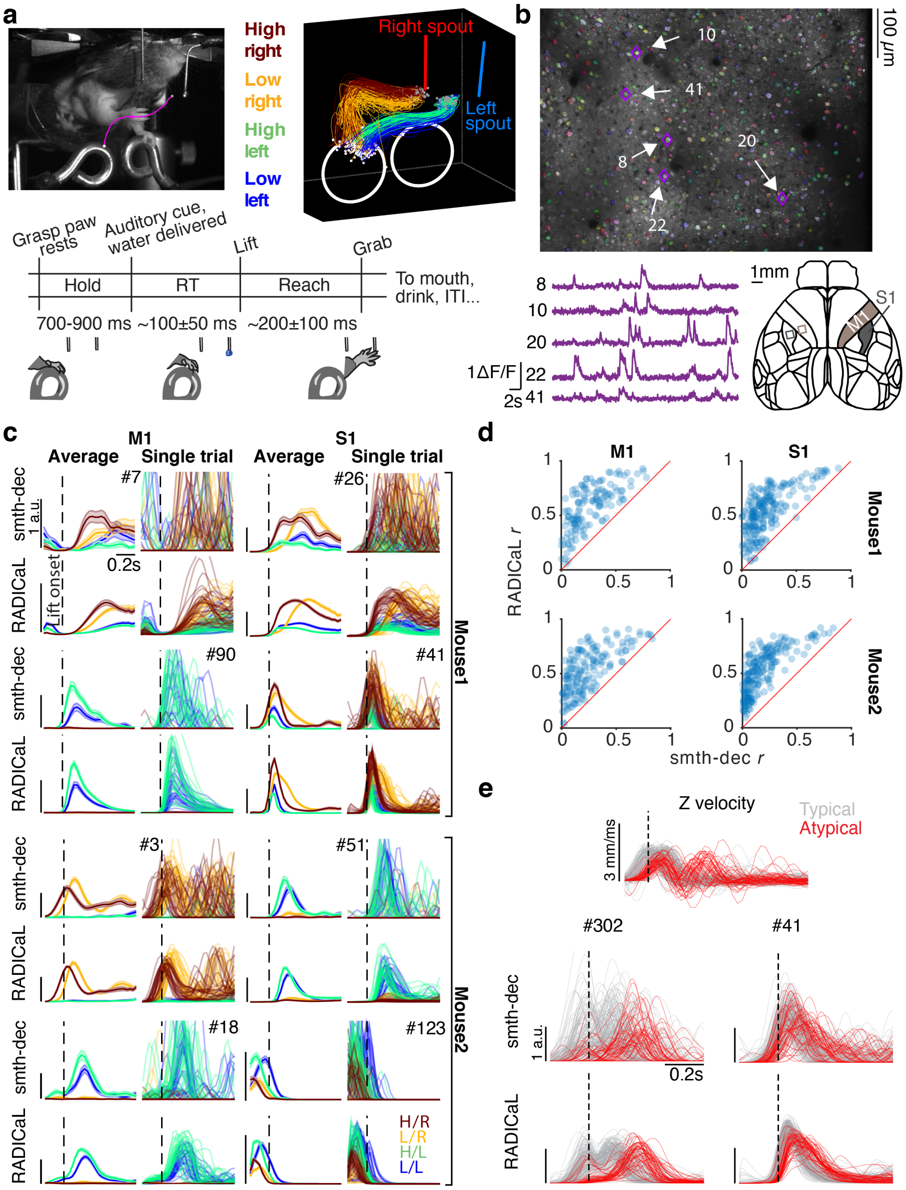


Figure 3.3: Application of RADICaL to real two-photon calcium imaging of a water grab task. (a) Task. Top left: Mouse performing the water grab task. Pink trace shows paw centroid trajectory. Bottom: Event sequence/task timing. RT: reaction time. ITI: inter-trial interval. Top right: Individual reaches colored by subgroup identity. (b) Top: an example field of view (FOV), identified neurons colored randomly. Bottom left: dF/F from a single trial for 5 example neurons. Bottom right: Allen Atlas M1/S1 brain regions imaged. (c) Comparison of trial-averaged (left) and single-trial (right) rates for 8 individual neurons for two different brain areas (left vs. right) and two different mice (top half vs. bottom half) for smth-dec and RADICaL (alternating rows). Left: each trace represents a different reach subgroup (4 in total) with error bars indicating s.e.m. Right: each trace represents an individual trial (same color scheme as trial-averaged panels). Odd rows: smth-dec event rates (Gaussian kernel: 40 ms s.d.). Even rows: RADICaL-inferred event rates. Horizontal scale bar represents 200ms. Vertical scale bar denotes event rate (a.u.). Vertical dashed line denotes lift onset time. (d) Performance of RADICaL and smth-dec in capturing the empirical PSTHs on single trials. Correlation coefficient r was computed between the inferred single-trial event rates and empirical PSTHs. Each point represents an individual neuron. (e) Kinematic profiles and neural representations of atypical trials. Top: Z-dimension of hand velocity profile. Each trace represents an individual trial, colored by typical vs. atypical. Atypical trials are identified as the trials that have a second peak in Z-dimension of the hand velocity that is larger than 50% of the first peak. Middle and Bottom: Comparison of single-trial rates for 2 example neurons (data from Mouse1/S1) for smth-dec (middle row) and RADICaL (bottom row). Each trace represents an individual trial (same color scheme as top row). Horizontal scale bar represents 200ms. Vertical scale bar denotes event rate (a.u.). Vertical dashed line denotes lift onset time.

We next tested whether the population activity inferred by RADICaL also showed meaningful structure on individual trials. We used principal component analysis (PCA) to produce low-dimensional visualizations of the population's activity (detailed in Methods). The low-D trajectories computed from the RADICaL-inferred rates showed consistent, clear single-trial structure that corresponded to behavioral conditions and subgroups for all four experiments (**Fig. 3.4a**, top row; **Fig. A.6**, top row), despite RADICaL receiving no direct information about which trials belonged to which subgroup, or even the kinematics used to define the subgroups. In comparison, low-D trajectories computed from the smth-dec rates showed noisy single-trial structure with little correspondence to behavioral subgroups (**Fig. 3.4a**, bottom row; **Fig. A.6**, bottom row). To provide a quantitative summary, we measured the distance of the low-D trajectories between each trial and other trials across subgroups (d_{across}) vs. within the same subgroup (d_{within}) for any given time and computed the distance ratio (detailed in Methods). The distance ratio (i.e., d_{across}/d_{within}) of RADICaL-derived trajectories was higher than smth-dec-derived trajectories across time points, which was also consistent across four experiments (**Fig. 3.4b**).

3.3.4 RADICaL captures dynamics that improve behavioral prediction

We next tested whether the RADICaL-inferred event rates were closely linked to behavior by decoding forepaw positions and velocities from the inferred event rates using cross-validated ridge regression (**Fig. 3.5a**; **Fig. A.7**). Decoding using RADICaL-inferred rates significantly outperformed results from smth-dec rates, or from the AutoLFADS-inferred rates (**Fig. 3.5b**; position: average R^2 of 0.91 across all experiments, versus 0.75 and 0.85 for smth-dec and AutoLFADS, respectively; velocity: average R^2 of 0.62 across the mice/areas, versus 0.37 and 0.51 for smth-dec and AutoLFADS, respectively; $p<0.05$ for position and velocity for all individual exper-

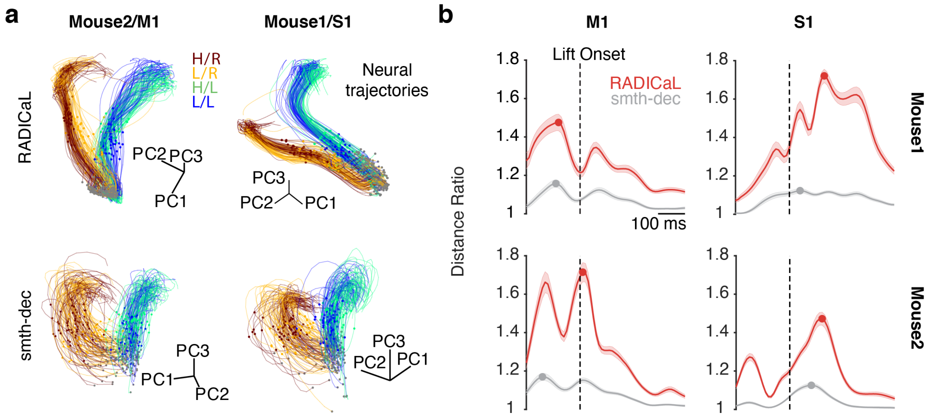


Figure 3.4: RADICaL produces neural trajectories reflecting trial subgroup identity in an unsupervised manner. (a) Single-trial neural trajectories derived from RADICaL rates (top row) and smth-dec rates (bottom row) for two experiments (left: Mouse2/M1; right: Mouse1/S1), colored by subgroups. Each trajectory is an individual trial, plotting from 200 ms before to 400 ms after lift onset. Lift onset times are indicated by the dots in the same colors with the trajectories. Grey dots indicate 200 ms prior to lift onset time. Neural trajectories from additional experiments are shown in **Fig. A.6**. (b) Performance of RADICaL and smth-dec in revealing distinct subgroups in single-trial neural trajectories. The ratio of the cross-group distance to the within-group distance was computed for each individual time point in a window from 200 ms before to 400 ms after lift onset. Horizontal scale bar represents 100ms. Vertical dashed line denotes lift onset time. Error bar indicates the s.e.m. across individual trials. Dots indicate the maximum ratio for each method.

iments, paired, one-sided t-test, detailed in Methods). Improvements achieved by RADICaL were shown on most trials (**Fig. B.6**). Importantly, the performance advantage was not achieved by simply predicting the mean event rates for all trials of a given condition: RADICaL also outperformed AutoLFADS and smth-dec in decoding the kinematic residuals (i.e., the single-trial deviations from the mean; **Fig. B.7**). To assess how decoding improvements were distributed as a function of frequency, we computed the coherence between the true and decoded positions and velocities for each method (**Fig. 3.5c**). RADICaL predictions showed higher coherence with behavior than predictions from smth-dec or AutoLFADS across a wide range of frequencies, and the difference in coherence between RADICaL and AutoLFADS widened (especially for position) at higher frequencies (5-15 Hz). This argues that RADICaL improved decoding particularly because it improved recovery of higher-frequency features of the neural activity. Notably, decoding was improved due to both innovations in RADICaL (i.e., modeling events with a ZIG distribution, and SBTT), and the combination of the two innovations significantly improved performance over each innovation alone (**Fig. B.8**).

We next tested whether RADICaL could capture meaningful trial-to-trial variability by predicting reaction time (RT) from the inferred event rates using cross-validated logistic regression (Kaufman et al., 2016) (detailed in Methods). The RT in a trial is defined as the time between water presentation and movement onset. RTs predicted

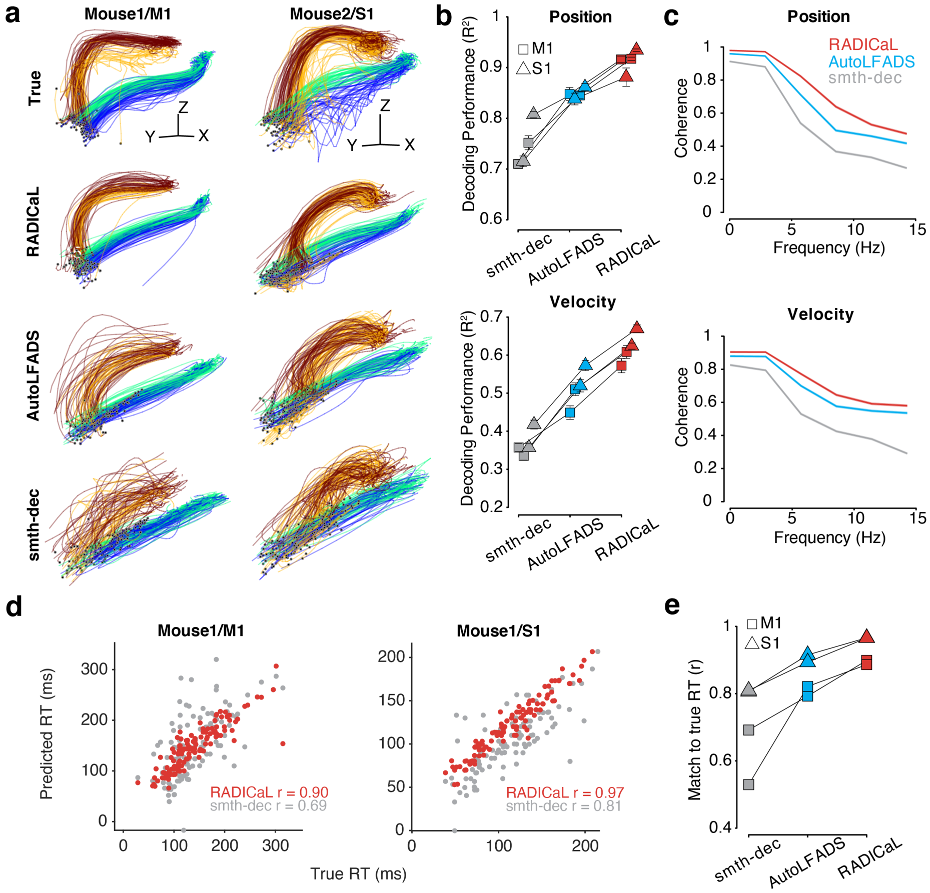


Figure 3.5: RADICaL improves prediction of behavior. (a) Decoding hand kinematics using ridge regression. Each column shows an example mouse/area. Row 1: true hand position trajectories, colored by subgroups. Rows 2–4: predicted hand positions using ridge regression applied to the event rates inferred by RADICaL or AutoLFADS, or smth-dec rates (Gaussian kernel: 40 ms s.d.). Hand positions from additional experiments are shown in **Fig. A.7**. (b) Decoding accuracy was quantified by measuring variance explained (R^2) between the true and decoded position (top) and velocity (bottom) across all trials across each of the 4 datasets (2 mice for M1, denoted by squares, and 2 mice for S1, denoted by triangles), for all 3 techniques. Error bar indicates the s.e.m. across 5 folds of test trials. (c) Quality of reconstructing the kinematics across frequencies was quantified by measuring coherence between the true and decoded position (top) and velocity (bottom) for individual trials across all 4 datasets, for all 3 techniques. (d) Predicting single-trial reaction times using RADICaL or smth-dec rates. Each dot represents an individual trial, color-coded by event rate inference method. Correlation coefficient r was computed between the true and predicted reaction times. Prediction of single-trial reaction times from additional experiments are shown in **Fig. A.8**. (e) Performance of predicting single-trial reaction times across each of the 4 datasets (2 mice for M1, denoted by squares, and 2 mice for S1, denoted by triangles), for all 3 techniques.

from RADICaL-inferred rates showed high correlation with the true RTs (**Fig. 3.5d**), and outperformed results from smth-dec rates, or from the AutoLFADS-inferred rates (**Fig. 3.5e**; **Fig. A.8**; average r of 0.93 across all experiments, versus 0.71 and 0.86 for smth-dec and AutoLFADS, respectively).

3.3.5 RADICaL retains high performance with reduced neuron counts

To evaluate RADICaL's performance as a function of population size, we gradually reduced the number of neurons used in training RADICaL or AutoLFADS, either in a random fashion (**Fig. 3.6**), or in a FOV-shrinking fashion (**Fig. A.9**). In both cases, RADICaL retained relatively high decoding performance as the population size was reduced. Decoding performance declined gradually, with a steeper slope for velocity. Notably, however, performance when only 25% of the neurons were used for training RADICaL was similar to that of AutoLFADS - and higher than for smth-dec - when those methods were applied to the full population of neurons. These results provide an avenue to retain information when scanning sparser populations (such as when a cell type of interest is in the minority), smaller areas when imaging deep structures with a limited FOV due to a relay (GRIN) lens, or using smaller FOVs to capture multiple layers or regions while retaining overall frame rate (see Discussion).

3.4 Discussion

2p imaging is a widely-used method for interrogating neural circuits, with the potential to monitor vast volumes of neurons and provide new circuit insights that elude electrophysiology. To date, however, it has proven challenging to precisely infer network state from imaging data, due in large part to the inherent noise, indicator dynamics, and low temporal resolution associated with 2p imaging. RADICaL

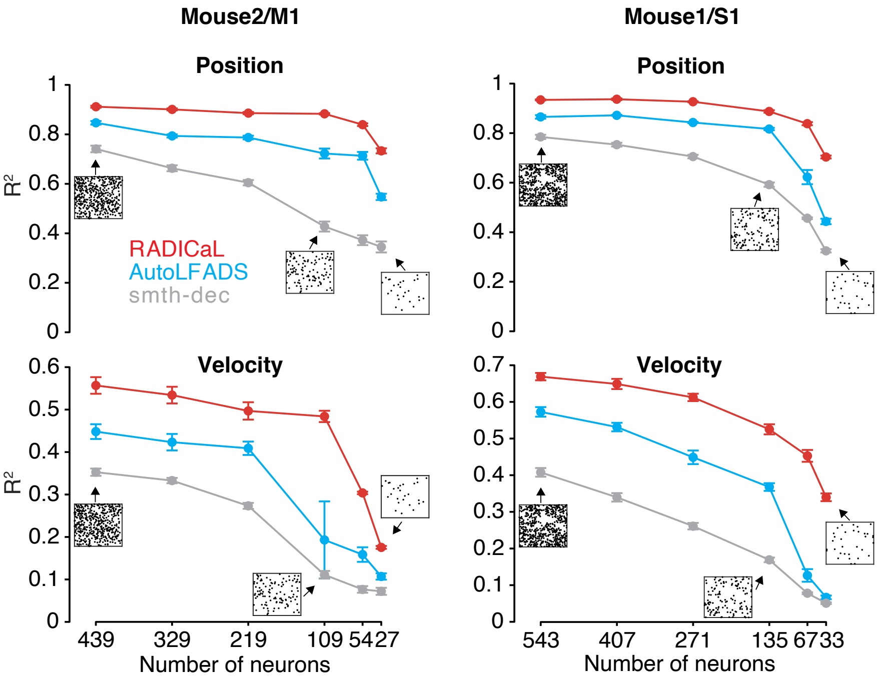


Figure 3.6: RADICaL retains high decoding performance in a neuron downsampling experiment. Decoding performance was measured as a function of the number of neurons used in each technique (top: Position; bottom: Velocity). Data are from Mouse2/M1 (left) and Mouse1/S1 (right). Performance was quantified using variance explained (R^2). Figure insets indicate the selected neurons in the FOV for the full population of neurons and examples for different subsets. Error bar indicates the s.e.m. across 5 folds of test trials. Each black dot in the insets represents a neuron. Analyses were robust to the seed used for selecting different random subsets of neurons (**Fig. B.9**).

bridges this gap. RADICaL is tailored specifically for 2p imaging, with a noise emissions model that is appropriate for deconvolved calcium events, and a novel network training strategy (SBTT) that takes advantage of the specifics of 2p laser scanning to achieve substantially higher temporal resolution. Through synthetic tests, we demonstrated that RADICaL accurately infers network state and substantially outperforms alternate approaches in uncovering high-frequency fluctuations. Then, through careful validation on real 2p data, we demonstrated that RADICaL infers network state trajectories that are closely linked to single-trial behavioral variability, even on fast timescales. Finally, we demonstrated that RADICaL maintains high-quality inference of network state even as the neural population size is reduced substantially.

The ability to de-noise neural activity on single trials is highly valuable. First, de-noising improves the ability to decode behavioral information from neural activity, allowing subtle relationships between neural activity and behavior to be revealed (**Fig. 3.5**). Second, de-noising on single trials reduces the dependence on the stereotyped behaviors needed for de-noising through trial-averaging, which could allow greater insight in experiments with animals such as mouse and marmoset, where powerful experimental tools are available but highly repeatable behaviors are challenging to achieve. A move away from trial-averaging could also enable better interpretability of more complex or naturalistic behaviors (Keshtkaran et al., 2021; Hatsopoulos et al., 2007; Krakauer et al., 2017; Whishaw et al., 2017; Wiltschko et al., 2020). Third, this de-noising capability will enable greater insight into processes that fundamentally differ from trial to trial, such as learning from errors (Herzfeld et al., 2018; Vyas et al., 2020b), variation in internal states such as arousal (Steinmetz et al., 2019; Stringer et al., 2019b), or paradigms in which tuning to uninstructed movements contaminates measurement of the task-related behavioral variables of interest (Musall et al., 2019). Finally, this de-noising greatly improves inference of network state (**Fig. 3.2**), mitigating some of the known distortions of neural activity introduced by calcium imaging (Wei et al., 2020b). Importantly, electrophysiology and calcium imaging have distinct advantages and disadvantages, and both provide biased information about the underlying neural population (Siegle et al., 2021). Whereas LFADS has served as a powerful tool for denoising electrophysiology data and accurately inferring network state, no similar method existed for the complementary technique of calcium imaging; RADICaL fills this gap.

In recent years, a variety of computational methods have been developed to analyze 2p imaging data (Pnevmatikakis, 2019b). 2p preprocessing pipelines (Pachitariu et al., 2017; Giovannucci et al., 2019) normally include methods that correct for brain motion, localize and demix neurons' fluorescence signals, and infer event rates from fluorescence traces. Several studies have applied deep learning in attempts to improve spike inference (Hoang et al., 2020; Rupprecht et al., 2021; Sebastian et al., 2021), while a few others have focused on uncovering population-level structure (Dechery and MacLean, 2018; Kirschbaum et al., 2019; Mackevicius et al., 2019; Triplett et al., 2020; Williams et al., 2018; Wu et al., 2018) or locally linear dynamics underlying population activity, in particular via switching linear dynamical systems-based methods (Costa et al., 2019; Glaser et al., 2020). Here we built RADICaL on the AutoLFADS architecture, which leverages deep learning and large-scale distributed training. This enables the integration of more accurate observation models (ZIG) and powerful optimization strategies (SBTT), while potentially inheriting the high performance and generalized applicability previously demonstrated for AutoLFADS (Keshtkaran et al., 2021).

Many behaviors are performed on fast timescales (e.g., saccades, reaches, movement correction, etc.), and thus previous work has made steps in overcoming the limits of modest 2p frame rates in attempts to infer the fast changes in neural firing rates that relate to these fast behaviors. Efforts to chip away at this barrier have relied on regularities imposed by repeated stimuli or highly stereotyped behavior (Picardo et al., 2016; Mano et al., 2019), or jittered inferred events on sub-frame timescales to minimize the reconstruction error of the associated fluorescence (Hoang et al., 2020). RADICaL takes a different approach. In particular, it links sub-frame timing to neural population dynamics, representing a more powerful and generalizable approach that does not require stereotypy in the behavior or neural response and which could therefore be applied to datasets with more naturalistic or flexible behaviors. Broadly speaking, this approach provides a solution to the spatiotemporal tradeoff that is inherent to any scanning technique, enabling retention of temporal resolution while increasing the spatial area of sampling.

As shown in our simulated experiments, deconvolution places an upper bound on RADICaL's performance, limiting its potential in slow sampling regimes (i.e., 2 Hz) with fast indicators or in more challenging inference cases (e.g., higher-frequency latent content, higher noise levels, etc). To mitigate these limitations, future work could build an end-to-end model that integrates the generative rates-to-fluorescence process and operates on the fluorescence traces directly. Complementary work has begun exploring in this direction (Prince et al., 2021), but our unique innovation of SBTT presents an opportunity to greatly improve the quality of recovering high-frequency features when the sampling rate is limited. More broadly, as benchmarking efforts are an invaluable resource for systematically comparing methods and building on advances from various different developers (Pei et al., 2021), carefully-designed benchmarking efforts for network state inference from 2p data could accelerate progress in this field.

The ability to achieve high-quality network state inference despite limited neuronal population size opens the door to testing new choices about how to perform the experiments themselves. For example, it could enable understanding the role of an uncommon neuronal subtype, or the single-trial outputs of an area by imaging projection neurons that are sparsely distributed throughout that area. With subcortical structures that require relay lenses, it could extract more information from a smaller FOV, permitting the use of a smaller relay lens that causes less damage to overlying brain structures. Or, when hopping between different layers (Chen et al., 2013a, 2015) or brain areas (Minderer et al., 2019; Sofroniew et al., 2016), fewer lines could be imaged per FOV to retain a higher overall frame rate while achieving good inference from each FOV. When the number of neurons within each FOV is limited, one further advantage that RADICaL inherits from LFADS is that it allows for multi-session stitching16, which could provide an avenue to combine data from different sessions to improve inference of the underlying dynamics for each FOV.

In sum, RADICaL provides a framework to push back the limits of the space-time tradeoff in 2p calcium imaging, enabling accurate inference of population dynamics in vast populations and with identified neurons. Future work will explore how best to exploit these capabilities for different experimental paradigms, and to link the power of dynamics with the anatomical detail revealed with calcium imaging.

3.5 Methods

3.5.1 AutoLFADS and RADICaL architecture and training

The core model that AutoLFADS and RADICaL build on is LFADS. A detailed overview of the LFADS model is given in (Sussillo et al., 2016; Pandarinath et al., 2018b). Briefly, LFADS is a sequential application of a variational auto-encoder (VAE). A pair of bidirectional RNNs (the initial condition and controller input encoders) operate on the spike sequence and produce initial conditions for the generator RNN and time-varying inputs for the controller RNN. All RNNs were implemented using gated recurrent unit (GRU) cells. At each time step, the generator state evolves with input from the controller and the controller receives delayed feedback from the generator. The generator states are linearly mapped to factors, which are mapped to the firing rate of the neurons using a linear mapping followed by an exponential nonlinearity. The optimization objective is to maximize a lower bound on the likelihood of the observed spiking activity given the rates produced by the generator network, and includes KL and L2 regularization penalties. During training, network weights are optimized using stochastic gradient descent and backpropagation through time.

Identical network sizes were used for both AutoLFADS and RADICaL runs and for both simulation and real 2p data. The dimension of initial condition encoder, controller input encoder, and controller RNNs was 64. The dimension of the generator RNN was 100. The generator was provided with 64-dimensional initial conditions and 2-dimensional controller outputs (i.e., inferred inputs u(t)) and linearly mapped to 100-dimensional factors. The initial condition prior distribution was Gaussian with a trainable mean that was initialized to 0 and a variance that was fixed to

0.1. The minimum allowable variance of the initial condition posterior distribution was set to 1e-4. The controller output prior was autoregressive with a trainable autocorrelation tau and noise variance, initialized to 10 and 0.1, respectively. The Adam optimizer (epsilon: 1e-8; beta1: 0.9; beta2: 0.99; initial learning rate: 1e-3) was used to control weight updates. The loss was scaled by a factor of 1e4 prior to computing the gradients for numerical stability. To prevent potential pathological training, the GRU cell hidden states were clipped at 5 and the global gradient norm was clipped at 300.

