UNIVERSITY OF CALIFORNIA SAN DIEGO

Probing neural circuitry and large-scale brain dynamics underlying cognitive deficits associated with schizophrenia

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Neurosciences

by

Robert Kim

Committee in charge:

Terrence J. Sejnowski, Chair
Maksim Bazhenov
Eric Halgren
Claudia Lainscsek
Gregory A. Light
Tatyana Sharpee

2020

The dissertation of Robert Kim is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____

Chair

University of California San Diego

2020

DEDICATION

To my beloved family, colleagues, and friends

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Throughout my graduate training, I was fortunate to have support and guidance from my mentors, lab members, friends, and family members. First and most importantly, I offer my deepest gratitude to my parents, Prof. Hyo Kim and Gi-Bun Lee, for all their sacrifices so that my brother (John Kim) and I could achieve our dreams. I also thank my brother, who I consider to be my lifetime friend, for supporting me countlessly. I thank my close friend and colleague, Debha Amatya, for always asking thought-provoking questions and his constant encouragement.

I cannot thank enough my advisor, Prof. Terrence J. Sejnowski, who was always willing to listen to any ideas, no matter how small they may be. His open-mindedness and invaluable advice throughout my graduate training allowed me to explore and master many different areas in computational neuroscience. I also thank Prof. Claudia Lainscsek, who closely mentored me and taught me topics related to nonlinear dynamical systems. I owe thanks to other lab members in the Computational Neurobiology Laboratory and at the Salk Institute for helpful discussions and feedback on my projects. Especially, I would like to thank Aaron Sampson, Chris Gonzalez, Nuttida Rungratsameetaweemana, Ben Tsuda, Yusi Chen, Sarah Schoch, Ilya Verzhbinsky, Yinghao Li, Jack Knickrehm, Oliver Ernst, Mohammad Samavat, Ben Huh, Gerald Pao, Xin Wang, Mariam Ordyan, Sara Sameni, Jason Fleischer, David Peterson, Tom Bartol, Don Spencer, and Lyle Muller. I am grateful to Jorge Aldana for his technical expertise and assistance with computing resources.

I would like to thank Prof. Gregory A. Light for his willingness to share his expertise in schizophrenia and psychiatry and for taking the time to teach me. His guidance and encouragement throughout my graduate years fueled my passion for studying schizophrenia. In addition, I sincerely appreciate all the feedback and advice that I received from the rest of my committee members: Profs. Maksim Bazhenov, Eric Halgren, and Tatyana Sharpee.

During my tenure as a graduate student, I was fortunate to receive the following grants that made the works in my dissertation possible: Harold R. Schwalenberg Medical Scholar-

ship, Burnand-Partridge Foundation Scholarship, National Institute of Health (NIH) training grant (T32MH020002-17), and NIH Ruth L. Kirschstein National Research Service Award (F30MH115605-01A1). I would also like to acknowledge the support of NVIDIA Corporation with the donation of the Quadro P6000 graphics processing unit.

The work presented in Sections 2.3.1 and 2.3.2 is currently being prepared for publication: Claudia Lainscsek, Robert Kim, Gregory A. Light, and Terrence J. Sejnowski. Dynamical clustering and reconstruction of nonlinear state distributions from noisy signals using DDA.

The work presented in Section 2.3.3 is currently being prepared for publication: Robert Kim, Claudia Lainscsek, Aaron L. Sampson, Michael L. Thomas, The COGS Investigators, Neal R. Swerdlow, David L. Braff, Terrence J. Sejnowski, and Gregory A. Light. Defining subtypes of schizophrenia using unifying dynamical-systems biomarkers.

Chapter 3, in full, is a reprint of the material as it appears in: Robert Kim, Yinghao Li, and Terrence J. Sejnowski. Simple framework for constructing functional spiking recurrent neural networks. *Proceedings of the National Academy of Sciences*, 116(45):22811–22820, 2019. The dissertation author was the primary investigator and author of this paper.

Chapter 4, in full, is a reprint of the material as it has been written in a manuscript that has been submitted for publication: Robert Kim and Terrence J. Sejnowski. Strong inhibitory signaling underlies stable temporal dynamics and working memory in spiking neural networks. Preprint at `https://www.biorxiv.org/content/10.1101/2020.02.11.944751v1` (2020). The dissertation author was the primary investigator and author of this paper.

VITA

2012	B. S. in Biomedical Engineering, Applied Mathematics and Statistics, Johns Hopkins University
2020	Ph. D. in Neurosciences, University of California San Diego

PUBLICATIONS

Kim R*, Lainscsek C*, Sampson AL, Thomas ML, The COGS Investigators, Swerdlow NR, Braff DL, Sejnowski TJ, and Light GA. Defining subtypes of schizophrenia using unifying dynamical-systems biomarkers. *In preparation*.
*Equal contribution

Lainscsek C*, Kim R*, Light GA, and Sejnowski TJ. Dynamical clustering and reconstruction of nonlinear state distributions from noisy signals using DDA. *In preparation*.
*Equal contribution

Li Y, Kim R, and Sejnowski TJ. Synaptic and membrane dynamics important for working memory in spiking neural networks. *In preparation*.

Kim R and Sejnowski TJ. Strong inhibitory signaling underlies stable temporal dynamics and working memory in spiking neural networks. Preprint at `https://www.biorxiv.org/content/10.1101/2020.02.11.944751v1` (2020).

Kim R, Li Y, and Sejnowski TJ. Simple framework for constructing functional spiking recurrent neural networks. *PNAS*, 116(45):22811–22820, 2019.

Lainscsek C, Sampson AL, Kim R, Thomas ML, Man K, Lainscsek X, The COGS Investigators, Swerdlow NR, Braff DL, Sejnowski TJ, and Light GA. Nonlinear Dynamics Underlying Sensory Processing Dysfunction in Schizophrenia. *PNAS*, 116(9):3847–3852, 2019.

Guo JN, Kim R, Chen Y, Negishi M, Jhun S, Weiss S, Ryu JH, Bai X, Xiao W, Feeney E, Rodriguez-Fernandez J, Mistry H, Crunelli V, Crowley MJ, Mayes LC, Constable RT, and Blumenfeld H. Impaired consciousness in patients with absence seizures investigated by functional MRI, EEG, and behavioural measures: a cross-sectional study. *Lancet Neurology*, 15(13):1336–1345, 2016.

Li W, Motelow JE, Zhan Q, Hu Y-C, Kim R, Chen WC, and Blumenfeld H. Cortical network switching: possible role of the lateral septum and cholinergic arousal. *Brain Stimulation*, 8(1):3–41, 2015.

Kim R, Cingolani O, Wittstein I, McLean R, Han L, Cheng K, Robinson E, Brinker J, Schulman SS, Berger RD, Henrikson CA, and Tereshchenko LG. Mechanical alternans is associated with mortality in acute hospitalized heart failure: prospective mechanical alternans study (MAS). *Circulation: Arrhythmia and Electrophysiology*, 7(2):259-266, 2014.

Kim R, Hyder F, and Blumenfeld H. Physiological Basis of BOLD fMRI Decreases. In Mingrui Zhao, Hongtao Ma, and Theodore H. Schwartz, editors, *Neurovascular Coupling Methods*, volume 88 of *Neuromethods*, pages 221–236. Springer New York, 2014.

Kim R, Kulkarni P, and Hannenhalli S. Derepression of cancer/testis antigens in cancer is associated with distinct patterns of DNA hypomethylation. *BMC Cancer*, 13(1):144, 2013.

Kim JJ, Buchbinder N*, Ammanuel S*, Kim R*, Moore E, O'Donnell N, Lee JK, Kulikowicz E, Acharya S, Allen RH, and Johnston MV. Cost-effective therapeutic hypothermia treatment device for hypoxic ischemic encephalopathy. *Medical Devices: Evidence and Research*, 6(1):1–10, 2013. *Equal contribution

Kulkarni P, Shiraishi T, Rajagopalan K, Kim R, Mooney SM, and Getzenberg R. Cancer/testis antigens and urological malignancies. *Nature Reviews Urology*, 9(7):386–396, 2012.

ABSTRACT OF THE DISSERTATION

Probing neural circuitry and large-scale brain dynamics underlying cognitive deficits associated with schizophrenia

by

Robert Kim

Doctor of Philosophy in Neurosciences

University of California San Diego, 2020

Professor Terrence J. Sejnowski, Chair

Schizophrenia is a complex neuropsychiatric disorder characterized by a wide range of clinical manifestations. Even though the etiology of schizophrenia is not known, the heterogeneous nature of the disorder strongly suggests that multiple pathways and brain areas are affected by a combination of internal and external factors. These factors include and not limited to: psychological, genetic, social, and environmental determinants. Cognitive impairment is one of the commonly observed clinical manifestations of schizophrenia. Working memory, which is an ability to encode and hold information over a short period, is severely impaired in schizophrenia. (1) Characterizing how such deficits manifest in large-scale dynamics and (2) understanding

the pathophysiology and circuit mechanisms behind working memory deficits associated with schizophrenia are the two main questions that I address in my dissertation.

To answer the first question, I employed a method based on nonlinear systems theory to quantify large-scale dynamical states of time-series data and to identify dynamically distinct subgroups (Chapter 2). I demonstrate that the method, which utilizes delay differential analysis (DDA), can effectively extract features reflective of significant state changes and detect subgroups with similar features. Applying the method to brain signals obtained from a large cohort of schizophrenia patients further revealed subgroups with distinct dynamical characteristics aligned with neurophysiological and clinical parameters.

To answer the second question, I first developed a biologically realistic computational model based on spiking recurrent neural networks (RNNs) capable of learning cognitive tasks that involve working memory (Chapter 3). By taking advantage of a close relationship between continuous and spike RNNs that emerges under certain conditions, the method provides an extremely simple platform that can be utilized to investigate how power-efficient network dynamics lead to complex cognitive computations. By employing the framework, I uncover and characterize important circuit mechanisms critical for working memory maintenance in Chapter 4. The uncovered microcircuitry underscores the importance of disinhibitory gating exerted by specific subtypes of inhibitory interneurons, further confirming recent experimental findings.

Overall, my dissertation provides important computational tools for probing both micro- and macro-scale circuit dynamics associated with cognitive deficits in schizophrenia.

Chapter 1

Introduction

1.1 Schizophrenia pathophysiology

1.1.1 Background

Schizophrenia is often referred to as a heterogeneous disorder due to its wide spectrum of clinical manifestations along with many underlying etiologies, most of which are not yet known. The dopamine hypothesis, which postulates that overstimulation of dopaminergic receptors (mainly D_2 receptors) leads to psychotic symptoms, is one of the first hypotheses to attempt to unravel the neural basis of schizophrenia. However, this hypothesis is largely based on the observation that antipsychotic medications prescribed for psychosis display D_2 receptor antagonism activities, and does not explain other schizophrenia-related phenotypes such as social memory deficits and cognitive impairment.

In order to explain the clinical features that cannot be explained by the dopamine hypothesis, a more plausible hypothesis, known as "glutamate hypothesis," was developed. This hypothesis stipulates that hypofunction of glutamatergic signaling leads to decreased excitatory input to parvalbumin (PV) inhibitory interneurons. The decreased excitatory input then results in underdevelopment of PV interneurons leading to reduced overall inhibition mediated

by γ-aminobutyric acid (GABA) and enhanced glutamate-mediated excitation. The resulting excitation-inhibition imbalance disrupts synchronous oscillations that are associated with attention, memory, and cognition. Even though the glutamate hypothesis requires decreased glutamate receptor activities, it is plausible that intrinsic abnormalities in PV interneurons without glutamatergic dysfunction could also lead to cognitive deficits.

In this section, I review several studies that illustrate the importance of PV interneurons in not only generating synchronous network activity (in both cortex and hippocampus) but also establishing normal development of cognitive functioning.

1.1.2 PV interneurons are key mediators of the intrinsic hippocampal theta rhythm

Although numerous studies have shown that GABAergic interneurons play an important role in generating theta oscillations intrinsic to the hippocampus [SER$^+$13], which type of interneurons contributes to the induction of the hippocampal theta rhythm and the effect of inactivation of these interneurons on the theta oscillations have not been studied extensively. A recent study by [AHM$^+$] employed optogenetics to selectively activate or silence interneurons that express PV and somatostatin (SOM) in the CA1 region [AHM$^+$]. Adeno-associated viral vector with genes encoding for either archaerhodopsin (ArchT) or modified channelrhodopsin (ChETA) was injected into the intact hippocampus of PV-Cre and SOM-Cre mouse lines. The group observed that a continuous stimulation (using a continuous light pulse lasting 10 seconds) of PV interneurons expressing ChETA was able to drive the spontaneous oscillation frequency to the theta range (close to 8 Hz). On the other hand, when a continuous light pulse lasting 30 seconds was employed to inactivate PV interneurons expressing ArchT (which alters pH to inhibit synaptic transmission [EGZW$^+$16]), the frequency and power of the ongoing theta rhythm decreased significantly, and recovered to the theta range again when the inactivating light pulse was terminated.

When SOM interneurons expressing ChETA were activated by light pulses (fixed duration of 50 ms) at various frequency settings (2 - 20 Hz), [AHM$^+$] observed only minor changes in the baseline spontaneous oscillation frequency (4.7 $\pm$ 0.2 Hz) and power. Interestingly, a significant increase in the oscillation strength (*not* in the theta range) was observed when the pulse frequency was close to the baseline oscillation frequency. This suggests that SOM interneurons, when stimulated at the ongoing rhythm of the hippocampal network, have the ability to modulate network oscillations. Silencing the SOM interneurons (via ArchT and 30 seconds of continuous light) exerted a small effect on the ongoing theta oscillations.

1.1.3 PV interneurons drive cortical gamma oscillations

[CCM$^+$09] employed optogenetics to investigate how inhibitory neurons in the cortex modulate cortical gamma rhythm by selectively activating FS-PV neurons and regular-spiking (RS) excitatory cells in the somatosensory cortex *in vivo* [CCM$^+$09]. By injecting adeno-associated viral vector expressing channelrhodopsin-2 (ChR2) in the barrel cortex of PV-Cre and αCamKII-Cre (which targets RS cells), the group was able to demonstrate reliably light-dependent activation of these cells and observed prominent inhibitory postsynaptic potentials (IPSPs) in RS cells during light activation of fast-spiking PV (FS-PV) neurons.

Next, [CCM$^+$09] stimulated FS-PV cells (in PV-Cre line) and RS cells (in αCamKII-Cre line) with 1-ms light pulses at various frequencies ranging from 8 Hz to 200 Hz, and measured local field potential (LFP), an indirect measure of local network synchronicity. When the FS-PV interneurons were activated by light pulses in the gamma range (20 - 80 Hz), a strong increase in gamma LFP power was observed. On the contrary, gamma LFP power enhancement could not be observed when the RS cells were driven by gamma range light pulses. Instead, the RS cells responded to a lower frequency stimulation (8 - 24 Hz). These findings suggest that FS-PV inhibitory neurons, not excitatory cells, are critical for generating cortical gamma oscillations.

In order to investigate the role of cortical gamma rhythm in information processing,

[CCM$^+$09] performed whisker stimulation during gamma oscillations (established by 40 Hz light pulses in the PV-Cre mice) and measured the number of spikes by RS cells. The RS cell spikes were significantly reduced during gamma-induced LFP peaks and increased during LFP troughs indicating that cortical gamma oscillations gate sensory responses. This study, combined with the study discussed in the previous section (Section 1.1.2), provides strong evidence that inhibitory interneurons are indispensable for creating neural network synchronicity in both cortex and hippocampus.

1.1.4 Loss of hippocampal PV interneurons disrupts synchronous activities and results in social memory deficits

Both studies discussed in the previous sections (Sections 1.1.2 and 1.1.3) explored the role of PV interneurons in generating hippocampal theta and cortical gamma oscillations. However, these studies did not look into the behavioral effects that might result from the inactivation/loss of PV interneurons. Using a mouse model of 22q11.2 deletion syndrome (also known as DiGeorge syndrome), [PND$^+$16] characterized the loss of PV interneurons limited to CA2 of the hippocampus in this mouse model and how this loss of inhibitory cells translated to impaired social memory. The 22q11.2 deletion is accompanied by a wide variety of clinical symptoms (heart abnormalities, immune system dysfunction, and renal anomalies), and it has also been linked to a significant increase in risk for psychiatric disorders such as schizophrenia and bipolar disorder. The mouse model of the 22q11.2 deletion, *Df(16)A*$^{+/-}$, has also been characterized by similar neuropsychiatric phenotypes (learning deficits and working memory impairment) [SXB$^+$08].

[PND$^+$16] first quantified the density of PV interneurons in different areas of the hippocampus from *Df(16)A*$^{+/-}$ mice using immunohistochemical staining. Interestingly, a significant decrease in PV interneuron density, relative to the wild-type (WT) control mice, in area CA2 of 8-week old *Df(16)A*$^{+/-}$ mice was observed. However, the loss of interneurons was

not observed in younger (4-week old) *Df(16)A*$^{+/-}$ mice, implying an age-dependent reduction in PV interneurons (i.e., the number of interneurons decreased over time in the mutant mice). To characterize the effects of CA2 PV interneurons at the circuit level, the group measured postsynaptic potentials (PSPs) in CA2 pyramidal neurons, which receive excitatory inputs from Schaeffer collaterals (SCs) in CA3. Since SCs are regulated by CA2 PV interneurons, the excitatory postsynaptic potential (EPSP) amplitude of the CA2 pyramidal cells in response to stimulation of SCs was significantly larger in *Df(16)A*$^{+/-}$ mice compared to the amplitude seen in WT control mice. These findings are summarized in Fig. 1.1. Furthermore, the addition of GABA antagonists increased EPSPs for *both Df(16)A*$^{+/-}$ and WT mice (no significant difference in peak EPSP amplitude).

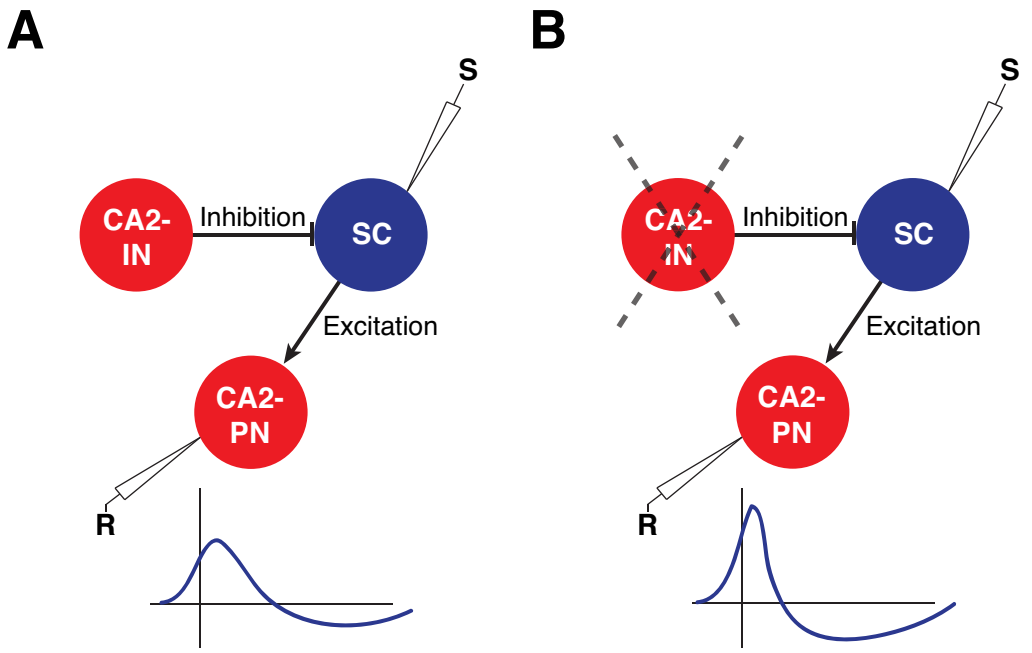


Figure 1.1: [PND^{+}16] stimulated Schaeffer collaterals and measured EPSPs at CA2-pyramidal cells. The EPSP peak amplitude was higher in *Df(16)A*$^{+/-}$ mice (**B**) compared to the amplitude observed in control mice (**A**). The lack of inhibitory drive from the CA2-interneurons to the Schaeffer collaterals resulted in greater excitatory drive to the CA2-pyramidal cells. CA2-IN = CA2-interneuron; SC = Schaeffer collaterals; CA2-PN = CA2-pyramidal neuron; S = stimulation site; R = recording site.

Because *Df(16)A*$^{+/-}$ mice lack inhibitory drive from CA2 PV interneurons, one might

hypothesize that stimulation of Schaeffer collaterals would result in more frequent firing of CA2 pyramidal cells. However, [PND$^+$16] observed an opposite trend: no action potentials were detected in *Df(16)A*$^{+/-}$ CA2 pyramidal cells after stimulation. The group also observed that the resting membrane potential of the pyramidal neurons in *Df(16)A*$^{+/-}$ mice was hyperpolarized. These findings indicate that the hyperpolarized resting potential of the *Df(16)A*$^{+/-}$ CA2 pyramidal neurons was able to override and suppress the overall increase in excitatory signaling that resulted from the loss of PV interneurons. This finding is important in that reduction in inhibitory interneurons does not necessarily result in enhanced excitability and can alter the intrinsic properties of the excitatory neurons (i.e., hyperpolarized resting potential).

Next, the group found a strong relationship between social memory impairment and silenced pyramidal neurons (caused by loss of PV interneurons). When a *Df(16)A*$^{+/-}$ mouse was re-exposed to the same mouse that it was exposed to in the previous trial, the mutant mouse spent more time exploring the "familiar" mouse. On the other hand, the exploration time was significantly decreased in WT mice. Based on these results combined with a previous study that also observed social memory deficits in mice whose CA2 pyramidal cells were completely silenced [HS14], the study concluded that the decreased density of PV interneurons in CA2 led to silencing of CA2 pyramidal neurons and impaired social learning.

1.1.5 NMDA receptors in PV interneurons are important for cortical gamma rhythm and cognitive behaviors

Previous sections covered the importance of PV interneurons in promoting synchronous network activity and normal cognitive function. Although dysfunction of inhibitory neurons has been suggested to be directly involved with cognitive impairment [PND$^+$16, LHV05], how this dysfunction arises in the first place is not yet fully understood.

Motivated by the previous works on the effect of NMDA receptor (NMDAR) blockers on cortical gamma rhythm, [CMS$^+$12] created a mouse line (PV-Cre $\times$ NR1f/f) that lacks

the NMDAR subunit NR1 in PV-expressing cells to study the role of NMDAR in the normal development of PV interneurons [CMS$^+$12]. The group confirmed the deletion of NMDAR by whole-cell recordings of PV cells in hippocampal slices of PV-Cre/NR1f/f mice. The deletion did not affect the number of PV interneurons implying that NMDARs are not required for survival of PV interneurons.

In order to characterize the functional features of the PV interneurons in PV-Cre/NR1f/f mice, [CMS$^+$12] introduced ChR2 into FS-PV cells. When light pulses in the gamma frequency range were used, a significant increase in gamma LFP power in layers 2/3 and 4 of the primary somatosensory cortex was observed in control mice (anesthetized). However, the gamma LFP enhancement was considerably reduced in PV-Cre/NR1f/f mice (anesthetized), suggesting that FS-PV neurons deficient of NMDA receptors lack the ability to induce cortical gamma oscillations. Interestingly, the loss of cortical gamma rhythm was *not* observed in both genotypes of awake behaving mice. Next, the group performed various behavioral tests (open field test, prepulse inhibition (PPI), T-maze test, and contextual/cued fear conditioning) to investigate how FS-PV cells without NMDAR contribute to behavioral/cognitive impairment. Of all the behavioral tests that they ran, only the contextual and cued fear conditioning test revealed a significant learning deficit in PV-Cre/NR1f/f mice.

1.1.6 Metabotropic glutamate receptors are critical for PV interneuron development

[CMS$^+$12] demonstrated that NMDA receptor dysfunction can alter intrinsic properties of PV interneurons and disrupt cortical gamma rhythm [CMS$^+$12]. To further probe the effect of glutamatergic signaling on inhibitory neurons, a more recent study by [BPDK$^+$15] investigated the importance of metabotropic glutamate receptor-5 (mGluR5).

The group first generated PV-Cre/mG5f/f mice that lacked mGluR5 in PV-expressing neurons. Then, the number of PV interneurons in the hippocampus (CA1, CA3, and dentate

gyrus), prelimbic cortex, and caudate putamen of the mutant and control mice were quantified. Surprisingly, the group found a region-specific decrease in PV interneurons in PV-Cre/mG5f/f mice: the number of PV neurons was significantly reduced in CA3, prelimbic cortex, and putamen. In addition, the number of synaptic contacts between interneurons and putative pyramidal cells was also significantly diminished. To study how this loss of PV neurons leads to neural network dysfunction, [BPDK$^+$15] recorded auditory event-related potentials (ERPs) using electrocorticography (ECoG). The average ERP waveform from the mutant mice (compared to the waveform from the control mice) revealed a distinct decrease in amplitude at 40 ms post-stimulus. Similar auditory ERP abnormalities have been reported from previous human clinical studies on schizophrenia and bipolar disorder. These findings suggest that mGluR5 is crucial for maturation and possibly survival of PV interneurons along with normal neural network development.

Behavioral and social phenotypes of PV-Cre/mG5f/f mice were next characterized via various tests (three-chamber novel object recognition, three-chamber social recognition, PPI, and Barnes maze test). PV-Cre/mG5f/f mice displayed impaired social recognition as evidenced by decreased exploration time during three-chamber novel object/social recognition tests. In order to find out if other memory modalities were similarly impaired, the group also assessed spatial memory using the Barnes maze, and discovered no spatial memory deficits in the mutant mice. This unique pattern of phenotypes displayed by PV-Cre/mG5f/f mice closely mimics the clinical features of neuropsychiatric disorders such as schizophrenia. In addition, the mutant mice also showed markedly elevated repetitive behaviors and PPI.

The above study paints a convincing picture of the role of glutamatergic pathway in interneuron development. The findings that mGluR5 is directly involved with maturation of interneurons suggest that NMDA signaling (potentiated by mGluR5) does contribute to development of normal PV inhibitor cells as proposed by [CMS$^+$12]. However, based on the severity of behavioral phenotypes observed in mice deficient of mGluR5, the study showed that mGluR5 does more than simply enhancing NMDA receptors and plays other roles (not yet known) critical

for PV interneuron development.

1.2 Microscale computational method: recurrent neural network model

1.2.1 Background

Previous studies have shown that working memory deficits often observed in schizophrenia could be attributed to excitation and inhibition (E/I) imbalance [FG18, Keh08]. As discussed in the previous section (Section 1.1), studies employing animal models have identified hypofunction of NMDA receptors on PV interneurons as one of the possible etiological components of schizophrenia [FG18, NZJ$^+$12, PND$^+$16, MSB$^+$18]. Furthermore, a recent study by [ZBC$^+$18] underscored the importance of spike timing in maintaining working memory. A computational RNN model that incorporates excitatory neurons and different subtypes of inhibitory interneurons can be a promising tool for elucidating how dysfunction of a subpopulation of inhibitory neurons could lead to working memory impairment.

1.2.2 Continuous-variable rate RNN

Continuous-variable rate RNNs, where recurrently connected units communicate via continuous signals (instead of discrete action potentials), have been widely utilized to uncover circuit mechanisms critical for performing various cognitive tasks [MSSN13, SYW16, Mic17]. Units in a continuous rate RNN are usually governed by the following set of equations:

$$\boldsymbol{\tau}\frac{d\boldsymbol{x}}{dt} = -\boldsymbol{x}(t) + \boldsymbol{w}\boldsymbol{r}(t) + \boldsymbol{w}_{in}\boldsymbol{u}(t) + \boldsymbol{\xi} \tag{1.1a}$$

$$\boldsymbol{r}(t) = \sigma(\boldsymbol{x}(t)) = \frac{1}{1+\exp(-\boldsymbol{x}(t))} \tag{1.1b}$$

$$\boldsymbol{o}(t) = \boldsymbol{w}_{out}\boldsymbol{r}(t) + b \tag{1.1c}$$

where $\boldsymbol{\tau} \in \mathbb{R}^{1\times N}$ refers to the synaptic time constants, $\boldsymbol{x} \in \mathbb{R}^{N\times T}$ represents the synaptic current variable from N units across T time-points, $\boldsymbol{r} \in \mathbb{R}^{N\times T}$ is the firing rates estimated by passing the synaptic current values ($\boldsymbol{x}$) through a nonlinear activation function (sigmoid in this case), $\boldsymbol{w}_{in} \in \mathbb{R}^{N\times N_{in}}$ defines connection weights from the time-varying inputs ($\boldsymbol{u} \in \mathbb{R}^{1\times T}$) to the network, and $\boldsymbol{w} \in \mathbb{R}^{N\times N}$ contains connection weights between N units. The output of the network ($\boldsymbol{o} \in \mathbb{R}^{1\times T}$) is a linear combination of all the firing rates specified by the output connection weight matrix, $\boldsymbol{w}_{out} \in \mathbb{R}^{1\times N}$, and the bias term, b. A schematic diagram illustrating a network producing a positive output signal upon receiving an input pulse is shown in Fig. 1.2.

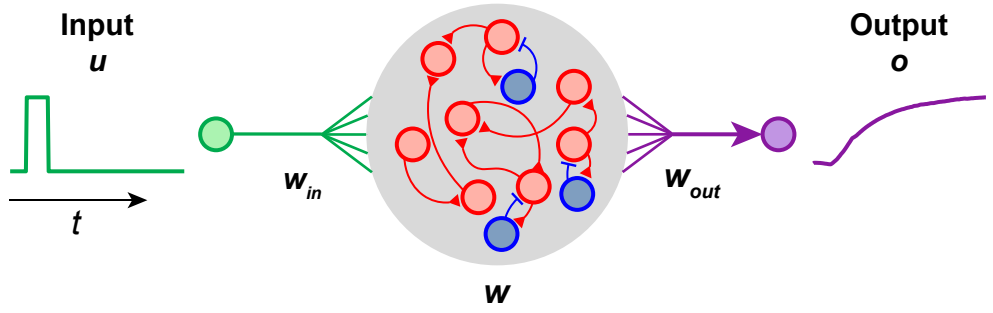


Figure 1.2: Schematic diagram illustrating a continuous RNN receiving a brief input pulse (green). The network consists of excitatory (red) and inhibitory (blue) units connected to one another (connection patterns specified by $\boldsymbol{w}$). The network output (purple) is a linear combination of the unit activities.

Rate RNNs can be trained to produce output signals ($\boldsymbol{o}$) closely resembling target signals ($\boldsymbol{z} \in \mathbb{R}^{1\times T}$) associated with a specific task. A loss function ($\mathcal{L}$), which measures how close

the RNN output signals are to the target signals, is employed to train and assess if the RNNs successfully learned the task. For example, the mean squared error (MSE) can be used to define the loss function:

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T}(z(t) - o(t))^2 = \frac{1}{T}\sum_{t=1}^{T}(z(t) - (\boldsymbol{w}_{out}\boldsymbol{r}(t) + b))^2 \tag{1.2}$$

where T is the total number of time points in a single trial.

Since the set of equations that govern the units (Eq. (1.1)) are continuous and differentiable, a gradient-descent supervised method, known as backpropagation through time (BPTT; [Wer90]), is often used to train rate RNNs to perform cognitive tasks [SYW16, YJS+19]. Given a set of n model parameters ($\boldsymbol{\theta} \in \{\theta_1, \theta_2, \ldots, \theta_n\}$), the gradient-descent algorithm tunes and optimizes the parameters to minimize the loss function ($\mathcal{L}$) in an iterative manner:

$$\theta_j^{(i+1)} = \theta_j^{(i)} - \eta\left(\frac{\partial \mathcal{L}^{(i)}}{\partial \theta_j^{(i)}}\right) \tag{1.3}$$

In Eq. (1.3), $\theta_j^{(i)}$ is the j-th model parameter at iteration i, and η is the learning rate, which controls the rate of the convergence of the gradient descent. The model parameters include the synaptic time constants ($\boldsymbol{\tau}$), recurrent connectivity structure ($\boldsymbol{w}$), readout weights ($\boldsymbol{w}_{out}$), and bias (b). Therefore, we have $\boldsymbol{\theta} \in \{\boldsymbol{\tau}, \boldsymbol{w}, \boldsymbol{w}_{out}, b\}$.

For each model parameter, BPTT needs to compute the gradient (partial differential in Eq. (1.3)). For the readout weights ($\boldsymbol{w}_{out}$), the gradient can be computed by simply differentiating the loss function (Eq. (1.2)) with respect to $\boldsymbol{w}_{out}$:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{w}_{out}} = \frac{1}{T}\sum_t (-2)\cdot(z(t) - \boldsymbol{w}_{out}\boldsymbol{r}(t) - b)\cdot \boldsymbol{r}(t) \tag{1.4a}$$

$$= \frac{2}{T}\sum_t \boldsymbol{r}(t)\cdot(\boldsymbol{w}_{out}\boldsymbol{r}(t) + b - z(t)) \tag{1.4b}$$

$$= \frac{2}{T}\sum_t \boldsymbol{r}(t)\cdot(o(t) - z(t)) \tag{1.4c}$$

Similarly, differentiating the loss function with respect to the bias term (b) leads to the same gradient as Eq. (1.4c).

For the recurrent connections ($\boldsymbol{w}$), computing the gradient is more involved. First, the gradient of the loss function at time t ($\mathcal{L}_t$) is defined as

$$\frac{\partial \mathcal{L}_t}{\partial w_{ik}} = \frac{\partial \mathcal{L}_t}{\partial r_t^{(i)}}\frac{\partial r_t^{(i)}}{\partial x_t^{(i)}}\frac{\partial x_t^{(i)}}{\partial w_{ik}} \tag{1.5}$$

where $r_t^{(i)}$ is the rate activity of unit i at time t, $x_t^{(i)}$ refers to the synaptic activity of unit i at time t, and $w_{ik} \in \boldsymbol{w}$ is the synaptic weight from unit k to unit i. For the first gradient on the righthand side (i.e., $\partial \mathcal{L}_t / \partial r_t^{(i)}$), the loss function needs to be first rewritten as:

$$\mathcal{L}_t = \frac{1}{T}\left(z_t - \left(\sum_{j=1}^{N} w_{out}^{(j)} r_t^{(j)} + b\right)\right)^2 \tag{1.6a}$$

$$= \frac{1}{T}\left(z_t - \left(w_{out}^{(i)} r_t^{(i)} + \sum_{\substack{j=1 \\ j\neq i}}^{N} w_{out}^{(j)} r_t^{(j)} + b\right)\right)^2 \tag{1.6b}$$

Differentiating Eq. (1.6b) with respect to $r_t^{(i)}$ results in:

$$\frac{\partial \mathcal{L}_t}{\partial r_t^{(i)}} = -\frac{2}{T} \cdot w_{out}^{(i)} \cdot \left(z_t - \left(w_{out}^{(i)} r_t^{(i)} + \sum_{\substack{j=1 \\ j \neq i}}^{N} w_{out}^{(j)} r_t^{(j)} + b \right) \right) \tag{1.7a}$$

$$= -\frac{2}{T} \cdot w_{out}^{(i)} \cdot (z_t - o_t) \tag{1.7b}$$

The second gradient on the righthand side of Eq. (1.5) (i.e., $\partial r_t^{(i)} / \partial x_t^{(i)}$) can be computed using Eq. (1.1b):

$$\frac{\partial r_t^{(i)}}{\partial x_t^{(i)}} = \sigma(x_t^{(i)}) \cdot (1 - \sigma(x_t^{(i)})) \tag{1.8}$$

Lastly, the third gradient on the righthand side of Eq. (1.5) (i.e., $\partial x_t^{(i)} / \partial w_{ik}$) can be computed by first substituting $\alpha = \Delta t / \boldsymbol{\tau}$ into the discretized version of Eq. (1.1a) (using Euler approximation method):

$$\boldsymbol{x}_t = (1 - \alpha)\, \boldsymbol{x}_{t-1} + \alpha(\boldsymbol{w}\boldsymbol{r}_{t-1} + \boldsymbol{w}_{in}\boldsymbol{u}_{t-1}) + \boldsymbol{\xi} \tag{1.9}$$

Focusing only on unit i leads to:

$$x_t^{(i)} = (1 - \alpha) x_{t-1}^{(i)} + \alpha \left(\sum_{j=1}^{N} w_{ij} \cdot r_{t-1}^{(j)} + \boldsymbol{w}_{in}^{(i)} \boldsymbol{u}_{t-1} \right) + \boldsymbol{\xi} \tag{1.10}$$

Applying the chain rule results in:

$$\frac{\partial x_t^{(i)}}{\partial w_{ik}} = \sum_{t'=1}^{t} \left(\frac{\partial x_t^{(i)}}{\partial x_{t'}^{(i)}} \frac{\partial x_{t'}^{(i)}}{\partial w_{ik}} \right) \tag{1.11}$$

Due to the recurrent nature, $\partial x_t^{(i)} / \partial x_{t'}^{(i)}$ can be expressed as

$$\frac{\partial x_t^{(i)}}{\partial x_{t'}^{(i)}} = \prod_{q=t'+1}^{t} \frac{\partial x_q^{(i)}}{\partial x_{q-1}^{(i)}} \tag{1.12}$$

Substituting Eq. (1.12) into Eq. (1.11) leads to

$$\frac{\partial x_t^{(i)}}{\partial w_{ik}} = \sum_{t'=1}^{t} \left\{ \left(\prod_{q=t'+1}^{t} \frac{\partial x_q^{(i)}}{\partial x_{q-1}^{(i)}} \right) \frac{\partial x_{t'}^{(i)}}{\partial w_{ik}} \right\} \tag{1.13}$$

The gradient of the loss function with respect to w_{ik} can be calculated by summing over all the time points:

$$\frac{\partial \mathcal{L}}{\partial w_{ik}} = \sum_{t=1}^{T} \frac{\partial \mathcal{L}_t}{\partial w_{ik}} = \sum_{t=1}^{T} \left(\frac{\partial \mathcal{L}_t}{\partial r_t^{(i)}} \frac{\partial r_t^{(i)}}{\partial x_t^{(i)}} \frac{\partial x_t^{(i)}}{\partial w_{ik}} \right) \tag{1.14}$$

1.3 Macroscale computational method: delay differential analysis

1.3.1 Background

Unlike RNN models which focus on local neural circuits, delay differential analysis (DDA) attempts to characterize collective, large-scale neural dynamics to provide an integrated view of the systems that give rise to the emergent phenomena of behavior and cognition. For example, DDA was able to identify abrupt changes in brain signals when unexpected, deviant auditory tones were given to test participants. Interestingly, these changes were significantly attenuated in patients diagnosed with schizophrenia ([LSK$^+$19]). In addition, DDA was able to detect dynamical state changes on electrocorticography (ECoG) signals 1-2 hours preceding idiopathic generalized seizures for a few patients (Fig. 1.3; [LWCS17]). More importantly, some of the seizure onset times marked by clinicians were several minutes later than the nonlinear dynamical changes detected by DDA (Seizure #1 in Fig. 1.3). These findings suggest that DDA might provide a more consistent method to detect seizures.

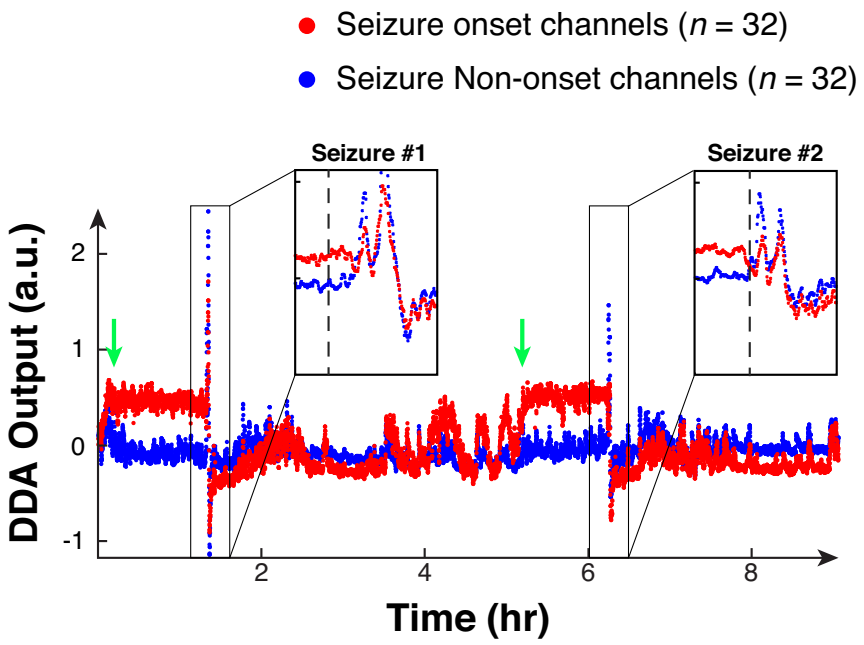


Figure 1.3: DDA output values from the seizure onset ECoG channels (determined by clinicians) increase significantly (green arrows) ˜1 hour before each clinical seizure onset. Dotted vertical lines (inset) indicate clinical seizure onset times.

1.3.2 Delay differential analysis (DDA)

DDA is based on delay differential equations (DDEs), which relate a derivative signal $\dot{x}(t)$ to a signal $x(t-\tau_n)$ non-uniformly delayed in time (Fig. 1.4). For brain signals, such as electroencephalography (EEG) and ECoG, the following three-term, second-order DDA model have been shown to be effective at capturing important nonlinear dynamics:

$$\dot{x}(t) = a_1 x_{\tau_1} + a_2 x_{\tau_2} + a_3 x_{\tau_1}^2 = f(a, x_{\tau_1}, x_{\tau_2}) \tag{1.15}$$

In Eq. (1.15), x_{τ_i} indicates a delayed version of the original time-series (i.e., $x(t-\tau_i)$). Supervised structure selection performed on various EEG signals ([LHW$^+$13, LWH$^+$13]) via repeated random subsampling cross-validation resulted in selection of Eq. (1.15) and helped determine τ_1 and τ_2. Therefore, the three estimated parameters (a_1, a_2, and a_3) along with the least square mean error ($\rho = \sqrt{\sum(\dot{x} - f(a, x_{\tau_1}, x_{\tau_2}))^2}$) are referred to as DDA features.

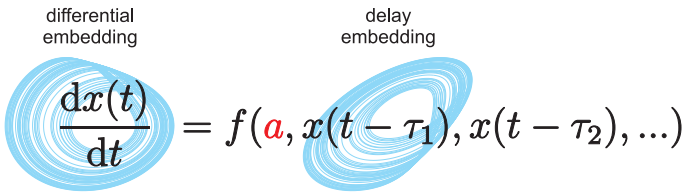


Figure 1.4: DDA performs a functional mapping between a differential embedding and a delay embedding. $\tau \in \mathbb{N}_0$ is the time delay and $\mathbf{a} \in \mathbb{R}$ are DDA features

In a classification framework, the DDA features can be computed for each time-series signal in a dataset with two data classes. Due to the sparsity of the model, a simple linear algebra operation, such as singular value decomposition (SVD), can be employed to find the optimal hyperplane that separates the two classes in the four dimensional DDA feature space. For each data sample, the shortest distance from the optimal hyperplane is then computed. Using these distance values, the classification performance can be computed via the area under the receiver operating characteristic (ROC) curve [KK97].

Bibliography

[AHM+] Bénédicte Amilhon, Carey Y. L. Huh, Frédéric Manseau, Guillaume Ducharme, Heather Nichol, Antoine Adamantidis, and Sylvain Williams. Parvalbumin interneurons of hippocampus tune population activity at theta frequency. *Neuron*, 86(5):1277–1289, 2016/11/19.

[BPDK+15] S A Barnes, A Pinto-Duarte, A Kappe, A Zembrzycki, A Metzler, E A Mukamel, J Lucero, X Wang, T J Sejnowski, A Markou, and M M Behrens. Disruption of mGluR5 in parvalbumin-positive interneurons induces core features of neurodevelopmental disorders. *Mol Psychiatry*, 20(10):1161–72, Oct 2015.

[CCM+09] Jessica A. Cardin, Marie Carlén, Konstantinos Meletis, Ulf Knoblich, Feng Zhang, Karl Deisseroth, Li-Huei Tsai, and Christopher I. Moore. Driving fast-spiking cells induces gamma rhythm and controls sensory responses. *Nature*, 459(7247):663–667, Apr 2009.

[CMS+12] M Carlén, K Meletis, J H Siegle, J A Cardin, K Futai, D Vierling-Claassen, C Ruhlmann, S R Jones, K Deisseroth, M Sheng, C I Moore, and L-H Tsai. A critical role for NMDA receptors in parvalbumin interneurons for gamma rhythm induction and behavior. *Mol Psychiatry*, 17(5):537–548, 05 2012.

[EGZW+16] Mohamady El-Gaby, Yu Zhang, Konstantin Wolf, Christof J. Schwiening, Ole Paulsen, and Olivia A. Shipton. Archaerhodopsin selectively and reversibly silences synaptic transmission through altered pH. *Cell Reports*, 16(8):2259 – 2268, 2016.

[FG18] Brielle R. Ferguson and Wen-Jun Gao. PV interneurons: Critical regulators of E/I balance for prefrontal cortex-dependent behavior and psychiatric disorders. *Frontiers in Neural Circuits*, 12:37, 2018.

[HS14] Frederick L. Hitti and Steven A. Siegelbaum. The hippocampal CA2 region is essential for social memory. *Nature*, 508(7494):88–92, 04 2014.

[Keh08] Colin Kehrer. Altered excitatory-inhibitory balance in the NMDA-hypofunction model of schizophrenia. *Frontiers in Molecular Neuroscience*, 1, 2008.

[KK97] M. N. Kremliovsky and J. B. Kadtke. Using delay differential equations as dynamical classifiers. In J. B. Kadtke and A. Bulsara, editors, *American Institute of Physics Conference Series*, volume 411 of *American Institute of Physics Conference Series*, pages 57–62, May 1997.

[LHV05] David A. Lewis, Takanori Hashimoto, and David W. Volk. Cortical inhibitory neurons and schizophrenia. *Nat Rev Neurosci*, 6(4):312–324, 04 2005.

[LHW+13] Claudia Lainscsek, Manuel E Hernandez, Jonathan Weyhenmeyer, Terrence J Sejnowski, and Howard Poizner. Non-linear dynamical analysis of EEG time series distinguishes patients with Parkinson's disease from healthy individuals. *Front Neurol*, 4:200, 2013.

[LSK+19] Claudia Lainscsek, Aaron L. Sampson, Robert Kim, Michael L. Thomas, Karen Man, Xenia Lainscsek, , Neal R. Swerdlow, David L. Braff, Terrence J. Sejnowski, and Gregory A. Light. Nonlinear dynamics underlying sensory processing dysfunction in schizophrenia. *Proceedings of the National Academy of Sciences*, 116(9):3847–3852, 2019.

[LWCS17] Claudia Lainscsek, Jonathan Weyhenmeyer, Sydney S. Cash, and Terrence J. Sejnowski. Delay differential analysis of seizures in multichannel electrocorticography data. *Neural Computation*, 29(12):3181–3218, 2017.

[LWH+13] Claudia Lainscsek, Jonathan Weyhenmeyer, Manuel E Hernandez, Howard Poizner, and Terrence J Sejnowski. Non-linear dynamical classification of short time series of the Rössler system in high noise regimes. *Front Neurol*, 4:182, 2013.

[Mic17] Thomas Miconi. Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *eLife*, 6:e20899, 2017.

[MSB+18] Thomas Marissal, Rodrigo F. Salazar, Cristina Bertollini, Sophie Mutel, Mathias De Roo, Ivan Rodriguez, Dominique Müller, and Alan Carleton. Restoring wild-type-like ca1 network dynamics and behavior during adulthood in a mouse model of schizophrenia. *Nature Neuroscience*, 2018.

[MSSN13] Valerio Mante, David Sussillo, Krishna V. Shenoy, and William T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, 2013.

[NZJ+12] Kazu Nakazawa, Veronika Zsiros, Zhihong Jiang, Kazuhito Nakao, Stefan Kolata, Shuqin Zhang, and Juan E. Belforte. GABAergic interneuron origin of schizophrenia pathophysiology. *Neuropharmacology*, 62(3):1574 – 1583, 2012.

[PND+16] Rebecca A. Piskorowski, Kaoutsar Nasrallah, Anastasia Diamantopoulou, Jun Mukai, Sami I. Hassan, Steven A. Siegelbaum, Joseph A. Gogos, and Vivien

Chevaleyre. Age-dependent specific changes in area CA2 of the hippocampus and social memory deficit in a mouse model of the 22q11.2 deletion syndrome. *Neuron*, 89(1):163 – 176, 2016.

[SER+13] Eran Stark, Ronny Eichler, Lisa Roux, Shigeyoshi Fujisawa, Horacio G. Rotstein, and György Buzsáki. Inhibition-induced theta resonance in cortical circuits. *Neuron*, 80(5):1263 – 1276, 2013.

[SXB+08] Kimberly L Stark, Bin Xu, Anindya Bagchi, Wen-Sung Lai, Hui Liu, Ruby Hsu, Xiang Wan, Paul Pavlidis, Alea A Mills, Maria Karayiorgou, and Joseph A Gogos. Altered brain microRNA biogenesis contributes to phenotypic deficits in a 22q11-deletion mouse model. *Nat Genet*, 40(6):751–760, 06 2008.

[SYW16] H. Francis Song, Guangyu R. Yang, and Xiao-Jing Wang. Training excitatory-inhibitory recurrent neural networks for cognitive tasks: A simple and flexible framework. *PLOS Computational Biology*, 12(2):e1004792, 2016.

[Wer90] P. J. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, 1990.

[YJS+19] Guangyu Robert Yang, Madhura R. Joglekar, H. Francis Song, William T. Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, 22(2):297–306, 2019.

[ZBC+18] Jennifer L. Zick, Rachael K. Blackman, David A. Crowe, Bagrat Amirikian, Adele L. DeNicola, Theoden I. Netoff, and Matthew V. Chafee. Blocking NMDAR disrupts spike timing and decouples monkey prefrontal circuits: Implications for activity-dependent disconnection in schizophrenia. *Neuron*, 98(6):1243–1255, 2018.

Chapter 2

Defining subtypes of schizophrenia using unifying dynamical-systems biomarkers

This chapter focuses on characterizing large-scale systems dynamics related to information processing and higher-order cognitive functions. Using a method based on nonlinear dynamical systems theory, I present a new algorithm capable of capturing large-scale dynamical states from complex time-series signals and identifying dynamically distinct subgroups. By applying the algorithm to a large dataset of brain signals obtained from schizophrenia (SZ) patients, I also characterize nonlinear dynamical features associated with psychosocial and cognitive dysfunction associated with SZ.

The work presented in Sections 2.3.1 and 2.3.2 is currently being prepared for publication: Claudia Lainscsek, Robert Kim, Gregory A. Light, and Terrence J. Sejnowski. Dynamical clustering and reconstruction of nonlinear state distributions from noisy signals using DDA.

The work presented in Section 2.3.3 is currently being prepared for publication: Robert Kim, Claudia Lainscsek, Aaron L. Sampson, Michael L. Thomas, The COGS Investigators, Neal R. Swerdlow, David L. Braff, Terrence J. Sejnowski, and Gregory A. Light. Defining subtypes of schizophrenia using unifying dynamical-systems biomarkers.

2.1 Introduction

Electroencephalography (EEG) is a highly utilized method to measure electrical signals generated from brain processes that exhibit complex and nonlinear dynamics. Due to its non-invasive nature along with high temporal resolution, EEG is especially useful for monitoring abnormal changes in the brain's cortical dynamics. Even though many previous studies have investigated EEG signals associated with neuropsychiatric disorders (Alzheimer's disease [LKJ+07, Jeo04, SAE15, GRA15]; Parkinson's disease [YM16, MLP+01, PJR01]; epilepsy [YZLC11, YZLW12, LCC+15]), identifying disease-specific nonlinear signatures in these signals and relating these markers to cognitive impairment still remain as a challenge to be addressed.