AutoLFADS is a recent implementation of the population based training (PBT) approach (Jaderberg et al., 2017) on LFADS to perform automatic, large-scale hyperparameter (HP) search. A detailed overview of AutoLFADS is in (Keshtkaran et al., 2021; Keshtkaran and Pandarinath, 2019). Briefly, PBT distributes training across dozens of models in parallel, and uses evolutionary algorithms to tune HPs over many generations. To do so, trials were first split into training and validation sets. At the beginning of training, the value of the searchable HPs was randomly drawn from an initial range for each individual model. At the end of each generation, a selection process was performed to choose models with higher performance (i.e., lower negative log likelihood, or NLL) on the validation set and replace the poor models with the higher performing models. The HPs of the higher performing models were perturbed before the next generation to increase the HP search space.

Training and hyperparameter search varies in the number of generations needed to converge (typically 70 - 150 generations), depending on the data and hardware used (number and type of GPUs). With our data and hardware (10x NVIDIA GeForce RTX 2080 Ti GPUs), a run of RADICaL typically converges in 3 - 5 hours. RADICaL was built in Python 2 and TensorFlow 1.14, and cloud implementations of RADICaL on Google Cloud Platform and NeuroCAAS are also being made available.

For the PBT approach, 20 single models were trained in parallel for both AutoL-

FADS and RADICaL runs and for both simulation and real 2p data. Generations consisted of 50 epochs, and KL and L2 regularization penalties were linearly ramped for the first 80 epochs of training during the first generation. Training was stopped when there was no improvement in performance after 25 generations. The HPs optimized by PBT were the model's learning rate and six regularization HPs: scaling weights for the L2 penalties on the generator, controller, and initial condition encoder RNNs, scaling weights for the KL penalties on the initial conditions and controller outputs, and two dropout probabilities ("keep ratio" for coordinated dropout (Keshtkaran and Pandarinath, 2019); and RNN network dropout probability). Coordinated dropout is a regularization technique which prevents pathological overfitting by forcing the network to model only structure that is shared across neurons. The magnitudes of the HP perturbation were controlled by weights and specified for different HPs (a weight of 0.3 results in perturbation factors between 0.7 and 1.3). The learning rate and dropout probabilities were restricted to their specified search ranges and were sampled from uniform distributions. The KL and L2 HPs were sampled from log-uniform distributions and could be perturbed outside of the initial search ranges. Identical hyperparameter settings were used for both RADICaL and AutoLFADS and for both synthetic datasets and real 2p datasets.

RADICaL is an adaptation of AutoLFADS for 2p calcium imaging. RADICaL operates on sequences of deconvolved calcium events $x(t)$. $x(t)$ are modeled as a noisy observation of an underlying time-varying Zero-Inflated Gamma (ZIG) distribution (Wei et al., 2020a):

$$x_n(t) \sim (1 - q_n(t)) \cdot \delta(0) + q_n(t) \cdot gamma(\alpha_n(t), k_n(t), loc_n), \tag{3.1}$$

where $x_n(t)$ is the distribution of observed deconvolved events, $\alpha_n(t)$, $k_n(t)$, and loc_n are the scale, shape, and location parameters, respectively, of the gamma distribution, and $q_n(t)$ denotes the probability of non-zeros, for neuron n at time t. loc_n was fixed as the minimum nonzero deconvolved event (s_{min}). In the original AutoLFADS model, factors were mapped to a single time-varying parameter for each neuron (the Poisson firing rate) via a linear transformation followed by an exponential nonlinearity. RADICaL instead infers the three time-varying parameters for each neuron, $\alpha_n(t)$, $k_n(t)$, and $q_n(t)$, by linearly transforming the factors followed by a trainable scaled sigmoid nonlinearity (sig_n). sig_n is a positive parameter that scales the outputs of the sigmoid to be in a range between 0 and sig_n, and is optimized alongside network weights. An L2 penalty is applied between sig_n and a PBT-searchable prior to prevent extreme values. The training objective is to minimize the negative log-likelihood of the deconvolved events given the inferred parameters:

$$\prod p(x_n(t)|\mathrm{ZIG}(\hat{\alpha_n}(t), \hat{k_n}(t), \hat{q_n}(t))) \tag{3.2}$$

The event rate for neuron n at time t was taken as the time-varying mean of the inferred ZIG distribution:

$$\hat{r_n}(t) = \hat{q_n}(t) \cdot (\hat{k_n}(t) \cdot \hat{\alpha_n}(t) + s_{min}) \tag{3.3}$$

In AutoLFADS, the instantaneous intensity parameter of the Poisson process completely specifies the spike count distribution for a neuron, while in RADICaL, the ZIG distribution requires three parameters. The RADICaL generator RNN can therefore produce features that may not directly correspond to the biological network's activity to produce the time-varying, three-parameter distribution for each neuron at its output. To avoid analyzing these parameters, rather than using the intermediate factors representation as an estimate of the biological network's state, we used the inferred event rates for the neuronal population. Doing so for both RADICaL and AutoLFADS allowed us to compare methods as directly as possible.

RADICaL uses an SBTT training strategy to achieve sub-frame modeling resolution. RADICaL operates on binned deconvolved calcium events, with bin size smaller than the frame timebase of imaging. Bins where the neurons were sampled were filled with the corresponding event rates, while bins where the neurons were not sampled were filled with *NaNs*. Choosing the sub-frame bin width involves a trade-off. Finer bins improve the possible temporal resolution, but if the data are binned too finely, there may be very few neurons in certain bins, leading to uncertainty about the estimated latent states. It is important to choose the sub-frame bin size to ensure a reasonable number of neurons in each bin. We recommend a neuron count greater than 20 per sub-frame bin based on the results from our neuron downsampling experiments.

The networks output the time-varying ZIG distribution at each sub-frame timestep; however, a mask was applied to the timesteps where the *NaN* samples were to prevent the cost computed from these timesteps being backpropagated during gradient calculation. As a result, the model weights were only updated based on the cost at the sampled timesteps. The reconstruction cost also excluded the cost calculated at the non-sampled timesteps so the PBT model selection was not affected by the cost computed from the non-sampled timesteps.

3.5.2 Simulation experiments

Generating spike trains from an underlying Lorenz system

Synthetic data were generated using the Lorenz system as described in the original LFADS work (Sussillo et al., 2016; Pandarinath et al., 2018b). Lorenz parameters were set to standard values (σ: 10, ρ: 28, and β: 8/3), and δt was set to 0.01. Datasets with different speeds of dynamics were generated by downsampling the original generated Lorenz states by different factors. The speed of the Lorenz dynamics was quantified based on the peak location of the power spectra of the Lorenz Z dimension, with a sampling frequency of 100 Hz. The downsampling factors were 3, 5, 7, 9, 11 and 14 for speeds 4, 7, 10, 13, 15 and 20 Hz, respectively. Each dataset/speed consisted of 32 conditions, with 60 trials per condition. Each condition was obtained by starting the Lorenz system with a random initial state vector and running it for 900 ms. The trial length for the 4 Hz dataset was longer (1200 ms) than that of other datasets (900 ms) to ensure that all conditions had significant features to be modeled - with shorter windows, the extremely low frequency oscillations caused the Lorenz states for some conditions to have little variance across the entire window, making it trivial to approximate the essentially flat firing rates. We simulated a population of 278 neurons with firing rates given by linear readouts of the Lorenz state variables using random weights, followed by an exponential nonlinearity. Scaling factors were applied so the baseline firing rate for all neurons was 3 spikes/sec. Each bin represents 10 ms and an arbitrary frame time was set to be 30 ms (i.e., one “imaging frame” takes 3 bins). Spikes from the firing rates were then generated by a Poisson process.

Generating fluorescence signals from synthetic spike trains

Realistic fluorescence signals were generated from the spike trains by convolving them with a kernel for an autoregressive process of order 2 and passing the results through a nonlinearity that matched values extracted from the literature for the calcium indicator GCaMP6f (Wei et al., 2020b; Dana et al., 2019) (**Fig. A.2a & b**). Three noise sources were added to reproduce variability present in real data (Art, 2006; Starck et al., 1998; Vogelstein et al., 2010): Gaussian noise to the size of the calcium spike, and Gaussian and Poisson noise to the final trace (**Fig. A.2a & b**). This fluorescence generation process was realized as follows: First, spike trains $s(t)$ were generated from the Lorenz system as mentioned above. Independent Gaussian noise ($sd = 0.1$) was added to each spike in the spike train to model the variability in spike amplitude. Next, we modeled the calcium concentration dynamics $c(t)$ as an autoregressive process of order 2:

$$c(t) = \gamma_1 c(t-1) + \gamma_2 c(t-2) + s(t) \tag{3.4}$$

with $s(t)$ representing the number of spikes at time t. The autoregressive coefficients and were computed based on the rise time, decay time ($\tau_{rise} = 20$ ms, $\tau_{decay} = 400$ ms for GCaMP6f) of the calcium indicators, and the sampling frequency. Note that while there is substantial variability in taus across neurons in real data (Wei et al., 2020b), selecting and mimicking this variability was not relevant in our work, because we compared the methods (i.e., RADICaL, AutoLFADS, and smth-dec) after deconvolution. The calcium concentration dynamics were further normalized so that the peak height of the calcium dynamics generated from a single spike equalled one, regardless of the sampling frequency. Subsequently, we computed the noiseless fluorescence signals by passing the calcium dynamics through a nonlinear transformation estimated from the literature (Dana et al., 2019) for the calcium indicator GCaMP6f (**Fig. A.2c & d**). After the nonlinear transformation, the relationship between spike size and trace size was corrupted, and therefore we assumed the baseline of fluorescence signals to be zero and the signals were rescaled to the range in [0,1] using min-max normalization. Finally, Gaussian noise ($\sim N(0, sn)$) and Poisson noise (simulated as gaussian with mean 0 and variance proportional to the signal amplitude at each time point via a constant d) were added to the normalized traces. The resulting fluorescence traces had the same sampling frequency as the synthetic spike trains (100 Hz).

A crucial parameter is the noise level associated with each fluorescence trace. High noise levels lead to very poor spike detection and very low noise levels enable a near-perfect reconstruction of the spike train. In order to select a realistic level of noise we matched the correlations between real and inferred spike trains of the simulated data to those observed in a recent benchmarking study (Berens et al., 2018). We found that a truncated normal distribution of noise level for Gaussian and Poisson noise best matched the correlations. More specifically, for each neuron, $sn = d$ was sampled independently from a truncated normal distribution N(0.12, 0.02) with the tail below 0.06 removed. With the above noise setting, the mean correlation coefficient r between the deconvolved events and ground truth spikes was 0.32, which is consistent with the standard results reported in the "spikefinder" paper (Berens et al., 2018) for OASIS. In our additional tests of model tolerance to spike inference noise, the Gaussian noise added to the fluorescence traces was increased by 2x or 4x. It is worth stressing that real data feature a broad range of noise levels that depend on the imaging conditions, depth, expression level, laser power and other factors. Here we did not attempt to investigate all possible noise conditions, but instead, we aimed to create a simulation with known latent variables (i.e., low-dimensional factors and event rates) that reasonably approximated realistic signal-to-noise levels, in order to provide a tractable test case to compare RADICaL to other methods before attempting comparisons on real data.

Recreating variability in sampling times due to 2p laser scanning

The fluorescence traces were simulated at 100 Hz as mentioned above. A subsampling step was then performed with sampling times for each neuron staggered in time to simulate the variability in sampling times due to 2p laser scanning (as in **Fig. 3.1e**). This produced fluorescence traces where individual neurons were sampled at 33.3 Hz, with phases of 0, 11, 22 ms based on each neuron's location (top, middle and bottom of the FOV, respectively). To break this down, each neuron was sparsely sampled every three time points and the relative sampled times between neurons were fixed. For example, in trial 1, neuron 1 was sampled at time points 1, 4, 7, . . . and neuron 2 was sampled at time points 2, 5, 8, . . . ; in trial 2, neuron 1 was sampled at time points 2, 5, 8, . . . and neuron 2 was sampled at time points 3, 6, 9, ... Thus, the sampling frequency for each individual neuron was 33.3 Hz, while the sampling frequency for the population was retained at 100 Hz by filling the non-sampled time points with

NaNs. The resulting 33.3 Hz simulated fluorescence signals for each individual neuron (i.e., with NaNs excluded) were deconvolved using OASIS (Friedrich et al., 2017) (as implemented in CaImAn (Giovannucci et al., 2019)) using an auto-regressive model of order 1 with s_{min} of 0.1. For experiments with slower imaging speeds, the same steps were repeated but the simulated 100 Hz fluorescence signals were subsampled at different rates (i.e., 16 Hz, 8 Hz and 2 Hz).

Data preparation for each method

Four methods (RADICaL, AutoLFADS, smth-dec and smth-sim-fluor) were compared by their performance on recovering the ground truth latent states across different datasets/speeds. Trials (480 total for each simulated dataset) were split into 80/20 training and validation sets for modeling AutoLFADS and RADICaL. To prepare data for non-RADICaL methods, non-sampled bins were removed so all the sampled bins were treated as if they were sampled at the same time and each bin then represented 30 ms (i.e., sampling frequency = 33.3 Hz). Preparing the data for AutoLFADS required discretizing the deconvolved events into spike count estimates, because AutoLFADS was primarily designed to model discrete spiking data. In the discretizing step, if the event rate was 0, it was left as 0; if the event rate was between 0 and 2, it was cast to 1 (to bias toward the generally higher probability of fewer spikes). If the event rate was greater than 2, it was rounded down to the nearest integer. We note that this is one of many possible patches to convert continuously-valued event intensities to natural numbers for compatibility with the Poisson distribution and AutoLFADS; a more principled solution would be to modify the network to use the ZIG distribution, as we have done in RADICaL. With smth-dec, the deconvolved events were smoothed by convolution with a Gaussian filter (6 ms s.d.) to produce event rates. With smth-sim-fluor, the generated fluorescence signals were smoothed by convolution with a Gaussian filter (6 ms s.d.) to produce event rates. The choice of filter width was optimized by sweeping values ranging from 3 to 40 ms. Smoothing with a 6 ms s.d. filter gave the highest performance in recovering the ground truth Lorenz states for experiments with higher Lorenz frequencies (i.e., >= 10 Hz). The event rates produced from RADICaL had a sampling frequency of 100 Hz, while the event rates produced from the non-RADICaL methods had a sampling frequency of 33.3 Hz. The non-RADICaL rates were then resampled at 100 Hz using linear interpolation.

Mapping to ground truth Lorenz states

Since our goal was to quantify modeling performance by estimating the underlying Lorenz states, we trained a mapping from the output of each model (i.e., the event rates) to the ground truth Lorenz states using ridge regression. First, we split the trials into training (80%) and test (20%) sets. We used the training set to optimize the regularization coefficient using 5-fold cross-validation, and used the optimal regularization coefficient to train the mapping on the full training set. We then quantified state estimation performance by applying this trained mapping to the test set and calculating the coefficient of determination (R^2) between the true and predicted Lorenz states. We repeated the above procedure five times with train/test splits drawn from the data in a complementary fashion. We reported the mean R^2 across the repeats, such that all reported numbers reflect held-out performance. We tested whether the difference of R^2 between each pair of methods was significant by performing a paired, one-sided Student's t-test on the distribution of R^2 across the five folds of predictions. In our simulations we observed a delay caused by deconvolution, where the deconvolved events came systematically later than the true spikes, consistent with findings in a recent study (Rupprecht et al., 2021). We swept across different lags between the event rates and the true latent states in the latent mapping analysis and chose to include a 30 ms lag correlation which gave the highest latent recovery performance empirically.

Additional tests of deconvolution using MLspike

To test whether RADICaL works on deconvolved events that have a spike-time-like structure, we tested MLspike (Deneux et al., 2016) as an alternative for deconvolution. Calcium traces were generated using the identical steps as described above. For MLspike, the cubic polynomial model was chosen as the nonlinearity model consistent with GCaMP6f. The drift parameter was set to 0.001. The decay time constant tau was set to 0.4s. We did not use auto calibration in MLspike because it produced inconsistent results in our tests. Instead, to give MLspike the best chance at high performance, we manually tuned the remaining parameters in MLspike by reducing the error rates for inferred spikes compared to ground truth spikes using a small subset of neurons. Transient amplitude was set to 1 and the noise parameter sigma was set to 0.15. Spikes inferred by MLspike were then prepared for AutoLFADS and RADICaL as described above. Note that the discretizing step was omitted here when preparing data for AutoLFADS.

3.5.3 Real 2p experiments

Subjects and surgical procedures

All procedures were approved by the University of Chicago Animal Care and Use Committee. Two male Ai148D transgenic mice (TIT2L-GC6f-ICL-tTA2, stock 030328; Jackson Laboratory) were used. Mice were individually housed in a reverse 12-hour light/dark cycle, with an ambient temperature of 71.5 degree fahrenheit and a humidity of 58%. Experiments were conducted during the animal's dark cycle. Each mouse underwent a single surgery. Mice were injected subcutaneously with dexamethasone (8 mg/kg) 24 hours and 1 hour before surgery. Mice were anesthetized with 2-2.5% inhaled isoflurane gas, then injected intraperitoneally with a ketamine-medetomidine solution (60 mg/kg ketamine, 0.25 mg/kg medetomidine), and maintained on a low level of supplemental isoflurane (0-1%) if they showed any signs that the depth of anesthesia was insufficient. Meloxicam was also administered subcutaneously (2 mg/kg) at the beginning of the surgery and for 1-3 subsequent days. The scalp was shaved, cleaned, and resected, the skull was cleaned and the wound margins glued to the skull with tissue glue (VetBond, 3M), and a 3 mm circular craniotomy was made with a 3 mm biopsy punch centered over the left CFA/S1 border. The coordinates for the center of CFA were taken to be 0.4 mm anterior and 1.6 mm lateral of bregma. The craniotomy was cleaned with SurgiFoam (Ethicon) soaked in phosphate-buffered solution (PBS), then virus (AAV9-CaMKII-Cre, stock 2.1*1013 particles/nL, 1:1 dilution in PBS, Addgene) was pressure injected (NanoJect III, Drummond Scientific) at two or four sites near the target site, with 140 nL injected at each of two depths per site (250 and 500 µm below the pia) over 5 minutes each. The craniotomy was then sealed with a custom cylindrical glass plug (3 mm diameter, 660 µm depth; Tower Optical) bonded (Norland Optical Adhesive 61, Norland) to a 4 mm #1 round coverslip (Harvard Apparatus), glued in place first with tissue glue (VetBond) and then with cyanoacrylate glue (Krazy Glue) mixed with dental acrylic powder (Ortho Jet; Lang Dental). A small craniotomy was also made using a dental drill over right CFA at 0.4 mm anterior and 1.6 mm lateral of bregma, where 140 nL of AAVretro-tdTomato (stock 1.02*1013 particles/nL, Addgene) was injected at 300 µm below the pia. This injection labeled cells in left CFA projecting to the contralateral CFA. Here, this labeling was used solely for stabilizing the imaging plane (see below). The small craniotomy was sealed with a drop of Kwik-Cast (World Precision Instruments). Two layers of MetaBond (C & B) were applied, then a custom laser-cut titanium head bar was affixed to the skull with black dental acrylic. Animals were awoken by administering atipamezole via intraperitoneal injection and allowed to recover at least 3 days before water restriction.

Behavioral task

The behavioral task (Fig. 3a) was a variant of the water reaching task of (Galiñanes et al., 2018) which we term the “water grab” task. This task was performed by water-restricted, head-fixed mice, with the forepaws beginning on paw rests (eyelet screws) and the hindpaws and body supported by a custom 3D printed clear acrylic tube enclosure. After holding the paw rests for 700-900 ms, a tone was played by stereo speakers and a 2-3 µL droplet of water appeared at one of two water spouts (22 gauge, 90-degree bent, 1” blunt dispensing needles, McMaster) positioned on either side of the snout. The pitch of the tone indicated the location of the water, with a 4000 Hz tone indicating left and a 7000 Hz tone indicating right, and it lasted 500 ms or until the mouse made contact with the correct water spout. The mouse could grab the water droplet and bring it to its mouth to drink any time after the tone began. Both the paw rests and spouts were wired with capacitive touch sensors (Teensy 3.2, PJRC). Good contact with the correct spout produced an inter-trial interval of 3-6 s, while failure to make contact (or insufficiently strong contact) with the spout produced an inter-trial interval of 20 s. Because the touch sensors required good contact from the paw, this setup encouraged complex contacts with the spouts. The mice were trained to make all reaches with the right paw and to keep the left paw on the paw rest during reaching. Training took approximately two weeks, though the behavior continued to solidify for at least two more weeks. Data presented here were collected after 6-8 weeks’ experience with the task. Control software was custom written in MATLAB R2018a using PsychToolbox 3.0.14, and for the Teensy. Touch event monitoring and task control were performed at 60 Hz.

Behavior was also recorded using a pair of cameras (BFS-U3-16S2M-CS, FLIR; varifocal lenses COZ2813CSIR2, Computar) mounted 150 mm from the right paw rest at 10° apart to enable 3D triangulation. Infrared illuminators enabled behavioral imaging while performing 2p imaging in a darkened microscope enclosure. Cameras were synchronized and recorded at 150 frames per second with real-time image cropping and JPEG compression, and streamed to one HDF5 file per camera (areaDetector module of EPICS, CARS). The knuckles and wrist of the reaching paw were tracked in each camera using DeepLabCut (Mathis et al., 2018) and triangulated into 3D using camera calibration parameters obtained from the MATLAB Stereo Camera Calibration toolbox (Heikkila and Silvén, 1997; Zhang, 2000). To screen the tracked markers for quality we created distributions of all inter-marker distances in 3D across every labeled frame and identified as problematic frames with any inter-marker distance exceeding the 99.9th percentile of its respective distribution. Trials with more than one problematic frame in the period of -200 ms to 800 ms after the raw reach onset were discarded (where reach onset was taken as the first 60 Hz tick after the paw rest touch sensor fell below contact threshold). The kinematics of all trials that passed this screening procedure were visualized to confirm quality. Centroid marker kinematics were obtained by averaging the kinematics of all paw markers, locking them to behavioral events and then smoothing using a Gaussian filter (15 ms s.d.). To obtain velocity and acceleration, centroid data was numerically differentiated with MATLAB's diff function and then smoothed again using a Gaussian filter (15 ms s.d.).

Two-photon imaging

Calcium imaging was performed with a Neurolabware two-photon microscope running Scanbox 4.1 and a pulsed Ti:sapphire laser (Vision II, Coherent). Depth stability of the imaging plane was maintained using a custom plugin that acquired an image stack at the beginning of the session (1.4 µm spacing), then compared a registered rolling average of the red-channel data to each plane of the stack. If sufficient evidence indicated that a plane not at the center of the stack was a better match to the image being acquired, the objective was automatically moved to compensate. This typically resulted in a slow and steady upward (outward) movement of the objective over the course of the session. This plane drift is probably due to ETL warming, as it occurred when imaging slides at high power but not low power. The power range used in imaging was approximately 50-65 mW average power, including the net power reduction due to end-of-line blanking.

Offline, images were run through Suite2p to perform motion correction, region-of-interest (ROI) detection, and fluorescence extraction from both ROIs and neuropil. ROIs were manually curated using the Suite2p GUI to retain only those corresponding to somas. We then subtracted the neuropil signal scaled by 0.77. Neuropil-subtracted ROI fluorescence was then detrended by performing a running 10th percentile operation, smoothing with a Gaussian filter (20 s s.d.), then subtracting the result from the trace. This result was fed into OASIS (Friedrich et al., 2017) using the 'thresholded' method, AR1 event model, and limiting the tau parameter to be between 300 and 800 ms. Neurons were discarded if they did not meet a minimum signal-to-noise (SNR) criterion. To compute SNR, we took the fluorescence at each time point when OASIS identified an "event" (non-zero), computed (fluorescence - neuropil) / neuropil, and computed the median of the resulting distribution. ROIs were excluded if this value was less than 0.05. To put events on a more useful scaling, for each ROI we found the distribution of event sizes, smoothed the distribution (ksdensity in MATLAB, with an Epanechnikov kernel and log transform), found the peak of the smoothed distribution, and divided all event sizes by this value. This rescales the peak of the distribution to have a value of unity. Data from two mice and two brain areas (4 sessions in total) were used (Mouse1/M1: 510 neurons, 560 trials; Mouse1/S1: 543 neurons, 506 trials; Mouse2/M1: 439 neurons, 475 trials; Mouse2/S1: 509 neurons, 421 trials).

Data preparation for modeling with RADICaL and AutoLFADS

To prepare data for RADICaL, the deconvolved events were normalized by the s_{min} value output by OASIS so that the minimal event size was 0.1 across all neurons. The deconvolved events for individual neurons had a sampling rate equal to the frame rate (31.08 Hz). For modeling with RADICaL, the deconvolved events were assigned into 10ms bins using the timing of individual measurements for each neuron to achieve sub-frame resolution (i.e., 100 Hz). The non-sampled bins were filled with NaNs. To prepare data for AutoLFADS, the deconvolved events were rescaled using the distribution-scaling method described above, and casted using the casting step described in the simulation section. For both AutoLFADS and smth-dec, the deconvolved events were assigned into a single time bin per frame (i.e., 32.17 ms bins) to mimic standard processing of 2p imaging data, where the sub-frame timing of individual measurements is discarded. Trials were created by aligning the data to 200 ms before and 800 ms after reach onset (100 time points per trial for RADICaL, and 31 time points per trial for AutoLFADS and smth-dec). An individual RADICaL model and AutoLFADS model were trained for each dataset (4 total). Failed trials (latency to contact with correct spout $>$ 15 s for Mouse1, 20 s for Mouse2), or trials where the grab to the incorrect spout occurred before the grab to the correct spout, were discarded. For each dataset, trials (Mouse1/M1: 552 total; Mouse1/S1: 500 total; Mouse2/M1: 467 total; Mouse2/S1: 413 total) were split into 80/20 training and validation.

Trial grouping

PSTH analysis and low dimensional neural trajectory visualization were performed based on subgroups of trials. Trials were sorted into two subgroups per spout based on the Z dimension (height) of hand position. The hand position was obtained by smoothing the centroid marker position with a Gaussian filter (40 ms s.d.). Time windows where the height of hand was used to split trials were hand-selected to present a good separation between subgroups of hand trajectories. For Mouse1/M1, a window of 30 ms to 50 ms after reach onset was used to split left condition trials and a window of 180 ms to 200 ms after reach onset was used to split right condition trials; for Mouse1/S1, a window of 140 ms to 160 ms after reach onset was used to split both left and right condition trials; for both Mouse2/M1 and Mouse2/S1, a window of 30 ms to 50 ms after reach onset was used to split both left and right condition trials. For both left or right conditions and for all mice/areas, 55 trials with the lowest and highest heights were selected as group 1 and group 2, respectively; trials with middle-range heights were discarded.

PSTH analysis and comparing RADICaL and AutoLFADS single-trial rates

RADICaL was first validated by comparing the PSTHs computed using RADICaL inferred event rates and the empirical PSTHs. Empirical PSTHs were computed by trial-averaging smth-dec rates (40 ms kernel s.d., 32.17 ms bins) within each of the 4 subgroups of trials. RADICaL inferred rates were first downsampled from 100 Hz to 31.08 Hz with an antialiasing filter applied, to match the sampling frequency (i.e., the frame rate) of the original deconvolved signals. RADICaL PSTHs were computed by similarly averaging RADICaL rates. Single-trial inferred rates were then compared to the empirical PSTHs to assess how well each method recapitulated the empirical PSTHs on single trials. The correlation coefficient r was computed between inferred single-trial event rates and the corresponding empirical PSTHs in a cross-validated fashion, i.e., each trial's inferred event rate was compared against an empirical PSTH computed using all other trials within the subgroup. r was assessed for the time window spanning 200 ms before to 800 ms after reach onset, and computed by concatenating all trials across the four subgroups, yielding one r for each neuron. Neurons that had fewer than 40 nonzero events within this time window (across all trials) were excluded from the analysis.

Low-D analysis

To visualize the low-dimensional neural trajectories that RADICaL produced, principal component analysis (PCA) was performed on RADICaL inferred rates and smth-dec event rates. RADICaL or smth-dec rates (aligned to 200 ms before and 800 ms after reach onset) were log-transformed (with $1e-4$ added to prevent numerical precision issues) and normalized to have zero mean and unit standard deviation for each neuron. PCA was applied to the trial-averaged rates and the projection matrix was then used to project the log-transformed and normalized single-trial rates (aligned to 200 ms before and 400 ms after reach onset) onto the top 3 PCs.

Subgroup distance ratio analysis

To quantitatively measure how informative RADICaL was about the subgroup identity of each trial, a subgroup distance ratio analysis was performed in the inferred rate space. For each trial at each time point, we measured the Euclidean distances to the corresponding time point of each other trials within the same subgroup as well as the distances to the corresponding time point of each trial from the other subgroup of the same condition. The distance ratio was computed as the ratio of the mean across-subgroup differences to the mean within-subgroup distances. A distance ratio greater than one indicates that the trial is more closely grouped with the trials within the same subgroup compared to the other subgroup. An averaged distance ratio was computed across all trials for each time point.

Decoding analysis

RADICaL-inferred rates, AutoLFADS-inferred rates, and smth-dec (Gaussian kernel 40 ms s.d.) rates were used to decode hand position and velocity using ridge regression. The hand position and velocity were obtained as described above and binned at 10 ms (i.e., 100 Hz). The non-RADICaL rates were retained to a sampling frequency of 100 Hz using linear interpolation. For simplicity, we did not include a lag between the neural data and kinematics. Trials with an interval between water presentation and reach onset that was longer than a threshold were discarded due to potential variations in behavior (e.g., inattention). The threshold was selected arbitrarily for different sessions based on the actual distribution of the intervals in the session (Mouse1/M1: 500 ms; Mouse1/S1: 600 ms; Mouse2/M1: 400 ms; Mouse2/S1: 600 ms). The data were aligned to 50 ms before and 350 ms after reach onset. The decoder was trained and tested using cross-validated ridge regression. First, we split the trials into training (80%) and test (20%) sets. We used the training set to optimize the regularization coefficient using 5-fold cross-validation, and used the optimal regularization coefficient to train the decoder on the full training set. This trained decoder was applied to the test set, and the coefficient of determination (R^2) was computed and averaged across x-, y- and z- kinematics. We repeated the above procedure five times with train/test splits drawn from the data in an interleaved fashion. We reported the mean R^2 across the repeats, such that all reported numbers reflect held-out performance. We tested whether the difference of R^2 between each pair of methods was significant by performing paired, one-sided Student's t-Tests on the distribution of R^2 across the five folds of predictions.

One possible concern is that RADICaL improves decoding not because the single-trial traces are better denoised, but instead because they for some reason result in learning a better decoder. To address this, we performed a "cross-decoder" analysis where the decoder trained with smth-dec rates was applied to the RADICaL inferred rates. Note that it is not guaranteed that the cross-decoder would give better performance even if RADICaL's rates are better denoised, because this is also a task of generalization - during training, the decoder did not see the RADICaL rates which might have different distributions of signal-to-noise across neurons or might require a different level of regularization. Despite this being a difficult task, the cross-decoder analysis shows improved performance over the original smth-dec decoding (**Fig. B.10**). This suggests that the improvement seen in **Fig. 3.5a & b** does not merely reflect the training performance of the decoder but also demonstrates the higher quality of the inferred rates themselves.

Coherence analysis

Coherence was computed between the true and predicted kinematics (window: 200 ms before and 500 ms after reach onset) across all trials and across all x-, y- and z-dimensions using magnitude-squared coherence (MATLAB: mscohere). The power spectral density estimation parameters within mscohere were specified to ensure a robust calculation on the single trial activity: Hanning windows with 35 timesteps (i.e., 350 ms) for the FFT and window size, and 25 timesteps (i.e., 250 ms) of overlap between windows.

Although the coherence analysis presents the performance of each method as a function of frequency (**Fig. 3.5c**), the values are not directly comparable to the latent recovery analysis in simulation (**Fig. 3.2c**). In the simulations, the known, true underlying latent states can be used to directly measure success. In contrast, with real data the true underlying latent states are unknown and the behavioral measurements (hand position and velocity) are indirect correlates. The coherence metric therefore includes other sources of error such as muscle and tracking noise. Both the quicker drop as frequency increases, and the smaller difference between methods, could potentially be explained by the limitations of indirect measurement. In addition, the relationship between neural activity and hand position/velocity may be nonlinear or history-dependent, while our decoding was linear and instantaneous.

Reaction time prediction analysis

RADICaL-inferred rates, AutoLFADS-inferred rates, and smth-dec (Gaussian kernel 40 ms s.d.) rates were used to predict reaction time (RT) using logistic regression. This analysis follows the same procedure used in ref. 30. Reaction time was defined as the interval from water presentation to movement onset. Movement onset was defined as the time when the speed of the paw centroid exceeded 20% of this trial's peak speed. Single-trial rates by the three methods were first aligned to movement onset, then projected into the top 10-PC space. Data were binned into a "premovement" time point (100ms before to movement onset) and a "movement" time point (movement onset to 100ms after). Trials were split into training (75%) and test (25%) sets. A logistic regression classifier was trained using the training set and returned a projection dimension that best discriminated between premovement and movement data. The projection returned by logistic regression was then used to project the test trials binned at original bin size (i.e., 100 Hz). The RT was predicted as the time when the projected activity crossed a 50% threshold. The correlation coefficient r was computed between the true and predicted RTs for the test trials, such that the reported numbers reflect held-out performance.

t-SNE analysis on the weights mapping from factors to ZIG parameters

RADICaL relies on sub-frame bins in which neurons are grouped based on their spatial locations within the FOV. Because this strategy results in consistent neuron grouping, it could potentially result in different groups of neurons corresponding to different latent factors. To test whether such an artifact existed, we visualized the transformation from latents to neurons by using t-SNE to reduce the 300-dimensional weights vector (100 factors * 3 ZIG parameters) into a 2-D t-SNE space for each individual neuron (510 neurons total) (**Fig. B.11**). We did not observe a relationship between neurons' position within the field of view (i.e., top, middle, and bottom) and the underlying factors. This suggested that the model did not use distinct factors for sets of neurons that were sampled with different phases, despite neurons in distant portions of the FOV never being grouped in the same bin.

Neuron downsampling

Two neuron downsampling experiments were performed with different procedures to test the methods' tolerance to low neuron counts. The first procedure was designed to mimic scanning a sparse population of neurons. To do so, the number of neurons included when training RADICaL or AutoLFADS was gradually reduced by randomly dropping a subset of neurons from the previous subset, with a fraction kept of 1, 3/4, 1/2, 1/4, 1/8 or 1/16. This results in 439, 329, 219, 109, 54 or 27 neurons kept for the Mouse2/M1 dataset, and 543, 407, 271, 135, 67 or 33 kept for the Mouse1/S1 dataset.

One RADICaL model and one AutoLFADS model were trained for each number of neurons. Decoding was performed using ridge regression (see above).

The other procedure was designed to emulate scanning a smaller field of view, such as when using a relay lens to image deep structures., Here, the number of neurons included when training RADICaL or AutoLFADS was gradually reduced by limiting the area of FOV that the neurons were sampled from. The area was shrunk from the entire FOV with an area-to-FOV ratio of 1, 25/36, 9/16, 1/4, and 1/9, resulting in the number of included neurons being 439, 321, 262, 121 or 59 for Mouse2/M1. An individual RADICaL model and AutoLFADS model were trained for each number of neurons. Decoding was performed using ridge regression (see above). Note that this analysis represents a lower bound on performance: for this proof-of-concept, we simply artificially excluded data from outside the restricted FOVs, which resulted in substantial time periods that lacked data entirely (e.g., 2/3 of the total sampling time for the smallest FOV considered). In a real application, those time periods that were artificially excluded could instead be used to monitor other brain areas or layers, or to monitor the same neurons with higher sampling rates, either of which might be expected to provide additional information.

Chapter 4

Inferring fast structures from calcium imaging at slow sampling rates using deconvolution-free dynamics modeling

4.1 Abstract

Two-photon (2p) calcium imaging is a powerful tool to monitor the activity of large neuronal populations and probe network-scale computations. However, greater spatial sampling results in lower temporal resolution due to the bandwidth limit created by raster scanning of the laser. A deep learning-based method, namely RADICaL, was recently developed to tackle the space-time trade-off and infer latent dynamics from 2p imaging with sub-frame temporal resolution. Yet, RADICaL relies on deconvolution, which sets an upper bound on the performance of inferring latent dynamics, especially in regimes where sampling rates are low (e.g., < 4Hz). Here we demonstrate that it is possible to remove the deconvolution step by integrating an Autoregressive

(AR) process into RADICaL to approximate the calcium dynamics for each observed neuron. Our novel regularization strategy, neuron Coordinated Dropout (nCD), allows the network to better separate the inference of population dynamics from the per-neuron, calcium dynamics. We first demonstrate that nCD provides an advantage in inferring population dynamics and reconstructing the underlying rates of the neurons. We next test deconvolution-free RADICaL (DfRAD), with nCD integrated, in slower sampling regimes. DfRAD retains a high performance in estimating high-frequency features ($>$ 7Hz) in the latent states as the sampling rate was lowered to 2Hz.

4.2 Introduction

Recent advances in neural interfaces have enabled access to the activity of large neuronal populations, allowing neuroscientists to study how computations underlying motor, sensory, and cognitive processes are implemented at the level of neural populations (Vyas et al., 2020a). Among these techniques, two-photon (2p) calcium imaging offers the revolutionary ability to monitor the activity of millions of neurons while identifying cell types and layers where the neurons are located (Demas et al., 2021; Pachitariu et al., 2017; Peron et al., 2015b; Chen et al., 2013a, 2015). Thus, 2p imaging provides an avenue to link population-level computations to biological structures.

However, with 2p imaging, neurons are serially scanned by a laser that traverses the field of view (FOV), which creates a fundamental trade-off between the size of the FOV (or the number of planes within the FOV), the sampling frequency, and the signal-to-noise (SNR) with which each neuron is sampled (Demas et al., 2021). With a larger FOV or more planes within a FOV, more neurons can be monitored, but the temporal frequency of sampling for each individual neuron is reduced (**Fig. 4.1a**).

A recently developed deep learning-based method, Recurrent Autoencoder for Discovering Imaged Calcium Latents (RADICaL), offers a principled approach to probe the space-time trade-off created by the bandwidth limit of 2p imaging (Zhu et al., 2021a). It improves temporal resolution at the level of population by incorporating the information about the subframe sample times of the neurons into the modeling of neural population dynamics (Zhu et al., 2021b,a) (detailed in 4.3). In applications to synthetic and real 2p data from sensorimotor areas, RADICaL achieves state-of-the-art performance in capturing high-frequency dynamics and predicting fast, ongoing behaviors (Zhu et al., 2021a).

Like many other computational approaches for analyzing calcium imaging data, RADICaL relies on deconvolution as a preprocessing step to obtain spike-like, deconvolved calcium events. As an indirect measurement of the neuronal activity, calcium imaging produces fluorescence traces that are a low-pass filtered version of the underlying spiking activity, with the rise and decay dynamics dictated by the time constants of the calcium indicators (Wei et al., 2020b; Pnevmatikakis, 2019b). Deconvolution is therefore commonly used to undo the effect of calcium indicators and reveal deconvolved events that improve downstream studies (Berens et al., 2018; Pnevmatikakis, 2019b). One key limitation of deconvolution, however, is that the performance of spike inference drops when the sampling frequency is reduced (**Fig. 4.1b**) (Zhu et al., 2021a) because the sparse sampling of the fluorescence is not sufficient to capture the indicator dynamics. Experiments with slow sampling rates are common; with imaging settings that allow monitoring of tens of thousands to millions of neurons, the sampling frequency is limited to 2 - 3 Hz (Demas et al., 2021; Pachitariu et al., 2017), where deconvolution completely breaks down (**Fig. 4.1b**) (Zhu et al., 2021a). Slow indicators such as GCaMP6s may be used as a remedy for slow sampling frequencies to capture more signals related to underlying spikes on coarse timescales, but they set a upper bound on the performance of deconvolution on fine timescales. Thus, deconvolution fundamentally prevents RADICaL from being generalized to massive-population / slow-sampling regimes, limiting the use of its solution to space-time trade-off to uncover fast structures in some of the most exciting, massive datasets in systems neuroscience.

Here we address this challenge by introducing deconvolution-free RADICaL (DfRAD), a method to model population dynamics from 2p data without the need of deconvolution by integrating Autoregressive (AR) models into RADICaL to account for the indicator dynamics (**Fig. 4.1c**). Section 4.3 provides a review of the background and related work. Section 4.4 details DfRAD and a novel regularization strategy, neuron Coordinated Dropout (nCD), to prevent the model from errorneously using the shared variability across the population to explain the individual neuron-level variability resulted from the randomness of spiking activity and indicator dynamics. In Section 4.5, we first demonstrate the effectiveness of nCD in achieving higher-fidelity inference of neural population dynamics from 2p data. With experiments sweeping across different frequencies of the underlying features and across different sampling rates, we show that DfRAD maintains high performance in inferring fast structures at extremely slow sampling frequencies (e.g., 2 Hz), outperforming alternate methods.

4.3 Background

Recent years have witnessed a burst of activity in developing models to uncover latent structures underlying neural population activity monitored by calcium imaging. Dimensionality reduction methods have been developed to identify repeated sequences (Mackevicius et al., 2019), temporal factors that slowly change across trials (Williams et al., 2018), or a latent manifold of odor representations (Wu et al., 2018), or to decouple evoked and spontaneous activities (Triplett et al., 2020). Additional work extracts recurring firing motifs directly from calcium imaging videos using a varia-

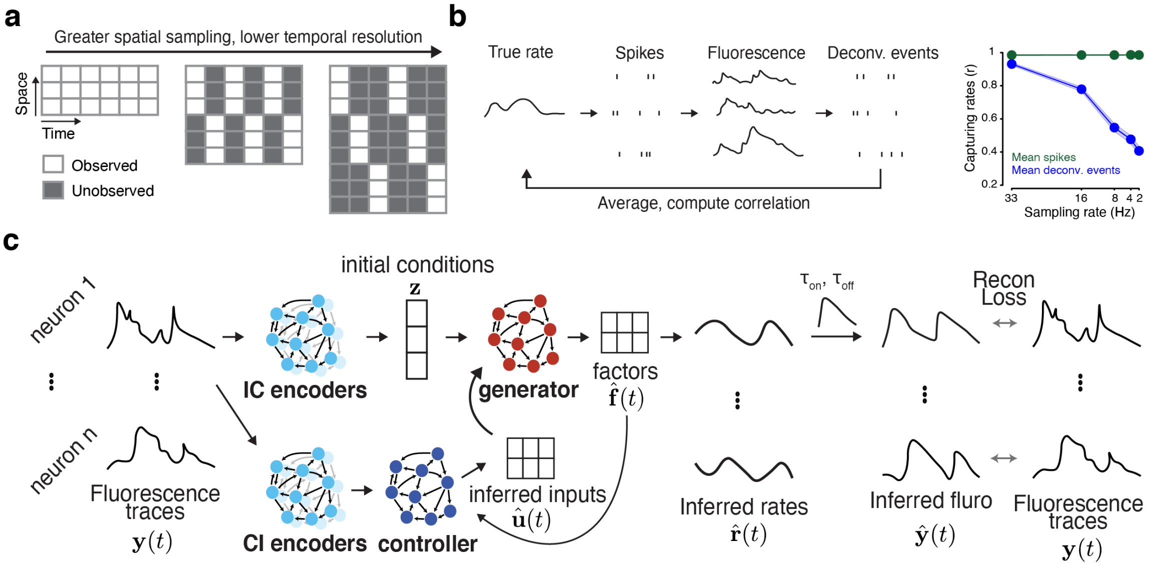


Figure 4.1: Eliminating deconvolution as a limitation at slow sampling rates. (a) In 2p calcium imaging, imaging a greater space results in sampling more neurons but the sampling rate of each individual neuron is lower. Raster scanning of the laser creates staggered sample times across the neurons. (b) To illustrate how deconvolution performs across sampling frequencies, we measured how well averaged, deconvolved events across repeated trials captured the true underlying rate for individual, simulated neurons, which is an effective way to test whether deconvolution irreversibly loses information about the underlying rates (left; simulation pipeline detailed in 4.5.1). Performance in capturing the ground truth firing rates as a function of Lorenz oscillation frequency was quantified by correlation coefficient r between the trial-averaged spikes or deconvolved events and the true rates (right). Error bars indicate the variability across simulated neurons. (c) The DfRAD architecture for inferring latent dynamics from neural population activity.

tional autoencoder framework (Kirschbaum et al., 2019). Switching linear dynamical systems (SLDS) -based methods are further considered to uncover locally linear dynamics underlying the population activity (Costa et al., 2019; Glaser et al., 2020). However, these studies do not utilize the subframe timing information, and thus, are limited in the ability to extract fast structures when the sampling rate is low.

Previous studies have made an effort to preserve subframe timing and overcome the limits of modest 2p frame rates. These studies make varying assumptions on the neural response and the behavior (Picardo et al., 2016), or trial structure (Mano et al., 2019). Rather than relying on these assumptions, RADICaL takes an alternative approach, which is to link subframe sample times of individual neurons to neural population dynamics. Briefly, RADICaL models the neural population activity as an input-driven dynamical system:

$$\dot{\boldsymbol{x}}_t = f(\boldsymbol{x}_t + \boldsymbol{u}_t), \tag{4.1}$$

where the latent state $\boldsymbol{x}_t \in \mathbb{R}^D$ evolves according to dynamics capture by a nonlinear function f, allows inputs $\boldsymbol{u}_t \in \mathbb{R}^K$ to perturb the system, and is seeded by initial condition $\boldsymbol{x}_0$. The observed activity (i.e., the deconvolved calcium events) $\boldsymbol{y}_t \in \mathbb{R}^N$ is a noisy reflection of the latent state of the dynamical system. RADICaL incorporates subframe timing information into the dynamics model by first rebinning the data into finer, subframe bins to reach desired temporal resolution and assigning the observed activity into these subframe bins based on the exact times when each individual neurons are sampled. In this way, a sparse data matrix with neurons' activity placed in stagger time bins was created, and the problem of improving temporal resolution is recasted as a missing data problem. The inherent traits of a dynamical system (i.e., D typically far smaller than N; f imposing certain temporal structure) become helpful constraints for inferring latent states $\boldsymbol{x}_t$ with even partially observed input data. RADICaL then uses selective backpropagation through time (SBTT) to infer dynamics from the partially observed data. SBTT is a neural network training method that allows to compute gradient using only the valid data and ignore the missing samples. With SBTT, RADICaL infers $\boldsymbol{x}_t$ at subframe temporal resolution and provides an avenue to link calcium imaged neural activity to fast, ongoing behavior.

One of the preprocessing steps for RADICaL is deconvolution, which is a critical step to extract estimates of the underlying spiking activity. There is a long and rich literature on deconvolution methods that exploits biophysical models of spike-to-fluorescence generation to detect spikes. These methods include algorithms that explicitly estimate the number of spikes using sequential Monte-Carlo (Vogelstein et al., 2009), Bayesian models (Pnevmatikakis et al., 2013; Deneux et al., 2016), supervised learning approach (Theis et al., 2016), and variational autoencoders (Speiser et al., 2017), or estimate the amplitudes of the calcium transients (i.e., calcium events, a correlate of firing rate) by casting the deconvolution problem as a convex optimization problem (Vogelstein et al., 2010; Pnevmatikakis et al., 2016; Jewell and Witten, 2018) with fast, online versions available (Friedrich et al., 2017). More recently, deep learning-based methods were developed to further improve deconvolution (Hoang et al., 2020; Sebastian et al., 2021; Rupprecht et al., 2021). However, as illustrated above in 4.2, a fundamental challenge of deconvolution is that the performance breaks down as the sampling rate is reduced.

A potential solution to avoid deconvolution and enable the application of RADICaL to slow sampling regimes is to integrate the biophysical model of fluorescence generation into the dynamics model itself. Hints in this direction include works that incorporate an Autoregressive (AR) model into a variational ladder autoencoder architecture (Prince et al., 2021) or a linear dynamical system (LDS) (Koh et al., 2022) to explain the indicator dynamics while modeling the neural population-level dynamics. However, these methods demonstrate limited performance in inferring dynamics from calcium imaging data, and do not offer a mechanism to utilize subframe timing to improve temporal resolution. Here, we take a similar approach to relieve RADICaL of deconvolution. By integrating AR models into RADICaL and applying a novel regularization strategy nCD, we present the first demonstration of precisely inferring fast dynamics from 2p imaging at extremely slow sampling rates.

4.4 Deconvolution-free RADICaL

4.4.1 Overview

Deconvolution-free RADICaL (DfRAD) is an extension of RADICaL for extracting fast neural population dynamics from 2p imaging data at slow sampling frequencies. The backbone of DfRAD is a sequential variational autoencoder (SVAE) that reconstructs the fluorescence traces given the underlying time-varying rates that reflect the latent dynamics underlying the neural population. The goal is to infer the rates and the latent dynamics given the observed fluorescence traces. DfRAD models the rate-to-fluorescence transformation as a generative AR model, and applies nCD, a regularization strategy, to force the network to only capture the shared structure between neurons as the population dynamics. The benefits of nCD in combination of the AR-integrated SVAE enable the model to see through the per-neuron-level calcium dynamics and precisely extract population-level dynamics at subframe temporal resolution when sampling rate is limited.