Methods that employ linear analysis have been widely used to analyze signals obtained from EEG. However, the underlying central nervous system that generates the brain signals is considered to be a network of many interconnected nonlinear dynamical systems. Therefore, linearization in an attempt to approximate these systems may discard relevant nonlinear information. In order to better capture large-scale dynamical dysfunction and systems-level changes present in neuropsychiatric diseases, computational methods based on nonlinear dynamics and systems theory recently emerged as a promising tool in neuroscience [Bre17]. One such method is delay differential analysis (DDA). DDA can be used to extract nonlinear dynamical features from time-series data via a functional mapping where a derivative embedding is expressed in terms of a delay embedding. The first-order derivative is related to a nonlinear function of non-uniformly delayed versions of the time-series data. An embedding maps a low dimensional time-series to a high dimensional object that contains information about all the state variables of an unknown, underlying dynamical system without having access to all the systems variables [Tak81]. In layman's terms, DDA seeks to find out how past events (delay embedding) contribute to changes that are occurring in the present (derivative embedding). The features that connect these two embeddings are then used to characterize the underlying nonlinear dynamics.

Previous studies have shown that DDA can be used to extract disease-specific dynamical features (Parkinson's movement data [LRS+12, LHW+13a], electrocardiogram recordings [LS13], and electrocorticogram (ECoG) data associated with epilepsy [LWCS17b]). More recently, we have shown that DDA can be used to identify unique nonlinear dynamical architectures hidden in a large dataset of EEG signals obtained from SZ patients [LSK+19]. SZ is a debilitating psychiatric disorder characterized by a wide range of clinical manifestations [SSM+97, JS15], and developing non-invasive biomarkers that can be utilized to identify subgroups within the heterogeneous disorder remains challenging. Motivated by this problem, we developed a new algorithm that employs DDA features to identify dynamically distinct subgroups present in time-series signals. Given a time-series dataset with varying degrees of nonlinear dynamics, our algorithm can not only identify dynamically distinct subgroups, but also reconstruct the distribution of the dynamical states in the data. In this chapter, I first present the clustering method and the simulation experiments that I designed to validate the algorithm. Next, I present the results that I obtained from applying the method to the Consortium on the Genetics of Schizophrenia (COGS-2) dataset containing EEG signals from non-psychiatric comparison (NC) and SZ participants.

2.2 Materials and methods

2.2.1 Delay differential analysis (DDA)

Given a time-series signal ($x(t)$), DDA employs delay differential embeddings that relate a derivative signal ($\dot{x}(t)$) to signals non-uniformly delayed in time ($x(t-\tau_n)$):

$$\dot{x} = \sum_{i=1}^{I} a_i \prod_{n=1}^{N} x(t-\tau_n)^{m_{n,i}} \tag{2.1}$$

where τ_n is the delay time constant, I specifies the number of terms in the DDA model, and $\sum_n m_{n,i}$ is the maximum order of the model. The derivative signal ($\dot{x}$) was estimated using central difference approximations [MM05].

Throughout this chapter, the following three-term, second-order DDA model will be used to compute DDA features:

$$\dot{x}(t) = f(a, x_{\tau_1}, x_{\tau_2}) = a_1 x_{\tau_1} + a_2 x_{\tau_2} + a_3 x_{\tau_1}^2 \tag{2.2}$$

where $x_{\tau_i} = x(t - \tau_i)$. Unsupervised structure selection performed on various time-series data via repeated random subsampling cross-validation resulted in selection of Eq. (2.2) and helped determine the two delay time constants, τ_1 and τ_2 [LHW$^+$13b, LWCS17a, LMP$^+$14]. Singular value decomposition (SVD) was performed to determine the coefficients (a_1, a_2, and a_3) in Eq. (2.2) [PTVF92]. The three estimated parameters (a_1, a_2, and a_3) along with the least squares error ($\rho = \sqrt{\sum (\dot{x} - f(a, x_{\tau_1}, x_{\tau_2}))^2}$) are what we refer to as DDA features.

2.2.2 Dynamical clustering algorithm

In order to identify subgroups with similar dynamical states, we first extracted DDA features from time-series signals using the model mentioned above (Eq. (2.2)). Here, we assume that the signals originate from two classes (red and blue in Fig. 2.1A). Next, we developed a genetic algorithm, a global search method based on the process of natural selection [Gol89], to identify non-representative samples from the two groups. Our genetic algorithm method is designed to identify local minima instead of global minima. A sample in a group was deemed non-representative if its DDA features were closer to the DDA features of the other class.

More formally, our genetic algorithm utilizes SVD to identify non-representative samples. If the hyperplane computed from a small subset of the samples leads to poor classification performance when applied to the rest of the data, then our algorithm returns these samples (orange

and cyan in Fig. 2.1A) as non-representative samples. We applied our genetic algorithm 10,000 times, each time identifying ten non-representative samples (five from each group). Lastly, we counted how many times each sample was selected as non-representative. The samples were then sorted by these count values, which we refer to as dynamic index (DI) values (Fig. 2.1B). Thus, low DI subgroups (red and blue in Fig. 2.1B) have DDA features that are well separated from each other, while the DDA features from the high DI subgroups (orange and cyan in Fig. 2.1B) are not distinguishable. Therefore, DI values can be used to not only reconstruct the underlying distribution of the dynamical states, but also identify subgroups with similar dynamical properties.

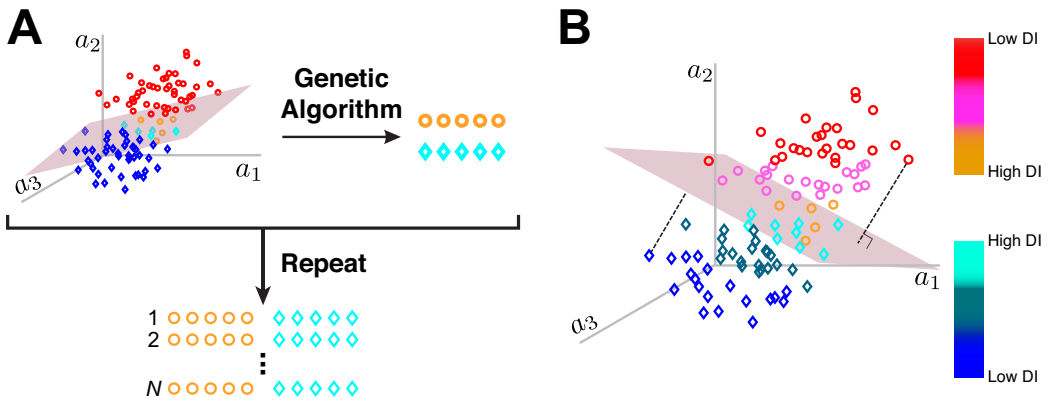


Figure 2.1: Schematic illustration of the algorithm to compute DI values. (A) Genetic algorithm is applied to DDA features extracted from time-series signals with two class labels (red and blue) to identify non-representative samples (orange and cyan). The process is repeated $N = 10000$ times. (B) For each sample, a DI value (the number of times it was selected as non-representative) is computed. The higher the DI value, the more non-representative the sample is.

2.2.3 Simulation experiment overview

In order to validate our method, we performed two simulation experiments. For each experiment, we generated two groups of time-series signals with a wide range of nonlinear dynamics by varying the experiment-specific dynamical parameter (Fig. 2.2). The main objective of each experiment was to use the DI values extracted from the time-series signals alone to (1) estimate and reconstruct the distribution of the dynamical parameter values and (2) identify

subgroups with similar dynamical features.

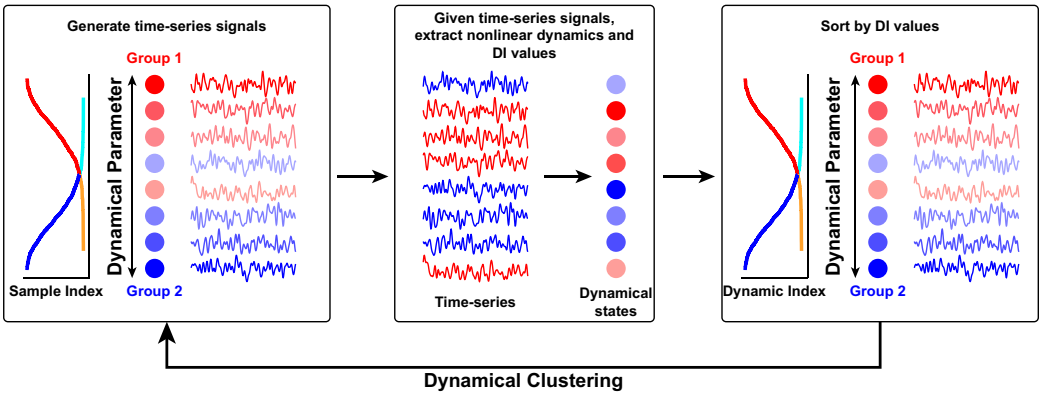


Figure 2.2: Schematic illustration of the simulated experiments. Experiment-specific dynamical parameter is varied in a manner that produces time-series signals with diverse nonlinear dynamics with two class labels (left). DDA features and DI values are computed from the time-series signals (center). The relationship between the computed DI values and the dynamical parameter values can be characterized (right).

2.2.4 EEG dataset

After we validated our algorithm on simulation datasets, we applied our method to a large dataset of EEG signals obtained from both NC and SZ participants. The dataset, the Consortium on the Genetics of Schizophrenia (COGS-2), contains continuous EEG data collected from NC ($n = 753$) and SZ subjects ($n = 877$) recruited from five COGS-2 study centers (University of California San Diego, University of California Los Angeles, University of Washington, University of Pennsylvania, and Mount Sinai School of Medicine) [LST+15]. EEG recordings (1 kHz sampling rate) from a single electrode at the vertex (CZ) along with an auditory oddball sequence were used to elicit mismatch negativity (MMN). The auditory sequence contained standard tones (90% of stimuli, 50-ms) and deviant tones (10% of stimuli, 100-ms) generated in a pseudorandom manner such that a minimum of two standard tones presented between deviant stimuli. EEG data from each subject were segmented into 150 trials with duration of 550 ms (100 ms pre-tone and 450 ms post-tone periods).

2.3 Results

2.3.1 Simulation experiment: chaotic Rössler system

For the first simulation study, the Rössler system was used to generate a large group of time-series signals. The Rössler system is a simple nonlinear system whose output can be either periodic or chaotic based on its three system parameters [R$\ddot{7}$6]. The system is defined by the following set of ordinary differential equations (ODEs):

$$\begin{aligned} \dot{x} &= -y - z \\ \dot{y} &= x + ay \\ \dot{z} &= b - xz - cz \end{aligned} \tag{2.3}$$

The Rössler system has been shown to be chaotic at $a = 0.2$, $c = 5.7$, and $0.37 \leq b \leq 0.47$ (Fig. 2.3A). Therefore, to generate two dynamically distinct groups of chaotic time-series signals, we first fixed a and c at 0.2 and 5.7, respectively. Next, we generated two groups (G1 and G2) of b values (1000 values in each group) randomly drawn from two overlapping chaotic regions: $0.37 \leq b \leq 0.46$ and $0.39 \leq b \leq 0.47$. The distributions of the two groups of b values followed beta distributions such that G1 contained mostly low values (between 0.37 and 0.42) with a few outliers (i.e., high b values), and G2 contained mostly high values (between 0.42 and 0.47) with outliers in the low range. The distribution of the b values from each group is shown in Fig. 2.3B. Lastly, for each b value, the Rössler system was numerically solved to extract $x(t)$ as a chaotic time-series signal with the integration step size of 0.04. For each b value, multiple "trials" (150 trials) of signals were generated using different random initial conditions. Each trial contained 570 time-points with random Gaussian noise (signal-to-noise ratio of 5 dB). Example time-series signals generated in this manner are shown in Fig. 2.3B. Therefore, the dataset for this experiment

contained $2,000 \times 150 = 300,000$ chaotic signals (150,000 from each group).

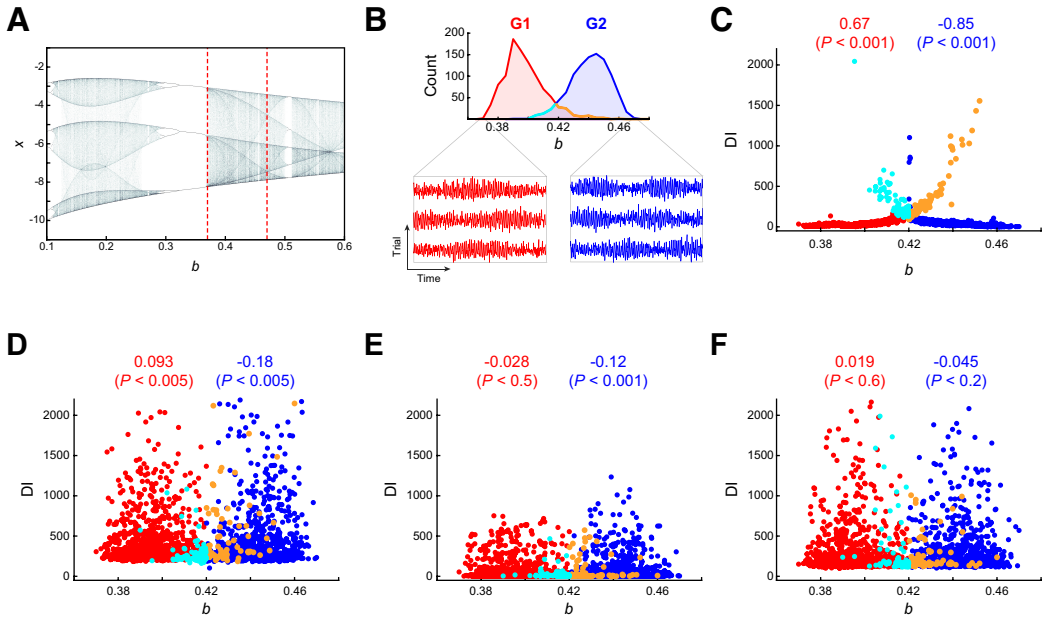


Figure 2.3: (A) Bifurcation diagram for the Rössler system with $a = 0.2$ and $c = 5.7$. The dynamical parameter (b) was varied between 0.37 and 0.47 (red dashed lines). (B) Distribution of the dynamical parameter (b) values for each group and example output signals for the Rössler experiment. (C) DI values extracted from the DDA features of the chaotic time-series signals were strongly correlated with the b values. (D–F) DI values extracted from the signal amplitude values (D), frequency features (E), and Lyapunov exponent estimates (F) were not correlated with the b values. Spearman rank correlation values shown. Note that the correlation coefficient for the second group (blue) is negative, because the DI values for low b values are considered non-representative for this group.

In order to reconstruct the distribution of the dynamical parameter values (i.e., distribution of the b values in Fig. 2.3B) using the chaotic signals alone, we first computed the four DDA features from each signal using Eq. (2.2). Running our genetic algorithm on the computed DDA features revealed that the samples from the overlapping region in the dynamic parameter space had high DI values (i.e., likely to be selected as non-representative samples). In addition, the DI values were significantly correlated with the actual b values (Spearman rank correlation, $r = 0.67$, $P < 0.0001$ for G1; Spearman rank correlation, $r = -0.85$, $P < 0.0001$ for G2; Fig. 2.3C), confirming that DDA features along with DI values could indeed capture dynamical states underlying chaotic time-series signals. On the other hand, employing linear features (amplitude

or frequency changes over time) to compute the DI values did not lead to accurate reconstruction of the distributions (Fig. 2.3D and E). The amplitude features were composed of the minimum, maximum, and mean values. For the frequency features, we considered 10 frequency bands range from 0 Hz to 100 Hz. In addition, the DI values computed using the Lyapunov exponent, a commonly used nonlinear feature for assessing if a dynamical system is chaotic [WSSV85], did not recover the underlying distribution of the b values (Fig. 2.3F).

2.3.2 Simulation experiment: coupled dynamical systems

To test if our algorithm can reconstruct dynamical states from more complex and biological systems, we generated another dataset derived from a spiking neural network model. A network of Izhikevich spiking neurons governed by the following equations [Izh03] was employed to simulate coupled dynamical systems:

$$\begin{aligned} \dot{v} &= 0.04v^2 + 5v + 140 - u + I \\ \dot{u} &= a \cdot (bv - u) \end{aligned}$$

where v and u refer to the membrane potential and a recovery variable, respectively. When the membrane voltage (v) reaches the action potential threshold (35 mV), then a spike is recorded and v is reset to c. The recovery variable (u), which takes into account both inactivation of sodium channels and activation of potassium channels, is reset to d when an action potential occurs. Therefore, the model contains four constant parameters (i.e., a, b, c, and d). The variable I can be used to deliver external currents or synaptic currents. The four parameters can be varied to create neurons with distinct activity patterns. For example, a can be set to a low value to produce slow membrane recovery, and d can be varied to control the amount of negative feedback from u to v.

The spiking neural network used in this chapter was composed of two types of neurons: regular spiking (RS) excitatory neurons and fast spiking (FS) inhibitory neurons. The network

contained 450 neurons (369 RS neurons and 81 FS neurons) sparsely connected to one another ($p = 0.20$ connectivity probability for each unit). For the RS neurons, we used $a = 0.01$ and $d = 8$ to generate low firing activities. For FS neurons, $a = 0.1$ along with a small negative feedback ($d = 2$) was used to simulate faster spiking activities often observed in cortical inhibitory neurons. The parameters b and c were fixed at 0.13 and -65 mV, respectively (for both RS and FS neurons).

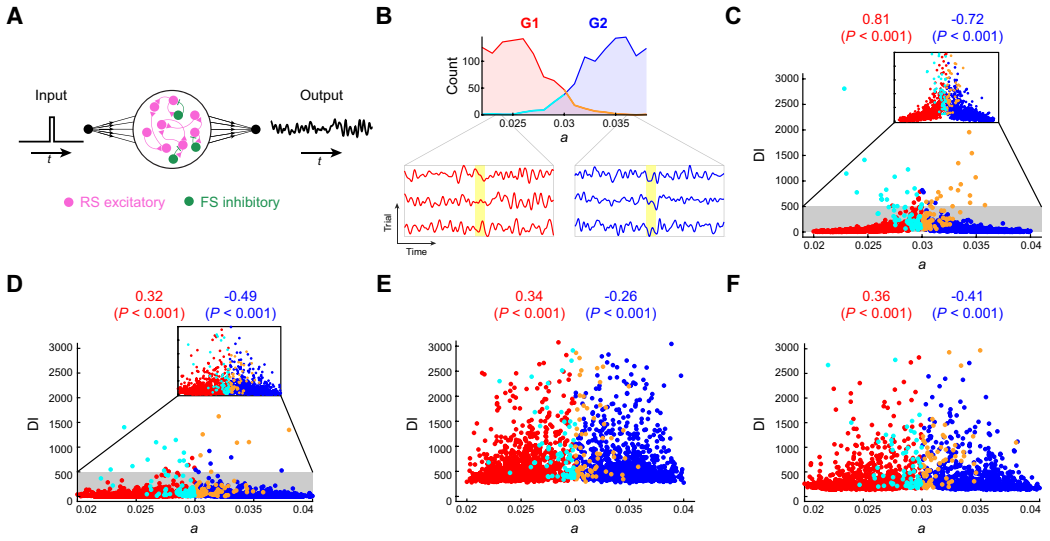


Figure 2.4: (A) Schematic diagram illustrating the spiking neural network simulation. A brief input pulse hyperpolarizes all the excitatory units in a network composed of 450 neurons (369 RS neurons and 81 FS neurons). The mean activity of the excitatory population constitutes the network output signal. (B) Distribution of the dynamical parameter (a) values for each group and example output signals. Yellow shades indicate the input pulse window. (C) DI values extracted from the DDA features of the chaotic time-series signals were strongly correlated with the a values. (D–F) DI values extracted from the signal amplitude values (D), frequency features (E), and Lyapunov exponent estimates (F) were not strongly correlated with the a values. Spearman rank correlation values shown. Note that the correlation coefficient for the second group (blue) is negative, because the DI values for low a values are considered non-representative for this group.

For this experiment, a brief input pulse was delivered to inhibit or hyperpolarize all the excitatory units in a network (Fig. 2.4A). The average activity of the excitatory units was used to compute the network output signal. The parameter a was chosen as the dynamical parameter. For the RS neurons only, a was varied between 0.02 and 0.04 to create networks with varying degrees of excitability. Similar to the first experiment, two groups of a values were generated such

that the first group (G1) contained less excitable networks ($n = 1000$), while the second group (G2) contained networks that were more excitable. Again, the two distributions of the a values overlapped to create non-representative or outlier samples (Fig. 2.4B). For each spiking network, we injected a short hyperpolarizing (i.e., inhibitory) input pulse to inhibit the excitatory units and measured the output signal, estimated as the mean firing rate changes of the excitatory population over time (Fig. 2.4C). This specific paradigm was used to simulate event-related potential (ERP) signals commonly used to measure brain electrical signals in response to brief sensory stimuli.

Applying our algorithm to the output signals revealed that the DI values were highly correlated with the dynamical parameter values (i.e., a values; Spearman rank correlation, $r = 0.81$, $P < 0.0001$ for G1; Spearman rank correlation, $r = -0.72$, $P < 0.0001$ for G2; Fig. 2.4D). These findings suggest that DDA and DI values can capture nonlinear dynamics that are often discarded by linear measures.

2.3.3 Functionally and dynamically distinct subgroups in the COGS-2 dataset

Both MMN and P3a (positive ERP component peaking at 250–300 ms after a stimulus onset) have been shown to be significantly attenuated in SZ [LB05, LST$^+$15, JS15]. Recently, [TGH$^+$17] showed that early auditory processing deficits (reflected by MMN, P3a, and reorienting negativity) in SZ led to impaired cognitive and psychosocial functioning. The nonlinear DDA features (a_3 in Eq. (2.2)) extracted from individual subject deviant minus standard waveform averages revealed that a_3 averages along with MMN and P3a amplitude values were significantly lower across the SZ subjects (mean t value = 10.8, mean Cohen's d = 0.55; Fig. 2.5; [LSK$^+$19]). Averaging across subjects for a_3 and ERP signals highlighted two distinct areas corresponding to the two deviance-detection ERP components (i.e., MMN and P3a). Interestingly, the timing of the two DDA peaks occurred before their corresponding ERP components (Fig. 2.5). For example, the peak group difference in area 1 in Fig. 2.5B occurred 70 ms before the peak

group difference in the ERP MMN window in Fig. 2.5A. Therefore, DDA captured significant group differences preceding previously established ERP biomarkers associated with auditory information processing.

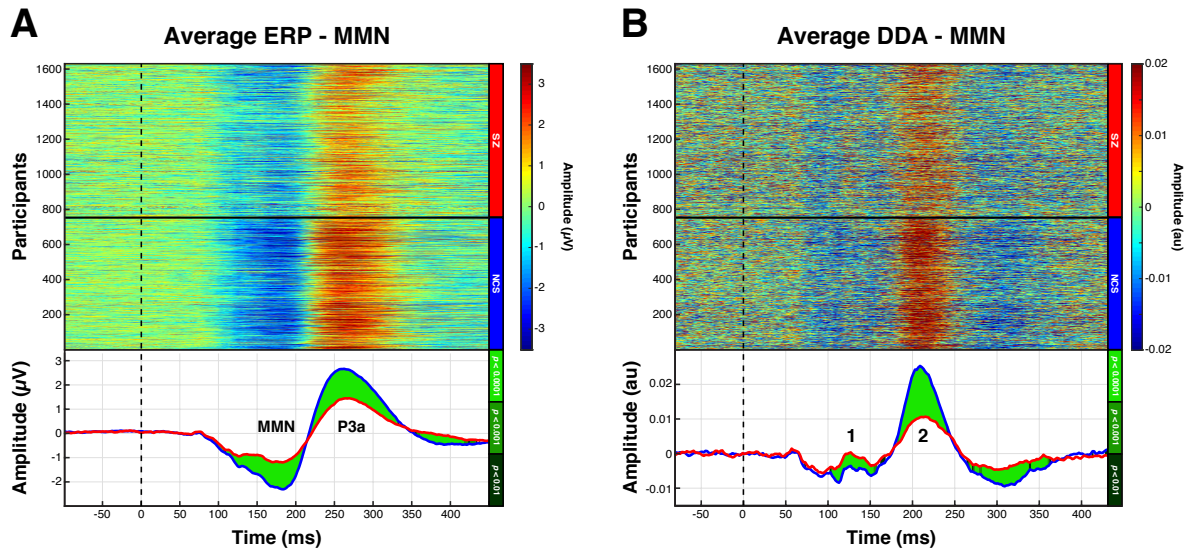


Figure 2.5: DDA identified dynamical state changes preceding each of the auditory deviance response complex components. **A.** NC subjects demonstrated robust MMN and P3a components as shown in the heatmap of the individual subject difference (deviant tone ERP - standard tone ERP) average signals (top panel). MMN and P3a can be appreciated in the group-level average signals (bottom panel). **B.** DDA a_3 coefficient values averaged within each subject revealed significantly decreased a_3 in the schizophrenia participants (top panel). As with the ERP results, the DDA group averages displayed two components with homologous waveform morphology and severity of deficits in schizophrenia, but the DDA components (numbered in the bottom panel) preceded their corresponding ERP peaks identified in **A** by 71 and 54 ms, respectively. The shaded regions in the group average signals represent statistically significant group differences. SZ, schizophrenia; NC, non-psychiatric comparison subjects.

Applying the clustering algorithm (see Section 2.2.2) to the DDA features revealed the distribution of the DI values that gave rise to three dynamically distinct subgroups within the NC and SZ cohorts (Fig. 2.6). For each cohort (NC and SZ subjects), there were highly representative subjects (hNC and hSZ; dark dots in Fig. 2.6). The DI values of these participants were low as their DDA features were indicative of their group dynamics. In addition, there were subjects with high DI values (rNC and rSZ; light dots in Fig. 2.6) suggesting that the nonlinear features of the subjects were non-representative of their group dynamics: DDA features of the SZ subjects

with high DI values were similar to the DDA features of the NC cohort, while the nonlinear dynamics of the NC subjects with high DI values were close to the DDA features of the SZ group. Lastly, each cohort contained a "representative" subgroup whose DI values fell in between the highly representative and non-representative subgroups (rNC and rSZ; Fig. 2.6). Therefore, our dynamical clustering algorithm revealed three subgroups from each cohort.

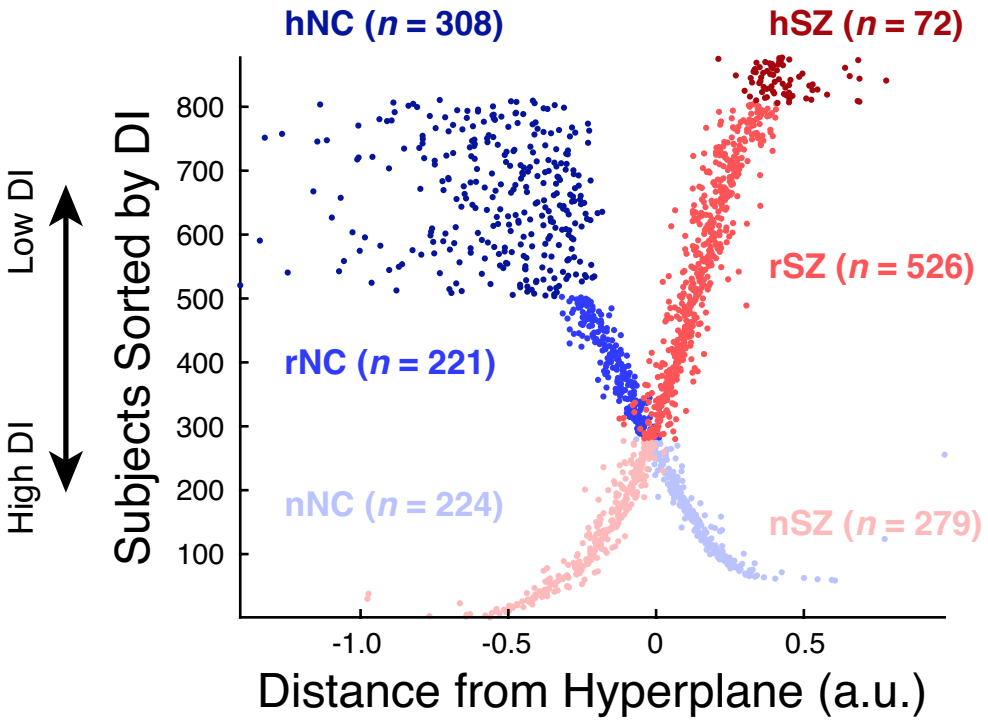


Figure 2.6: Distribution of the DI values computed from the COGS-2 dataset. Based on the DI values, three dynamically distinct subgroups within each cohort were identified. Dark red and dark blue groups indicate subgroups with low DI values, while the light red and light blue dots represent subgroups with high DI values. The high DI subgroups contained non-representative subjects. hNC, highly representative non-psychiatric comparison subjects; rNC, representative non-psychiatric comparison subjects; nNC, non-representative non-psychiatric comparison subjects; hSZ, highly representative schizophrenia; rSZ, representative schizophrenia; nSZ, non-representative schizophrenia.

Next, we wanted to investigate if the dynamic subgroups identified by our algorithm were distinct in the functional and clinical domains. Computing the average deviant minus standard waveforms for each subgroup revealed that the DI values closely tracked the MMN and P3a measures (Fig. 2.7). For the NC cohort, the highly representative subgroup (hNC) had the

highest MMN and P3a amplitudes, while the MMN and P3a were significantly diminished for the non-representative subgroup (Fig. 2.7A). For the SZ cohort, the waveform average of the highly representative subgroup revealed almost non-existent MMN and P3a components (Fig. 2.7B). Interestingly, the non-representative SZ subjects manifested prominent MMN and P3a signals (Fig. 2.7B). These results indicate that DI values are strongly aligned with previously established neurophysiological biomarkers (i.e., MMN and P3a).

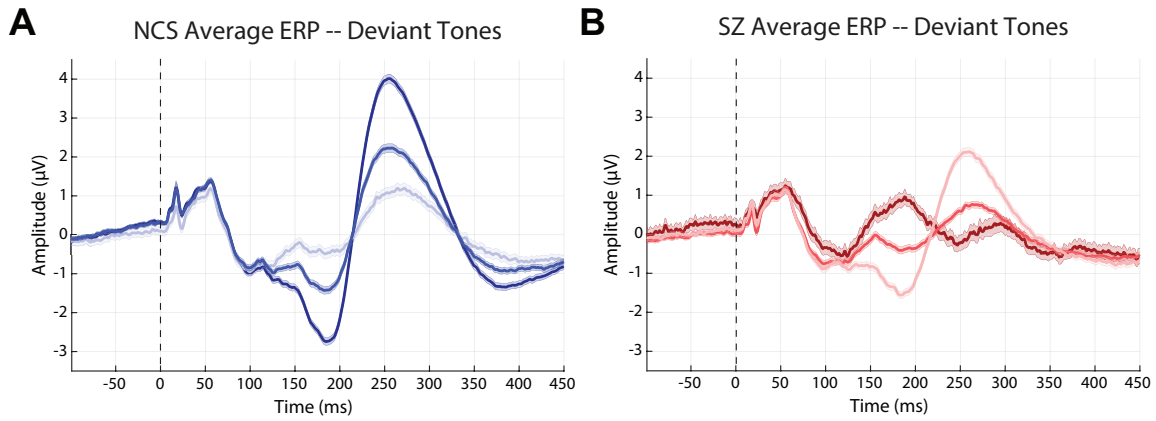


Figure 2.7: The average deviant minus standard waveform signals from the three dynamic subgroups identified within the non-psychiatric comparison (**A**) and schizophrenia (**B**) cohorts. SZ, schizophrenia; NC, non-psychiatric comparison subjects.

Does the highly representative subgroup within the SZ cohort have poor psychosocial and cognitive functioning? Does the non-representative subgroup within the NC group carry an increased risk for SZ? To address these questions, we investigated the cognitive test scores from the computerized neurocognitive battery (CNB) developed by [GRH$^+$10]. The battery is designed to assess several important neurocognitive domains, including working memory, attention, and mental flexibility [GRH$^+$10, IBR$^+$12]. More specifically, we analyzed a total of ten tests: abstraction and mental flexibility, attention, verbal memory, face memory, spatial memory, working memory, spatial processing, emotion identification, sensorimotor processing speed, and motor speed. For each test (excluding sensorimotor processing speed and motor speed), both accuracy and response reaction time measures were included for analysis. For sensorimotor processing speed and motor speed measures, only response time was analyzed. Each measure

was z-scored based on age-matched NC participants. Averaging across all the measures for each participant within each dynamic subgroup revealed a downward trend going from the hNC subgroup to the hSZ subgroup (Fig. 2.8). Within each cohort, the average z-scores were not significantly different among the three subgroups, while the scores from the NC subgroups were significantly greater than those from the SZ subgroups (Kruskal-Wallis test, $H = 133.05$, $P < 0.001$; Dunn's multiple comparison test with $P < 0.05$).

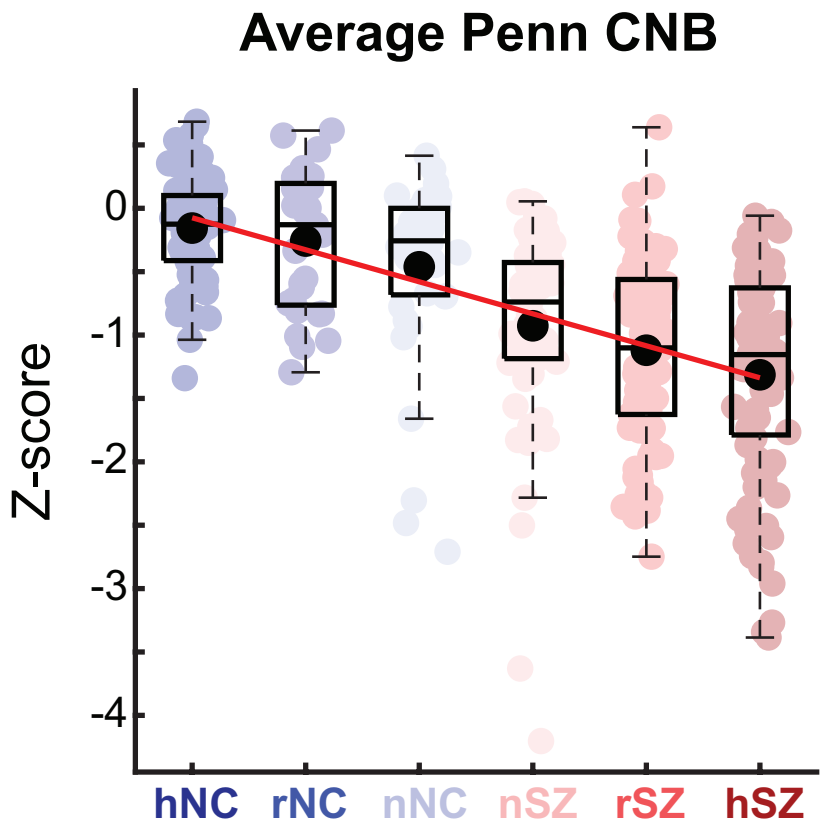


Figure 2.8: Average computerized neurocognitive battery (CNB) measures from the dynamic subgroups. Boxplot central lines, median; black circles, mean; bottom and top edges, lower and upper quartiles; whiskers, 1.5*interquartile range; outliers not plotted; red line, linear fit through the average values. hNC, highly representative non-psychiatric comparison subjects; rNC, representative non-psychiatric comparison subjects; nNC, non-representative non-psychiatric comparison subjects; hSZ, highly representative schizophrenia; rSZ, representative schizophrenia; nSZ, non-representative schizophrenia.

2.4 Discussion

Understanding how multiple brain areas work together to produce emergent behavior and cognition is one of the fundamental challenges that exist in the field of neuroscience. The brain's ability to precisely coordinate these areas in the face of a constantly changing environment is important for maintaining normal psychosocial and cognitive functioning. SZ affects multiple interacting dynamical systems (i.e., sensory processing and attention, working memory) subserving cognition. The degree to which each system is affected may be related to the varying degrees of cognitive impairment observed in these disorders. Therefore, the main focus of this chapter was to characterize these degrees of impairment in the building blocks of cognition by relating large-scale nonlinear state changes detected from brain signals to cognitive functioning.

I have shown that DDA is effective at extracting important nonlinear features which, in turn, could be utilized to reconstruct the distribution of the nonlinear states underlying time-series signals. Using two simulation experiments (Sections 2.3.1 and 2.3.2) containing time-series signals with a wide range of nonlinear dynamics, I also demonstrated that our method based on DDA can accurately estimate the ground truth distributions of the dynamical parameter values. Furthermore, the findings from the second simulation experiment (Section 2.3.2), modeled after local neural microcircuits, suggest that our method could be applied to brain signals empirically measured via EEG or ECoG.

Motivated by our simulation experiments, we applied DDA and our clustering algorithm to an EEG dataset obtained from large cohorts of NC and SZ participants (COGS-2). DDA features computed from the EEG signals revealed important nonlinear dynamical state changes corresponding to ERP components (MMN and P3a) associated with auditory information processing. The DDA features were appreciable immediately in response to auditory stimuli (both standard and deviant tones). Interestingly, DDA detected significant nonlinear activities preceding MMN and P3a components, but the functional importance of these dynamical state changes

remains to be investigated. Applying our clustering method to the dataset revealed previously unidentified subgroups within each cohort. Within each cohort, we characterized three subgroups based on the DI values computed from the DDA features. By probing their neurophysiological and neurocognitive measures, we further demonstrated that our algorithm can detect subgroups that are distinct in not only nonlinear dynamics but in functional and clinical domains.

The work presented in Sections 2.3.1 and 2.3.2 is currently being prepared for publication: Claudia Lainscsek, Robert Kim, Gregory A. Light, and Terrence J. Sejnowski. Dynamical clustering and reconstruction of nonlinear state distributions from noisy signals using DDA.

The work presented in Section 2.3.3 is currently being prepared for publication: Robert Kim, Claudia Lainscsek, Aaron L. Sampson, Michael L. Thomas, The COGS Investigators, Neal R. Swerdlow, David L. Braff, Terrence J. Sejnowski, and Gregory A. Light. Defining subtypes of schizophrenia using unifying dynamical-systems biomarkers.

Bibliography

[Bre17] Michael Breakspear. Dynamic models of large-scale brain activity. *Nature Neuroscience*, 20(3):340–352, Feb 2017.

[Gol89] David E. Goldberg. *Genetic Algorithms in Search, Optimization and Machine Learning*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1st edition, 1989.

[GRA15] Parham Ghorbanian, Subramanian Ramakrishnan, and Hashem Ashrafiuon. Stochastic non-linear oscillator models of EEG: the Alzheimer's disease case. *Front Comput Neurosci*, 9:48, 2015.

[GRH+10] Ruben C. Gur, Jan Richard, Paul Hughett, Monica E. Calkins, Larry Macy, Warren B. Bilker, Colleen Brensinger, and Raquel E. Gur. A cognitive neuroscience-based computerized battery for efficient measurement of individual differences: Standardization and initial construct validation. *Journal of Neuroscience Methods*, 187(2):254 – 262, 2010.

[IBR+12] Farzin Irani, Colleen M. Brensinger, Jan Richard, Monica E. Calkins, Paul J. Moberg, Waren Bilker, Raquel E. Gur, and Ruben C. Gur. Computerized neurocognitive test performance in schizophrenia: A lifespan analysis. *The American Journal of Geriatric Psychiatry*, 20(1):41 – 52, 2012.

[Izh03] Eugene M. Izhikevich. Simple model of spiking neurons. *Trans. Neur. Netw.*, 14(6):1569–1572, November 2003.

[Jeo04] Jaeseung Jeong. EEG dynamics in patients with Alzheimer's disease. *Clin Neurophysiol*, 115(7):1490–505, Jul 2004.

[JS15] Daniel C. Javitt and Robert A. Sweet. Auditory dysfunction in schizophrenia: integrating clinical and basic features. *Nat Rev Neurosci*, 16(9):535–550, 09 2015.

[LB05] Gregory A. Light and David L. Braff. Mismatch Negativity Deficits Are Associated With Poor Functioning in Schizophrenia Patients. *Archives of General Psychiatry*, 62(2):127–136, 02 2005.

[LCC+15] Wen-Yu Lu, Jyun-Yu Chen, Chi-Feng Chang, Wen-Chin Weng, Wang-Tso Lee, and Jiann-Shing Shieh. Multiscale entropy of electroencephalogram as a potential predictor for the prognosis of neonatal seizures. *PLoS One*, 10(12):e0144732, 2015.

[LHW+13a] Claudia Lainscsek, Manuel E. Hernandez, Jonathan Weyhenmeyer, Terrence J. Sejnowski, and Howard Poizner. Non-linear dynamical analysis of EEG time series distinguishes patients with Parkinson's disease from healthy individuals. *Front Neurol*, 4:200, 2013.

[LHW+13b] Claudia Lainscsek, Manuel E. Hernandez, Jonathan Weyhenmeyer, Terrence J Sejnowski, and Howard Poizner. Non-linear dynamical analysis of EEG time series distinguishes patients with Parkinson's disease from healthy individuals. *Frontiers in Neurology*, 4(200), 2013.

[LKJ+07] Christoph Lehmann, Thomas Koenig, Vesna Jelic, Leslie Prichep, Roy E John, Lars-Olof Wahlund, Yadolah Dodge, and Thomas Dierks. Application and comparison of classification algorithms for recognition of Alzheimer's disease in electrical brain activity (EEG). *J Neurosci Methods*, 161(2):342–50, Apr 2007.

[LMP+14] Claudia Lainscsek, Valerie Messager, Adriana Portman, Terrence J. Sejnowski, and Christophe Letellier. Automatic sleep scoring from a single electrode using delay differential equations. In J. Awrejcewicz, editor, *Applied Non-Linear Dynamical Systems, Springer Proceedings in Mathematics & Statistics*, volume 93, pages 371–382. Springer, 2014.

[LRS+12] C. Lainscsek, P. Rowat, L. Schettino, D. Lee, D. Song, C. Letellier, and H. Poizner. Finger tapping movements of Parkinson's disease patients automatically rated using nonlinear delay differential equations. *Chaos*, 22(1):013119, Mar 2012.

[LS13] Claudia Lainscsek and Terrence J. Sejnowski. Electrocardiogram classification using delay differential equations. *Chaos*, 23(2):023132, Jun 2013.

[LSK+19] Claudia Lainscsek, Aaron L. Sampson, Robert Kim, Michael L. Thomas, Karen Man, Xenia Lainscsek, , Neal R. Swerdlow, David L. Braff, Terrence J. Sejnowski, and Gregory A. Light. Nonlinear dynamics underlying sensory processing dysfunction in schizophrenia. *Proceedings of the National Academy of Sciences*, 116(9):3847–3852, 2019.

[LST+15] Gregory A Light, Neal R Swerdlow, Michael L Thomas, Monica E Calkins, Michael F Green, Tiffany A Greenwood, Raquel E Gur, Ruben C Gur, Laura C Lazzeroni, Keith H Nuechterlein, Marlena Pela, Allen D Radant, Larry J Seidman, Richard F Sharp, Larry J Siever, Jeremy M Silverman, Joyce Sprock, William S Stone, Catherine A Sugar, Debby W Tsuang, Ming T Tsuang, David L Braff, and Bruce I Turetsky. Validation of mismatch negativity and P3a for use in multi-site studies of Schizophrenia: characterization of demographic, clinical, cognitive, and functional correlates in COGS-2. *Schizophr Res*, 163(1-3):63–72, Apr 2015.

[LWCS17a] C. Lainscsek, J. Weyhenmeyer, S. S. Cash, and T. J. Sejnowski. Delay differential analysis of seizures in multichannel electrocorticography data. *Neural Computation*, 29:3181–3281, 2017.

[LWCS17b] Claudia Lainscsek, Jonathan Weyhenmeyer, Sydney S. Cash, and Terrence J. Sejnowski. Delay differential analysis of seizures in multichannel electrocorticography data. *Neural Computation*, 29(12):3181–3218, Dec 2017.

[MLP+01] Viktor Müller, Werner Lutzenberger, Friedemann Pulvermüller, Bettina Mohr, and Niels Birbaumer. Investigation of brain dynamics in Parkinson's disease by methods derived from nonlinear dynamics. *Exp Brain Res*, 137:103–110, 2001.

[MM05] E. Miletics and G. Molnárka. Implicit extension of Taylor series method with numerical derivatives for initial value problems. *Computers & Mathematics with Applications*, 50(7):1167–1177, 2005.

[PJR01] L. Pezard, R. Jech, and E. Růzicka. Investigation of non-linear properties of multichannel EEG in the early stages of Parkinson's disease. *Clin Neurophysiol*, 112(1):38–45, Jan 2001.

[PTVF92] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, USA, second edition, 1992.

[R$\ddot{7}$6] O. E. Rössler. An equation for continuous chaos. *Phys Lett A*, 57:397, 1976.

[SAE15] Samantha Simons, Daniel Abasolo, and Javier Escudero. Classification of Alzheimer's disease from quadratic sample entropy of electroencephalogram. *Healthc Technol Lett*, 2(3):70–3, Jun 2015.

[SSM+97] Karen A. Spindler, Edith V. Sullivan, Vinod Menon, Kelvin O. Lim, and Adolf Pfefferbaum. Deficits in multiple systems of working memory in schizophrenia. *Schizophrenia Research*, 27(1):1 – 10, 1997.

[Tak81] Floris Takens. *Detecting strange attractors in turbulence*, pages 366–381. Springer Berlin Heidelberg, Berlin, Heidelberg, 1981.

[TGH+17] Michael L. Thomas, Michael F. Green, Gerhard Hellemann, Catherine A. Sugar, Melissa Tarasenko, Monica E. Calkins, Tiffany A. Greenwood, Raquel E. Gur, Ruben C. Gur, Laura C. Lazzeroni, Keith H. Nuechterlein, Allen D. Radant, Larrry J. Seidman, Alexandra L. Shiluk, Larry J. Siever, Jeremy M. Silverman, Joyce Sprock, William S. Stone, Neal R. Swerdlow, Debby W. Tsuang, Ming T. Tsuang, Brece I. Turetsky, David L. Braff, and Gregory A. Light. Modeling deficits from early auditory information processing to psychosocial functioning in schizophrenia. *JAMA Psychiatry*, 74(1):37–46, 2017.

[WSSV85] Alan Wolf, Jack B. Swift, Harry L. Swinney, and John A. Vastano. Determining lyapunov exponents from a time series. *Physica D: Nonlinear Phenomena*, 16(3):285 – 317, 1985.

[YM16] R. Yuvaraj and M. Murugappan. Hemispheric asymmetry non-linear analysis of EEG during emotional responses from idiopathic Parkinson's disease patients. *Cogn Neurodyn*, 10(3):225–34, Jun 2016.

[YZLC11] Qi Yuan, Weidong Zhou, Shufang Li, and Dongmei Cai. Epileptic EEG classification based on extreme learning machine and nonlinear features. *Epilepsy Res*, 96(1-2):29–38, Sep 2011.

[YZLW12] Qi Yuan, Weidong Zhou, Yinxia Liu, and Jiwen Wang. Epileptic seizure detection with linear and nonlinear features. *Epilepsy Behav*, 24(4):415–21, Aug 2012.

Chapter 3

Simple framework for constructing functional spiking recurrent neural networks

A biologically plausible computational model could help elucidate neural mechanisms required for performing higher-order cognitive functions. In this chapter, I present a simple framework to construct biologically realistic spiking recurrent neural networks (RNNs) capable of learning a wide range of cognitive tasks. The work presented here is reproduced and adapted from: Robert Kim, Yinghao Li, and Terrence J. Sejnowski. Simple framework for constructing functional spiking recurrent neural networks. *Proceedings of the National Academy of Sciences*, 116(45):22811–22820, 2019.

3.1 Abstract

Cortical microcircuits exhibit complex recurrent architectures that possess dynamically rich properties. The neurons that make up these microcircuits communicate mainly via discrete

spikes, and it is not clear how spikes give rise to dynamics that can be used to perform computationally challenging tasks. In contrast, continuous models of rate-coding neurons can be trained to perform complex tasks. Here, we present a simple framework to construct biologically realistic spiking RNNs capable of learning a wide range of tasks. Our framework involves training a continuous-variable rate RNN with important biophysical constraints and transferring the learned dynamics and constraints to a spiking RNN in a one-to-one manner. The proposed framework introduces only one additional parameter to establish the equivalence between rate and spiking RNN models. We also study other model parameters related to the rate and spiking networks to optimize the one-to-one mapping. By establishing a close relationship between rate and spiking models, we demonstrate that spiking RNNs could be constructed to achieve similar performance as their counterpart continuous rate networks.

3.2 Introduction

Dense recurrent connections common in cortical circuits suggest their important role in computational processes [GR95, FsSY$^+$02, Wan08]. Network models based on RNNs of continuous-variable rate units have been extensively studied to characterize network dynamics underlying neural computations [SCS88, SA09, LB13, MSSN13, KC18, MO18]. Methods commonly used to train rate networks to perform cognitive tasks can be largely classified into three categories: recursive least squares (RLS)-based, gradient-based, and reward-based algorithms. The First-Order Reduced and Controlled Error (FORCE) algorithm, which utilizes RLS, has been widely used to train RNNs to produce complex output signals [SA09] and to reproduce experimental results [LB13, EPQD16, RHT16]. Gradient descent-based methods, including Hessian-free methods, have been also successfully applied to train rate networks in a supervised manner and to replicate the computational dynamics observed in networks from behaving animals [MSSN13, BSR$^+$13, SYW16]. Unlike the previous two categories (i.e., RLS-based and

gradient-based algorithms), reward-based learning methods are more biologically plausible and have been shown to be as effective in training rate RNNs as the supervised learning methods [SYW17, Mic17, WKNK$^+$18, ZCL$^+$18]. Even though these models have been vital in uncovering previously unknown computational mechanisms, continuous rate networks do not incorporate basic biophysical constraints such as the spiking nature of biological neurons.

Training spiking network models where units communicate with one another via discrete spikes is more difficult than training continuous rate networks. The non-differentiable nature of spike signals prevents the use of gradient descent-based methods to train spiking networks directly, although several differentiable models have been proposed [HS18, LDP16]. Due to this challenge, FORCE-based learning algorithms have been most commonly used to train spiking recurrent networks. While recent advances have successfully modified and applied FORCE training to construct functional spike RNNs [ADM16, DCA16, TUKM16, NC17, KC18], FORCE training is computationally inefficient and unstable when connectivity constraints, including separate populations for excitatory and inhibitory populations (Dale's principle) and sparse connectivity patterns, are imposed [DCA16].

Due to these limitations, computational capabilities of spiking networks that abide by biological constraints have been challenging to explore. For instance, it is not clear if spiking RNNs operating in a purely rate-coding regime can perform tasks as complex as the ones rate RNN models are trained to perform. If such spiking networks can be constructed, then it would be important to characterize how much spiking-related noise not present in rate networks affects the performance of the networks. Establishing the relationship between these two types of RNN models could also serve as a good starting point for designing power-efficient spiking networks that can incorporate both rate and temporal coding.

To address the above questions, we present a computational framework for directly mapping rate RNNs with basic biophysical constraints to leaky integrate-and-fire (LIF) spiking RNNs without significantly compromising task performance. Our method introduces only one

additional parameter to place the spiking RNNs in the same dynamic regime as their counterpart rate RNNs, and takes advantage of the previously established methods to efficiently optimize network parameters while adhering to biophysical restrictions. These previously established methods include training a continuous-variable rate RNN using a gradient descent-based method [Wer90, MS11, PMB13, BBP13] and connectivity weight matrix parametrization method to impose Dale's principle [SYW16]. The gradient descent learning algorithm allowed us to easily optimize many parameters including the connectivity weights of the network and the synaptic decay time constant for each unit. The weight parametrization method proposed by [SYW16] was utilized to enforce Dale's principles and additional connectivity patterns without significantly affecting computational efficiency and network stability.