4.4.2 The AR-integrated SVAE architecture

Like RADICaL, DfRAD models the single-trial population dynamics by learning the dynamical rule f, the initial condition $\boldsymbol{x}_0$, and inputs $\boldsymbol{u}_t$, as detailed in 4.3. The SVAE consists of two main components, an encoder and a generator. A bidirectional RNN encoder ($\mathrm{RNN}^{\mathrm{z}}$) takes as input the observed fluorescence traces $\boldsymbol{y}_t$ and produces a conditional distribution over initial condition $\boldsymbol{z}$, $Q(\boldsymbol{z}|\boldsymbol{y}_t)$, with means and diagonal covariance matrices taken as a linear transformation of the final states of the bidirectional RNNs, $\mathbf{E}^{\mathrm{z}}$.

$$\boldsymbol{\mu}^{\mathrm{z}} = \mathbf{W}^{\mu^{\mathrm{z}}}(\mathbf{E}^{\mathrm{z}}) \tag{4.2}$$

$$\boldsymbol{\sigma}^{\mathrm{z}} = \exp\left(\frac{1}{2}\mathbf{W}^{\sigma^{\mathrm{z}}}(\mathbf{E}^{\mathrm{z}})\right). \tag{4.3}$$

A Kullback-Leibler (KL) divergence loss is applied to penalize deviation of $Q(\boldsymbol{z}|\boldsymbol{y}_t)$ from an uninformative Gaussian prior $P(\boldsymbol{z})$. The initial conditions $\hat{\boldsymbol{z}}$ is then dawn from $Q(\boldsymbol{z}|\boldsymbol{y}_t)$ and mapped to the initial state of the generator RNN, $\mathbf{g}_0$.

$$\hat{\boldsymbol{z}} \sim \mathcal{N}\left(\boldsymbol{z} \mid \boldsymbol{\mu}^{\mathrm{z}}, \boldsymbol{\sigma}^{\mathrm{z}}\right) \tag{4.4}$$

$$\mathbf{g}_0 = \mathbf{W}^{\mathbf{g}_0}(\hat{\boldsymbol{z}}) \tag{4.5}$$

The generator RNN, $\mathrm{RNN}^{\mathrm{gen}}$ learns to approximate the dynamical rules f. To allow the SVAE to model data observed from a non-autonomous dynamical system, a controller RNN, $\mathrm{RNN}^{\mathrm{con}}$, is used to learn a set of time-varying "inferred inputs", $\boldsymbol{u}_t$. A second bidirectional RNN encoder, $\mathrm{RNN}^{\mathrm{ce}}$, operates on the fluorescence $\boldsymbol{y}_t$ and produces time-varying cell state $\mathbf{E}_t^{\mathrm{con}}$. At each time step t, the controller takes as input a delayed feedback from the generator, which is the state of a linear readout of the generator at the previous time step, $\mathbf{f}_{t-1}$, and $\mathbf{E}_t^{\mathrm{con}}$, such that the cell state of $\mathrm{RNN}^{\mathrm{con}}$, $\boldsymbol{con}_t$, is updated as:

$$\boldsymbol{con}_t = \mathrm{RNN}^{\mathrm{con}}\left(\boldsymbol{con}_{t-1}, [\mathbf{E}_t^{\mathrm{con}}, \mathbf{f}_{t-1}]\right). \tag{4.6}$$

Similar to $\hat{\boldsymbol{z}}$, the inferred input $\hat{\boldsymbol{u}}_t$ is then drawn from diagonal Gaussian distributions $Q(\boldsymbol{u}_t|\boldsymbol{y}_t)$, where mean and log-variance are given by a linear transformation of $\boldsymbol{con}_t$.

$$\hat{\boldsymbol{u}}_t \sim \mathcal{N}\left(\boldsymbol{u}_t \left| \boldsymbol{\mu}_t^{\mathrm{u}}, \boldsymbol{\sigma}_t^{\mathrm{u}}\right.\right) \tag{4.7}$$

where

$$\boldsymbol{\mu}_t^{\mathrm{u}} = \mathbf{W}^{\mu^{\mathrm{u}}}(\boldsymbol{con}_t) \tag{4.8}$$

$$\boldsymbol{\sigma}_t^{\mathrm{u}} = \exp\left(\frac{1}{2}\mathbf{W}^{\sigma^{\mathrm{u}}}(\boldsymbol{con}_t)\right). \tag{4.9}$$

A second KL penalty is applied between an autoregressive Gaussian prior $P(\boldsymbol{u}_t)$ and $Q(\boldsymbol{u}_t|\boldsymbol{y}_t)$. The inferred input $\boldsymbol{u}_t$ is then injected into the generator. At each time step, the generator states, $\mathbf{g}_t$, is updated as:

$$\mathbf{g}_t = \mathrm{RNN}^{\mathrm{gen}}\left(\mathbf{g}_{t-1}, \hat{\boldsymbol{u}}_t\right) \tag{4.10}$$

$\mathbf{g}_t$ are then linearly mapped to the factors, $\hat{\mathbf{f}}_t$, which are in turn linearly mapped to the neuronal dimensions followed by an exponential nonlinearity to produce the underlying rates of the neurons.

$$\hat{\mathbf{f}}_t = \mathbf{W}^{\mathrm{fac}}(\mathbf{g}_t) \tag{4.11}$$

$$\hat{\mathbf{r}}_t = \exp\left(\mathbf{W}^{\mathrm{rate}}\left(\hat{\mathbf{f}}_t\right)\right) \tag{4.12}$$

The rate for a given neuron, (r_t) is an abstract quantity that fluctuates in a coordinated way with the rates of other neurons in the population, reflecting the shared, population-level dynamics. We assume the observed fluorescence trace for a given neuron is driven by the calcium concentration dynamics dictated by the indicator, and is also influenced by the underlying time-varying rate r_t as the input to the calcium concentration dynamics. We approximate the calcium concentration (c_t) dynamics using an AR process of order p (**Fig. 4.1c**):

$$c_t = \sum_{k=1}^{p} \gamma_k c_{t-k} + r_t \tag{4.13}$$

where γ_k are trainable parameters that capture the time constants (decay: τ_d, rise: τ_r) of the indicator by:

$$\gamma_1 = \exp\left(-1/\tau_d\right), \text{ if p = 1, or} \tag{4.14}$$

$$\gamma_1 = \exp\left(-1/\tau_d\right) + \exp\left(-1/\tau_r\right), \tag{4.15}$$

$$\gamma_2 = \exp\left(-1/\tau_d\right) * \exp\left(-1/\tau_r\right), \text{ if p = 2} \tag{4.16}$$

We test both AR1 and AR2 models. We use a fast indicator in our experiments to better recover fast population dynamics. Because the fast indicators have τ_r that is small compared to the length of the time bin, we present results using the AR1 model which is sufficient to capture instantaneous increase of calcium concentration in response to a positive input. The fluorescence is related to the calcium concentration as:

$$y_t = a\left(c_t + b\right) \tag{4.17}$$

where a is a nonnegative scalar, b is the baseline concentration, both being trainable parameters. The observed fluorescence $\boldsymbol{y}_t$ is modeled as noisy samples from a time-varying Gaussian distribution, where the mean is $\hat{\boldsymbol{y}}_t$ and the standard deviation $\hat{\sigma^{\mathrm{u}}}$ is a vector of trainable, time-invariant parameters, one for each neuron.

The training objective function is defined as the log likelihood of the data, $\sum_{\mathbf{y}} \log P(\mathbf{y}_{1:T})$, optimized in a VAE setting by maximizing a variational lower bound, $\mathcal{L}$, on the log likelihood,

$$\log P(\mathbf{y}_{1:T}) \geq \mathcal{L} = \mathcal{L}^{y} - \mathcal{L}^{KL}, \tag{4.18}$$

where $\mathcal{L}^{y}$ is the reconstruction loss which is the log likelihood of $\boldsymbol{y}_t$ given the inferred parameters:

$$\mathcal{L}^{y} = \left\langle \sum_{t=1}^{T} \log \Big(\mathrm{Gaussian}(\boldsymbol{y}_t | \hat{\boldsymbol{y}}_t, \hat{\sigma^{\mathrm{u}}}) \Big) \right\rangle_{\boldsymbol{z}, \mathbf{u}_{1:T}} \tag{4.19}$$

And $\mathcal{L}^{KL}$ is the KL penalty described above. Additional $L2$ regularization penalties on the weights of the recurrent networks are also applied. During training, network weights and learnable parameters described above are optimized using stochastic gradient descent and backpropagation through time. Hyperparameters are tuned using population-based training inherited from RADICaL (Zhu et al., 2021a) and AutoLFADS (Keshtkaran et al., 2021) to achieve reliably high-performing models.

4.4.3 Neuron Coordinated Dropout

In the literature of biophysical models for calcium concentration dynamics, the AR process is directly perturbed by spikes (Pnevmatikakis et al., 2016). Spikes are in turn considered as noisy samples drawn from the time-varying rates via a Poisson process.

Such randomness in spike generation creates a spiking variability that is accumulated in time due to the slow decay of the calcium dynamics. Further, because spikes reflect the underlying rates which represent the shared population dynamics (detailed in 4.3), they also contain variability of the coordinated patterns of activity across the population. Thus, the observed fluorescence at any given time carries a mixture of the accumulated spiking variability at the individual-neuron level and the shared variability at the population level. We formulate the concepts of population-level variability and per-neuron-level variability with **Fig. 4.2a**. Consider a dynamical system where the latent states are driven by f. Different initial conditions result in distinct trajectories in the latent space (**Fig. 4.2a**, left) and distinct rates in the high-dimensional neuron space (not shown), creating the population-level variability. Taking one condition, different spiking patterns are generated even with the same underlying rate (**Fig. 4.2a**, right). The spiking variability propagates to the fluorescence traces through the AR calcium dynamics model, creating the per-neuron-level variability.

Our DfRAD generative model passes the rates directly to the AR process, which does not explain the per-neuron-level variability. As a result, it is possible for the model to learn erroneous population-level variability to explain the per-neuron-level variability. Building an appropriate model to explain the accumulated spiking variability is not trivial. Prince et al. (2021) build a recognition network to explicitly infer spikes to explain the calcium dynamics as part of a hierarchy of dynamical systems. Ganmor et al. (2016) compute the likelihood of observing the fluorescence traces based on the rates by integrating all possible spike counts in each time bin and use the observed fluorescence at each time step to approximate the calcium concentration for updating the AR process to the next time step. Koh et al. (2022) integrate deconvolution with LDS and approximate the spiking variability using a Gaussian noise term that can be naturally solved with the EM algorithm. However, these studies either scale up the complexity of the model leading to harder HP optimization (Prince et al., 2021), did not demonstrate SOTA performance (Prince et al., 2021; Koh et al., 2022) due to simplifying assumptions about the underlying dynamics (Koh et al., 2022), or rely on densely sampled fluorescence traces at a fast rate which are not available in slower sampling regimes (Ganmor et al., 2016).

With the goal of precisely inferring the population dynamics, rather than explaining the per-neuron-level variability, here we take a simple approach to better separate the inference of population dynamics from the inference of calcium dynamics. Similar to Coordinated Dropout (CD) (Keshtkaran and Pandarinath, 2019), our approach, nCD, makes a reasonable assumption that the observed neuronal activity is from a lower-dimensional, latent state. With this assumption, for the observed activity of a given neuron, the portion that is part of the population dynamics is shared with and can be inferred by activity of other neurons in the population, and the portion that reflects the spiking variability and calcium dynamics is independent to each neuron and not informative from activity of other neurons. At each training step, nCD regularizes the flow of information through the network by applying a random mask to dropout a proportion of channels (i.e., neurons) at the input and using the complement of that mask to block the gradients of the rest of the channels that were seen by the network as input (**Fig. 4.2b**). In this way, we ensure that the neurons' rates can only be reconstructed using the information that comes from other neurons (i.e., information shared across the population) and receive no information about the per-neuron-level activity (due to spikes and calcium dynamics) of themselves. Thus, this simple strategy better regularizes the learning of the population dynamics and prevents the network predicting erroneous population-level variability to explain the per-neuron-level variability.

4.5 Experiments and Results

4.5.1 Simulation pipeline

To test the performance of DfRAD and the effectiveness of nCD on estimating latent states, we need a fluorescence dataset for which ground truth latent states and neural firing rates were known. Real neural data does not have measurable, ground truth latent states or firing rates. Behaviors as an indirect measurement are often slow and sampled at low temporal resolution in experiments with slow imaging rates that are suitable for testing DfRAD. Therefore, we use realistic simulations of fluorescence traces to test DfRAD across a variety of experimental conditions. We adopt the simulation pipeline described in Zhu et al. (2021a). Briefly, artificial datasets were generated by simulating a population of neurons (278 neurons) whose firing rates are tied to the state of a Lorenz system (Zhao and Park, 2017; Sussillo et al., 2016). We simulated 32 conditions, each obtained by starting the Lorenz system with a random initial state vector. 60 spike trains (i.e., trials) were drawn from the firing rates for each condition via a Poisson process. We generated decay time constants across neurons from a distribution characterized for GCaMP6f in Wei et al. (2020b). Fluorescence traces were then simulated by passing the spike trains through an AR model ($p = 1$) with the decay time constants. Various sources of noise were injected to the fluorescence traces (Zhu et al., 2021a).

4.5.2 Inferring higher-quality latent states with nCD

We first tested a baseline model with nCD disabled. As detailed in 4.4.3, because the model sees the activity of each neuron and reconstructs the activity of the same neuron, it is possible for the model to explain the per-neuron-level variability by predicting erroneous variability in the inferred rates. Here we examined the rates reconstruction within each condition, where the rates were consistent across trials and the trial-to-trial variability in the fluorescence traces was exclusively resulted from the per-neuron-level spiking variability and calcium dynamics. Rates predicted by the baseline model showed a significant amount of trial-to-trial variability that was not present in the ground truth rates (**Fig. 4.2c**, left). To understand whether the inferred initial conditions were informative about the true conditions, we performed t-SNE on the high-dimensional vector of the inferred rates at the trial start. Trials belonging to the same condition should be clustered closely with each other in the state space. However, low-dimensional embeddings of the inferred initial conditions were tangled and not informative of what true conditions they belonged to (**Fig. 4.2d**, left).

We next tested the full DfRAD with nCD enabled. The inferred rates showed little erroneous, trial-to-trial variability and correlated closely with the ground truth rates (**Fig. 4.2c**, right). The low dimensional embeddings of the inferred initial conditions showed clean clusters of trials belonging to the same conditions (**Fig. 4.2d**, right). We measured the correlation coefficient r between the true and inferred rates to quantitatively assess the quality of the rate inference. DfRAD-inferred rates demonstrated higher correlation with the ground truth rates across the majority of the neurons compared to the baseline model with nCD disabled. To evaluate the performance in recovering the ground truth Lorenz states, we trained a cross-validated ridge regression to map the inferred rates to the ground truth Lorenz states and measured the R^2 between the true and inferred latent states. **Fig. 4.2f** shows the true and predicted Lorenz states for an example condition. DfRAD outperformed the baseline model in capturing the ground truth latent states.

4.5.3 Improved inference at extremely slow sampling rates

With DfRAD, the promise is to improve inference of latent dynamics at slower sampling rates. We tested this by evaluating the performance of inference across a wide

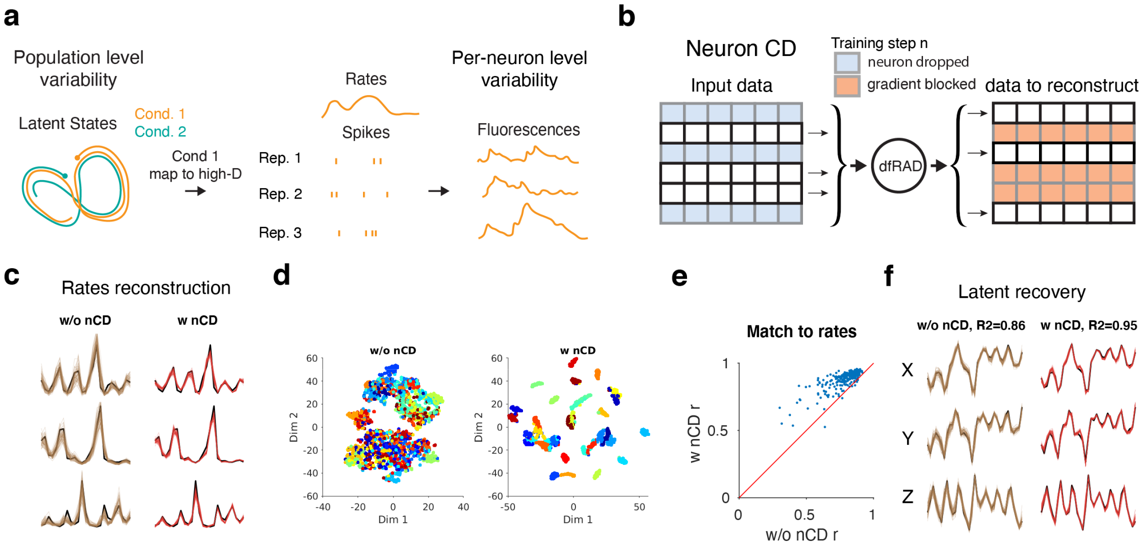


Figure 4.2: Neuron Coordinated Dropout (nCD) improves inference of population-level dynamics. (a) Illustration of population-level variability and per-neuron-level variability. (b) Illustration of nCD for a single training example. (c) True and inferred rates for three example neurons in a single example condition of a Lorenz system. Black: ground truth. Colored: inferred. Each colored trace represents a trial. (d) 2-dimensional t-SNE space representation of the inferred initial conditions. Each dot represents a trial. (e) Performance comparison in capturing the ground truth rates. Correlation coefficient r was computed between the true and inferred single-trial rates. Each point represents an individual neuron. (f) True and inferred Lorenz latent states (X/Y/Z dimensions) for a single example condition. Black: ground truth. Colored: inferred. Each colored trace represents a trial.

range of the frequency of the underlying features and across different sampling rates.

The Lorenz states, rates, spikes and fluorescence traces were generated at a sampling frequency of 33 Hz. In real 2p experiments, the laser traverses the FOV, and the fluorescence trace of a given neuron is sparsely sampled at times when the laser hits, depending on where the neuron is located in the FOV. We first set a 2p sampling rate (i.e., the frame rate) at which each neuron's fluorescence trace was sampled. We then randomly assigned a location to each simulated neuron, such that each neuron was sampled at the 2p sampling rate but the sampling times for different neurons were staggered to simulate 2p laser scanning sampling times. We next reduced the 2p sampling rate to simulate the slowly sampled 2p datasets.

To assess the performance of inferring latent dynamics with different underlying frequencies, we also varied the speed of the Lorenz dynamics. Different speeds of dynamics were generated by downsampling the original generated Lorenz states by different factors. For each speed, we report the peak location of the Z dimension power spectrum, which contains the most concentrated and highest frequencies. Faster dynamics are harder to capture. Here we specifically tested whether DfRAD could improve our ability to infer fast dynamics at slow sampling rates.

We tested 3 sampling frequencies. At a fast, 33 Hz sampling rate, all time points of the fluorescence traces were sampled, leading to a fully observed dataset (**Fig. 4.3a**, top). At slower sampling rates (4 Hz and 2 Hz), each fluorescence trace was heavily sparse-sampled, leading to partially observed data (**Fig. 4.3a**, middle and bottom on the right; 87% missing for 4 Hz and 93% missing for 2 Hz). As described in 4.2, lower sampling rates can enable imaging larger (or more) FOVs and therefore more neurons in real applications. We therefore increased the number of neurons for simulations at slower sampling rates (**Fig. 4.3b**, bottom). We compared the performance of DfRAD to the standard processing methods in the field, deconvolving and applying Gaussian smoothing ("smth-dec") or using the fluorescence traces directly ("fluor"). While SBTT allows DfRAD to output inferred rates at the original, 33 Hz sampling frequency at which the ground truth data (i.e., Lorenz states and rates) was generated, outputs from smth-dec and fluor are at the frame rate (i.e., 33 Hz, 4 Hz, or 2 Hz depending on the 2p sampling rate). Thus, outputs from these latter methods were linearly interpolated to 33 Hz to examine how well they uncovered the ground truth.

We first evaluated the performance of each method in inferring the rates. **Fig. 4.3b** shows an example trial's inferred rates by smth-dec and DfRAD across different sampling rates. The underlying Lorenz system had a Z frequency peak at 7 Hz. At a 2p sampling rate of 33 Hz, both smth-dec and DfRAD revealed structures that corresponded closely with the true rates (**Fig. 4.3b**, top). At slower sampling rates (4 Hz and 2 Hz), DfRAD precisely captured the structures in the rates, while smth-dec completely failed (**Fig. 4.3b**, middle and bottom). We again measured the correlation between the true and inferred rates from different methods. At 33

Hz sampling, DfRAD-inferred rates showed superior performance in matching the ground truth rates across a variety of underlying oscillation frequencies, compared to smoothed deconvolved calcium events or fluorescences (**Fig. 4.3c**, top). This suggests that DfRAD provides powerful denoising benefits even in the fast sampling regimes. At slower sampling rates (4 Hz and 2 Hz), the performance of smth-dec and fluor dropped substantially ($r < 0.2$), while DfRAD maintained a reasonable performance in capturing the rates of datasets with fast Lorenz dynamics (**Fig. 4.3c**, middle and bottom; mean $r >= 0.7$ for 7 Hz Lorenz oscillations and > 0.4 for 9 Hz Lorenz oscillations). We next quantified the performance of predicting latent states using cross-validated ridge regression. **Fig. 4.3d** shows the true and predicted latent states for an example trial across 2p sampling rates. DfRAD retained a high performance in recovering fast latent structure ($R^2 >= 0.8$ for 7 Hz Lorenz oscillations and > 0.5 for 9 Hz Lorenz oscillations), while smth-dec and fluor failed ($R^2 < 0.1$) at slow sampling frequencies.

4.6 Discussion

We introduced DfRAD, a novel approach for learning fast latent dynamics from slowly sampled 2p imaging data. It inherits from RADICaL the ability to infer single-trial dynamics at subframe temporal resolution, eliminates the need of deconvolution as a preprocessing step to enable generalizability in slow sampling regimes, and applies a novel regularization technique to better separate the inference of population dynamics from the per-neuron calcium dynamics. When applied to datasets with varying 2p sampling rates, we show that DfRAD precisely recovers fast structures (>= 7Hz) of the underlying latent states and rates at slow sampling rates (<= 4 Hz), outperforming standard methods in the field. Taken together, DfRAD provides a promising avenue for uncovering rich and complex dynamics from massive neural populations

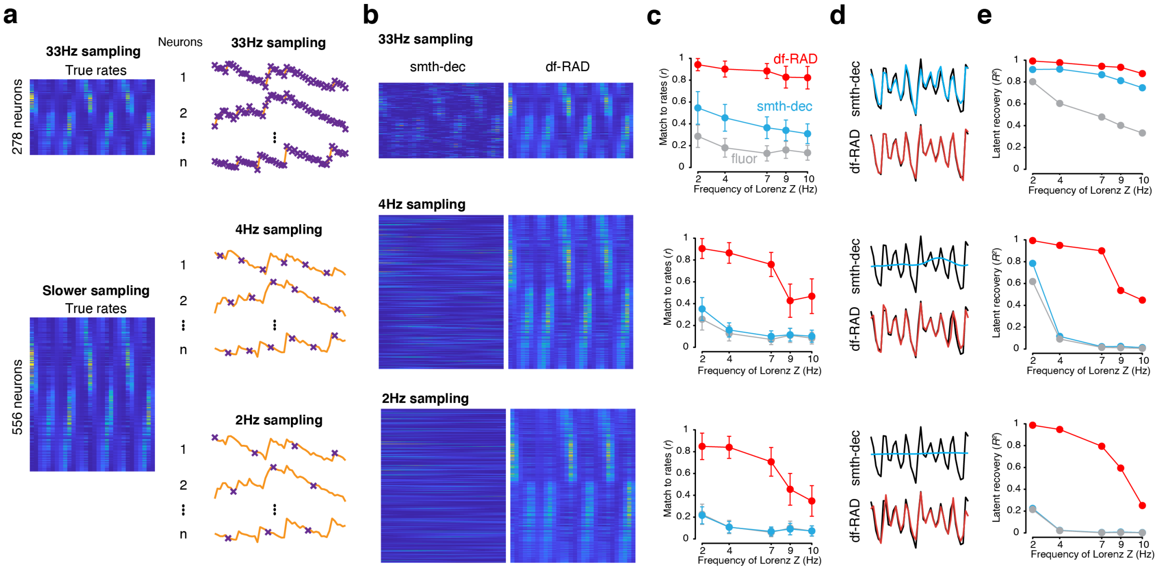


Figure 4.3: DfRAD improves inference of fast dynamics at extremely slow sampling rates. (a) Example ground truth firing rates from a Lorenz system simulated with 7 Hz oscillations (left). Illustration of 2p sampling at different sampling rates (right). Orange traces: simulated fluorescence traces. Purple crosses: sampled data points. (b) Example DfRAD inferred rates and smoothed deconvolved calcium signals. (c) Performance in capturing ground truth firing rates as a function of the underlying oscillations frequency was quantified by r for all 3 methods across sampling rates (top: 33 Hz; middle: 4 Hz; bottom: 2 Hz). (d) True and inferred Lorenz latent states (Z dimension) across sampling rates for a single example trial from Lorenz systems simulated with 7 Hz oscillations. Black: true. Colored: inferred. (e) Performance in estimating the Lorenz Z dimension as a function of Lorenz oscillation frequencies was quantified by variance explained (R^2) for all 3 methods across sampling rates.

monitored by 2p imaging.

Though we made an effort to cover a wide range of experimental conditions (e.g., different sampling rates, frequencies of the underlying structures), the parameter space of possible experiments is very large (various indicators, noise levels, etc). It remains untested how this technique would perform in other settings. In our experiments at slow sampling rates, we increased the number of neurons because slower sampling allows for imaging more neurons. However, we only doubled the neuron counts to provide a proof-of-concept test. In reality, reducing the sampling rate from 33 Hz to e.g., 2 Hz can allow sampling 16x more neurons with potentially richer and more complex dynamics. DfRAD makes it possible to infer such dynamics at high temporal resolution. Yet, increased neuron counts represent a computational challenge as the dimensionality of the datasets scales up quickly. More efficient computational strategies (e.g., sparse matrix storage, etc) will need to be developed to enable wider usability of DfRAD.

Chapter 5

Dissertation summary and future directions

5.1 Dissertation summary

My research and dissertation center on developing computational methods to break through the limits imposed by space-time trade-offs in 2p calcium imaging. In the first study (chapter 2), We recognize that the space-time trade-off due to bandwidth limits is a fundamental challenge that modern neural interfaces are faced with. We aim to provide a principled solution to this problem, which can be applicable for a variety of recording techniques. We first recast the problem as a missing data problem, where the set of neurons being monitored changes dynamically at short intervals. We then tackle this challenge through the lens of dynamical systems and develop a machine learning innovation, SBTT, a neural network training strategy that trains deep generative models of latent dynamics from neural data that are partially observed. We demonstrated the effectiveness of this method with data from both electrophysiology and 2p calcium imaging. In the second study (chapter 3), we took a deeper dive on the 2p imaging applications. We develop innovations tailored specifically for 2p data, integrate SBTT to model dynamics at subframe temporal resolution, and pack our methods as a open-source framework, RADICaL, that can be used by the broad neuroscience community. We validate RADICaL extensively. We first demonstrate that RADICaL greatly improves the ability to infer high-frequency features through carefully designed simulations. We then test RADICaL's performance on real data across multiple subjects and brain areas. RADICaL precisely uncovers single-trial latent dynamics that corresponds closely with behavior on a moment-by-moment basis. We also provide thorough analyses on when RADICaL will succeed or fail and its performance in low-neuron count situations. In the third study (chapter 4), we recognize that 2p imaging enables access to massive number of neurons, potentially to millions of neuron, but the sampling frequency is limited to extremely low. However, even RADICaL have a mechanism to improve temporal resolution, it cannot be generalized to slower sampling regimes because it relies on deconvolution which breaks down at slow sampling rates. We therefore developed machine learning innovations to integrate the generative AR process into RADICaL and regularize the network to better capture the population dynamics. We show that our method, namely DfRAD, is capable of inferring fast dynamics ($>$ 7 Hz) when sampling rate is as low as 2 Hz. Together, methods developed by my work bypass the modest temporal sampling of 2p imaging and provide an avenue towards modeling fast, complex dynamics from massive number of neurons and understanding how population-level computations are tied with anatomical and circuit details reveled via 2p calcium imaging.

5.2 Future directions

One promise of my work is to enable the modeling of latent dynamics from massive number of neurons. However, increased number of neurons (from hundreds to tens of thousands to a million) imposes great computational challenges. More efficient strategies to handle the data need to be developed before applying our tools to massive datasets. Due to the complexity of the model, in-depth investigation of how the memory cost is distributed across different parts of the model is needed. Sparse matrices can be potentially utilized to handle the data to reduce memory cost. Related, the current models are written and tested in TensorFlow 1.14. Migration to TensorFlow 2 or PyTorch is a necessary next step to flexibly develop new features as more resources are available at the new frameworks.