Combining these two existing methods with correct parameter values enabled us to directly map rate RNNs trained with backpropagation to LIF RNNs in a one-to-one manner. The parameters critical for mapping to succeed included the network size, the nonlinear activation function employed for training rate RNNs, and a constant factor for scaling down the connectivity weights of the trained rate RNNs. Here, we investigated these parameters along with other LIF parameters and identified the range of values required for the mapping to be effective. We demonstrate that when these parameters are set to their optimal values, the LIF models constructed from our framework can perform the same tasks the rate models are trained to perform equally well.

3.3 Materials and methods

The implementation of our framework and the codes to generate all the figures in this chapter are available at `https://github.com/rkim35/spikeRNN`. The repository also contains implementation of other tasks including autonomous oscillation and delayed match-to-sample (DMS) tasks.

3.3.1 Continuous rate network structure

The continuous rate RNN model contains N units recurrently connected to one another. The dynamics of the model is governed by

$$\boldsymbol{\tau}^d \frac{d\boldsymbol{x}}{dt} = -\boldsymbol{x} + W^{rate}\boldsymbol{r}^{rate} + \boldsymbol{I}_{ext} \tag{3.1}$$

where $\boldsymbol{\tau}^d \in \mathbb{R}^{1\times N}$ corresponds to the synaptic decay time constants for the N units in the network (see Section 3.3.3 on how these are initialized and optimized), $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ is the synaptic current variable, $W^{rate} \in \mathbb{R}^{N\times N}$ is the synaptic connectivity matrix, and $\boldsymbol{r}^{rate} \in \mathbb{R}^{1\times N}$ is the output of the units. The output of each unit, which can be interpreted as the firing rate estimate, is obtained by applying a nonlinear transfer function to the synaptic current variable ($\boldsymbol{x}$) elementwise:

$$\boldsymbol{r}^{rate} = \phi(\boldsymbol{x})$$

We use a standard logistic sigmoid function for the transfer function to constrain the firing rates to be non-negative:

$$\phi(\boldsymbol{x}) = \frac{1}{1+\exp(-\boldsymbol{x})} \tag{3.2}$$

The connectivity weight matrix (W^{rate}) is initialized as a random, sparse matrix drawn from a normal distribution with zero mean and a standard deviation of $1.5/\sqrt{N \cdot P_c}$ where $P_c = 0.20$ is the initial connectivity probability.

The external currents ($\boldsymbol{I}_{ext}$) include task-specific input stimulus signals (see Section 3.6) along with a Gaussian white noise variable:

$$\boldsymbol{I}_{ext} = W_{in}\boldsymbol{u} + \mathcal{N}(0, 0.01)$$

where the time-varying stimulus signals ($\boldsymbol{u} \in \mathbb{R}^{N_{in}\times 1}$) are fed to the network via $W_{in} \in \mathbb{R}^{N\times N_{in}}$,

a Gaussian random matrix with zero mean and unit variance. N_{in} corresponds to the number of input signals associated with a specific task, and $\mathcal{N}(0, 0.01) \in \mathbb{R}^{N\times 1}$ represents a Gaussian random noise with zero mean and variance of 0.01.

The output of the rate RNN at time t is computed as a linear readout of the population activity:

$$o^{rate}(t) = W_{out}^{rate} \boldsymbol{r}^{rate}(t)$$

where $W_{out}^{rate} \in \mathbb{R}^{1\times N}$ refers to the readout weights.

Eq. (3.1) is discretized using the first-order Euler approximation method:

$$\begin{aligned} \boldsymbol{x}_t &= \left(1 - \frac{\Delta t}{\boldsymbol{\tau}^d}\right) \boldsymbol{x}_{t-1} + \frac{\Delta t}{\boldsymbol{\tau}^d} (W^{rate} \boldsymbol{r}_{t-1}^{rate} + W_{in} \boldsymbol{u}_{t-1}) \\ &+ \mathcal{N}(0, 0.01) \end{aligned}$$

where $\Delta t = 5$ ms is the discretization time step size used throughout this study.

3.3.2 Spiking network structure

For our spiking RNN model, we considered a network of leaky integrate-and-fire (LIF) units governed by

$$\tau_m \frac{d\boldsymbol{v}}{dt} = -\boldsymbol{v} + W^{spk} \boldsymbol{r}^{spk} + \boldsymbol{I}_{ext} \tag{3.3}$$

In the above equation, $\tau_m = 10$ ms is the membrane time constant shared by all the LIF units, $\boldsymbol{v} \in \mathbb{R}^{1\times N}$ is the membrane voltage variable, $W^{spk} \in \mathbb{R}^{N\times N}$ is the recurrent connectivity matrix, and $\boldsymbol{r}^{spk} \in \mathbb{R}^{1\times N}$ represents the spike trains filtered by a synaptic filter. Throughout the study, the

double exponential synaptic filter was used to filter the presynaptic spike trains:

$$\begin{aligned} \frac{dr_i^{spk}}{dt} &= -\frac{r_i^{spk}}{\tau_i^d} + s_i \\ \frac{ds_i}{dt} &= -\frac{s_i}{\tau_r} + \frac{1}{\tau_r \tau_i^d} \sum_{t_i^k < t} \delta(t - t_i^k) \end{aligned}$$

where $\tau_r = 2$ ms and τ_i^d refer to the synaptic rise time and the synaptic decay time for unit i, respectively. The synaptic decay time constant values ($\tau_i^d \in \boldsymbol{\tau}^d$) are trained and transferred to our LIF RNN model (see Section 3.3.3). The spike train produced by unit i is represented as a sum of Dirac δ functions, and t_i^k refers to the k-th spike emitted by unit i.

The external current input ($\boldsymbol{I}_{ext}$) is similar to the one used in our continuous model (see Section 3.3.1). The only difference is the addition of a constant background current set near the action potential threshold (see below).

The output of our spiking model at time t is given by

$$o^{spk}(t) = W_{out}^{spk} \boldsymbol{r}^{spk}(t)$$

Other LIF model parameters were set to the values used by [NC17]. These include the action potential threshold (-40 mV), the reset potential (-65 mV), the absolute refractory period (2 ms), and the constant bias current (-40 pA). The parameter values for the LIF and the quadratic integrate-and-fire (QIF) models are listed in Section 3.6, Table 3.1.

3.3.3 Training details

In this study, we only considered supervised learning tasks. A task-specific target signal ($\boldsymbol{z}$) is used along with the rate RNN output ($\boldsymbol{o}^{rate}$) to define the loss function ($\mathcal{L}$), which our rate RNN model is trained to minimize. Throughout the study, we used the root mean squared error

(RMSE) defined as

$$\mathcal{L} = \sqrt{\left(\sum_{t=1}^{T}(z(t) - o^{rate}(t))^2\right)} \tag{3.4}$$

where T is the total number of time points in a single trial.

In order to train the rate model to minimize the above loss function (Eq. 3.4), we employed Adaptive Moment Estimation (ADAM) stochastic gradient descent algorithm. The learning rate was set to 0.01, and the TensorFlow default values were used for the first and second moment decay rates. The gradient descent method was used to optimize the following parameters in the rate model: synaptic decay time constants ($\boldsymbol{\tau}^d$), recurrent connectivity matrix (W^{rate}), and readout weights (W^{rate}_{out}).

Here we describe the method to train synaptic decay time constants ($\boldsymbol{\tau}^d$) using back-propagation. First, the time constants are initialized with random values within the specified range:

$$\boldsymbol{\tau}^d = \sigma(\mathcal{N}(0,1)) \cdot \tau_{step} + \tau^d_{min}$$

where $\sigma(\cdot)$ is the sigmoid function (identical to Eq. 3.2) used to constrain the time constants to be non-negative. The time constant values are also bounded by the minimum (τ^d_{min}) and the maximum ($\tau^d_{max} = \tau^d_{min} + \tau_{step}$) values. The error computed from the loss function (Eq. 3.4) is then backpropagated to update the time constants at each iteration:

$$\frac{\partial \mathcal{L}}{\partial \boldsymbol{\tau}^d} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{r}} \cdot \frac{\partial \boldsymbol{r}}{\partial \boldsymbol{x}} \cdot \frac{\partial \boldsymbol{x}}{\partial \boldsymbol{\tau}^d}$$

The method proposed by [SYW16] was used to impose Dale's principle and create separate excitatory and inhibitory populations. Briefly, the recurrent connectivity matrix (W^{rate}) in the rate model is parametrized by

$$W^{rate} = [W^{rate}]_+ \cdot D \tag{3.5}$$

where the rectified linear operation ($[\cdot]_+$) is applied to the connectivity matrix at each update step. The diagonal matrix ($D \in \mathbb{R}^{N\times N}$) contains +1's for excitatory units and -1's for inhibitory units in the network. Each unit in the network is randomly assigned to one group (excitatory or inhibitory) before training, and the assignment does not change during training (i.e., D stays fixed).

To impose specific connectivity patterns, we apply a binary mask ($M \in \mathbb{R}^{N\times N}$) to Eq. 3.5:

$$W^{rate} = \left([W^{rate}]_+ \cdot D\right) \odot M$$

where $\odot$ refers to the Hadamard operation (elementwise multiplication). Similar to the diagonal matrix (D), the mask matrix stays fixed throughout training. For example, the following mask matrix can be used to create a subgroup of inhibitory units (Group A) that do not receive synaptic inputs from the rest of the inhibitory units (Group B) in the network (Fig. 3.10):

$$m_{ij} = \begin{cases} 0 & i \in \text{Group A}, j \in \text{Group B} \\ 1 & \text{otherwise} \end{cases}$$

where $m_{ij} \in M$ establishes (if $m_{ij} = 1$) or removes (if $m_{ij} = 0$) the connection from unit j to unit i.

3.3.4 Transfer learning from a rate model to a spiking model

In this section, we describe the method that we developed to perform transfer learning from a trained rate model to a LIF model. Once the rate RNN model is trained using the gradient descent method, the rate model parameters are transferred to a LIF network in a one-to-one manner. First, the LIF network is initialized to have the same topology as the trained rate RNN. Next, the input weight matrix (W_{in}) and the synaptic decay time constants ($\boldsymbol{\tau}^d$) are transferred to the spiking RNN without any modification. Lastly, the recurrent connectivity matrix (W^{rate}) and the readout weights (W_{out}^{rate}) are scaled by a constant number, λ, and transferred to the spiking

network.

If the recurrent connectivity weights from the trained rate model are transferred to a spiking network without any changes, the spiking model produces largely fluctuating signals (as illustrated in Fig. 3.2B), because the LIF firing rates are significantly larger than 1 (whereas the firing rates of the rate model are constrained to range between zero and one by the sigmoid transfer function).

To place the spiking RNN in the similar dynamic regime as the rate network, we first assume a linear relationship between the rate model connectivity weights and the spike model weights:

$$W^{spk} = \lambda \cdot W^{rate}$$

Using the above assumption, the synaptic drive (d) that unit i in the LIF RNN receives can be expressed as

$$\begin{aligned} d_i^{spk}(t) &= \sum_{j=1}^{N} w_{ij}^{spk} \cdot r_j^{spk}(t) \\ &\approx \sum_{j=1}^{N} (\lambda \cdot w_{ij}^{rate}) \cdot r_j^{spk}(t) \\ &= \sum_{j=1}^{N} w_{ij}^{rate} \cdot (\lambda \cdot r_j^{spk}(t)) \end{aligned} \tag{3.6}$$

where $w_{ij}^{spk} \in W^{spk}$ is the synaptic weight from unit j to unit i.

Similarly, unit i in the rate RNN model receives the following synaptic drive at time t:

$$d_i^{rate}(t) = \sum_{j=1}^{N} w_{ij}^{rate} \cdot r_j^{rate}(t) \tag{3.7}$$

If we set the above two synaptic drives (Eq. 3.6 and Eq. 3.7) equal to each other, we have:

$$\begin{aligned} d_i^{spk}(t) &= d_i^{rate}(t) \\ \sum_{j=1}^{N} w_{ij}^{rate} \cdot (\lambda \cdot r_j^{spk}(t)) &= \sum_{j=1}^{N} w_{ij}^{rate} \cdot r_j^{rate}(t) \end{aligned} \tag{3.8}$$

Generalizing Eq. 3.8 to all the units in the network, we have

$$\boldsymbol{r}^{rate}(t) = \lambda \cdot \boldsymbol{r}^{spk}(t)$$

Therefore, if there exists a constant factor (λ) that can account for the firing rate scale difference between the rate and the spiking models, the connectivity weights from the rate model (W^{rate}) can be scaled by the factor and transferred to the spiking model.

The readout weights from the rate model (W_{out}^{rate}) are also scaled by the same constant factor (λ) to have the spiking network produce output signals similar to the ones from the trained rate model:

$$\begin{aligned} o^{rate}(t) &= W_{out}^{rate} \cdot \boldsymbol{r}^{rate}(t) \\ &\approx W_{out}^{rate} \cdot (\lambda \cdot \boldsymbol{r}^{spk}(t)) \\ &= (\lambda \cdot W_{out}^{rate}) \cdot \boldsymbol{r}^{spk}(t) = o^{spk}(t) \end{aligned}$$

In order to find the optimal scaling factor, we developed a simple grid search algorithm. For a given range of values for $1/\lambda$ (ranged from 20 to 75 with a step size of 5), the algorithm finds the optimal value that maximizes the task performance.

3.4 Results

Here we provide a brief overview of the two types of recurrent neural networks (RNNs) that we employed throughout this study (more details in Section 3.3): continuous-variable firing rate RNNs and spiking RNNs. The continuous-variable rate network model consisted of N rate units whose firing rates were estimated via a nonlinear input-output transfer function [SCS88, SA09]. The model was governed by the following set of equations:

$$\tau_i^d \frac{dx_i}{dt} = -x_i + \sum_{j=1}^{N} w_{ij}^{rate} r_j^{rate} + I_{ext} \tag{3.9}$$

$$r_i^{rate} = \phi(x_i) \tag{3.10}$$

where τ_i^d is the synaptic decay time constant for unit i, x_i is the synaptic current variable for unit i, w_{ij}^{rate} is the synaptic strength from unit j to unit i, and I_{ext} is the external current input to unit i. The firing rate of unit i (r_i^{rate}) is given by applying a nonlinear transfer function ($\phi(\cdot)$) to the synaptic current variable. Since the firing rates in spiking networks cannot be negative, we chose the activation function for our rate networks to be a non-negative saturating function (standard sigmoid function) and parametrized the connectivity matrix ($w_{ij}^{rate} \in W^{rate}$) to enforce Dale's principle and additional connectivity constraints (see Section 3.3).

The second RNN model that we considered was a network composed of N spiking units. Throughout this study, we focused on networks of leaky integrate-and-fire (LIF) units whose membrane voltage dynamics were given by:

$$\tau_m \frac{dv_i}{dt} = -v_i + \sum_{j=1}^{N} w_{ij}^{spk} r_j^{spk} + I_{ext} \tag{3.11}$$

where τ_m is the membrane time constant (set to 10 ms throughout this study), v_i is the membrane voltage of unit i, w_{ij}^{spk} is the synaptic strength from unit j to unit i, r_j^{spk} represents the synaptic filtering of the spike train of unit j, and I_{ext} is the external current source. The discrete nature of

r_j^{spk} (see Section 3.3) has posed a major challenge for directly training spiking networks using gradient-based supervised learning. Even though the main results presented here are based on LIF networks, our method can be generalized to quadratic integrate-and-fire (QIF) networks with only few minor changes to the model parameters (Section 3.6, Table 3.1).

Continuous rate network training was implemented using the open-source software library TensorFlow in Python, while LIF/QIF network simulations along with the rest of the analyses were performed in MATLAB.

3.4.1 Training continuous rate networks

Throughout this study, we used a gradient-descent supervised method, known as Back-propagation Through Time (BPTT), to train rate RNNs to produce target signals associated with a specific task [Wer90, SYW16]. The method we employed is similar to the one used by previous studies ([MS11, BBP13, SYW16]; more details in Section 3.3) with one major difference in synaptic decay time constants. Instead of assigning a single time constant to be shared by all the units in a network, our method tunes a synaptic constant for each unit using BPTT (see Section 3.3). Although tuning of synaptic time constants may not be biologically plausible, this feature was included to model diverse intrinsic synaptic timescales observed in single cortical neurons [SKS$^+$13, WSB$^+$18, CTW$^+$18].

We trained rate RNNs of various sizes on a simple task modeled after a Go-NoGo task to demonstrate our training method (Fig. 3.1). Each network was trained to produce a positive mean population activity approaching +1 after a brief input pulse (Fig. 3.1A). For a trial without an input pulse (i.e., NoGo trial), the networks were trained to maintain the output signal close to zero. The units in a rate RNN were sparsely connected via W^{rate} and received a task-specific input signal through weights (W_{in}) drawn from a normal distribution with zero mean and unit

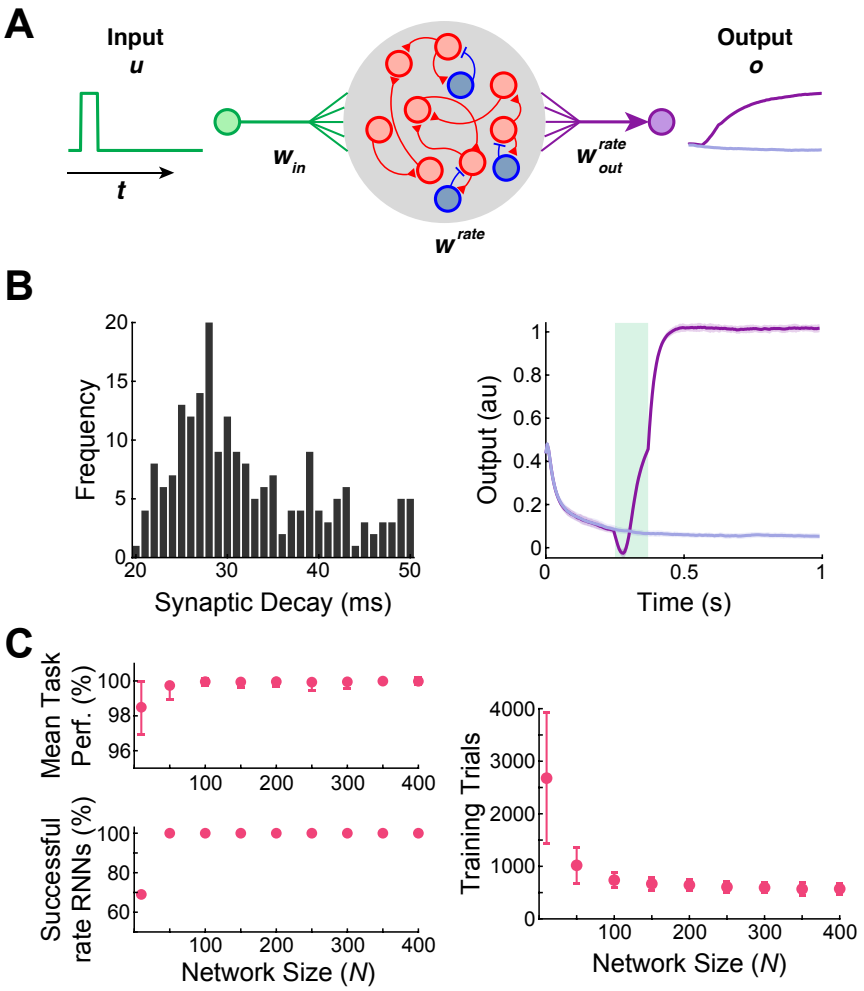


Figure 3.1: **Rate RNNs trained to perform the Go-NoGo task. A.** Schematic diagram illustrating a continuous rate RNN model trained to perform the Go-NoGo task. The rate RNN model contained excitatory (red circles) and inhibitory (blue circles) units. **B.** Distribution of the tuned synaptic decay time constants (Mean $\pm$ SD, 28.2 $\pm$ 9.4 ms; left) and the average trained rate RNN task performance (right) from an example rate RNN model. The mean $\pm$ SD output signals from 50 Go trials (dark purple) and from 50 NoGo trials (light purple) are shown. The green box represents the input stimulus given for the Go trials. The rate RNN contained 200 units (169 excitatory and 31 inhibitory units). **C.** Rate RNNs with different network sizes trained to perform the Go-NoGo task. For each network size, 100 RNNs with random initial conditions were trained. All the networks successfully trained performed the task almost perfectly (range 96–100%; left). As the network size increased, the number of training trials decreased (Mean $\pm$ SD shown; right).

variance. The network output (o^{rate}) was then computed using a set of linear readout weights:

$$o^{rate}(t) = W_{out}^{rate} \cdot \boldsymbol{r}^{rate}(t) \tag{3.12}$$

where W_{out}^{rate} is the readout weights and $\boldsymbol{r}^{rate}(t)$ is the firing rate estimates from all the units in the network at time t. The recurrent weight matrix (W^{rate}), the readout weights (W_{out}^{rate}), and the synaptic decay time constants ($\boldsymbol{\tau^d}$) were optimized during training, while the input weight matrix (W_{in}) stayed fixed (see Section 3.3).

The network size (N) was varied from 10 to 400 (9 different sizes), and 100 networks with random initializations were trained for each size. For all the networks, the minimum and the maximum synaptic decay time constants were fixed to 20 ms and 50 ms, respectively. As expected, the smallest rate RNNs ($N = 10$) took the longest to train, and only 69% of the rate networks with $N = 10$ were successfully trained (see Section 3.6 for training termination criteria; Fig. 3.1C).

3.4.2 One-to-one mapping from continuous rate networks to spiking networks

We developed a simple procedure that directly maps dynamics of a trained continuous rate RNN to a spiking RNN in a one-to-one manner.

In our framework, the three sets of the weight matrices (W_{in}, W^{rate}, and W_{out}^{rate}) along with the tuned synaptic time constants ($\boldsymbol{\tau^d}$) from a trained rate RNN are transferred to a network of LIF spiking units. The spiking RNN is initialized to have the same topology as the rate RNN. The input weight matrix and the synaptic time constants are simply transferred without any modification, but the recurrent connectivity and the readout weights need to be scaled by a constant factor (λ) in order to account for the difference in the firing rate scales between the rate model and the spiking model (see Section 3.3; Fig. 3.2A). The effects of the scaling factor is clear in an example LIF RNN model constructed from a rate model trained to perform the Go-NoGo task (Fig. 3.2B). With an appropriate value for λ, the LIF network performed the task with the same accuracy as the rate network, and the LIF units fired at rates similar to the "rates" of the continuous network units (Section 3.6, Fig. 3.8). In addition, the LIF network reproduced

the population dynamics of the rate RNN model as shown by the time evolution of the top three principal components extracted by the principal component analysis (Section 3.6, Fig. 3.9).

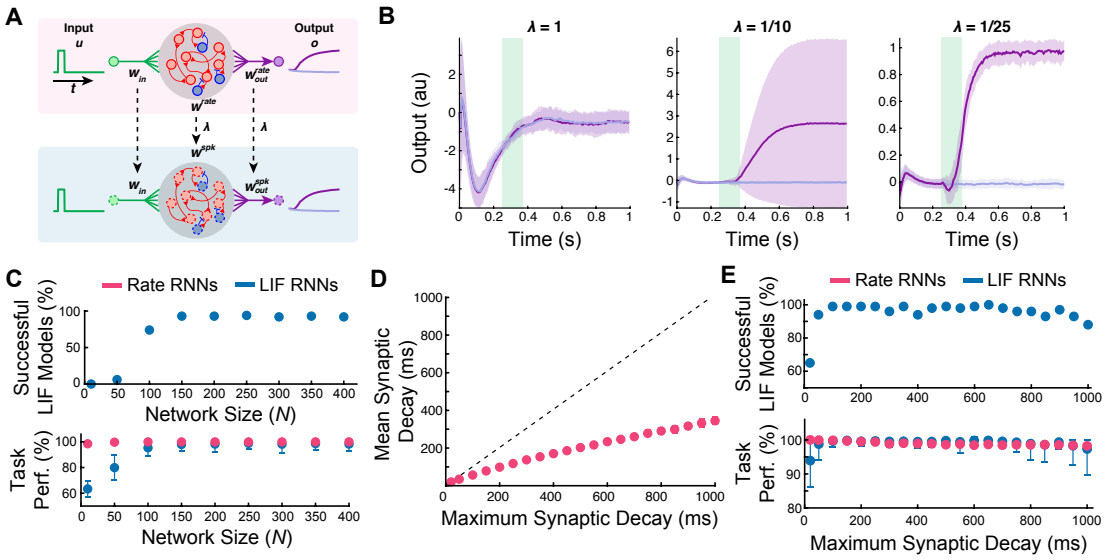


Figure 3.2: **Mapping trained rate RNNs to LIF RNNs for the Go-NoGo task. A.** Schematic diagram illustrating direct mapping from a continuous rate RNN model (top) to a spiking RNN model (bottom). The optimized synaptic decay time constants ($\boldsymbol{\tau^d}$) along with the weight parameters (W_{in}, W^{rate}, and W^{rate}_{out}) were transferred to a spiking network with LIF units (red and blue circles with a dashed outline). The connectivity and the readout weights were scaled by a constant factor, λ. **B.** LIF RNN performance on the Go-NoGo task without scaling ($\lambda = 1$; left), with insufficient scaling (middle), and with appropriate scaling (right). The network contained 200 units (169 excitatory and 31 inhibitory units). Mean $\pm$ SD over 50 Go and 50 NoGo trials. **C.** Successfully converted LIF networks and their average task performance on the Go-NoGo task with different network sizes. All the rate RNNs trained in Fig. 3.1 were converted to LIF RNNs. The network size was varied from $N = 10$ to 400. **D.** Average synaptic decay values for $N = 250$ across different maximum synaptic decay constants. **E.** Successfully converted LIF networks and their average task performance on the Go-NoGo task with fixed network size ($N = 250$) and different maximum synaptic decay constants. The maximum synaptic decay constants were varied from 20 ms to 1000 ms.