Tools from my work enable modeling dynamics from 2p data at subframe temporal resolution. There is trade-off on the desired bin size and the sparsity of the input data. As an extreme example, modeling data with a 2p sampling rate of 2 Hz at a modeling frequency of 500 Hz will result in an input matrix where 99.6% of the data is missing. It is unlikely the model could model the dynamics in such situations and is untested whether increasing the number of neurons (or how much increase) could rescue the model's performance. One potential solution is to utilize the idea of pretraining. Results in Chapter 2 show that pretraining a dynamics model with fully observed data and applying it to subsequent, partially observed data could improve the inference on the partially observed data. This strategy can be used combat the high-sparsity problem if high-bandwidth experimental sessions are available. Systematic tests are needed to validate this idea in such regimes. Related, one consequence of rebinning input data at subframe resolution is that the sequence length is substantially increased. Our models are based on RNNs, which are prone to problems related to long sequences, such as gradient vanishing. Long sequences also substantially increase computational cost and lead to long training time. Alternative architectures, such as Transformers can be considered.

Compared to applications to electrophysiology, modeling dynamics is a relatively new trend in calcium imaging. Our work resolves a series of specific challenges and opens the door for precise inference of dynamics from calcium imaging data. This enables a lot of exciting, new questions to be explored. First, calcium imaging allows the identification of cell types when monitoring a large network of neurons. Do populations of different cell types exhibit different dynamics? How do neurons with different cell types contribute to the overall dynamics of the network? How do different cell types communicate at the population level? These are a mixture of scientific and engineering questions. Future work can build dynamics models for different cell types and study the properties of the dynamics from different cell types. Alternatively, an interesting engineering challenge is to integrate cell type identities into the dynamics model of the whole network. These studies will facilitate the understanding of how a neural circuitry gives rise to function. Further, calcium imaging can image a large FOV, with spatial information available, or image multiple planes, with layer information available. How do populations across different spatial locations or layers contribute to the overall dynamics? How do neurons from different layers or brain areas communicate? Can we build dynamics model that incorporates in the information about spatial locations, layers or brain areas? Can we build dynamics model to explain the communication between different populations? These, again, are interesting scientific and engineering questions waited to be answered.

Appendix A

Extended data figures

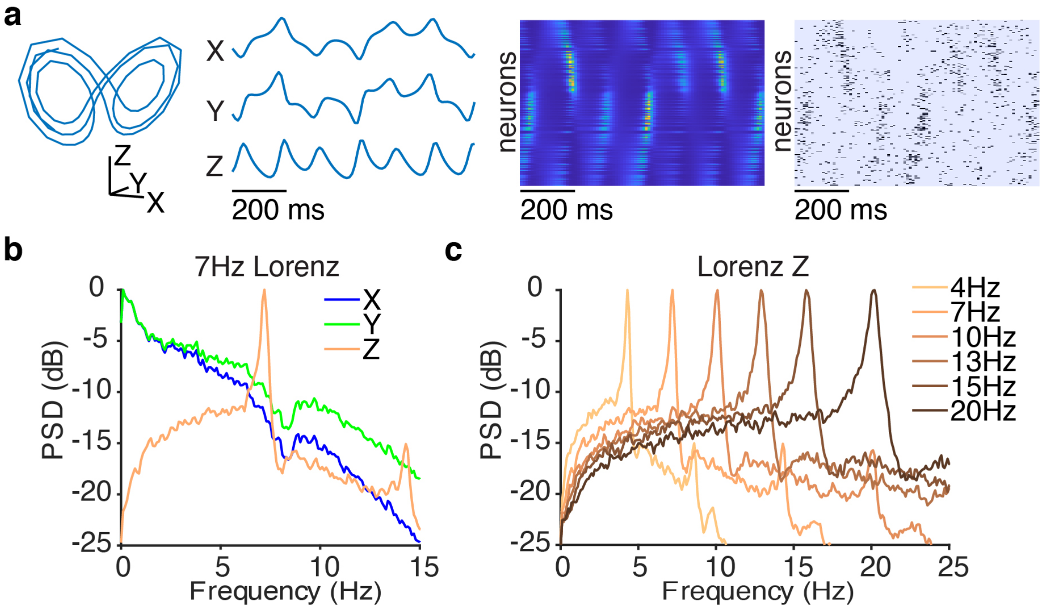


Figure A.1: Simulation of Lorenz system at different speeds. This figure illustrates the underlying dynamical system used for the simulation experiments. (a) An example Lorenz trajectory in a 3-dimensional state space (far left) and with three dynamic variables plotted as a function of time (middle left) for a system with Z-oscillation peak frequency of 7 Hz (i.e., the power spectrum of the Lorenz system's Z-dimension had a pronounced peak at 7 Hz). Firing rates for the simulated neurons were computed by a linear readout of the Lorenz variables followed by an exponential nonlinearity (middle right). Spikes from the firing rates were then generated by a Poisson process (far right). The example trial shown here is identical to "Trial 2" in **Fig. 3.2a**, but with a wider plotting window. (b) Power spectrum of the individual Lorenz variables for the system with a Z-oscillation peak frequency at 7 Hz. Because only the Z variable has a clear peak in the power spectrum, this variable was used exclusively for all further analyses in simulations except **Fig. B.1**. (c) Power spectrum of the Z dimension for Lorenz systems simulated with different Z-oscillation peak frequencies.

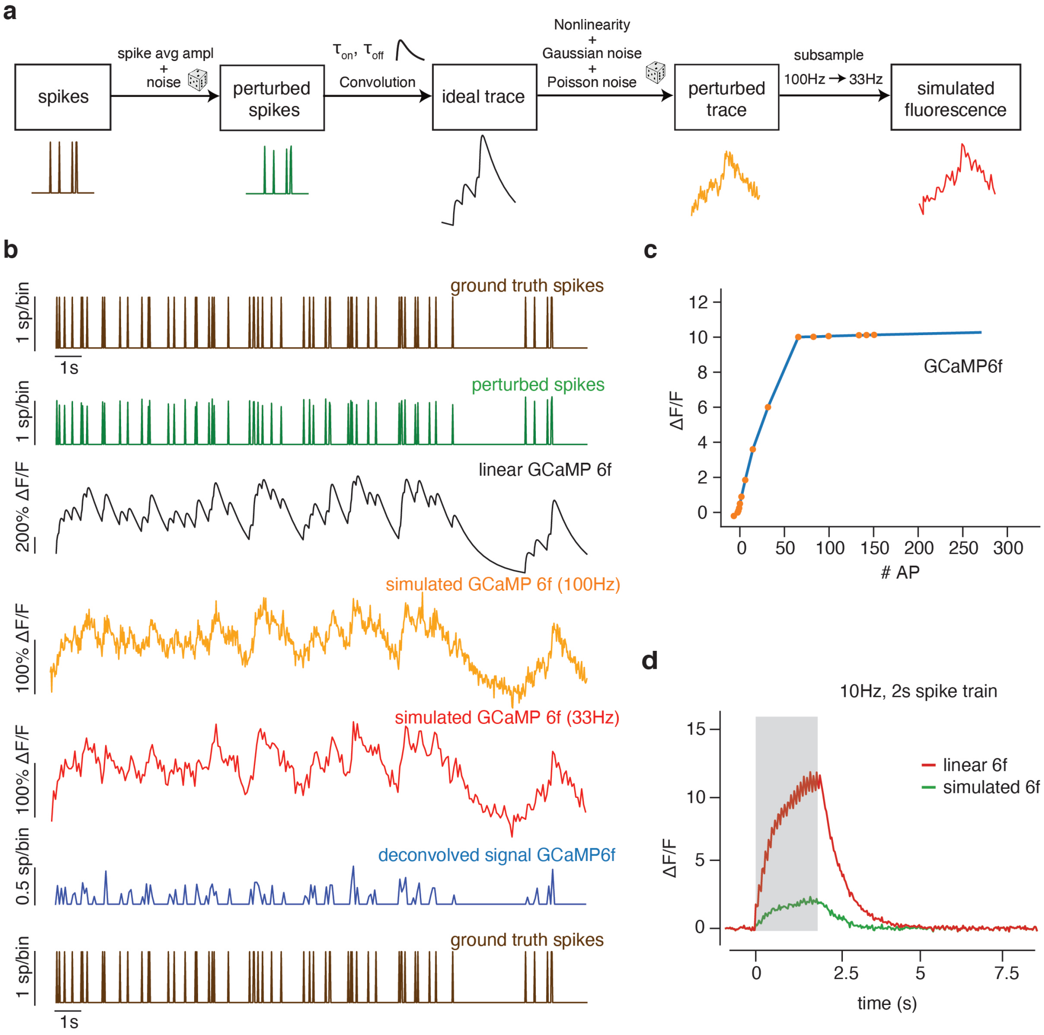


Figure A.2: Simulation pipeline to generate artificial fluorescence traces from the underlying Lorenz system. (a) This pipeline begins from the Poisson-random spikes generated in the far-right panel of **Fig. A.1**. Calcium traces were generated by first corrupting the spikes with amplitude noise, then modeling the dynamics of calcium indicators in response to a spike with an autoregressive process of order 2 transformed by a piecewise-linear non-linearity. Sources of noise corrupting this fluorescence trace were then added. The nonlinearity and noise sources were chosen to approximate the variability observed in real data. (b) Example ground truth and simulated data using a GCaMP6f model. From top to bottom: original ground truth spikes fed into the simulator, perturbed spikes, idealized calcium trace, fluorescence trace with nonlinearity and noise sources added, fluorescence trace after subsampling, deconvolved spikes, and finally original ground truth spikes fed into the simulator (shown again for comparison; same as top). (c) Estimated nonlinearities for GCaMP6f from (Dana et al., 2019). (d) Example traces generated by the simulator for a train of 10 Hz stimuli, with and without nonlinearity applied.

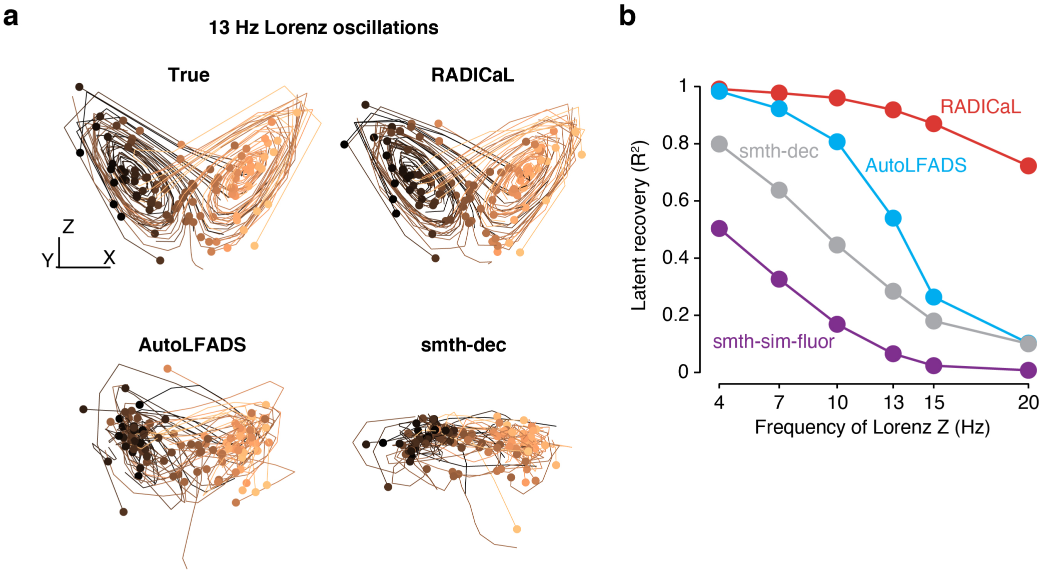


Figure A.3: RADICaL retains high latent recovery performance in a simulation experiment that lacks stereotyped conditions. This analysis was targeted at determining whether RADICaL simply 'memorized' the stereotyped trajectories for a limited number of conditions, or whether it could generalize to cases where each trial was more unique. To answer this question, we designed a "zero condition" simulation experiment, where each trial had its own unique Lorenz initial state and there were no repeated trials with the same underlying latent trajectories. (a) Example true (top left) and estimated Lorenz trajectories by RADICaL (top right), AutoLFADS (bottom left), and smth-dec (bottom right). Each trajectory is an individual trial, colored by the location of the initial state of the true Lorenz trajectory. The initial states of the trials are indicated by the dots in the same colors as the trajectories. (b) Performance in estimating the Lorenz Z dimension as a function of Lorenz oscillation frequency was quantified by variance explained (R^2) for all 4 methods.

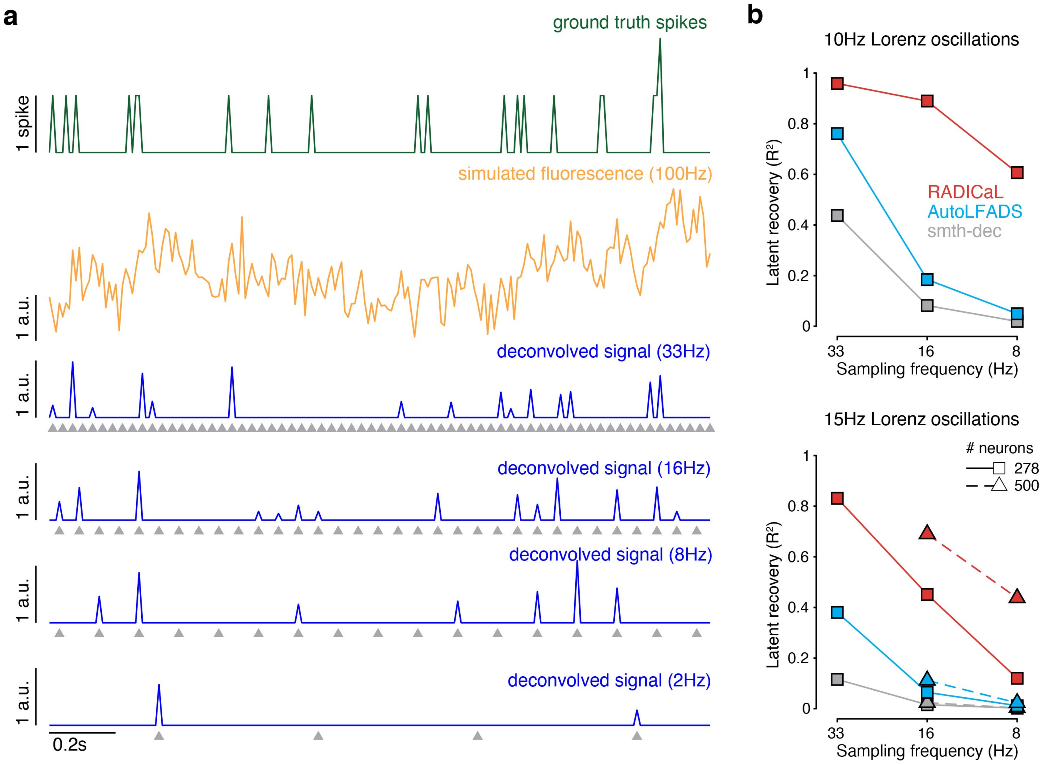


Figure A.4: RADICaL retains high latent recovery performance at slower imaging speeds, but there are limits to deconvolution with slower sampling. To understand the extent to which the model performance depends on imaging speeds, we simulated data at different sampling rates ranging from 2 Hz to 33.3 Hz. (a) Example ground truth spikes, simulated fluorescence, and deconvolved signals at different sampling rates. Sample times are denoted by gray triangles. Deconvolution performance degraded at slower sampling rates, particularly in regimes when transients could be missed entirely. In our simulation we used a GCaMP6f model with a decay time of 400ms (see Methods). At an imaging rate of 2 Hz, the majority of transients were missed and the estimate of the decay time constant tau was inaccurate (916.8 +/- 49.4ms, compared to the ground truth 400ms). Because deconvolution performs poorly at these sampling rates (i.e., ¡= 2 Hz) with fast indicators, we do not recommend using RADICaL under such circumstances. (b) Performance in estimating the Lorenz Z dimension as a function of sampling rate was quantified by variance explained (R^2) for all 3 methods, for Lorenz oscillation frequencies of 10Hz (top) and 15Hz (bottom). Squares with solid lines denote experiments with 278 neurons. Triangles with dashed lines denote experiments with 500 neurons. RADICaL retained high performance and outperformed AutoLFADS and smth-dec in recovering the latent states of a 10 Hz Lorenz system at moderately slow sampling rates (8 and 16 Hz; top). In real experiments, there may be benefits to slower sampling, e.g., one can image more neurons using a larger FOV. Increasing the number of neurons boosted RADICaL's performance, while AutoLFADS and smth-dec showed negligible improvement (bottom).

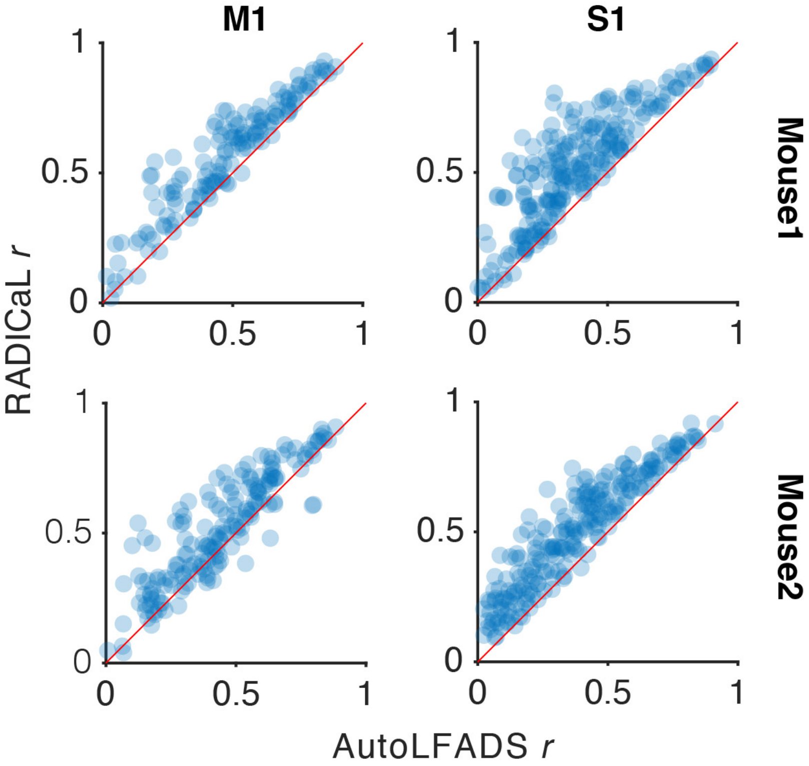


Figure A.5: Performance of RADICaL and AutoLFADS in capturing the empirical PSTHs on single trials in the mouse water grab experiments. This figure is related to **Fig. 3.3d**, but compares RADICaL with AutoLFADS instead of smth-dec. Correlation coefficient r was computed between the inferred single-trial event rates and empirical PSTHs. Each point represents an individual neuron. These results demonstrate that RADICaL captures the key features of individual neurons' responses from single-trial activity better than AutoLFADS in nearly every case.

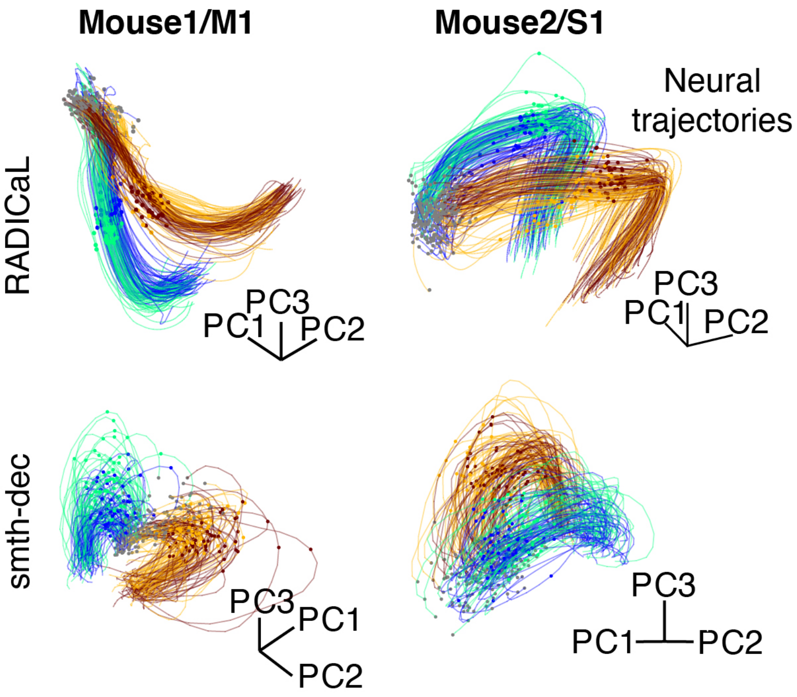


Figure A.6: Single-trial neural trajectories for additional mouse water grab experiments. This figure is related to **Fig. 3.4a**, and shows the remaining datasets. Single-trial, log-transformed event rates were projected into a subspace computed by applying PCA to the trial-averaged, log-transformed rates, colored by subgroups. Lift onset times are indicated by the dots in the same colors as the trajectories. Gray dots indicate 200 ms prior to lift onset time. Top row: single-trial neural trajectories derived from RADICaL rates; Bottom row: single-trial neural trajectories derived from smth-dec rates.

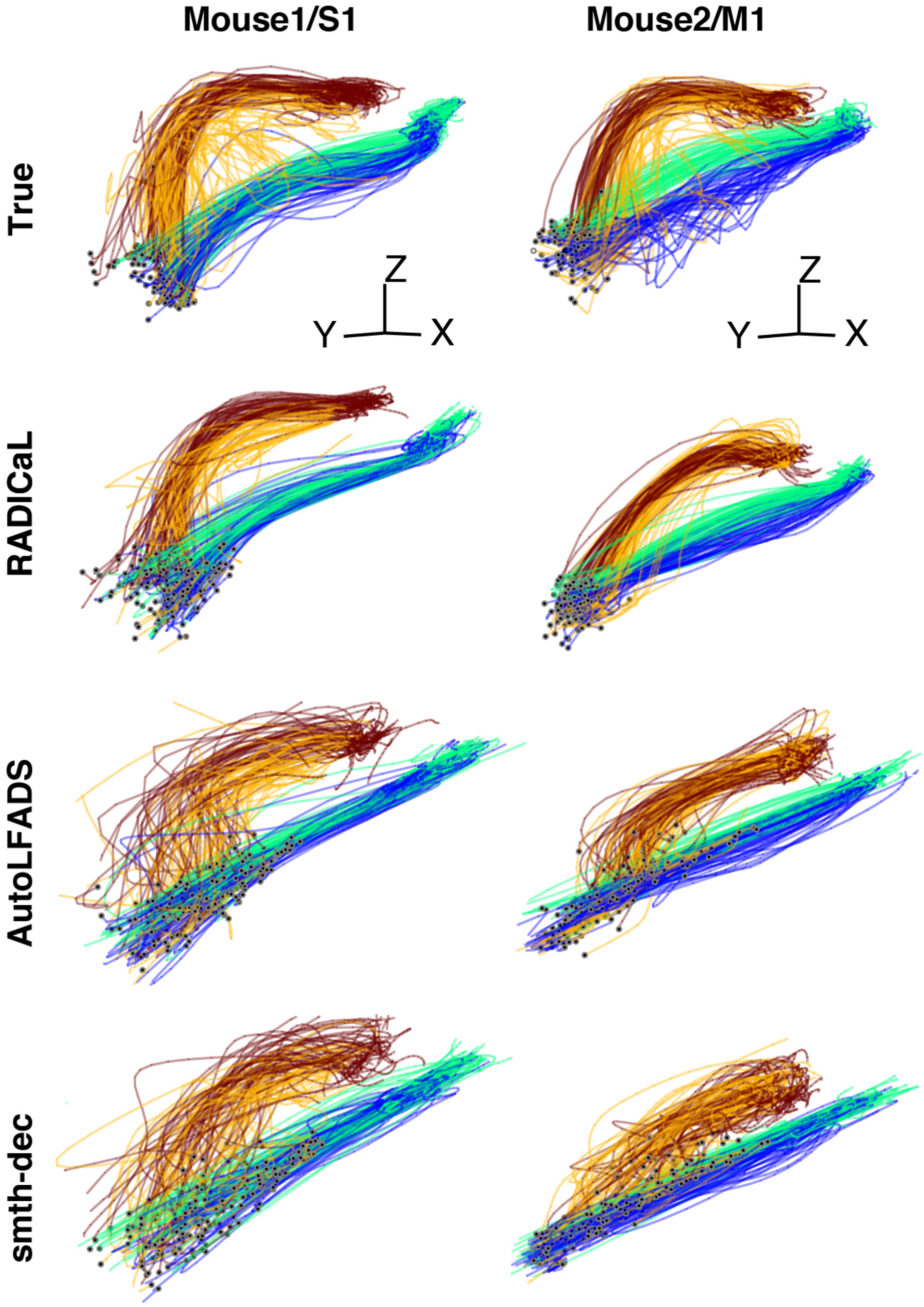


Figure A.7: Hand trajectories for additional mouse water grab experiments. This figure is related to Figure 5a, and shows the remaining datasets. True and decoded hand positions for Mouse1/S1 (left) and Mouse2/M1 (right).