Using the procedure outlined above, we converted all the rate RNNs trained in the previous section to spiking RNNs. Only the rate RNNs that successfully performed the task (i.e., training termination criteria met within the first 6000 trials) were converted. Fig. 3.2C characterizes the proportion of the LIF networks that successfully performed the Go-NoGo task ($\geq 95\%$ accuracy; same threshold used to train the rate models; see Section 3.6) and the average task performance of the LIF models for each network size group. For each conversion, the scaling factor (λ) was

determined via a grid search method (see Section 3.3). The LIF RNNs constructed from the small rate networks ($N = 10$ and $N = 50$) did not perform the task reliably, but the LIF model became more robust as the network size increased, and the performance gap between the rate RNNs and the LIF RNNs was the smallest for $N = 250$ (Fig. 3.2C).

In order to investigate the effects of the synaptic decay time constants on the mapping robustness, we trained rate RNNs composed of 250 units ($N = 250$) with different maximum time constants (τ^d_{max}). The minimum time constant (τ^d_{min}) was fixed to 20 ms, while the maximum constant was varied from 20 ms to 1000 ms. For the first case (i.e., $\tau^d_{min} = \tau^d_{max} = 20$ ms), the synaptic decay time constants were not trained and fixed to 20 ms for all the units in a rate RNN. For each maximum constant value, 100 rate RNNs with different initial conditions were trained, and only successfully trained rate networks were converted to spiking RNNs. For each maximum synaptic decay condition, all 100 rate RNNs were successfully trained. As the maximum decay constant increased, the average tuned synaptic decay constants increased sub-linearly (Fig. 3.2D). For the shortest synaptic decay time constant considered (20 ms), the average task performance was the lowest at $93.91 \pm 7.78\%$, and 65% of the converted LIF RNNs achieved at least 95% accuracy (Fig. 3.2E). The LIF models for the rest of the maximum synaptic decay conditions were robust. Although this might indicate that tuning of $\boldsymbol{\tau^d}$ is important for the conversion of rate RNNs to LIF RNNs, we further investigated the effects of the optimization of $\boldsymbol{\tau^d}$ in the last section (see Section 3.4.4).

Our framework also allows seamless integration of additional functional connectivity constraints. For example, a common cortical microcircuitry motif where somatostatin-expressing interneurons inhibit both pyramidal and parvalbumin-positive neurons can be easily implemented in our framework (see Section 3.3 and Section 3.6, Fig. 3.10). In addition, Dale's principle is not required for our framework (see Section 3.6, Fig. 3.11).

3.4.3 LIF networks for context-dependent input integration

The Go-NoGo task considered in the previous section did not require complex cognitive computations. In this section, we consider a more complex task and probe whether spiking RNNs can be constructed from trained rate networks in a similar fashion. The task considered here is modeled after the context-dependent sensory integration task employed by [MSSN13]. Briefly, [MSSN13] trained rhesus monkeys to integrate inputs from one sensory modality (dominant color or dominant motion of randomly moving dots) while ignoring inputs from the other modality. A contextual cue was also given to instruct the monkeys which sensory modality they should attend to. The task required the monkeys to utilize flexible computations as the same modality can be either relevant or irrelevant depending on the contextual cue. Previous works have successfully trained continuous rate RNNs to perform a simplified version of the task and replicated the neural dynamics present in the experimental data [MSSN13, SYW16, Mic17]. Using our framework, we constructed the first spiking RNN model to our knowledge that can perform the task and capture the dynamics observed in the experimental data.

For the task paradigm, we adopted a similar design as the one used by the previous modeling studies [MSSN13, SYW16, Mic17]. A network of recurrently connected units received two streams of noisy input signals along with a constant-valued signal that encoded the contextual cue (Fig. 3.3A; see Section 3.3). To simulate a noisy sensory input signal, a random Gaussian time-series signal with zero mean and unit variance was first generated. Each input signal was then shifted by a positive or negative constant ("offset") to encode evidence toward the (+) or (-) choice, respectively. Therefore, the offset value determined how much evidence for the specific choice was represented in the noisy input signal. The network was trained to produce an output signal approaching +1 (or -1) if the cued input signal had a positive (or negative) mean. For example, if the cued input signal was generated using a positive offset value, then the network should produce an output that approaches +1 regardless of the mean of the irrelevant input signal.

Rate networks with different sizes ($N = 10, 50, \ldots, 450, 500$) were trained to perform the

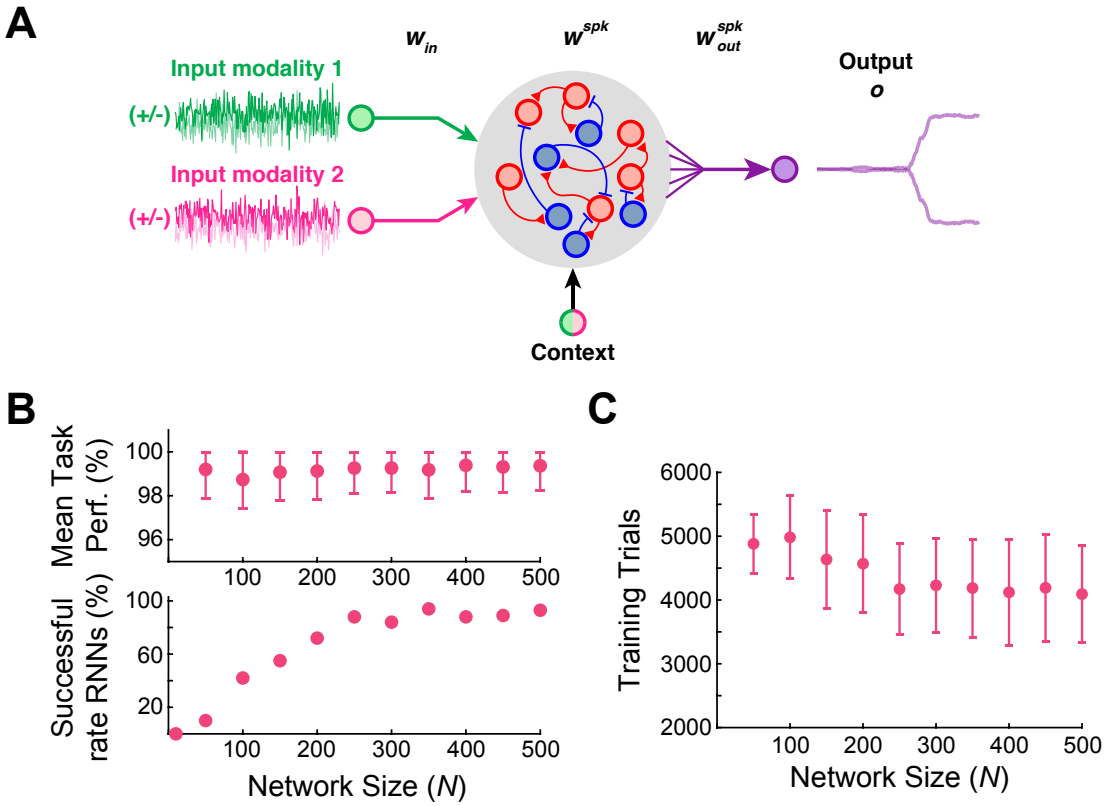


Figure 3.3: **Rate RNNs trained to perform the contextual integration task. A.** Diagram illustrating the task paradigm modeled after the context-dependent task used by [MSSN13]. Two streams of noisy input signals (green and magenta lines) along with a context signal were delivered to the LIF network. The network was trained to integrate and determine if the mean of the cued input signal (i.e., cued offset value) was positive ("+" choice) or negative ("-" choice). **B.** Rate RNNs with different network sizes trained to perform the contextual integration task. The network size was varied from $N = 10$ to 500. For each network size, 100 RNNs with random initial conditions were trained. The average task performance (top) and the proportion of the successful rate models (bottom) are shown. A model was successful if its mean task performance was $\geq 95\%$. **C.** Average number of training trials required for each network size. As the network size increased, the number of training trials decreased (Mean $\pm$ SD shown).

task. As this is a more complex task compared to the Go-NoGo task considered in the previous section, the number of units and trials required to train rate RNNs was larger than the models trained on the Go-NoGo task (Fig. 3.3B and 3.3C). The synaptic decay time constants were again limited to a range of 20 ms and 50 ms, and 100 rate RNNs with random initial conditions were trained for each network size. For the smallest network size ($N = 10$), the rate networks could not be trained to perform the task within the first 6000 trials (Fig. 3.3B).

Next, all the rate networks successfully trained for the task were transformed into LIF models. Example output responses along with the distribution of the tuned synaptic decay constants from a converted LIF model ($N = 250$, $\tau^d_{min} = 20$ ms, $\tau^d_{max} = 50$ ms) are shown in Fig. 3.4A and 3.4B. The task performance of the LIF model was 98% and comparable to the rate RNN used to construct the spiking model (Fig. 3.4C). In addition, the LIF network manifested population dynamics similar to the dynamics observed in the group of neurons recorded by [MSSN13] and rate RNN models investigated in previous studies [MSSN13, SYW16, Mic17]: individual LIF units displayed mixed representation of the four task variables (modality 1, modality 2, network choice, and context; see Section 3.6, Fig. 3.12A), and the network revealed the characteristic line attractor dynamics (Section 3.6, Fig. 3.12B).

Similar to the spiking networks constructed for the Go-NoGo task, the LIF RNNs performed the input integration task more accurately as the network size increased (Fig. 3.4D). Next, the network size was fixed to $N = 250$ and τ^d_{max} was gradually increased from 20 ms to 1000 ms. For $\tau^d_{min} = \tau^d_{max} = 20$ ms, all 100 rate networks failed to learn the task within the first 6000 trials. The conversion from the rate models to the LIF models did not lead to significant loss in task performance for all the other maximum decay constant values considered (Fig. 3.4E).

3.4.4 Analysis of the conversion method

Previous sections illustrated that our framework for converting rate RNNs to LIF RNNs is robust as long as the network size is not too small ($N \geq 200$), and the optimal size was $N = 250$

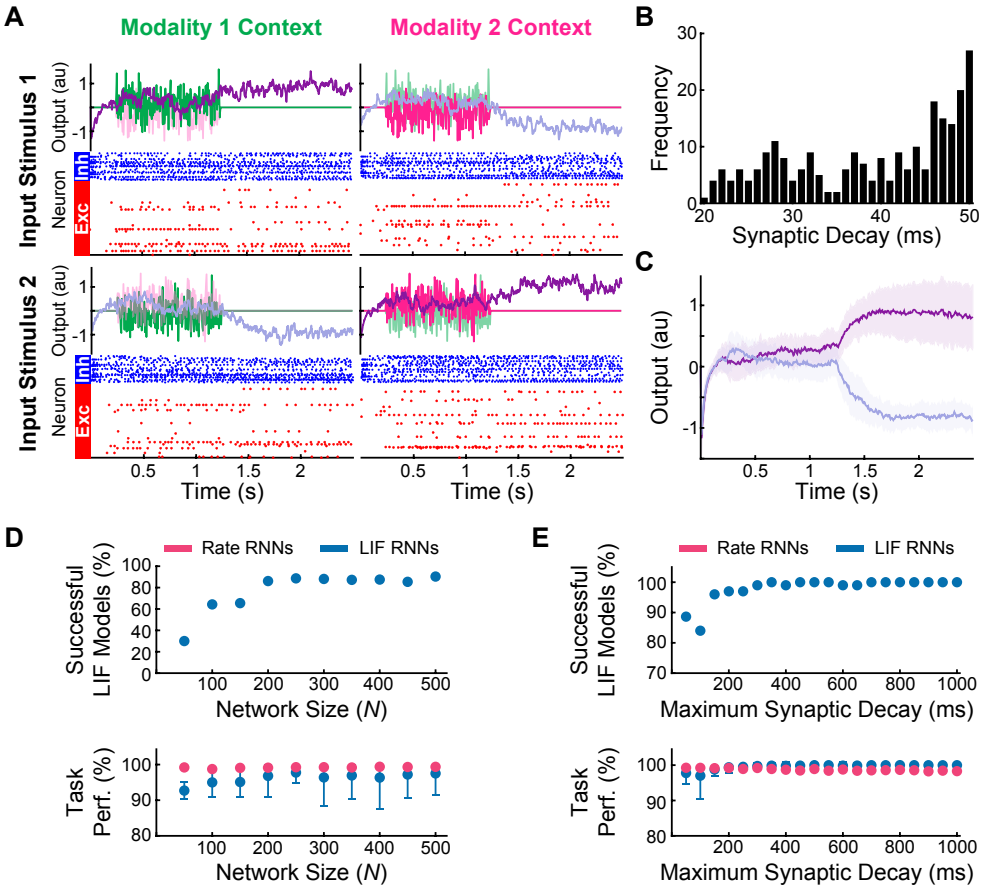


Figure 3.4: **LIF network models constructed to perform the contextual integration task.** **A.** Example output responses and spike raster plots from a LIF network model for two different input stimuli (rows) and two contexts (columns). The network contained 250 units (188 excitatory and 62 inhibitory units), and the noisy input signals were scaled by 0.5 vertically for better visualization of the network responses (purple lines). **B.** Distribution of the optimized synaptic decay time constants ($\boldsymbol{\tau}^{\boldsymbol{d}}$) for the example LIF network (Mean $\pm$ SD, 38.9 $\pm$ 9.3 ms). The time constants were limited to range between 20 ms and 50 ms. **C.** Average output responses of the example LIF network. Mean $\pm$ SD network responses across 100 randomly generated trials shown. **D.** Successfully converted LIF networks and their average task performance across different network sizes. The network size was varied from $N = 10$ to 500. The rate RNNs trained in Fig. 3.3 were used. **E.** Successfully converted LIF networks with $N = 250$ and their average task performance across different maximum synaptic decay constants (varied from 20 ms to 1000 ms).

for both tasks. When the network size is too small, it is harder to train rate RNNs and the rate models successfully trained do not reliably translate to spiking networks (Fig. 3.2D and Fig. 3.4D). In this section, we further investigate the relationship between rate and LIF RNN models and characterize other parameters crucial for the conversion to be effective.

Training synaptic decay time constants

As shown in Fig. 3.5, training the synaptic decay constants for all the rate units is not required for the conversion to work. Rate RNNs (100 models with different initial conditions) with the synaptic decay time constant fixed to 35 ms (average τ^d value for the networks trained with $\tau^d_{min} = 20$ ms and $\tau^d_{max} = 50$ ms) were trained on the Go-NoGo task and converted to LIF RNNs (Fig. 3.5). The task performance of these LIF networks was not significantly different from the performance of the spiking models with optimized synaptic decay constants bounded between 20 ms and 50 ms. The number of the successful LIF models with the fixed synaptic decay constant was also comparable to the number of the successful LIF models with the tuned decay constants (Fig. 3.5).

Other LIF parameters

We also probed how LIF model parameters affected our framework. More specifically, we focused on the refractory period and synaptic filtering. The LIF models constructed in the previous sections used an absolute refractory period of 2 ms and a double exponential synaptic filter (see Section 3.3). Rate models ($N = 250$ and $\tau^d_{max} = 100$ ms) trained on the sensory integration task were converted to LIF networks with different values of the refractory period. As the refractory period became longer, the task performance of the spiking RNNs decreased rapidly (Fig. 3.6A). When the refractory period was set to 0 ms, the LIF RNNs still performed the integration task with a moderately high average accuracy ($92.8 \pm 14.3\%$), but the best task performance was achieved when the refractory period was set to 2 ms (average performance, $97.0 \pm 6.6\%$; Fig. 3.6A inset).

We also investigated how different synaptic filters influenced the mapping process. We first fixed the refractory period to its optimal value (2 ms) and constructed 100 LIF networks ($N = 250$) for the integration task using a double synaptic filter (see Section 3.3; Fig. 3.6B light

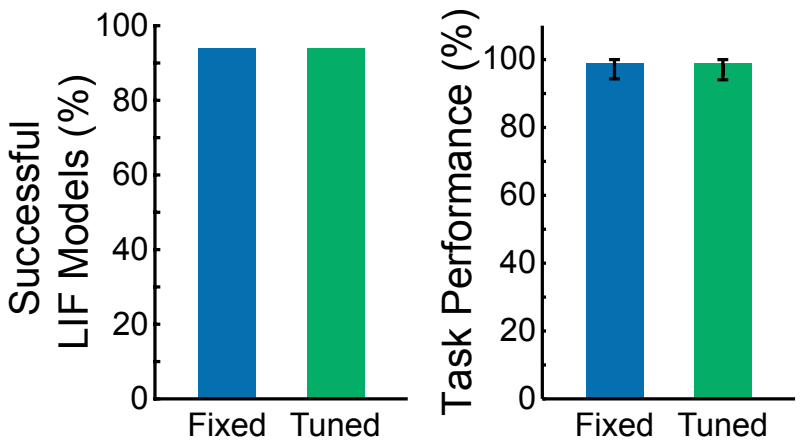


Figure 3.5: **Optimizing synaptic decay constants is not required for conversion of rate RNNs.** The Go-NoGo task performance of the LIF RNNs constructed from the rate networks with a fixed synaptic constant ($\tau^d = 35$ ms; blue) was not significantly different from the performance of the LIF RNNs with tuned synaptic decay time constants ($\tau^d_{min} = 20$ ms, $\tau^d_{max} = 50$ ms; green).

blue). Next, the synaptic filter was changed to the following single exponential filter:

$$\tau_i^d \frac{dr_i^{spk}}{dt} = -r_i^{spk} + \sum_{t_i^k < t} \delta(t - t_i^k)$$

where r_i^{spk} represents the filtered spike train of unit i and t_i^k refers to the k-th spike emitted by unit i. The task performance of the LIF networks with the above single exponential synaptic filter was $95.7 \pm 7.3\%$, and it was not significantly different from the performance of the double exponential synaptic LIF models ($97.0 \pm 6.6\%$; Fig. 3.6B).

Initial connectivity weight scaling

We considered the role of the connectivity weight initialization in our framework. In the previous sections, the connectivity weights (W^{rate}) of the rate networks were initialized as random, sparse matrices with zero mean and a standard deviation of $g/\sqrt{N \cdot P_c}$, where $g = 1.5$ is the gain term that controls the dynamic regime of the networks and $P_c = 0.20$ is the initial connectivity probability (see Section 3.3). Previous studies have shown that rate networks operating in a

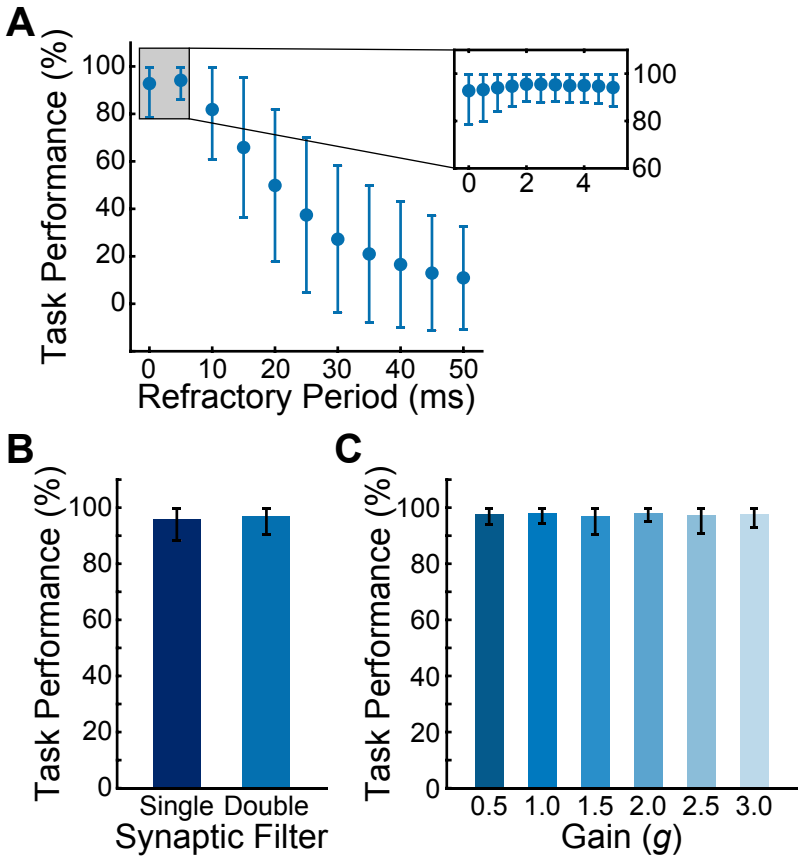


Figure 3.6: **Effects of the refractory period, synaptic filter, and rate RNN connectivity weight initialization. A.** Average contextual integration task performance of the LIF network models ($N = 250$) with different refractory period values. The refractory period was varied from 0 ms (i.e., no refractory period) to 50 ms. The inset shows the average task performance across finer changes in the refractory period. Mean $\pm$ SD shown. **B.** Average contextual integration task performance of the LIF network models ($N = 250$ and refractory period = 2 ms) with the single exponential synaptic filter (dark blue) and the double exponential synaptic filter (light blue). Mean $\pm$ SD shown. **C.** Average contextual integration task performance of the LIF network models ($N = 250$, refractory period = 2 ms, and double exponential synaptic filter) with different connectivity gain initializations. Mean $\pm$ SD shown.

high gain regime ($g > 1.0$) produce chaotic spontaneous trajectories, and this rich dynamics can be harnessed to perform complex computations [LB13, RHT16]. By varying the gain term, we determined if highly chaotic initial dynamics were required for successful conversion. We considered six different gain terms ranging from 0.5 to 3.5, and for each gain term, we constructed 100 LIF RNNs (from 100 rate RNNs with random initial conditions; Fig. 3.6C) to perform the contextual integration task. The LIF models performed the task equally well across all the gain

terms considered (no statistical significance detected).

Transfer function

One of the most important factors that determines whether rate RNNs can be mapped to LIF RNNs in a one-to-one manner is the nonlinear transfer function used in the rate models. We considered three non-negative transfer functions commonly used in the machine learning field to train rate RNNs on the Go-NoGo task: sigmoid, rectified linear, and softplus functions (Fig. 3.7A; see Section 3.6). For each transfer function, 100 rate models ($N = 250$ and $\tau^d_{max} = 50$ ms) were trained. Although all 300 rate models were trained to perform the task almost perfectly (Fig. 3.7B), the average task performance and the number of successful LIF RNNs were highest for the rate models trained with the sigmoid transfer function (Fig. 3.7C). None of the rate models trained with the rectified linear transfer function could be successfully mapped to LIF models, while the spiking networks constructed from the rate models trained with the softplus function were not robust and produced incorrect responses (Section 3.6, Fig. 3.13).

3.5 Discussion

In the current study, we presented a simple framework that harnesses the dynamics of trained continuous rate network models to produce functional spiking RNN models. We identified a set of parameters required to directly transform trained rate RNNs to LIF models, thus establishing a one-to-one correspondence between these two model types. Despite of additional spiking-related parameters, surprisingly only a single parameter (i.e., scaling factor) was required for LIF RNN models to closely mimic their counterpart rate models. Furthermore, this framework can flexibly impose functional connectivity constraints and heterogeneous synaptic time constants.

We investigated and characterized the effects of several model parameters on the stability of the transfer learning from rate models to spiking models. The parameters critical for the

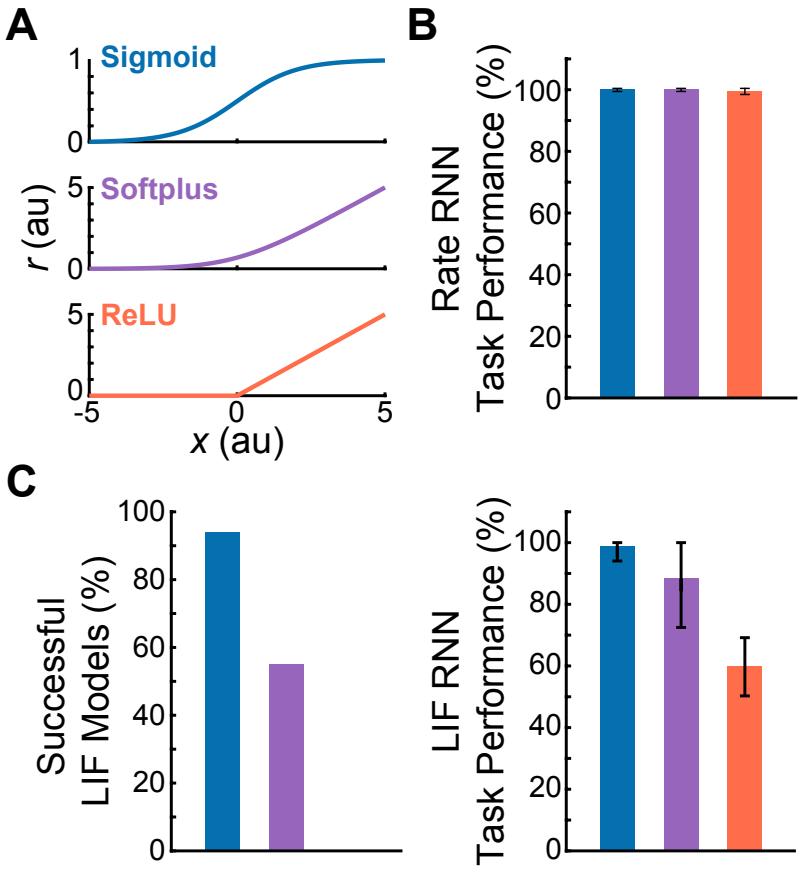


Figure 3.7: **Comparison of the LIF RNNs derived from the rate RNNs trained with three non-negative activation functions. A.** Three non-negative transfer functions were considered: sigmoid, softplus, and rectified linear (ReLU) functions. **B.** All 300 rate RNNs (100 networks per activation function) were successfully trained to perform the Go-NoGo task. **C.** Of the 100 sigmoid LIF networks constructed, 94 networks successfully performed the task. The conversion rates for the softplus and ReLU LIF models were 55% and 0%, respectively. Mean $\pm$ SD task performance: 98.8 $\pm$ 4.7% (sigmoid), 88.3 $\pm$ 15.8% (softplus), and 59.7 $\pm$ 9.5% (ReLU).

mapping to be robust included the network size, choice of activation function for training rate RNNs, and a constant factor to scale down the connectivity weights of the trained rate networks. Although the softplus and rectified linear activation functions are popular for training deep neural networks, we demonstrated that the rate networks trained with these functions do not translate robustly to LIF RNNs (Fig. 3.7). On the other hand, the rate models trained with the sigmoid function were transformed to LIF models with high fidelity.