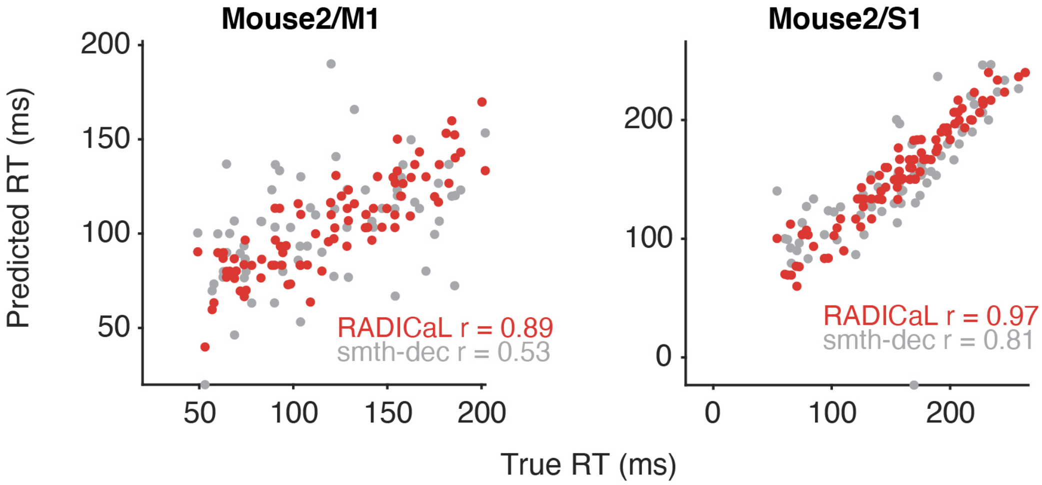


Figure A.8: Prediction of single-trial reaction times for additional mouse water grab experiments. This figure is like **Fig. 3.5d**, for the remaining datasets. Each dot represents an individual trial, color-coded by the technique. Correlation coefficient r was computed between the true and predicted reaction times. Data from Mouse2/M1 (left) and Mouse2/S1 (right).

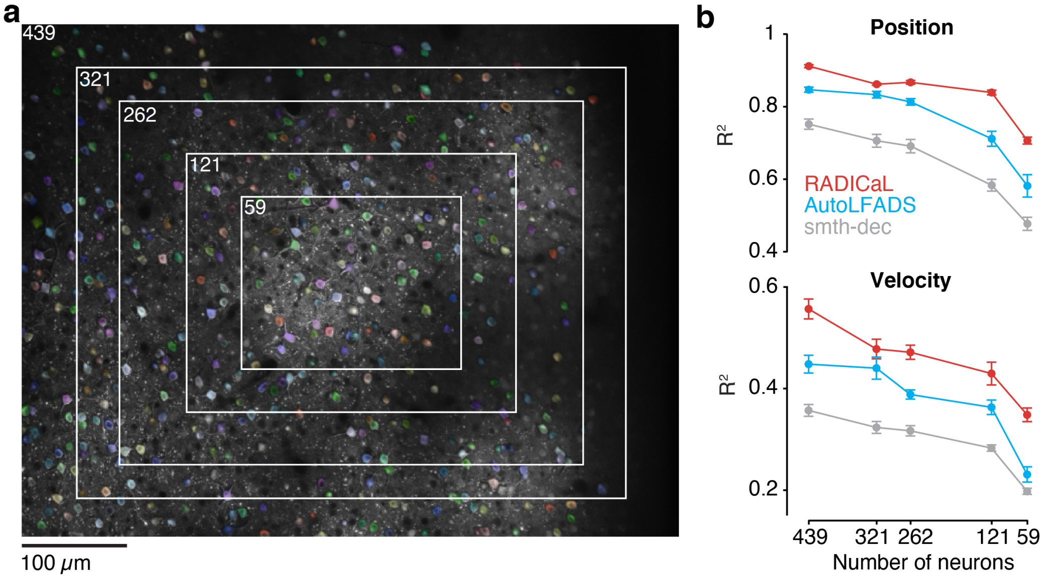


Figure A.9: RADICaL retains high decoding performance in an FOV-shrinking experiment. This is an alternative method for evaluating performance with reduced neuron counts to the method in **Fig. 3.6**. (a) The area selected to include was gradually shrunk to the center of the FOV to reduce the number of neurons included in training RADICaL or AutoLFADS. (b) Decoding performance measured using variance explained (R^2) as a function of the number of neurons used in each technique (top: Position; bottom: Velocity). Error bar indicates the s.e.m. across 5 folds of test trials. Data from Mouse2/M1.

Appendix B

Supplementary figures

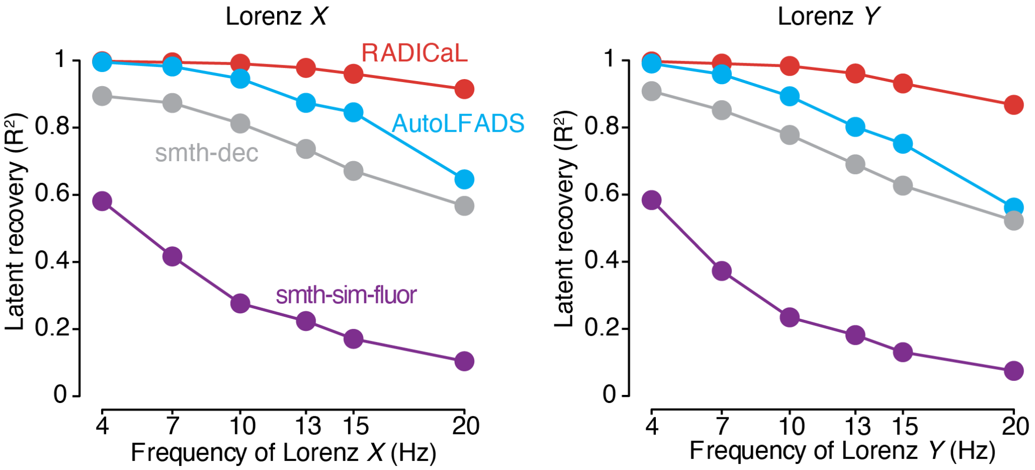


Figure B.1: Performance of estimating other Lorenz dimensions in the simulation experiments. Performance of estimating Lorenz X (left) and Y (right) dimensions as a function of simulation frequency was quantified by variance explained (R^2) for all 4 methods. Note that these variables are dominated by lower frequencies than the Z variable used in other figures, and therefore make for an easier challenge. We therefore used the Z variable for all other results.

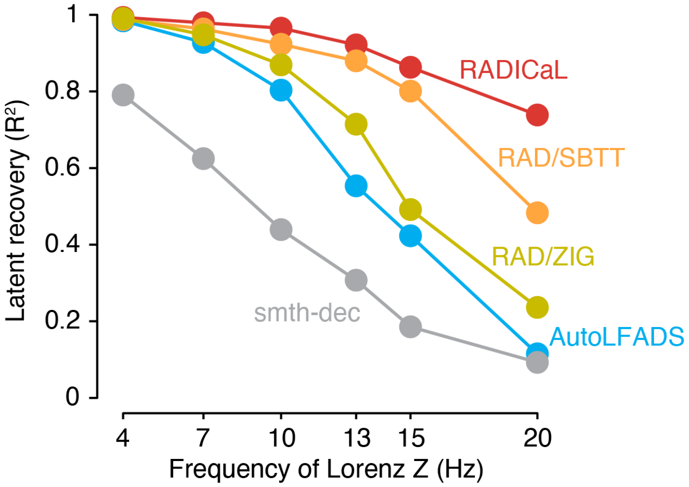


Figure B.2: Both SBTT and ZIG improve latent recovery performance separately. To understand the contributions of ZIG and SBTT independently in RADICaL's performance in latent recovery, we fit RADICaL to different Lorenz oscillation frequencies with only the ZIG emission model enabled (no SBTT; "RAD/ZIG") or only SBTT enabled (no ZIG; "RAD/SBTT"). Performance in estimating the Lorenz Z dimension as a function of Lorenz oscillation frequency was quantified by variance explained (R^2). RAD/ZIG performed a little better than AutoLFADs, while RAD/SBTT performs substantially better, but combining both (the full RADICaL model) performed substantially better still.

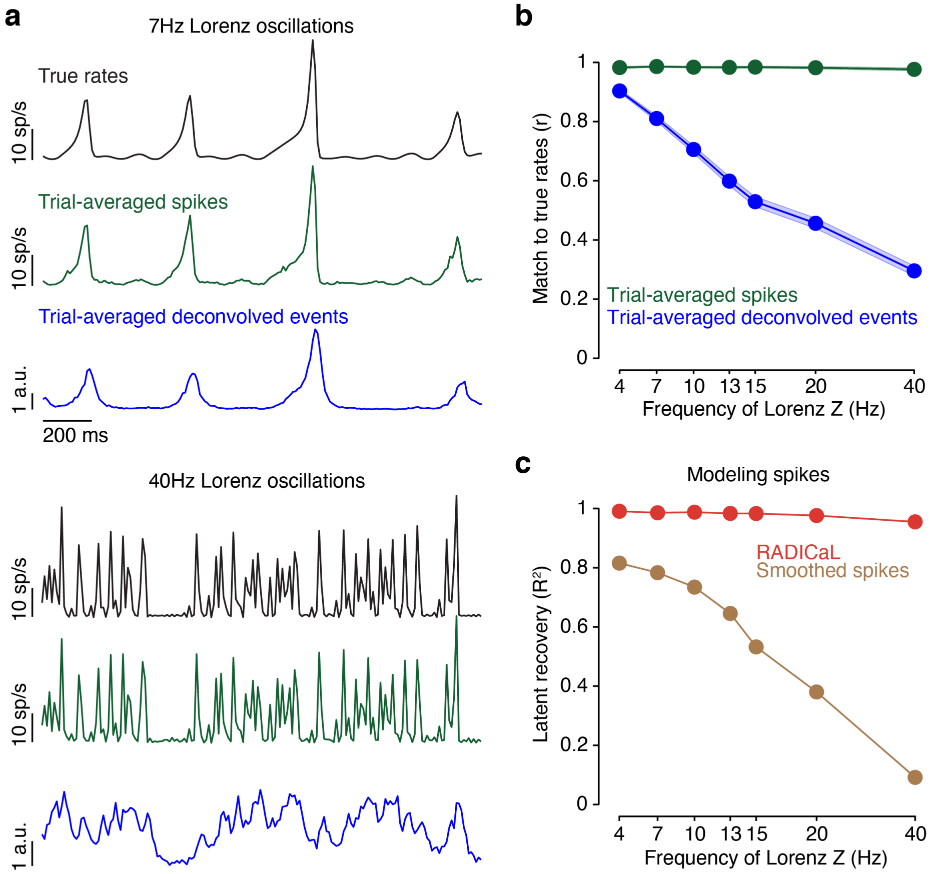


Figure B.3: Deconvolution places an upper bound on RADICaL's performance recovering higher-frequency features. To understand how deconvolution performs across Lorenz oscillation frequencies, we measured how well trial-averaged deconvolved events captured the true underlying rates for individual (simulated) neurons. Averaging deconvolved events across the noisy repeated trials that have the same true underlying rates is a straightforward way to test, on average, whether deconvolution irreversibly loses information about the underlying rates. There were three main steps in the rates-to-events generation process: Poisson sampling of spikes from the underlying rates, fluorescence generation and sub-sampling, and deconvolution (detailed in Methods). To specifically isolate the effect of fluorescence generation and deconvolution on rate recovery, we also tested recovery with those steps omitted, i.e., spikes generated in the rates-to-events process were sub-sampled from a sampling frequency of 100 Hz to 33.3 Hz as was done for the fluorescence traces, and were averaged across trials to quantify how well they captured the true underlying rates. (a) Example ground truth firing rates, averaged spikes across trials (3000 trials), and averaged deconvolved calcium events across trials (3000 trials), for Lorenz oscillation frequencies of 7Hz (top) and 40Hz (bottom). (b) Performance in capturing the ground truth firing rates as a function of Lorenz oscillation frequency was quantified by correlation coefficient r between the trial-averaged spikes or deconvolved events and the true rates. Error bars indicate the variability across simulated neurons. The correlation between the trial-averaged deconvolved events and the true rates dropped as the Lorenz oscillation frequency increased, suggesting that deconvolution fails at higher Lorenz oscillation frequencies. The correlation between the trial-averaged spikes and the true rates did not drop as the Lorenz oscillation frequency increased, suggesting that the drop seen for the deconvolved events was mainly due to deconvolution and not the Poisson sampling and sub-sampling steps. (c) To determine whether RADICaL's performance loss for high-frequency signals was purely due to deconvolution failure or might involve limitations of the model itself, we eliminated the fluorescence generation/deconvolution step and applied RADICaL directly to the sub-sampled spiking activity. In this test, we did not use RADICAL's ZIG observation model, but kept the SBTT approach and used a Poisson observation model. Performance in using ground truth spikes to estimate the Lorenz Z dimension as a function of Lorenz oscillation frequency was quantified by variance explained (R^2) for smoothing and RADICaL. RADICaL retained high performance in latent recovery across Lorenz oscillation frequencies from 4Hz to 40Hz, whereas smoothing showed a much faster degradation of latent recovery performance. Together, these analyses demonstrate that the degradation in RADICaL's performance at higher Lorenz oscillation frequencies is mainly due to inaccuracies in deconvolution, and not due to the model itself.

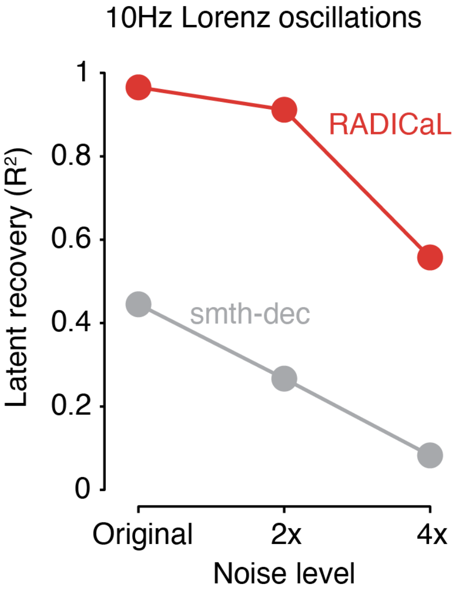


Figure B.4: Model tolerance to spike inference noise. In our simulations, we chose parameters so that the resulting signal-to-noise regime produced similar correlations between real and inferred spike trains as observed in a recent benchmarking study (Berens et al., 2018) (see Methods). However, the spike inference noise can vary in real experiments and could affect RADICaL's performance. To test how larger spike inference noise affects the performance of RADICaL and smth-dec, we raised the level of the Gaussian noise used in generating simulated fluorescence traces by 2x or 4x. Performance in estimating the Lorenz Z dimension as a function of the level of spike inference noise was quantified by variance explained (R^2) for RADICaL and smth-dec. Performance declined for both methods as the noise level increased. However, RADICaL retained high performance at the 2x noise level (R^2=0.91) and reasonable performance at the 4x noise level (R^2=0.56). Smth-dec had low performance across the board (R^2=0.27 and 0.08 for 2x and 4x noise, respectively). Notably, RADICaL performed better at the 4x noise level than smth-dec at the original noise level.

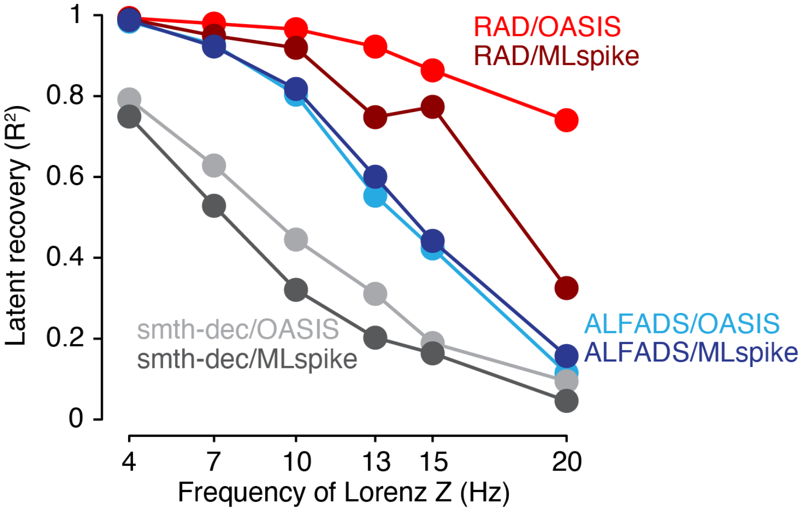


Figure B.5: RADICAL (with SBTT) improves latent recovery when using spikes inferred by MLspike, but does not perform as well as when using OASIS for deconvolution. To test whether RADICaL could be effective with deconvolution algorithms that infer spike times instead of event rates, we analyzed simulated data that had spike inference performed with MLspike (Deneux et al., 2016). Performance in estimating the Lorenz Z dimension as a function of Lorenz oscillation frequency was quantified by variance explained (R^2) for six methods. These included three methods in which the inputs were deconvolved events from OASIS: RADICaL ("RAD/OASIS"), AutoLFADS ("ALFADS/OASIS") and smoothing ("smth-dec/OASIS"); and three methods in which inputs were spikes inferred with MLspike: RADICaL ("RAD/MLspike"), AutoLFADS ("ALFADS/MLspike") and smoothing ("smth-dec/MLspike"). When pairing RADICaL with deconvolution methods that produce spike times as output, we can use a Poisson observation model (as one would use for spikes measured via electrophysiology) instead of ZIG, while retaining the SBTT approach for sub-frame sampling. RADICaL with a Poisson observation model (RAD/MLspike) was able to model MLspike output, and substantially outperformed AutoLFADS and smth-dec (ALFADS/MLspike and smth-dec/MLspike), but did not perform as well as RADICaL applied to OASIS-deconvolved events (RAD/OASIS). In addition, the parameter tuning required for MLspike is more involved and requires more expertise than OASIS (see Methods). Therefore, we recommend using OASIS as the deconvolution method for RADICaL.

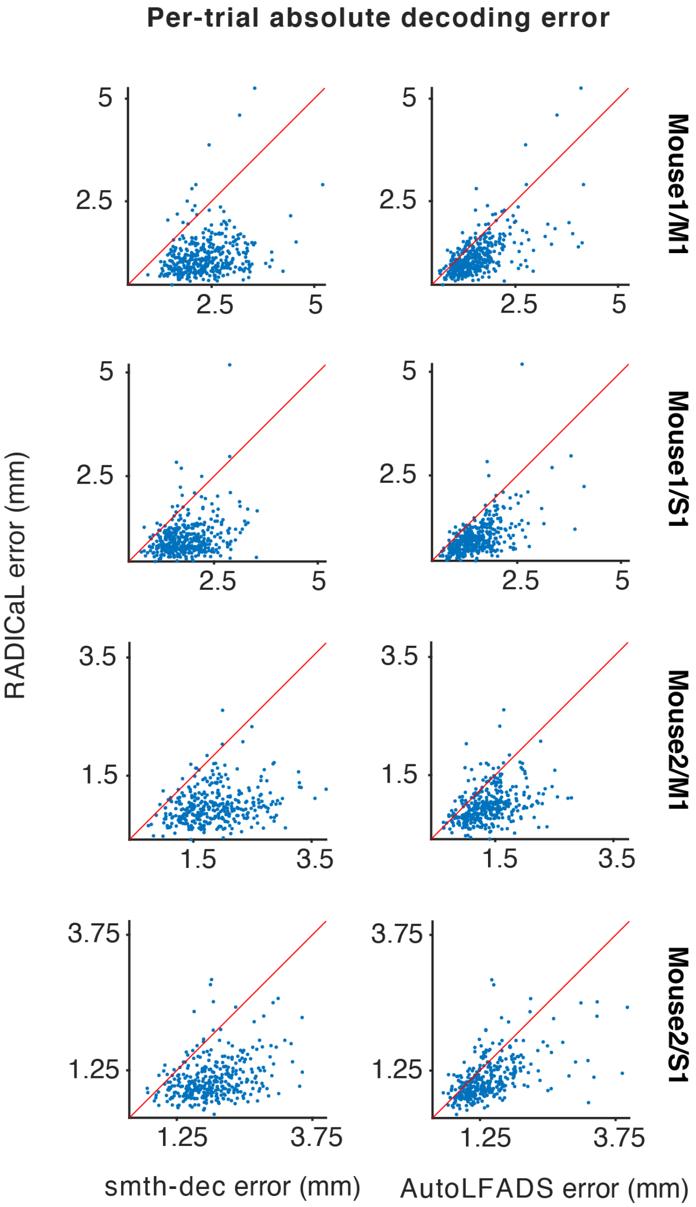


Figure B.6: RADICaL reduces decoding errors on the vast majority of single trials for all datasets. Single-trial decoding error was quantified by measuring the absolute difference between the true and decoded hand position for each individual trial. Each point represents an individual trial. Error was greatly reduced compared with both smth-dec (left) and AutoLFADS (right).

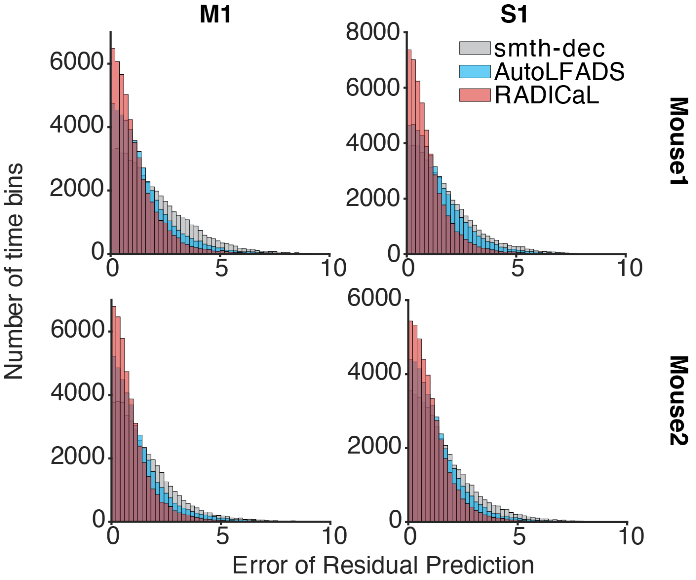


Figure B.7: RADICaL improves prediction of single-trial deviations from the mean of hand positions. This figure demonstrates that RADICaL does not simply learn a 'typical' trajectory for left-reach trials and another for right-reach trials, but instead reflects small deviations from the condition average better than other methods. The residuals of hand positions (i.e., single-trial deviations from the mean) were computed by subtracting the left-reach or right-reach trial-averaged hand positions from the single trials. Error of residual prediction was computed by taking the absolute value of the difference between true and predicted residuals of hand positions.

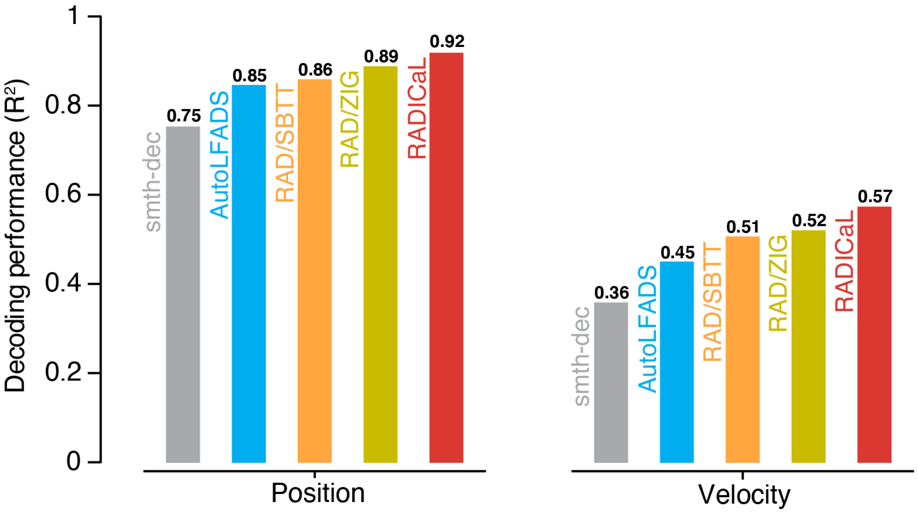


Figure B.8: Performance of ZIG-only (A-ZIG) and SBTT-only (A-SBTT) on decoding hand kinematics. To test whether the innovations of RADICaL contributed separately to the improved decoding performance, we performed an ablation study where we enabled solely the ZIG emissions model (RAD/ZIG) or SBTT (RAD/SBTT). Decoding accuracy was quantified by measuring variance explained (R^2) between the true and decoded position (left) and velocity (right) across all trials, for RAD/ZIG, RAD/SBTT and other techniques. Analyzed data are from Mouse2/M1. Note that in this test, RAD/ZIG outperformed RAD/SBTT, which is the opposite of the results from synthetic data shown in **Fig. B.2**. The discrepancy could potentially be due to different properties of the datasets, such as the frequency of the underlying features or noise properties. However, for both simulated and real data, either innovation (SBTT or ZIG) helps improve performance and combining them yields the highest performance.

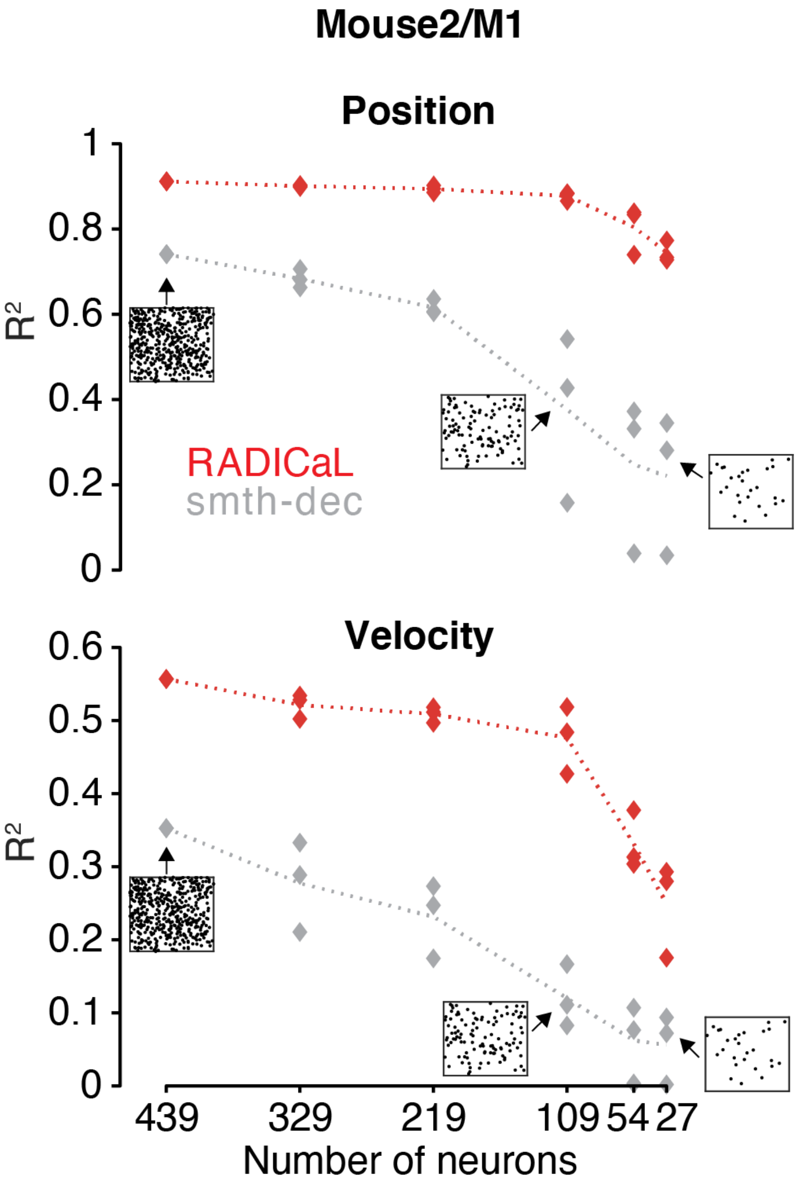


Figure B.9: RADICaL is robust to the random seed used in selecting subsets of neurons in a neuron downsampling experiment. This figure is related to Figure 6. Decoding performance measured using variance explained (R^2) as a function of the number of neurons used in each technique (top: Position; bottom: Velocity). For a given number of neurons (except the full population of 439 neurons), 3 random seeds were used, and each data point represents an individual random seed. The dotted line represents the mean performance across the three random seeds for each method. Data from Mouse2/M1. Figure insets indicate the selected neurons in the FOV for experiments of the full population and example subsets of the population.