Another important parameter was the constant scaling factor used to scale W^{rate} and

W_{out}^{rate} before transferring them to LIF networks. When the scaling factor was set to its optimal value (found via grid search), the LIF units behaved like their counterpart rate units, and the spiking networks performed the tasks the rate RNNs were trained to perform (Fig. 3.2). Another parameter that affected the reliability of the conversion was the refractory period parameter of the LIF network models. The LIF performance was optimal when the refractory was set to 2 ms (Fig. 3.6A). Training the synaptic decay time constants, choice of synaptic filter (between single and double exponential filter), and connectivity weight initialization did not affect the mapping procedure (Fig. 3.5 and Fig. 3.6B–C).

The type of approach used in this study (i.e., conversion of a rate network to a spiking network) has been previously employed in neuromorphic engineering to construct power-efficient deep spiking networks [CCK15, DNB+15, DZC+16, HE16, RLHP16, SYW+19]. These studies mainly employed feedforward multi-layer networks or convolutional neural networks aimed to accurately classify input signals or images without placing too much emphasis on biophysical limitations. The overarching goal in these studies was to maximize task performance while minimizing power consumption and computational cost. On the other hand, the main aim of the present study was to construct spiking recurrent network models that abide by important biological constraints in order to relate emerging mechanisms and dynamics to experimentally observed findings. To this end, we have carefully designed our continuous rate RNNs to include several biological features. These include (1) recurrent architectures, (2) sparse connectivity that respects Dale's principle, and (3) heterogeneous synaptic decay time constants.

For constructing spiking RNNs, recent studies have proposed methods that built on the FORCE method to train spiking RNNs [KC18, TUKM16, DCA16, ADM16]. Conceptually, our work is most similar to the work by [DCA16]. The method developed by [DCA16] also relies on mapping a trained continuous-variable rate RNN to a spiking RNN model. However, the rate RNN model used in their study was designed to provide dynamically rich auxiliary basis functions meant to be distributed to overlapping populations of spiking units. Due to this

reason, the relationship between their rate and spiking models is rather complex, and it is not straightforward to impose functional connectivity constraints on their spiking RNN model. An additional procedure was introduced to implement Dale's principle, but this led to more fragile spiking networks with considerably increased training time [DCA16]. The one-to-one mapping between rate and spiking networks employed in our method solved these problems without sacrificing network stability and computational cost: biophysical constraints that we wanted to incorporate into our spiking model were implemented in our rate network model first and then transferred to the spiking model.

While our framework incorporated the basic yet important biological constraints, there are several features that are also not biologically realistic in our models. The gradient-descent method employed to tune the rate model parameters, including the connectivity weights and the synaptic decay time constants, in a supervised manner is not biologically plausible. Although tuning of the synaptic time constants is not realistic and has not been observed experimentally, previous studies have underscored the importance of the diversity of synaptic time scales both *in silico* and *in vivo* [KC18, WSB$^+$18, CTW$^+$18]. In addition, other works have validated and uncovered neural mechanisms observed in experimental settings using RNN models trained with backpropagation [MSSN13, SYW16, CSFW17], thus highlighting that a network model can be biologically plausible even if it was constructed using non-biological means. Another limitation of our method is the lack of temporal coding in our LIF models. Since our framework involves rate RNNs that operate in a rate coding scheme, the spiking RNNs that our framework produces also employ rate coding by nature. Previous studies have shown that spike-coding can improve spiking efficiency and enhance network stability [ADM16, DM16, AMDS18], and recent studies emphasized the importance of precise spike coordination without modulations in firing rates [ZBC$^+$18, SAHD19]. Lastly, our framework does not model nonlinear dendritic processes which have been shown to play a significant role in efficient input integration and flexible information processing [UMBL15, YMW16, TUKM16]. Incorporating nonlinear dendritic processes into our

platform using the method proposed by [TUKM16] will be an interesting next step to further investigate the role of dendritic computation in information processing.

In summary, we provide an easy-to-use platform that converts a continuous recurrent network model with basic biological constraints to a spiking model. The tight relationship between rate and LIF RNN models under certain parameter values suggests that spiking networks could be put together to perform complex tasks traditionally employed to train and study continuous rate networks. Future work needs to focus on why and how such a tight relationship emerges. The framework along with the findings presented in this study lays the groundwork for discovering new principles on how neural circuits solve computational problems with discrete spikes and for constructing more power efficient spiking networks. Extending our platform to incorporate other commonly used neural network architectures could help design biologically plausible deep learning networks that operate at a fraction of the power consumption required for current deep neural networks.

Chapter 3, in full, is a reprint of the material as it appears in: Robert Kim, Yinghao Li, and Terrence J. Sejnowski. Simple framework for constructing functional spiking recurrent neural networks. *Proceedings of the National Academy of Sciences*, 116(45):22811–22820, 2019. The dissertation author was the primary investigator and author of this paper.

3.6 Appendix

3.6.1 Implementation of computational tasks and figure details

In this section, I describe the details of the parameters and methods used to generate all the main figures in the previous section.

Fig. 3.1

A rate RNN of $N = 200$ units (169 excitatory and 31 inhibitory units) was trained to perform the Go-NoGo task for Fig. 3.1B. Each trial lasted for 1000 ms (200 time steps with 5 ms step size). The minimum and the maximum synaptic decay time constants were set to 20 ms and 50 ms, respectively. An input stimulus with a pulse 125 ms in duration was given for a Go trial, while no input stimulus was given for a NoGo trial. The network was trained to produce an output signal approaching +1 after the stimulus offset for a Go trial. For a NoGo trial, the network was trained to maintain its output at zero. A trial was considered correct if the maximum output signal during the response window was above 0.7 for the Go trial type. For a NoGo trial, if the maximum response value was less than 0.3, the trial was considered correct. For training, 6000 trials were randomly generated, and the model performance was evaluated after every 100 trials. Training was terminated when the loss function value fell below 7 and the task performance reached at least 95%. The termination criteria were usually met at or before 2000 trials for this task.

For Fig. 3.1C, rate RNNs with 9 different sizes (N =10, 50, 100, 150, 200, 250, 300, 350, 400) were trained. For each network size, 100 rate RNNs with random initial conditions were trained on the Go-NoGo task.

Fig. 3.2

The rate RNN trained in Fig. 3.1B was converted to a LIF RNN using different scaling factor (λ) values for Fig. 3.2B. The double exponential synaptic filter was used, and the gain term (g) for the rate RNN initialization was set to 1.5. The LIF parameters listed in Table 3.1 were used for all the LIF network models constructed in Fig. 3.2.

Fig. 3.3

Rate RNNs with 11 different network sizes (N =10, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500) were trained on the contextual integration task. For each network size, 100 rate RNNs with random initial conditions were trained.

For the task design, the input matrix ($\boldsymbol{u} \in \mathbb{R}^{4\times 500}$) contained four stimuli channels across time (500 time steps with 5 ms step size). The first two channels corresponded to the modality 1 and modality 2 noisy input signals. These signals were modeled as white-noise signals (sampled from the standard normal distribution) with constant offset terms. The sign of the offset term modeled the evidence toward (+) or (-) choices, while the magnitude of the offset determined the strength of the evidence. The noisy signals were only present during the stimulus window (250 ms – 1250 ms). The last two channels of $\boldsymbol{u}$ represented the modality 1 and the modality 2 context signals. For instance, the third channel of $\boldsymbol{u}$ is set to one and the fourth channel is set to zero throughout the trial duration to model Modality 1 context.

For each trial used to train the rate model, the offset values for the two modality input signals were randomly set to -0.5 or +0.5. The context signals were randomly set such that either modality 1 (third input channel is set to 1) or modality 2 (fourth input channel is set to 1) was cued for each trial. If the offset term of the cued modality was +0.5 (or -0.5) for a given trial, the network was instructed to produce an output signal approaching +1 (or -1) after the stimulus window. The model performance was assessed after every 100 training trials, and the training termination conditions were same as the ones used for Fig. 3.1.

Fig. 3.4

A network of $N = 250$ LIF units (188 excitatory and 62 inhibitory units) were constructed from a rate RNN model trained to perform the context-dependent input integration task for Fig. 3.4A. The scaling factor (λ) was set to 1/60. The double exponential synaptic filter was used, and the gain term (g) for the rate RNN initialization was set to 1.5. The LIF parameters listed in

Table 3.1 were used for all the LIF network models constructed in Fig. 3.4.

Fig. 3.5

Rate RNNs ($N = 250$) were trained on the Go-NoGo task with and without optimizing the synaptic decay time constants ($\boldsymbol{\tau^d}$). For each condition, 100 rate RNNs were trained. For the fixed synaptic decay constant condition, τ^d was fixed to 35 ms. For the tuned synaptic decay condition, $\tau^d_{min} = 20$ ms and $\tau^d_{max} = 50$ ms.

Fig. 3.6

For Fig. 3.6A, all 100 rate RNNs ($N = 250$, $\tau^d_{min} = 20$ ms, $\tau^d_{max} = 100$ ms) trained in Fig. 3.4E were converted to LIF RNNs with different values of the refractory period. The following 20 refractory period values were considered: 0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 10, 15, 20, 25, 30, 35, 40, 45, 50 ms.

Fig. 3.7

The following softplus function was used:

$$r = \log(\exp(x) + 1)$$

For the networks trained with the softplus and ReLU activation functions, the following range of values for $1/\lambda$ was used for the grid search: 4 to 26 with a step size of 2.

3.6.2 Quadratic integrate-and-fire model

For the quadratic integrate-and-fire (QIF) model (Fig. 3.14), we considered a network of units governed by

$$\tau_m \frac{d\boldsymbol{v}}{dt} = \boldsymbol{v}^2 + W^{spk}\boldsymbol{r}^{spk} + \boldsymbol{I}_{ext}$$

The definitions of the variables are identical to the ones used for the LIF network model.

3.6.3 Code availability

The implementation of our framework and the codes to generate all the figures in this work are available at `https://github.com/rkim35/spikeRNN`. The repository also contains implementation of other tasks including autonomous oscillation and delayed match-to-sample (DMS) tasks.

3.6.4 Data availability

All the trained models used in the present study are available at the following repository: `https://osf.io/jd4b6`.

3.6.5 Supplementary figures

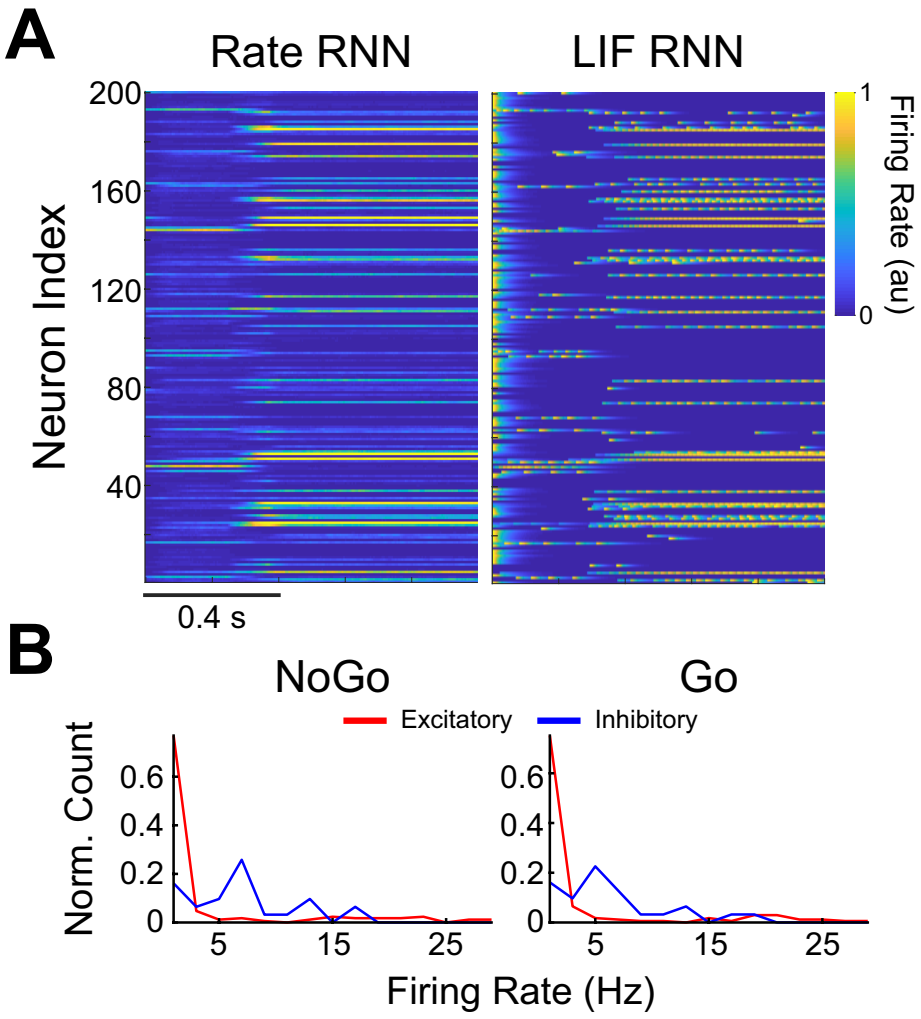


Figure 3.8: **Comparison of the time-varying rates of the continuous-variable rate units and the LIF units. A.** A single Go trial was used to extract the rates from the rate RNN trained in Fig. 3.1B. The firing rates of the LIF RNN constructed using the optimal scaling factor ($\lambda = 1/25$) are shown on the right. The firing rates of the LIF units were normalized to range from 0 to 1 for comparison. **B.** Distribution of the firing rates for a NoGo trial (left) and a Go trial (right).

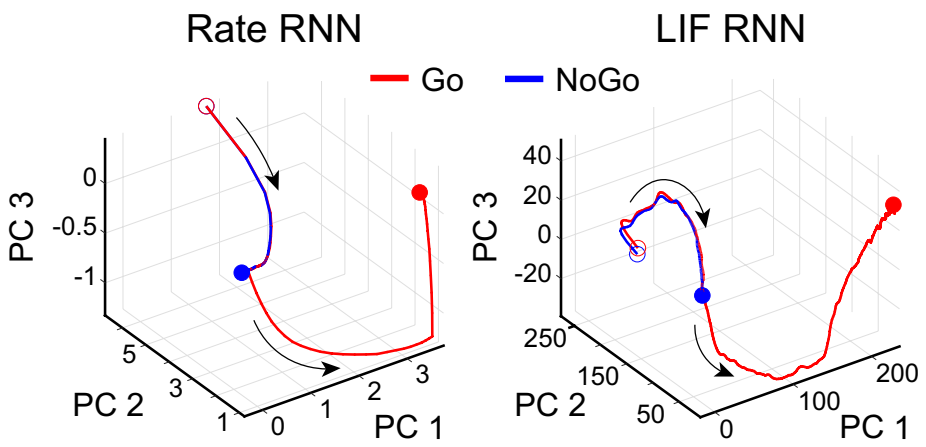


Figure 3.9: **Comparison of the top three PCs extracted from the network activities of the rate and LIF RNNs trained to perform the Go-NoGo task.** Principal component analysis (PCA) was performed on the firing rates derived from a rate RNN and a LIF RNN trained to perform the Go-NoGo task. The rate RNN contained 200 units (169 excitatory and 31 inhibitory units), and the LIF model was constructed from the rate model. The firing rates from 50 Go trials and 50 NoGo trials were obtained from the two RNN models. For both models, the top three principal components (PCs) captured 99% of the variance. Red and blue empty circles indicate the trial onset for the Go and the NoGo trials, respectively. Red and blue filled circles represent the end of the trial for the Go and the NoGo trials, respectively.

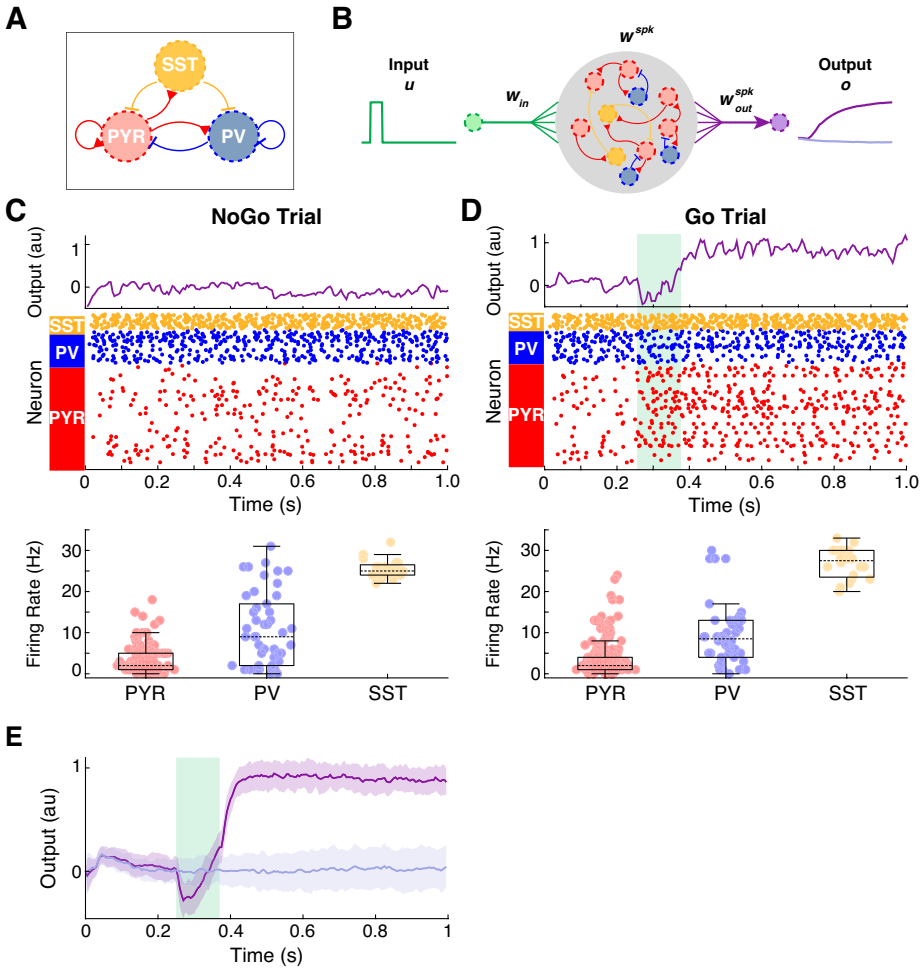


Figure 3.10: **Incorporation of additional functional connectivity constraints. A.** Common cortical microcircuit motif where somatostatin-expressing interneurons (SST; yellow circle) inhibit both pyramidal (PYR; red circle) and parvalbumin-expressing (PV; blue circle) neurons. **B.** Schematic illustrating the incorporation of the connectivity motif shown in **A** into a LIF network model. The connectivity pattern was imposed during training of a rate network model ($N = 200$) to perform the Go-NoGo task. There were 134 PYR, 46 PV, and 20 SST units. A spiking model was constructed using the trained rate model with $\lambda = 1/50$. **C.** Example output response and spikes from the LIF network model for a single NoGo trial. Mean $\pm$ SD firing rate for each population is also shown (PYR, 3.08 $\pm$ 3.29 Hz; PV, 10.80 $\pm$ 8.94 Hz; SST, 25.50 $\pm$ 2.33 Hz). **D.** Example output response and spikes from the LIF network model for a single Go trial. Mean $\pm$ SD firing rate for each population is also shown (PYR, 4.72 $\pm$ 5.89 Hz; PV, 9.30 $\pm$ 8.16 Hz; SST, 27.05 $\pm$ 3.98 Hz). Box plot central lines, median; bottom and top edges, lower and upper quartiles. **E.** LIF network model performance on 50 NoGo trials (light purple) and 50 Go trials (dark purple). Mean $\pm$ SD shown.

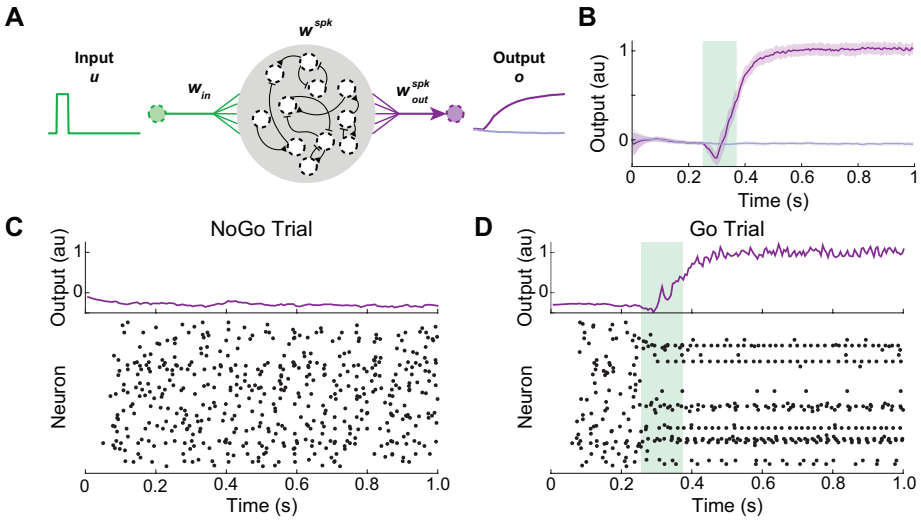


Figure 3.11: **Dale's principle constraint can be relaxed. A.** Schematic diagram showing a LIF network model without Dale's principle. A rate RNN model ($N = 200$) without Dale's principle was first trained to perform the Go-NoGo task. The scaling factor (λ) was set to 1/50. Note that each unit (black dotted circles) can exert both excitatory and inhibitory effects. **B.** LIF network model performance on 50 NoGo trials (light purple) and 50 Go trials (dark purple). Mean $\pm$ SD shown. **C.** Example output response (top) and spikes (bottom) from the LIF network model for a single NoGo trial. **D.** Example output response (top) and spikes (bottom) from the LIF network model for a single Go trial.

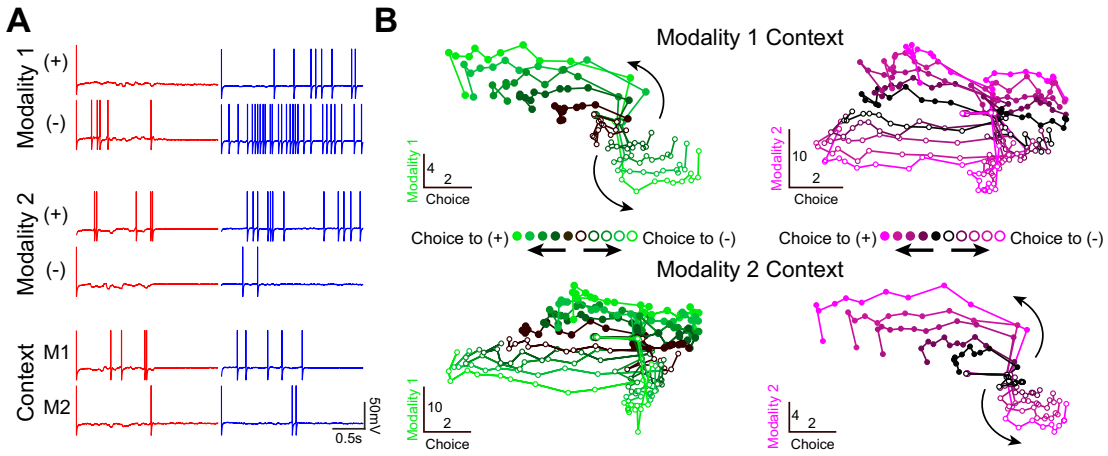


Figure 3.12: **The LIF network model employs mixed representations of the task variables.** **A.** Mixed representation of the task variables at the level of single units from a LIF network ($N = 400$; 299 excitatory and 101 inhibitory units; $\tau^d_{min} = 20$ ms and $\tau^d_{max} = 100$ ms). An excitatory unit (red) and an inhibitory unit (blue) with mixed representation of three task variables (modality 1, modality 2, and context) are shown as examples. The excitatory neuron preferred modality 1 input signals with negative offset values, modality 2 signals with positive offset values, and modality 1 context (left column). The inhibitory neuron also exhibited similar biases (right column). **B.** Average population responses projected to a low dimensional state space. The targeted dimensionality reduction technique (developed in [MSSN13]) was used to project the population activities to the state space spanned by the task-related axes. For the modality 1 context (top row), the population responses from the trials with various modality 1 offset values were projected to the choice and modality 1 axes (left). The same trials were sorted by the irrelevant modality (modality 2) and shown on the right. Similar conventions used for the modality 2 context (bottom row). The offset magnitude (i.e., amount of evidence toward "+" or "-" choice) increases from dark to light. Filled and empty circles correspond to "+" choice and "-" choice trials, respectively.