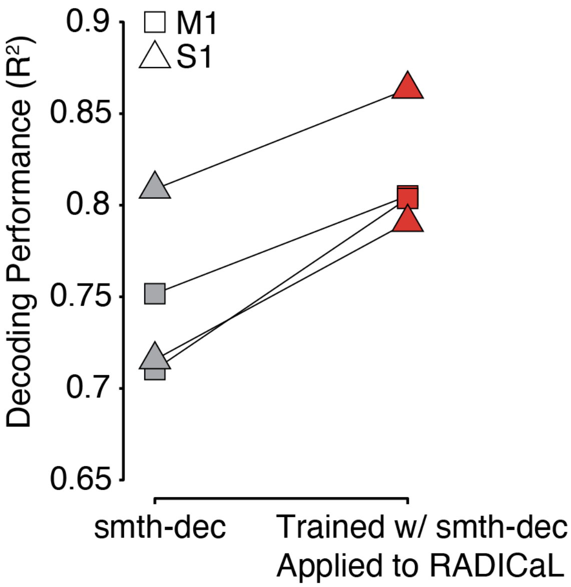


Figure B.10: RADICaL improves decoding performance using decoders trained with smth-dec rates. This analysis demonstrates that the decoding performance benefit due to RADICaL cannot be due to training a better decoder alone, but results from better denoising of the trajectories themselves. Decoding performance was quantified by measuring variance explained (R^2) between the true and decoded hand position across each of the 4 datasets (2 mice for M1, denoted by squares, and 2 mice for S1, denoted by triangles), for decoders trained with smth-dec rates and applied to smth-dec rates (gray) or applied to RADICaL rates (red).

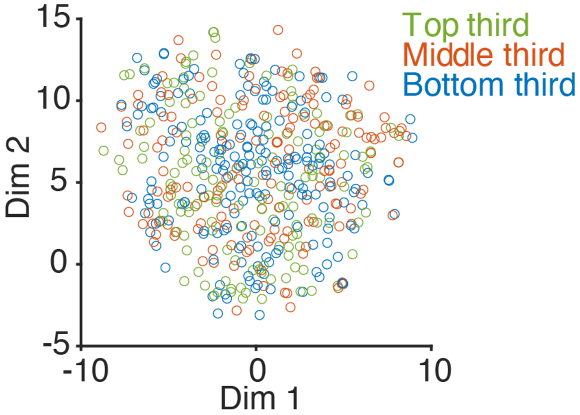


Figure B.11: Visualization of transformation from factors to neurons. This analysis demonstrates that the different bands of the image use the same factors and not segregated ones, despite being divided up into separate sub-bins for improving temporal resolution with SBTT. The plot shows a 2-dimensional t-SNE space representation of weights mapping from RADICaL factors to ZIG parameters for Mouse1/M1. Each point represents an individual neuron (510 neurons total). Neurons are color coded based on the neurons' position within the field of view (i.e., top, middle, and bottom). The interspersal of the points shows that neurons do not have systematically different relationships with the factors in RADICAL based on which band they are in.

Bibliography

K. C. Ames, S. I. Ryu, and K. V. Shenoy. Neural dynamics of reaching following incorrect or absent motor preparation. *Neuron*, 81(2):438–451, 2014.

N. Apthorpe, A. Riordan, R. Aguilar, J. Homann, Y. Gu, D. Tank, and H. S. Seung. Automatic Neuron Detection in Calcium Imaging Data Using Convolutional Networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/0771fc6f0f4b1d7d1bb73bbbe14e0e31-Paper.pdf`.

J. Art. Photon detectors for confocal microscopy. In *Handbook of biological confocal microscopy*, pages 251–264. Springer, 2006.

P. Berens, J. Freeman, T. Deneux, N. Chenkov, T. McColgan, A. Speiser, J. H. Macke, S. C. Turaga, P. Mineault, P. Rupprecht, and others. Community-based benchmarking improves spike rate inference from two-photon calcium imaging data. *PLoS computational biology*, 14(5):e1006157, 2018. Publisher: Public Library of Science.

D. A. Borton, M. Yin, J. Aceros, and A. Nurmikko. An implantable wireless neural interface for recording cortical circuit dynamics in moving primates. *Journal of neural engineering*, 10(2):026010, 2013.

F. Carnevale, V. de Lafuente, R. Romo, O. Barak, and N. Parga. Dynamic control of response criterion in premotor cortex during perceptual detection under temporal uncertainty. *Neuron*, 86(4):1067–1077, 2015.

J. L. Chen, S. Carta, J. Soldado-Magraner, B. L. Schneider, and F. Helmchen. Behaviour-dependent recruitment of long-range projection neurons in somatosensory cortex. *Nature*, 499(7458):336–340, 2013a. Publisher: Nature Publishing Group.

S. X. Chen, A. N. Kim, A. J. Peters, and T. Komiyama. Subtype-specific plasticity of inhibitory circuits in motor cortex during motor learning. *Nature neuroscience*, 18(8):1109–1115, 2015. Publisher: Nature Publishing Group.

T.-W. Chen, T. J. Wardill, Y. Sun, S. R. Pulver, S. L. Renninger, A. Baohan, E. R. Schreiter, R. A. Kerr, M. B. Orger, V. Jayaraman, et al. Ultrasensitive fluorescent proteins for imaging neuronal activity. *Nature*, 499(7458):295–300, 2013b.

M. M. Churchland and K. V. Shenoy. Temporal complexity and heterogeneity of single-neuron activity in premotor and motor cortex. *Journal of neurophysiology*, 97(6):4235–4257, 2007.

M. M. Churchland, J. P. Cunningham, M. T. Kaufman, S. I. Ryu, and K. V. Shenoy. Cortical preparatory activity: representation of movement or first cog in a dynamical machine? *Neuron*, 68(3):387–400, 2010.

M. M. Churchland, J. P. Cunningham, M. T. Kaufman, J. D. Foster, P. Nuyujukian, S. I. Ryu, and K. V. Shenoy. Neural population dynamics during reaching. *Nature*, 487(7405):51–56, 2012.

M. M. Churchland, M. T. Kaufman, J. P. Cunningham, J. D. Foster, P. Nuyujukian, S. I. Ryu, and K. V. Shenoy. Neural population dynamics during reaching dataset. https://dandiarchive.org/dandiset/000070/, 2021. [Data set]. DANDI archive.

A. C. Costa, T. Ahamed, and G. J. Stephens. Adaptive, locally linear models of complex dynamics. *Proceedings of the National Academy of Sciences*, 116(5):1501–1510, Jan. 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1813476116. URL http://www.pnas.org/lookup/doi/10.1073/pnas.1813476116.

J. P. Cunningham and B. M. Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, Nov. 2014. ISSN 1097-6256, 1546-1726. doi: 10.1038/nn.3776. URL http://www.nature.com/articles/nn.3776.

H. Dana, Y. Sun, B. Mohar, B. K. Hulse, A. M. Kerlin, J. P. Hasseman, G. Tsegaye, A. Tsang, A. Wong, R. Patel, J. J. Macklin, Y. Chen, A. Konnerth, V. Jayaraman, L. L. Looger, E. R. Schreiter, K. Svoboda, and D. S. Kim. High-performance calcium sensors for imaging activity in neuronal populations and microcompartments. *Nature Methods*, 16(7):649–657, July 2019. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-019-0435-6. URL http://www.nature.com/articles/s41592-019-0435-6.

M. A. Davenport and J. Romberg. An overview of low-rank matrix recovery from incomplete observations. *IEEE J. of Selected Topics in Signal Processing*, 10(4): 608–622, June 2016.

J. B. Dechery and J. N. MacLean. Functional triplet motifs underlie accurate predictions of single-trial responses in populations of tuned and untuned V1 neurons. *PLoS computational biology*, 14(5):e1006153, 2018. Publisher: Public Library of Science San Francisco, CA USA.

J. Demas, J. Manley, F. Tejera, H. Kim, K. Barber, F. M. Traub, B. Chen, and

A. Vaziri. Volumetric calcium imaging of 1 million neurons across cortical regions at cellular resolution using light beads microscopy. *bioRxiv*, 2021.

T. Deneux, A. Kaszas, G. Szalay, G. Katona, T. Lakner, A. Grinvald, B. Rózsa, and I. Vanzetta. Accurate spike estimation from noisy calcium signals for ultrafast three-dimensional imaging of large neuronal populations in vivo. *Nature Communications*, 7(1):12190, Nov. 2016. ISSN 2041-1723. doi: 10.1038/ncomms12190. URL `http://www.nature.com/articles/ncomms12190`.

M. Denker, A. Yegenoglu, and S. Grün. Collaborative HPC-enabled workflows on the HBP Collaboratory using the Elephant framework. In *Neuroinformatics 2018*, page P19, 2018. doi: 10.12751/incf.ni2018.0019. URL `https://abstracts.g-node.org/conference/NI2018/abstracts#/uuid/023bec4e-0c35-4563-81ce-2c6fac282abd`.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

L. Duncker, G. Bohner, J. Boussard, and M. Sahani. Learning interpretable continuous-time models of latent stochastic dynamical systems. In *International Conference on Machine Learning*, pages 1726–1734. PMLR, 2019.

G. F. Elsayed, A. H. Lara, M. T. Kaufman, M. M. Churchland, and J. P. Cunningham. Reorganization between preparatory and movement population responses in motor cortex. *Nature communications*, 7(1):1–15, 2016.

E. E. Fetz. Are movement parameters recognizably coded in the activity of single neurons? *Behavioral and brain sciences*, 15(4):679–690, 1992.

R. D. Flint, M. C. Tate, K. Li, J. W. Templer, J. M. Rosenow, C. Pandarinath, and

M. W. Slutzky. The representation of finger movement and force in human motor and premotor cortices. *Eneuro*, 7(4), 2020.

J. Freeman, N. Vladimirov, T. Kawashima, Y. Mu, N. J. Sofroniew, D. V. Bennett, J. Rosen, C.-T. Yang, L. L. Looger, and M. B. Ahrens. Mapping brain activity at scale with cluster computing. *Nature methods*, 11(9):941–950, 2014.

J. Friedrich, P. Zhou, and L. Paninski. Fast online deconvolution of calcium imaging data. *PLOS Computational Biology*, 13(3):e1005423, Mar. 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005423. URL `https://dx.plos.org/10.1371/journal.pcbi.1005423`.

G. L. Galiñanes, C. Bonardi, and D. Huber. Directional reaching for water as a cortex-dependent behavioral framework for mice. *Cell reports*, 22(10):2767–2783, 2018. Publisher: Elsevier.

J. A. Gallego, M. G. Perich, R. H. Chowdhury, S. A. Solla, and L. E. Miller. Long-term stability of cortical population dynamics underlying consistent behavior. *Nature neuroscience*, 23(2):260–270, 2020.

E. Ganmor, M. Krumin, L. F. Rossi, M. Carandini, and E. P. Simoncelli. Direct Estimation of Firing Rates from Calcium Imaging Data. *arXiv:1601.00364 [q-bio]*, Jan. 2016. URL `http://arxiv.org/abs/1601.00364`. arXiv: 1601.00364.

Y. Gao, E. Archer, L. Paninski, and J. P. Cunningham. Linear dynamical neural population models through nonlinear embeddings. *arXiv preprint arXiv:1605.08454*, 2016.

A. P. Georgopoulos, J. F. Kalaska, R. Caminiti, and J. T. Massey. On the relations between the direction of two-dimensional arm movements and cell discharge in primate motor cortex. *Journal of Neuroscience*, 2(11):1527–1537, 1982.

A. Giovannucci, J. Friedrich, M. Kaufman, A. Churchland, D. Chklovskii, L. Paninski, and E. A. Pnevmatikakis. OnACID: Online Analysis of Calcium Imaging Data in Real Time. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

A. Giovannucci, J. Friedrich, P. Gunn, J. Kalfon, B. L. Brown, S. A. Koay, J. Taxidis, F. Najafi, J. L. Gauthier, P. Zhou, B. S. Khakh, D. W. Tank, D. B. Chklovskii, and E. A. Pnevmatikakis. CaImAn an open source tool for scalable calcium imaging data analysis. *eLife*, 8:e38173, Jan. 2019.

J. Glaser, M. Whiteway, J. P. Cunningham, L. Paninski, and S. Linderman. Recurrent Switching Dynamical Systems Models for Multiple Interacting Neural Populations. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14867–14878. Curran Associates, Inc., 2020. URL `https://proceedings.neurips.cc/paper/2020/file/aa1f5f73327ba40d47ebce155e785aaf-Paper.pdf`.

S. Goering, E. Klein, L. S. Sullivan, A. Wexler, B. A. y Arcas, G. Bi, J. M. Carmena, J. J. Fins, P. Friesen, J. Gallant, et al. Recommendations for responsible development and application of neurotechnologies. *Neuroethics*, pages 1–22, 2021.

M. D. Golub, M. Y. Byron, and S. M. Chase. Internal models for interpreting neural population activity during sensorimotor control. *Elife*, 4:e10015, 2015.

M. D. Golub, P. T. Sadtler, E. R. Oby, K. M. Quick, S. I. Ryu, E. C. Tyler-Kabara, A. P. Batista, S. M. Chase, and B. M. Yu. Learning by neural reassociation. *Nature neuroscience*, 21(4):607–616, 2018.

M. Hardt, T. Ma, and B. Recht. Gradient descent learns linear dynamical systems.

Journal of Machine Learning Research, 19(29):1–44, 2018. URL `http://jmlr.org/papers/v19/16-465.html`.

C. D. Harvey, P. Coen, and D. W. Tank. Choice-specific sequences in parietal cortex during a virtual-navigation decision task. *Nature*, 484(7392):62–68, 2012.

N. G. Hatsopoulos, Q. Xu, and Y. Amit. Encoding of movement fragments in the motor cortex. *Journal of Neuroscience*, 27(19):5105–5114, 2007. Publisher: Soc Neuroscience.

E. Hazan, H. Lee, K. Singh, C. Zhang, and Y. Zhang. Spectral filtering for general linear dynamical systems. In *Adv. in Neural Information Processing Systems (NeurIPS)*, Montréal, Canada, Dec. 2018.

J. Heikkila and O. Silvén. A four-step camera calibration procedure with implicit image correction. In *Proceedings of IEEE computer society conference on computer vision and pattern recognition*, pages 1106–1112. IEEE, 1997.

D. Hernandez, A. K. Moretti, Z. Wei, S. Saxena, J. Cunningham, and L. Paninski. A novel variational family for hidden nonlinear markov models. *arXiv preprint arXiv:1811.02459*, 2018.

D. J. Herzfeld, Y. Kojima, R. Soetedjo, and R. Shadmehr. Encoding of error and learning to correct that error by the Purkinje cells of the cerebellum. *Nature neuroscience*, 21(5):736–743, 2018. Publisher: Nature Publishing Group.

B. Hidasi, A. Karatzoglou, L. Baltrunas, and D. Tikk. Session-based recommendations with recurrent neural networks. Nov. 2015.

H. Hoang, M.-a. Sato, S. Shinomoto, S. Tsutsumi, M. Hashizume, T. Ishikawa, M. Kano, Y. Ikegaya, K. Kitamura, M. Kawato, and K. Toyama. Improved hyperacuity estimation of spike timing from calcium imaging. *Scientific Reports*, 10

(1):17844, Dec. 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-74672-y. URL `http://www.nature.com/articles/s41598-020-74672-y`.

D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574, 1959.

C. Hurwitz, K. Xu, A. Srivastava, A. Buccino, and M. Hennig. Scalable spike source localization in extracellular recordings using amortized variational inference. *Advances in Neural Information Processing Systems*, 32:4724–4736, 2019.

H. Inan, M. A. Erdogdu, and M. Schnitzer. Robust Estimation of Neural Signals in Calcium Imaging. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/e449b9317dad920c0dd5ad0a2a2d5e49-Paper.pdf`.

M. Jaderberg, V. Dalibard, S. Osindero, W. M. Czarnecki, J. Donahue, A. Razavi, O. Vinyals, T. Green, I. Dunning, K. Simonyan, C. Fernando, and K. Kavukcuoglu. Population Based Training of Neural Networks. *arXiv:1711.09846 [cs]*, Nov. 2017. URL `http://arxiv.org/abs/1711.09846`. arXiv: 1711.09846.

M. Jas, T. Achakulvisut, A. Idrizović, D. Acuna, M. Antalek, V. Marques, T. Odland, R. Garg, M. Agrawal, Y. Umegaki, P. Foley, H. Fernandes, D. Harris, B. Li, O. Pieters, S. Otterson, G. D. Toni, C. Rodgers, E. Dyer, M. Hamalainen, K. Kording, and P. Ramkumar. Pyglmnet: Python implementation of elastic-net regularized generalized linear models. *Journal of Open Source Software*, 5(47):1959, 2020. doi: 10.21105/joss.01959. URL `https://doi.org/10.21105/joss.01959`.

S. Jewell and D. Witten. Exact spike train inference via 0 optimization. *The annals of applied statistics*, 12(4):2457, 2018.

S. W. Jewell, T. D. Hocking, P. Fearnhead, and D. M. Witten. Fast nonconvex deconvolution of calcium imaging data. *Biostatistics*, 21(4):709–726, Oct. 2020. ISSN 1465-4644, 1468-4357. doi: 10.1093/biostatistics/kxy083. URL `https://academic.oup.com/biostatistics/article/21/4/709/5310127`.

J. J. Jun, N. A. Steinmetz, J. H. Siegle, D. J. Denman, M. Bauza, B. Barbarits, A. K. Lee, C. A. Anastassiou, A. Andrei, Aydın, and others. Fully integrated silicon probes for high-density recording of neural activity. *Nature*, 551(7679):232–236, 2017. Publisher: Nature Publishing Group.

P. Kaifosh, J. D. Zaremba, N. B. Danielson, and A. Losonczy. SIMA: Python software for analysis of dynamic fluorescence imaging data. *Frontiers in Neuroinformatics*, 8, Sept. 2014. ISSN 1662-5196. doi: 10.3389/fninf.2014.00080. URL `http://journal.frontiersin.org/article/10.3389/fninf.2014.00080/abstract`.

J. C. Kao, P. Nuyujukian, S. I. Ryu, M. M. Churchland, J. P. Cunningham, and K. V. Shenoy. Single-trial dynamics of motor cortex and their applications to brain-machine interfaces. *Nature communications*, 6(1):1–12, 2015.

J. C. Kao, S. I. Ryu, and K. V. Shenoy. Leveraging neural dynamics to extend functional lifetime of brain-machine interfaces. *Scientific reports*, 7(1):1–16, 2017.

M. T. Kaufman, M. M. Churchland, S. I. Ryu, and K. V. Shenoy. Cortical activity in the null space: permitting preparation without movement. *Nature neuroscience*, 17(3):440–448, 2014.

M. T. Kaufman, J. S. Seely, D. Sussillo, S. I. Ryu, K. V. Shenoy, and M. M. Churchland. The largest response component in the motor cortex reflects movement timing but not movement type. *Eneuro*, 3(4), 2016. Publisher: Society for Neuroscience.

S. W. Keemink, S. C. Lowe, J. M. P. Pakan, E. Dylda, M. C. W. van Rossum, and N. L. Rochefort. FISSA: A neuropil decontamination toolbox for calcium imaging signals. *Scientific Reports*, 8(1):3493, Dec. 2018. ISSN 2045-2322. doi: 10.1038/s41598-018-21640-2. URL http://www.nature.com/articles/s41598-018-21640-2.

M. R. Keshtkaran and C. Pandarinath. Enabling hyperparameter optimization in sequential autoencoders for spiking neural data. In *Advances in Neural Information Processing Systems*, pages 15937–15947, 2019.

M. R. Keshtkaran, A. R. Sedler, R. H. Chowdhury, R. Tandon, D. Basrai, S. L. Nguyen, H. Sohn, M. Jazayeri, L. E. Miller, and C. Pandarinath. A large-scale neural network training framework for generalized estimation of single-trial population dynamics. *bioRxiv*, 2021. Publisher: Cold Spring Harbor Laboratory.

T. D. Kim, T. Z. Luo, J. W. Pillow, and C. D. Brody. Inferring latent dynamics underlying neural population activity via neural differential equations. In *Proceedings of the 38th International Conference on Machine Learning*, 2021.

E. Kirschbaum, M. Haußmann, S. Wolf, H. Sonntag, J. Schneider, S. Elzoheiry, O. Kann, D. Durstewitz, and F. A. Hamprecht. LeMoNADe: Learned Motif and Neuronal Assembly Detection in calcium imaging videos. *arXiv:1806.09963 [q-bio]*, Feb. 2019. URL http://arxiv.org/abs/1806.09963. arXiv: 1806.09963.

E. Kirschbaum, A. Bailoni, and F. A. Hamprecht. DISCo: Deep Learning, Instance Segmentation, and Correlations for Cell Segmentation in Calcium Imaging. In A. L. Martel, P. Abolmaesumi, D. Stoyanov, D. Mateus, M. A. Zuluaga, S. K. Zhou, D. Racoceanu, and L. Joskowicz, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, volume 12265, pages 151–162. Springer International Publishing, Cham, 2020. ISBN 978-3-030-59721-4 978-3-030-59722-1. doi: 10.1007/978-3-030-59722-1_15. URL http://link.springer.

com/10.1007/978-3-030-59722-1_15. Series Title: Lecture Notes in Computer Science.

E. Klein, T. Brown, M. Sample, A. R. Truitt, and S. Goering. Engineering the brain: ethical issues and the introduction of neural devices. *Hastings Center Report*, 45 (6):26–35, 2015.

T. H. Koh, W. E. Bishop, T. Kawashima, B. B. Jeon, R. Srinivasan, S. J. Kuhlman, M. B. Ahrens, S. M. Chase, and M. Y. Byron. Dimensionality reduction of calcium-imaged neuronal population activity. *bioRxiv*, 2022.

J. W. Krakauer, A. A. Ghazanfar, A. Gomez-Marin, M. A. MacIver, and D. Poeppel. Neuroscience needs behavior: correcting a reductionist bias. *Neuron*, 93(3):480–490, 2017. Publisher: Elsevier.

K. C. Lakshmanan, P. T. Sadtler, E. C. Tyler-Kabara, A. P. Batista, and B. M. Yu. Extracting low-dimensional latent structure from time series in the presence of delays. *Neural computation*, 27(9):1825–1856, 2015.

J. Lecoq, M. Oliver, J. H. Siegle, N. Orlova, and C. Koch. Removing independent noise in systems neuroscience data using deepinterpolation. *bioRxiv*, 2020.

H. Lee and C. Zhang. Robust guarantees for learning an autoregressive filter. In A. Kontorovich and G. Neu, editors, *Proceedings of the 31st International Conference on Algorithmic Learning Theory*, volume 117 of *Proceedings of Machine Learning Research*, pages 490–517, San Diego, California, USA, Feb. 2020. PMLR. URL http://proceedings.mlr.press/v117/lee20a.html.

S. Linderman, M. Johnson, A. Miller, R. Adams, D. Blei, and L. Paninski. Bayesian learning and inference in recurrent switching linear dynamical systems. In *Artificial Intelligence and Statistics*, pages 914–922. PMLR, 2017.

J. Lu, C. Li, J. Singh-Alvarado, Z. C. Zhou, F. Fröhlich, R. Mooney, and F. Wang. MIN1PIPE: A Miniscope 1-Photon-Based Calcium Imaging Signal Extraction Pipeline. *Cell Reports*, 23(12):3673–3684, June 2018. ISSN 22111247. doi: 10.1016/j.celrep.2018.05.062. URL https://linkinghub.elsevier.com/retrieve/pii/S221112471830826X.

J. H. Macke, L. Buesing, J. P. Cunningham, B. M. Yu, K. V. Shenoy, and M. Sahani. Empirical models of spiking in neural populations. In *Advances in Neural Information Processing Systems 24: 25th conference on Neural Information Processing Systems (NIPS 2011)*, pages 1350–1358, 2012.

E. L. Mackevicius, A. H. Bahle, A. H. Williams, S. Gu, N. I. Denisenko, M. S. Goldman, and M. S. Fee. Unsupervised discovery of temporal sequences in high-dimensional datasets, with applications to neuroscience. *eLife*, 8:e38471, Feb. 2019. ISSN 2050-084X. doi: 10.7554/eLife.38471. URL https://elifesciences.org/articles/38471.

O. Mano, M. S. Creamer, C. A. Matulis, E. Salazar-Gatzimas, J. Chen, J. A. Zavatone-Veth, and D. A. Clark. Using slow frame rate imaging to extract fast receptive fields. *Nature communications*, 10(1):1–13, 2019. Publisher: Nature Publishing Group.

V. Mante, D. Sussillo, K. V. Shenoy, and W. T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *nature*, 503(7474):78–84, 2013.

R. Maruyama, K. Maeda, H. Moroda, I. Kato, M. Inoue, H. Miyakawa, and T. Aonishi. Detecting cells using non-negative matrix factorization on calcium imaging data. *Neural Networks*, 55:11–19, July 2014. ISSN 08936080. doi: 10.1016/j.neunet.2014.03.007. URL https://linkinghub.elsevier.com/retrieve/pii/S0893608014000707.

A. Mathis, P. Mamidanna, K. M. Cury, T. Abe, V. N. Murthy, M. W. Mathis, and M. Bethge. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289, 2018. Publisher: Nature Publishing Group.

M. Minderer, K. D. Brown, and C. D. Harvey. The spatial structure of neural encoding in mouse posterior cortex during navigation. *Neuron*, 102(1):232–248, 2019. Publisher: Elsevier.

H. Miranda, V. Gilja, C. A. Chestek, K. V. Shenoy, and T. H. Meng. Hermesd: A high-rate long-range wireless transmission system for simultaneous multichannel neural recording applications. *IEEE Transactions on Biomedical Circuits and Systems*, 4 (3):181–191, 2010.

G. Mishne and A. S. Charles. Learning Spatially-correlated Temporal Dictionaries for Calcium Imaging. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1065–1069, Brighton, United Kingdom, May 2019. IEEE. ISBN 978-1-4799-8131-1. doi: 10.1109/ICASSP.2019.8683375. URL `https://ieeexplore.ieee.org/document/8683375/`.

G. Mishne, R. R. Coifman, M. Lavzin, and J. Schiller. Automated cellular structure extraction in biological images with applications to calcium imaging data. preprint, Neuroscience, May 2018. URL `http://biorxiv.org/lookup/doi/10.1101/313981`.

J. S. Montijn, G. T. Meijer, C. S. Lansink, and C. M. Pennartz. Population-level neural codes are robust to single-neuron variability from a multidimensional coding perspective. *Cell reports*, 16(9):2486–2498, 2016. Publisher: Elsevier.

A. S. Morcos and C. D. Harvey. History-dependent variability in population dynamics during evidence accumulation in cortex. *Nature neuroscience*, 19(12):1672–1681, 2016.

E. A. Mukamel, A. Nimmerjahn, and M. J. Schnitzer. Automated Analysis of Cellular Signals from Large-Scale Calcium Imaging Data. *Neuron*, 63(6):747–760, Sept. 2009. ISSN 08966273. doi: 10.1016/j.neuron.2009.08.009. URL `https://linkinghub.elsevier.com/retrieve/pii/S0896627309006199`.

S. Musall, M. T. Kaufman, A. L. Juavinett, S. Gluf, and A. K. Churchland. Single-trial neural dynamics are dominated by richly varied movements. *Nature neuroscience*, 22(10):1677–1686, 2019. Publisher: Nature Publishing Group.

E. Musk et al. An integrated brain-machine interface platform with thousands of channels. *Journal of medical Internet research*, 21(10):e16194, 2019.

M. Nonnenmacher, S. C. Turaga, and J. H. Macke. Extracting low-dimensional dynamics from multiple large-scale neural population recordings by learning to predict correlations. *arXiv preprint arXiv:1711.01847*, 2017.

S. Oymak and N. Ozay. Non-asymptotic identification of lti systems from a single trajectory. In *2019 American Control Conference (ACC)*, pages 5655–5661, 2019. doi: 10.23919/ACC.2019.8814438.

J. Oñativia, S. R. Schultz, and P. L. Dragotti. A finite rate of innovation algorithm for fast and accurate spike detection from two-photon calcium imaging. *Journal of Neural Engineering*, 10(4):046017, Aug. 2013. ISSN 1741-2560, 1741-2552. doi: 10.1088/1741-2560/10/4/046017. URL `https://iopscience.iop.org/article/10.1088/1741-2560/10/4/046017`.

M. Pachitariu, A. M. Packer, N. Pettit, H. Dalgleish, M. Hausser, and M. Sahani. Extracting regions of interest from biological images with convolutional sparse block coding. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper/2013/file/1f50893f80d6830d62765ffad7721742-Paper.pdf.

M. Pachitariu, C. Stringer, M. Dipoppa, S. Schröder, L. F. Rossi, H. Dalgleish, M. Carandini, and K. D. Harris. Suite2p: beyond 10,000 neurons with standard two-photon microscopy. *BioRxiv*, page 061507, 2017.

M. Pachitariu, C. Stringer, and K. D. Harris. Robustness of spike deconvolution for neuronal calcium imaging. *Journal of Neuroscience*, 38(37):7976–7985, 2018. Publisher: Soc Neuroscience.

C. Pandarinath, K. C. Ames, A. A. Russo, A. Farshchian, L. E. Miller, E. L. Dyer, and J. C. Kao. Latent Factors and Dynamics in Motor Cortex and Their Application to Brain–Machine Interfaces. *Journal of Neuroscience*, 38(44):9390–9401, Oct. 2018a. ISSN 0270-6474, 1529-2401. doi: 10.1523/JNEUROSCI.1669-18.2018. URL https://www.jneurosci.org/content/38/44/9390. Publisher: Society for Neuroscience Section: Symposium and Mini-Symposium.

C. Pandarinath, D. J. O'Shea, J. Collins, R. Jozefowicz, S. D. Stavisky, J. C. Kao, E. M. Trautmann, M. T. Kaufman, S. I. Ryu, L. R. Hochberg, J. M. Henderson, K. V. Shenoy, L. F. Abbott, and D. Sussillo. Inferring single-trial neural population dynamics using sequential auto-encoders. *Nature Methods*, 15(10):805–815, Oct. 2018b. ISSN 1548-7091, 1548-7105. doi: 10.1038/s41592-018-0109-9. URL http://www.nature.com/articles/s41592-018-0109-9.

L. Paninski. Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, 15(4):243–262, 2004.

F. Pei, J. Ye, D. Zoltowski, D. Zoltowski, A. Wu, R. Chowdhury, H. Sohn,

J. O' Doherty, K. V. Shenoy, M. Kaufman, M. Churchland, M. Jazayeri, L. Miller, J. Pillow, I. M. Park, E. Dyer, and C. Pandarinath. Neural Latents Benchmark '21: Evaluating latent variable models of neural population activity. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL `https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/file/979d472a84804b9f647bc185a877a8b5-Paper-round2.pdf`.

S. Peron, T.-W. Chen, and K. Svoboda. Comprehensive imaging of cortical networks. *Current opinion in neurobiology*, 32:115–123, 2015a.

S. P. Peron, J. Freeman, V. Iyer, C. Guo, and K. Svoboda. A cellular resolution map of barrel cortex activity during tactile behavior. *Neuron*, 86(3):783–799, 2015b.

B. Petreska, B. M. Yu, J. P. Cunningham, G. Santhanam, S. Ryu, K. V. Shenoy, and M. Sahani. Dynamical segmentation of single trials from population neural data. *Advances in neural information processing systems*, 24:756–764, 2011.

M. Picardo, J. Merel, K. Katlowitz, D. Vallentin, D. Okobi, S. Benezra, R. Clary, E. Pnevmatikakis, L. Paninski, and M. Long. Population-Level Representation of a Temporal Sequence Underlying Song Production in the Zebra Finch. *Neuron*, 90 (4):866–876, May 2016. ISSN 08966273. doi: 10.1016/j.neuron.2016.02.016. URL `https://linkinghub.elsevier.com/retrieve/pii/S0896627316001094`.

C. Piech, J. Bassen, J. Huang, S. Ganguli, M. Sahami, L. J. Guibas, and J. Sohl-Dickstein. Deep knowledge tracing. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28. Curran Associates, Inc., 2015. URL `https://proceedings.neurips.cc/paper/2015/file/bac9162b47c56fc8a4d2a519803d51b3-Paper.pdf`.

E. Pnevmatikakis, D. Soudry, Y. Gao, T. A. Machado, J. Merel, D. Pfau, T. Reardon, Y. Mu, C. Lacefield, W. Yang, M. Ahrens, R. Bruno, T. M. Jessell, D. Peterka, R. Yuste, and L. Paninski. Simultaneous Denoising, Deconvolution, and Demixing of Calcium Imaging Data. *Neuron*, 89(2):285–299, Jan. 2016. ISSN 08966273. doi: 10.1016/j.neuron.2015.11.037. URL `https://linkinghub.elsevier.com/retrieve/pii/S0896627315010843`.

E. A. Pnevmatikakis. Analysis pipelines for calcium imaging data. *Current opinion in neurobiology*, 55:15–21, 2019a.

E. A. Pnevmatikakis. Analysis pipelines for calcium imaging data. *Current Opinion in Neurobiology*, 55:15–21, Apr. 2019b. ISSN 09594388. doi: 10.1016/j.conb.2018.11.004. URL `https://linkinghub.elsevier.com/retrieve/pii/S0959438818300941`.

E. A. Pnevmatikakis, J. Merel, A. Pakman, and L. Paninski. Bayesian spike inference from calcium imaging data. In *2013 Asilomar Conference on Signals, Systems and Computers*, pages 349–353, Pacific Grove, CA, USA, Nov. 2013. IEEE. ISBN 978-1-4799-2390-8 978-1-4799-2388-5. doi: 10.1109/ACSSC.2013.6810293. URL `http://ieeexplore.ieee.org/document/6810293/`.

L. Y. Prince, S. Bakhtiari, C. J. Gillon, and B. A. Richards. Parallel inference of hierarchical latent dynamics in two-photon calcium imaging of neuronal populations. preprint, Neuroscience, Mar. 2021. URL `http://biorxiv.org/lookup/doi/10.1101/2021.03.05.434105`.

B. C. Raducanu, R. F. Yazicioglu, C. M. Lopez, M. Ballini, J. Putzeys, S. Wang, A. Andrei, V. Rochus, M. Welkenhuysen, N. v. Helleputte, et al. Time multiplexed active neural probe with 1356 parallel recording sites. *Sensors*, 17(10):2388, 2017.

D. Raposo, M. T. Kaufman, and A. K. Churchland. A category-free neural population supports evolving demands during decision-making. *Nature neuroscience*, 17(12): 1784–1792, 2014.

E. D. Remington, D. Narain, E. A. Hosseini, and M. Jazayeri. Flexible sensorimotor computations through rapid reconfiguration of cortical dynamics. *Neuron*, 98(5): 1005–1019, 2018.

S. Reynolds, T. Abrahamsson, R. Schuck, P. J. Sjöström, S. R. Schultz, and P. L. Dragotti. ABLE: An Activity-Based Level Set Segmentation Algorithm for Two-Photon Calcium Imaging Data. *eneuro*, 4(5):ENEURO.0012–17.2017, Sept. 2017. ISSN 2373-2822. doi: 10.1523/ENEURO.0012-17.2017. URL `http://eneuro.org/lookup/doi/10.1523/ENEURO.0012-17.2017`.

S. A. Romano, V. Pérez-Schuster, A. Jouary, J. Boulanger-Weill, A. Candeo, T. Pietri, and G. Sumbre. An integrated calcium imaging processing toolbox for the analysis of neuronal population dynamics. *PLOS Computational Biology*, 13(6):e1005526, June 2017. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1005526. URL `https://dx.plos.org/10.1371/journal.pcbi.1005526`.

P. Rupprecht, S. Carta, A. Hoffmann, M. Echizen, A. Blot, A. C. Kwan, Y. Dan, S. B. Hofer, K. Kitamura, F. Helmchen, and others. A database and deep learning toolbox for noise-optimized, generalized spike inference from calcium imaging. *Nature Neuroscience*, 24(9):1324–1337, 2021. Publisher: Nature Publishing Group.

A. A. Russo, S. R. Bittner, S. M. Perkins, J. S. Seely, B. M. London, A. H. Lara, A. Miri, N. J. Marshall, A. Kohn, T. M. Jessell, et al. Motor cortex embeds muscle-like commands in an untangled population response. *Neuron*, 97(4):953–966, 2018.

P. T. Sadtler, K. M. Quick, M. D. Golub, S. M. Chase, S. I. Ryu, E. C. Tyler-Kabara, B. M. Yu, and A. P. Batista. Neural constraints on learning. *Nature*, 512(7515): 423–426, 2014.

K. Sahasrabuddhe, A. A. Khan, A. P. Singh, T. M. Stern, Y. Ng, A. Tadić, P. Orel, C. LaReau, D. Pouzzner, K. Nishimura, et al. The argo: A high channel count recording system for neural recording in vivo. *Journal of Neural Engineering*, 18 (1):015002, 2021.

S. Saxena, I. Kinsella, S. Musall, S. H. Kim, J. Meszaros, D. N. Thibodeaux, C. Kim, J. Cunningham, E. M. C. Hillman, A. Churchland, and L. Paninski. Localized semi-nonnegative matrix factorization (LocaNMF) of widefield calcium imaging data. *PLOS Computational Biology*, 16(4):e1007791, Apr. 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007791. URL `https://dx.plos.org/10.1371/journal.pcbi.1007791`.

V. Scheuss, R. Yasuda, A. Sobczyk, and K. Svoboda. Nonlinear [ca2+] signaling in dendrites and spines caused by activity-dependent depression of ca2+ extrusion. *Journal of Neuroscience*, 26(31):8183–8194, 2006.

M. Schimel, T.-C. Kao, K. T. Jensen, and G. Hennequin. ilqr-vae: control-based learning of input-driven dynamics with applications to neural data. *bioRxiv*, pages 2021–10, 2022.

J. Sebastian, M. Sur, H. A. Murthy, and M. Magimai-Doss. Signal-to-signal neural networks for improved spike estimation from calcium imaging data. *PLOS Computational Biology*, 17(3):e1007921, Mar. 2021. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1007921. URL `https://dx.plos.org/10.1371/journal.pcbi.1007921`.

Q. She and A. Wu. Neural dynamics discovery via gaussian process recurrent neural networks. In *Uncertainty in Artificial Intelligence*, pages 454–464. PMLR, 2020.

K. V. Shenoy, M. Sahani, and M. M. Churchland. Cortical Control of Arm Movements: A Dynamical Systems Perspective. *Annual Review of Neuroscience*, 36(1):337–359, July 2013. ISSN 0147-006X, 1545-4126. doi: 10.1146/

annurev-neuro-062111-150509. URL `http://www.annualreviews.org/doi/10.1146/annurev-neuro-062111-150509`.

J. H. Siegle, P. Ledochowitsch, X. Jia, D. J. Millman, G. K. Ocker, S. Caldejon, L. Casal, A. Cho, D. J. Denman, S. Durand, et al. Reconciling functional differences in populations of neurons recorded with two-photon imaging and electrophysiology. *Elife*, 10:e69068, 2021.

M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht. Learning without mixing: Towards a sharp analysis of linear system identification. In *Proc. Conference on Learning Theory (COLT)*, volume 75 of *Proceedings of Machine Learning Research*, pages 439–473. PMLR, July 2018.

M. Simchowitz, R. Boczar, and B. Recht. Learning linear dynamical systems with semi-parametric least squares. In A. Beygelzimer and D. Hsu, editors, *Proc. Conference on Learning Theory (COLT)*, volume 99 of *Proceedings of Machine Learning Research*, pages 2714–2802, Phoenix, USA, June 2019. PMLR. URL `http://proceedings.mlr.press/v99/simchowitz19a.html`.

M. Simchowitz, K. Singh, and E. Hazan. Improper learning for non-stochastic control. In J. Abernethy and S. Agarwal, editors, *Proc. Conference on Learning Theory (COLT)*, volume 125 of *Proceedings of Machine Learning Research*, pages 3320–3436. PMLR, July 2020. URL `http://proceedings.mlr.press/v125/simchowitz20a.html`.

J. D. Simeral, T. Hosman, J. Saab, S. N. Flesher, M. Vilela, B. Franco, J. Kelemen, D. M. Brandman, J. G. Ciancibello, P. G. Rezaii, et al. Home use of a percutaneous wireless intracortical brain-computer interface by individuals with tetraplegia. *IEEE Transactions on Biomedical Engineering*, 2021.

N. J. Sofroniew, D. Flickinger, J. King, and K. Svoboda. A large field of view two-photon mesoscope with subcellular resolution for in vivo imaging. *Elife*, 5:e14472, 2016. Publisher: eLife Sciences Publications Limited.

S. Soltanian-Zadeh, K. Sahingur, S. Blau, Y. Gong, and S. Farsiu. Fast and robust active neuron segmentation in two-photon calcium imaging using spatiotemporal deep learning. *Proceedings of the National Academy of Sciences*, 116(17):8554–8563, Apr. 2019. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1812995116. URL `http://www.pnas.org/lookup/doi/10.1073/pnas.1812995116`.

A. Speiser, J. Yan, E. W. Archer, L. Buesing, S. C. Turaga, and J. H. Macke. Fast amortized inference of neural activity from calcium imaging data with variational autoencoders. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/5d6646aad9bcc0be55b2c82f69750387-Paper.pdf`.

J.-L. Starck, F. D. Murtagh, and A. Bijaoui. *Image processing and data analysis: the multiscale approach*. Cambridge University Press, 1998.

N. A. Steinmetz, P. Zatka-Haas, M. Carandini, and K. D. Harris. Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576(7786):266–273, 2019. Publisher: Nature Publishing Group.

N. A. Steinmetz, C. Aydin, A. Lebedeva, M. Okun, M. Pachitariu, M. Bauza, M. Beau, J. Bhagat, C. Böhm, M. Broux, et al. Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, 372(6539): eabf4588, 2021.

I. H. Stevenson and K. P. Kording. How advances in neural recording affect data analysis. *Nature neuroscience*, 14(2):139–142, 2011.

C. Stringer, M. Pachitariu, N. Steinmetz, M. Carandini, and K. D. Harris. High-dimensional geometry of population responses in visual cortex. *Nature*, 571(7765): 361–365, July 2019a. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-019-1346-5. URL `http://www.nature.com/articles/s41586-019-1346-5`.

C. Stringer, M. Pachitariu, N. Steinmetz, C. B. Reddy, M. Carandini, and K. D. Harris. Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, 364 (6437), 2019b. Publisher: American Association for the Advancement of Science.

D. Sussillo, R. Jozefowicz, L. Abbott, and C. Pandarinath. LFADS-latent factor analysis via dynamical systems. *arXiv preprint arXiv:1608.06315*, 2016.

L. Theis, P. Berens, E. Froudarakis, J. Reimer, M. Román Rosón, T. Baden, T. Euler, A. Tolias, and M. Bethge. Benchmarking Spike Rate Inference in Population Calcium Imaging. *Neuron*, 90(3):471–482, May 2016. ISSN 08966273. doi: 10.1016/j.neuron.2016.04.014. URL `https://linkinghub.elsevier.com/retrieve/pii/S0896627316300733`.

L. Tian, S. A. Hires, T. Mao, D. Huber, M. E. Chiappe, S. H. Chalasani, L. Petreanu, J. Akerboom, S. A. McKinney, E. R. Schreiter, et al. Imaging neural activity in worms, flies and mice with improved gcamp calcium indicators. *Nature methods*, 6 (12):875–881, 2009.

M. A. Triplett, Z. Pujic, B. Sun, L. Avitan, and G. J. Goodhill. Model-based decoupling of evoked and spontaneous neural activity in calcium imaging data. *PLOS Computational Biology*, 16(11):e1008330, Nov. 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008330. URL `https://dx.plos.org/10.1371/journal.pcbi.1008330`.

A. Tsiamis and G. J. Pappas. Finite sample analysis of stochastic system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3648–3654, 2019. doi: 10.1109/CDC40024.2019.9029499.

J. T. Vogelstein, B. O. Watson, A. M. Packer, R. Yuste, B. Jedynak, and L. Paninski. Spike Inference from Calcium Imaging Using Sequential Monte Carlo Methods. *Biophysical Journal*, 97(2):636–655, July 2009. ISSN 00063495. doi: 10.1016/j.bpj.2008.08.005. URL `https://linkinghub.elsevier.com/retrieve/pii/S0006349509003117`.

J. T. Vogelstein, A. M. Packer, T. A. Machado, T. Sippy, B. Babadi, R. Yuste, and L. Paninski. Fast Nonnegative Deconvolution for Spike Train Inference From Population Calcium Imaging. *Journal of Neurophysiology*, 104(6):3691–3704, Dec. 2010. ISSN 0022-3077, 1522-1598. doi: 10.1152/jn.01073.2009. URL `https://www.physiology.org/doi/10.1152/jn.01073.2009`.

S. Vyas, M. D. Golub, D. Sussillo, and K. V. Shenoy. Computation through neural population dynamics. *Annual Review of Neuroscience*, 43:249, 2020a.

S. Vyas, D. J. O'Shea, S. I. Ryu, and K. V. Shenoy. Causal role of motor preparation during error-driven learning. *Neuron*, 106(2):329–339, 2020b. Publisher: Elsevier.

X.-X. Wei, D. Zhou, A. Grosmark, Z. Ajabi, F. Sparks, P. Zhou, M. Brandon, A. Losonczy, and L. Paninski. A zero-inflated gamma model for deconvolved calcium imaging traces. *arXiv preprint arXiv:2006.03737*, 2020a.

Z. Wei, B.-J. Lin, T.-W. Chen, K. Daie, K. Svoboda, and S. Druckmann. A comparison of neuronal population dynamics measured with calcium imaging and electrophysiology. *PLoS computational biology*, 16(9):e1008198, 2020b.

I. Q. Whishaw, J. Faraji, J. Kuntz, B. M. Agha, M. Patel, G. A. Metz, and M. H. Mohajerani. Organization of the reach and grasp in head-fixed vs freely-moving mice provides support for multiple motor channel theory of neocortical organization. *Experimental brain research*, 235(6):1919–1932, 2017. Publisher: Springer.

A. H. Williams, T. H. Kim, F. Wang, S. Vyas, S. I. Ryu, K. V. Shenoy, M. Schnitzer, T. G. Kolda, and S. Ganguli. Unsupervised Discovery of Demixed, Low-Dimensional Neural Dynamics across Multiple Timescales through Tensor Component Analysis. *Neuron*, 98(6):1099–1115.e8, June 2018. ISSN 08966273. doi: 10.1016/j.neuron.2018.05.015. URL `https://linkinghub.elsevier.com/retrieve/pii/S0896627318303878`.

A. B. Wiltschko, T. Tsukahara, A. Zeine, R. Anyoha, W. F. Gillis, J. E. Markowitz, R. E. Peterson, J. Katon, M. J. Johnson, and S. R. Datta. Revealing the structure of pharmacobehavioral space through motion sequencing. *Nature neuroscience*, 23 (11):1433–1443, 2020. Publisher: Nature Publishing Group.

L. N. Wimalasena, J. F. Braun, M. R. Keshtkaran, D. Hofmann, J. Á. Gallego, C. Alessandro, M. C. Tresch, L. E. Miller, and C. Pandarinath. Estimating muscle activation from emg using deep learning-based dynamical systems models. *Journal of Neural Engineering*, 19(3):036013, 2022.

A. Wu, N. A. Roy, S. Keeley, and J. W. Pillow. Gaussian process based nonlinear latent structure discovery in multivariate spike train data. *Advances in neural information processing systems*, 30:3496, 2017a.

A. Wu, S. Pashkovski, S. R. Datta, and J. W. Pillow. Learning a latent manifold of odor representations from neural responses in piriform cortex. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper/2018/file/17b3c7061788dbe82de5abe9f6fe22b3-Paper.pdf`.

C.-Y. Wu, A. Ahmed, A. Beutel, A. J. Smola, and H. Jing. Recurrent recommender networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 495–503, New York, NY, USA, 2017b. Association for Computing Machinery. ISBN 9781450346757. doi: 10.1145/3018661.3018689.

L. Xu and M. Davenport. Dynamic matrix recovery from incomplete observations under an exact low-rank constraint. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29. Curran Associates, Inc., 2016. URL `https://proceedings.neurips.cc/paper/2016/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf`.

L. Xu and M. A. Davenport. Simultaneous recovery of a series of low-rank matrices by locally weighted matrix smoothing. In *2017 IEEE 7th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, pages 1–5, 2017. doi: 10.1109/CAMSAP.2017.8313104.

L. Xu and M. A. Davenport. Dynamic knowledge embedding and tracing. May 2020.

E. Yaksi and R. W. Friedrich. Reconstruction of firing rate changes across neuronal populations by temporally deconvolved Ca2+ imaging. *Nature Methods*, 3(5):377–383, May 2006. ISSN 1548-7091, 1548-7105. doi: 10.1038/nmeth874. URL `http://www.nature.com/articles/nmeth874`.

J. Ye and C. Pandarinath. Representation learning for neural population activity with neural data transformers. *arXiv preprint arXiv:2108.01210*, 2021.

B. M. Yu, J. P. Cunningham, G. Santhanam, S. I. Ryu, K. V. Shenoy, and M. Sahani. Gaussian-Process Factor Analysis for Low-Dimensional Single-Trial Analysis of Neural Population Activity. *Journal of Neurophysiology*, 102(1):614–635, July 2009. ISSN 0022-3077, 1522-1598. doi: 10.1152/jn.90941.2008. URL https://www.physiology.org/doi/10.1152/jn.90941.2008.

R. Yuste. From the neuron doctrine to neural networks. *Nature reviews neuroscience*, 16(8):487–497, 2015.

Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000. Publisher: IEEE.

Y. Zhao and I. M. Park. Variational latent gaussian process for recovering single-trial dynamics from population spike trains. *Neural computation*, 29(5):1293–1316, 2017. Publisher: MIT Press.

P. Zhou, S. L. Resendez, J. Rodriguez-Romaguera, J. C. Jimenez, S. Q. Neufeld, A. Giovannucci, J. Friedrich, E. A. Pnevmatikakis, G. D. Stuber, R. Hen, M. A. Kheirbek, B. L. Sabatini, R. E. Kass, and L. Paninski. Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. *eLife*, 7: e28728, Feb. 2018. ISSN 2050-084X. doi: 10.7554/eLife.28728. URL https://elifesciences.org/articles/28728.

F. Zhu, H. A. Grier, R. Tandon, C. Cai, A. Giovannucci, M. T. Kaufman, and C. Pandarinath. A deep learning framework for inference of single-trial neural population activity from calcium imaging with sub-frame temporal resolution. *bioRxiv*, 2021a.

F. Zhu, A. Sedler, H. A. Grier, N. Ahad, M. Davenport, M. Kaufman, A. Giovannucci, and C. Pandarinath. Deep inference of latent dynamics with spatio-temporal super-resolution using selective backpropagation through time. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. S. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 2331–2345. Curran Associates, Inc., 2021b. URL https://proceedings.neurips.cc/paper/2021/file/1325cdae3b6f0f91a1b629307bf2d498-Paper.pdf.

D. Zoltowski, J. Pillow, and S. Linderman. A general recurrent state space framework for modeling neural dynamics during decision-making. In *International Conference on Machine Learning*, pages 11680–11691. PMLR, 2020.