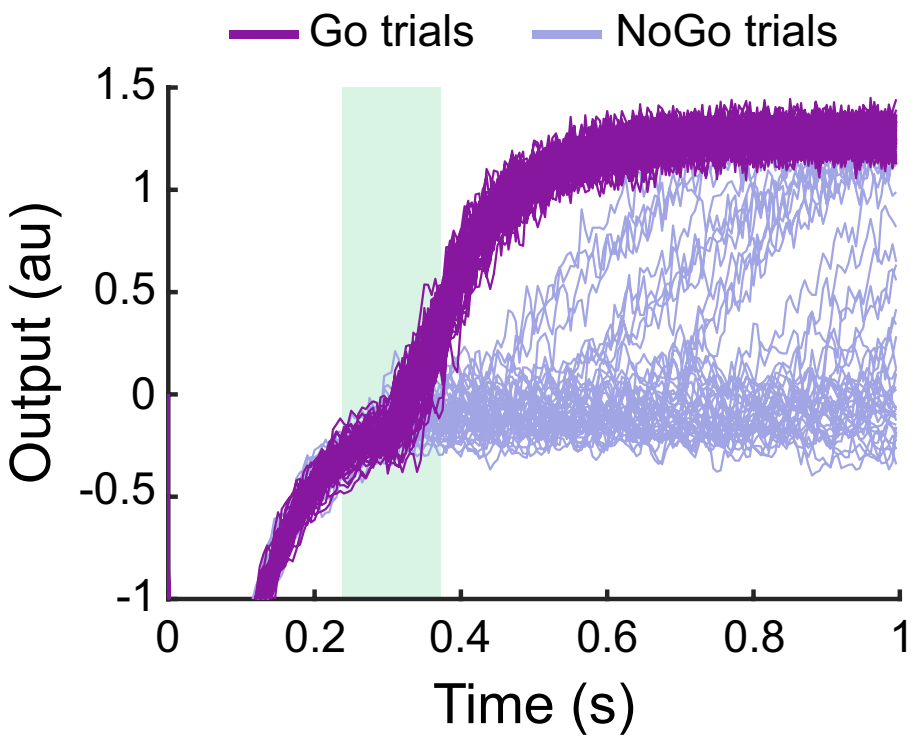


Figure 3.13: **Example output responses from a softplus LIF RNN constructed to perform the Go-NoGo task.** Individual output responses from 50 Go trials (dark purple) and 50 NoGo trials (light purple) are shown. The optimal scaling factor was 1/10, and the performance of the model was 78%.

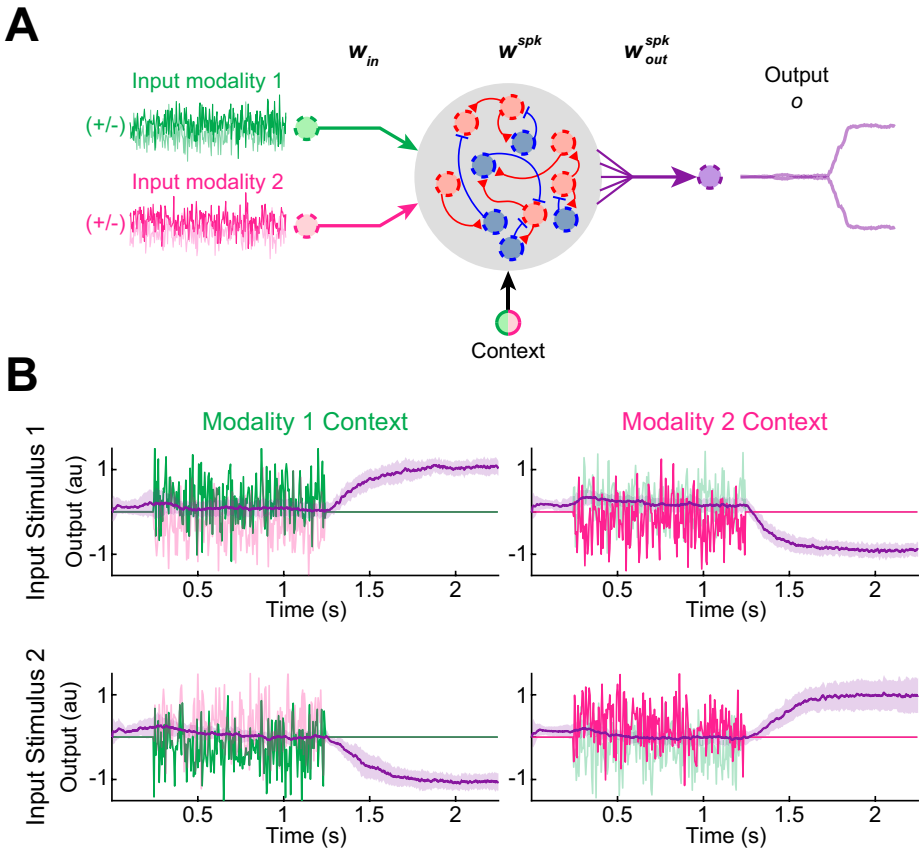


Figure 3.14: **Quadratic integrate-and-fire (QIF) model constructed to perform the context-dependent input integration task. A.** The task paradigm and the trained rate network model used for Fig. 3.12 were employed to build a QIF model. The QIF model parameter values are listed in Table 3.1. **B.** The QIF model successfully performed the task by integrating cued modality input signals. Example noisy input signals (scaled by 0.5 vertically for visualization; green and magenta lines) from a single trial are shown. Mean $\pm$ SD response signals (purple lines) across 50 trials for each trial type.

Table 3.1: Parameter values used to construct LIF and QIF networks

	LIF	QIF
Membrane time constant (τ_m)	10 ms	10 ms
Absolute refractory period	2 ms	2 ms
Synaptic rise time (τ_r)	2 ms	2 ms
Constant bias current	-40 pA	0 pA
Spike threshold	-40 mV	30 mV
Spike reset voltage	-65 mV	-65 mV

Bibliography

[ADM16] Larry F. Abbott, Brian DePasquale, and Raoul-Martin Memmesheimer. Building functional networks of spiking model neurons. *Nature Neuroscience*, 19(3):350–355, Mar 2016.

[AMDS18] Alireza Alemi, Christian K. Machens, Sophie Denéve, and Jean-Jacques E. Slotine. Learning nonlinear dynamics in efficient, balanced spiking networks using local plasticity rules. In *AAAI*, 2018.

[BBP13] Y. Bengio, Nicholas Boulanger-Lewandowski, and Razvan Pascanu. Advances in optimizing recurrent networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8624–8628, May 2013.

[BSR+13] Omri Barak, David Sussillo, Ranulfo Romo, Misha Tsodyks, and Larry F. Abbott. From fixed points to chaos: Three models of delayed discrimination. *Progress in Neurobiology*, 103:214 – 222, 2013.

[CCK15] Yongqiang Cao, Yang Chen, and Deepak Khosla. Spiking deep convolutional neural networks for energy-efficient object recognition. *Int. J. Comput. Vision*, 113(1):54–66, May 2015.

[CSFW17] Warasinee Chaisangmongkon, Sruthi K. Swaminathan, David J. Freedman, and Xiao-Jing Wang. Computing by robust transience: How the fronto-parietal network performs sequential, category-based decisions. *Neuron*, 93(6):1504–1517, Mar 2017.

[CTW+18] Sean E. Cavanagh, John P. Towers, Joni D. Wallis, Laurence T. Hunt, and Steven W. Kennerley. Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. *Nature Communications*, 9(1), Aug 2018.

[DCA16] Brian DePasquale, Mark M. Churchland, and Larry F. Abbott. Using firing-rate dynamics to train recurrent networks of spiking model neurons. *arXiv e-prints*, page arXiv:1601.07620, 2016.

[DM16] Sophie Denéve and Christian K Machens. Efficient codes and balanced networks. *Nature Neuroscience*, 19(3):375–382, Mar 2016.

[DNB+15] Peter U. Diehl, Daniel Neil, Jonathan Binas, Matthew Cook, Shih-Chii Liu, and Michael Pfeiffer. Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, July 2015.

[DZC+16] Peter U. Diehl, Guido Zarrella, Andrew Cassidy, Bruno U. Pedroni, and Emre Neftci. Conversion of artificial recurrent neural networks to spiking neural networks for low-power neuromorphic hardware. In *2016 IEEE International Conference on Rebooting Computing (ICRC)*, pages 1–8, Oct 2016.

[EPQD16] Pierre Enel, Emmanuel Procyk, René Quilodran, and Peter Ford Dominey. Reservoir computing properties of neural dynamics in prefrontal cortex. *PLOS Computational Biology*, 12(6):e1004967, Jun 2016.

[FsSY+02] Gidon Felsen, Yao song Shen, Haishan Yao, Gareth Spor, Chaoyi Li, and Yang Dan. Dynamic modification of cortical orientation tuning mediated by recurrent connections. *Neuron*, 36(5):945 – 954, 2002.

[GR95] Patricia S. Goldman-Rakic. Cellular basis of working memory. *Neuron*, 14(3):477 – 485, 1995.

[HE16] Eric Hunsberger and Chris Eliasmith. Training spiking deep networks for neuromorphic hardware. *CoRR*, abs/1611.05141, 2016.

[HS18] Dongsung Huh and Terrence J Sejnowski. Gradient descent for spiking neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 1433–1443. Curran Associates, Inc., 2018.

[KC18] Christopher M. Kim and Carson C. Chow. Learning recurrent dynamics in spiking networks. *eLife*, 7:e37124, Sep 2018.

[LB13] Rodrigo Laje and Dean V. Buonomano. Robust timing and motor patterns by taming chaos in recurrent neural networks. *Nature Neuroscience*, 16(7):925–933, May 2013.

[LDP16] Jun Haeng Lee, Tobi Delbruck, and Michael Pfeiffer. Training deep spiking neural networks using backpropagation. *Frontiers in Neuroscience*, 10:508, 2016.

[Mic17] Thomas Miconi. Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *eLife*, 6:e20899, feb 2017.

[MO18] Francesca Mastrogiuseppe and Srdjan Ostojic. Linking connectivity, dynamics, and computations in low-rank recurrent neural networks. *Neuron*, 99(3):609–623, Aug 2018.

[MS11] James Martens and Ilya Sutskever. Learning recurrent neural networks with hessian-free optimization. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pages 1033–1040, USA, 2011. Omnipress.

[MSSN13] Valerio Mante, David Sussillo, Krishna V. Shenoy, and William T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, Nov 2013.

[NC17] Wilten Nicola and Claudia Clopath. Supervised learning in spiking neural networks with force training. *Nature Communications*, 8:2208, Dec 2017.

[PMB13] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pages III–1310–III–1318, 2013.

[RHT16] Kanaka Rajan, Christopher D. Harvey, and David W. Tank. Recurrent network models of sequence generation and memory. *Neuron*, 90(1):128 – 142, 2016.

[RLHP16] Bodo Rueckauer, Iulia-Alexandra Lungu, Yuhuang Hu, and Michael Pfeiffer. Theory and tools for the conversion of analog to spiking convolutional neural networks, 2016.

[SA09] David Sussillo and Larry F. Abbott. Generating coherent patterns of activity from chaotic neural networks. *Neuron*, 63(4):544 – 557, 2009.

[SAHD19] Neda Shahidi, Ariana R. Andrei, Ming Hu, and Valentin Dragoi. High-order coordination of cortical spiking activity modulates perceptual accuracy. *Nature Neuroscience*, 22(7):1148–1158, May 2019.

[SCS88] Haim Sompolinsky, Andrea Crisanti, and Hans Juergen Sommers. Chaos in random neural networks. *Physcial Review Letters*, 61:259–262, Jul 1988.

[SKS+13] Mark G. Stokes, Makoto Kusunoki, Natasha Sigala, Hamed Nili, David Gaffan, and John Duncan. Dynamic coding for cognitive control in prefrontal cortex. *Neuron*, 78(2):364–375, Apr 2013.

[SYW16] H. Francis Song, Guangyu R. Yang, and Xiao-Jing Wang. Training excitatory-inhibitory recurrent neural networks for cognitive tasks: A simple and flexible framework. *PLOS Computational Biology*, 12(2):e1004792, Feb 2016.

[SYW17] H. Francis Song, Guangyu R. Yang, and Xiao-Jing Wang. Reward-based training of recurrent neural networks for cognitive and value-based tasks. *eLife*, 6:e21492, Jan 2017.

[SYW+19] Abhronil Sengupta, Yuting Ye, Robert Wang, Chiao Liu, and Kaushik Roy. Going deeper in spiking neural networks: Vgg and residual architectures. *Frontiers in Neuroscience*, 13:95, 2019.

[TUKM16] Dominik Thalmeier, Marvin Uhlmann, Hilbert J. Kappen, and Raoul-Martin Memmesheimer. Learning universal computations with spikes. *PLOS Computational Biology*, 12(6):e1004895, Jun 2016.

[UMBL15] Balázs B Ujfalussy, Judit K Makara, Tiago Branco, and Máté Lengyel. Dendritic nonlinearities are tuned for efficient spike-based computations in cortical circuits. *eLife*, 4:e10056, dec 2015.

[Wan08] Xiao-Jing Wang. Decision making in recurrent neuronal circuits. *Neuron*, 60(2):215–234, Oct 2008.

[Wer90] Paul J. Werbos. Backpropagation through time: what it does and how to do it. *Proceedings of the IEEE*, 78(10):1550–1560, Oct 1990.

[WKNK+18] Jane X. Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, 21(6):860–868, May 2018.

[WSB+18] D. F. Wasmuht, E. Spaak, T. J. Buschman, E. K. Miller, and M. G. Stokes. Intrinsic neuronal dynamics predict distinct functional roles during working memory. *Nature Communications*, 9:3499, Aug 2018.

[YMW16] Guangyu Robert Yang, John D. Murray, and Xiao-Jing Wang. A dendritic disinhibitory circuit mechanism for pathway-specific gating. *Nature Communications*, 7:12815, Sep 2016.

[ZBC+18] Jennifer L. Zick, Rachael K. Blackman, David A. Crowe, Bagrat Amirikian, Adele L. DeNicola, Theoden I. Netoff, and Matthew V. Chafee. Blocking NMDAR disrupts spike timing and decouples monkey prefrontal circuits: Implications for activity-dependent disconnection in schizophrenia. *Neuron*, 98(6):1243–1255, Jun 2018.

[ZCL+18] Zhewei Zhang, Zhenbo Cheng, Zhongqiao Lin, Chechang Nie, and Tianming Yang. A neural network model for the orbitofrontal cortex and task space acquisition during reinforcement learning. *PLOS Computational Biology*, 14(1):1–24, 01 2018.

Chapter 4

Strong inhibitory signaling underlies stable temporal dynamics and working memory in spiking neural networks

The computational framework introduced in the previous chapter is applied to train recurrent networks on working memory (WM) tasks in order to characterize network dynamics required for short-term memory maintenance. Impairment in WM has been consistently observed in patients diagnosed with schizophrenia [Man03, VGH$^+$16], and understanding circuit mechanisms underlying WM computations could shed light on the pathophysiology of the complex disorder.

The work presented here is reproduced and adapted from: Robert Kim and Terrence J. Sejnowski. Strong inhibitory signaling underlies stable temporal dynamics and working memory in spiking neural networks. Preprint at `https://www.biorxiv.org/content/10.1101/2020.02.11.944751v1` (2020).

4.1 Abstract

Cortical neurons process information on multiple timescales, and areas important for working memory (WM) contain neurons capable of integrating information over a long timescale. However, the underlying mechanisms for the emergence of neuronal timescales stable enough to support WM are unclear. By analyzing a spiking recurrent neural network (RNN) model trained on a WM task and activity of single neurons in the primate prefrontal cortex, we show that the temporal properties of our model and the neural data are remarkably similar. Dissecting our RNN model revealed strong inhibitory-to-inhibitory connections underlying a disinhibitory microcircuit as a critical component for long neuronal timescales and WM maintenance. We also found that enhancing inhibitory-to-inhibitory connections led to more stable temporal dynamics and improved task performance. Finally, we show that a network with such microcircuitry can perform other tasks without disrupting its pre-existing timescale architecture, suggesting that strong inhibitory signaling underlies a flexible WM network.

4.2 Introduction

Temporal receptive fields and neuronal timescales are hierarchically organized across the cortex [MBF$^+$14, CKG$^+$15]. Areas important for higher cognitive functions are capable of integrating and processing information in a robust manner and reside at the top of the hierarchy [MBF$^+$14, CKG$^+$15, CWKH16]. The prefrontal cortex (PFC) is a higher-order cortical region that supports a wide range of complex cognitive processes including WM, an ability to encode and maintain information over a short period of time [MED96, FA71]. However, the underlying circuit mechanisms that give rise to stable temporal receptive fields strongly associated with WM are not known and experimentally challenging to probe. A better understanding of possible mechanisms could elucidate not only how areal specialization in the cortex emerges but also how local cortical microcircuits carry out WM computations.

Previous experimental studies reported that baseline activities of single neurons in the primate PFC contained unique temporal receptive field structures. Using decay time constants of spike-count autocorrelation functions obtained from neurons at rest, these studies demonstrated that the primate PFC is mainly composed of neurons with large time constants or timescales [MBF$^+$14, FTMG17, CTW$^+$18, WSB$^+$18]. In addition, neurons with longer timescales carried more information during the delay period of a WM task compared to short timescale neurons [WSB$^+$18]. A large-scale computational model where heterogeneous timescales were naturally organized in a hierarchical manner that closely matched the hierarchy observed in the primate neocortex has been proposed [CKG$^+$15]. The framework utilized a gradient of recurrent excitation to establish varying degrees of temporal dynamics [CKG$^+$15]. Although their findings suggest that recurrent excitation is correlated with area-specific timescales, it is still unclear if recurrent excitation indeed directly regulates neuronal timescales and WM computations.

Recent experimental studies paint a different picture where diverse inhibitory interneurons form intricate microcircuits in the PFC to execute memory formation and retrieval [KJL$^+$16, KD17, XLT$^+$19, CC19, KPdA$^+$19]. Both somatostatin (SST) and vasoactive intestinal peptide (VIP) interneurons have been shown to form a microcircuit that can disinhibit excitatory cells via inhibition of parvalbumin (PV) interneurons [PXH$^+$13, TLR16]. Furthermore, SST and VIP neurons at the center of such disinhibitory microcircuitry were causally implicated with impaired associative and working memory via optogenetic manipulations [KJL$^+$16, KD17, CC19, KPdA$^+$19]. Consistent with these observations, the primate anterior cingulate cortex, which is at the top of the timescale hierarchy [MBF$^+$14], was found to contain more diverse and stronger inhibitory inputs compared to the lateral PFC [MGWL17]. A recent theoretical study also showed that inhibitory-to-inhibitory synapses, although much fewer in number compared to excitatory connections, is a critical component for implementing robust maintenance of memory patterns [MRL18].

In order to characterize how strong inhibitory signaling enables WM maintenance and

leads to slow temporal dynamics, we constructed a spiking RNN model to perform a WM task and compared the emerging timescales with the timescales derived from the prefrontal cortex of rhesus monkeys trained to perform similar WM tasks. Here, we show that both primate PFC and our RNN model utilize units with long timescales to sustain stimulus information. By analyzing and dissecting the RNN model, we illustrate that inhibitory-to-inhibitory synapses incorporated into a disinhibitory microcircuit tightly control both neuronal timescales and WM task performance. Finally, we show that the primate PFC exhibits signs that it is already equipped with strong inhibitory connectivity even before learning the WM task, implying that a gradient of recurrent inhibition could naturally result in functional specialization in the cortex. We confirm this with our model and show that the task performance of RNNs with short timescales can be enhanced via increased recurrent inhibitory signals. Overall, our work offers timely insight into the role of diverse inhibitory signaling in WM and provides a circuit mechanism that can explain previously observed experimental findings.

4.3 Materials and methods

4.3.1 Continuous rate RNN model

The spiking RNNs used in this chapter were generated by first training their counterpart continuous-variable rate RNNs using a gradient descent algorithm. After training, the continuous RNNs were converted to leaky integrate-and-fire (LIF) RNNs using the method that I presented in the previous chapter [KLS19]. The continuous RNN model contained $N = 200$ recurrently connected units that were governed by

$$\begin{aligned} \boldsymbol{\tau}^{\boldsymbol{d}} \frac{d\boldsymbol{x}}{dt} &= -\boldsymbol{x} + W^{rate}\boldsymbol{r}^{rate} + \boldsymbol{I}_{ext} \\ \boldsymbol{r}^{rate} &= \frac{1}{1+\exp(-\boldsymbol{x})} \end{aligned} \tag{4.1}$$

where 20 ms $\leq \boldsymbol{\tau^d} \leq$ 125 ms $\in \mathbb{R}^{1\times N}$ corresponds to the synaptic decay time constants for the N units in the network, $\boldsymbol{x} \in \mathbb{R}^{1\times N}$ is the synaptic current variable, $W^{rate} \in \mathbb{R}^{N\times N}$ is the synaptic connectivity matrix, and $\boldsymbol{r}^{rate} \in \mathbb{R}^{1\times N}$ refers to the firing rate estimates of the units. A standard logistic sigmoid function was used to estimate a firing rate of a neuron from its synaptic current (x).

The external currents ($\boldsymbol{I}_{ext}$) include task-specific input stimulus signals (see Section 4.3.2) along with a Gaussian white noise variable:

$$\boldsymbol{I}_{ext} = W_{in}\boldsymbol{u} + \mathcal{N}(0, 0.01)$$

where the time-varying, task-specific stimulus signals ($\boldsymbol{u} \in \mathbb{R}^{N_{in}\times 1}$) are given to the network via $W_{in} \in \mathbb{R}^{N\times N_{in}}$, a Gaussian random matrix with zero mean and unit variance. N_{in} corresponds to the number of input signals associated with a specific task, and $\mathcal{N}(0, 0.01) \in \mathbb{R}^{N\times 1}$ represents a Gaussian random noise with zero mean and variance of 0.01.

A linear readout of the population activity was used to define the output of the rate network:

$$o^{rate}(t) = W_{out}^{rate}\boldsymbol{r}^{rate}(t)$$

where $W_{out}^{rate} \in \mathbb{R}^{1\times N}$ refers to the readout weights.

Eq. (4.1) is discretized using the first-order Euler approximation method:

$$\begin{aligned} \boldsymbol{x}_t &= \left(1 - \frac{\Delta t}{\boldsymbol{\tau^d}}\right)\boldsymbol{x}_{t-1} + \frac{\Delta t}{\boldsymbol{\tau^d}}(W^{rate}\boldsymbol{r}_{t-1}^{rate} + W_{in}\boldsymbol{u}_{t-1}) \\ &+ \mathcal{N}(0, 0.01) \end{aligned}$$

where $\Delta t = 5$ ms is the discretization time step size used throughout this study.

4.3.2 Training details

Adam (adaptive moment estimation), a stochastic gradient descent algorithm, was used to update the synaptic decay variable (τ_s), recurrent connections (W^{rate}) and readout weights (W^{rate}_{out}). The learning rate was set to 0.01, and the TensorFlow default values were used for the first and second moment decay rates. In addition, Dale's principle (i.e., separate excitatory and inhibitory populations) was imposed using the method previously proposed [SYW16]. For re-training previously trained RNNs (Fig. 4.8), only the input weights (W_{in}) were trainable, and the recurrent weights and the readout weights were fixed to their trained values.

Two LIF RNN models were employed in this study by training rate RNNs on two different tasks: delayed match-to-sample (DMS) and 2-alternative forced choice (AFC) tasks.

DMS RNNs

For the DMS RNN model, the input matrix ($\boldsymbol{u} \in \mathbb{R}^{2\times 500}$) contained two input channels for two sequential stimuli (over 500 time steps with 5 ms step size). The first channel delivered the first stimulus (250 ms in duration) after 1 s (200 time steps) of fixation, while the second channel modeled the second stimulus (250 ms in duration), which began 50 ms after the offset of the first stimulus. The short delay (50 ms) allowed rate RNNs to learn the task efficiently, and the delay duration was increased after training (see below). During each stimulus window, the corresponding input channel was set to either -1 or +1. If the two sequential stimuli had the same sign (-1/-1 or +1/+1), the network was trained to produce an output signal approaching +1 after the offset of the second stimulus. If the stimuli had opposite signs (-1/+1 or +1/-1), then the network produced an output signal approaching -1. The training was stopped when the loss function fell below 7, and the task performance was greater than 95%. After the rate RNNs were successfully trained and converted to LIF networks, a subgroup of LIF RNNs that performed the actual DMS paradigm used in the main text (i.e., delay duration set to 750 ms) with accuracy greater than 95% were identified and analyzed. For Figs. 4.5, 4.6 and 4.7, a group of LIF RNNs

that performed the DMS task with accuracy between 60% and 80% was used.

AFC RNNs

The input matrix ($\boldsymbol{u} \in \mathbb{R}^{1\times 350}$) for the AFC paradigm was set to 0 for the first 200 time steps (i.e., 1 s fixation). A short stimulus (125 ms in duration) of either -1 or +1 was given after the fixation period. After the stimulus offset, the network was trained to produce an output signal approaching -1 for the "-1" stimulus and +1 for the "+1" stimulus. The training termination criteria were the same as those used for the DMS model above.

4.3.3 Spiking RNN model

For our spiking RNN model, we considered a network of leaky integrate-and-fire (LIF) units recurrently connected to one another. These units are governed by:

$$\tau_m \frac{dv_i(t)}{dt} = -v_i(t) + (x_i(t) + I_{ext}(t))R \tag{4.2}$$

where τ_m is the membrane time constant (10 ms), $v_i(t)$ is the membrane voltage of unit i at time t, $x_i(t)$ is the synaptic input current that unit i receives at time t, I_{ext} is the external input current, and R is the leak resistance (set to 1). The synaptic input current (x) is modeled using a double-exponential synaptic filter applied to the presynaptic spike trains:

$$\begin{aligned} x_i &= \sum_{j=1}^{N} W_{ij}^{spk} r_j^{spk} \\ \frac{dr_i^{spk}}{dt} &= -\frac{r_i^{spk}}{\tau_i^d} + s_i \\ \frac{ds_i}{dt} &= -\frac{s_i}{\tau_r} + \frac{1}{\tau_r \tau_i^d} \sum_{t_i^k < t} \delta(t - t_i^k) \end{aligned}$$

where W_{ij}^{spk} is the recurrent connection strength from unit j to unit i, $\tau_r = 2$ ms is the synaptic rise time and τ_i^d refers to the synaptic decay time for unit i. The synaptic decay time constant values and the recurrent connectivity matrix were transferred from the trained rate RNNs (more details described in the previous chapter and in [KLS19]). The spike train produced by unit i is represented as a sum of Dirac δ functions, and t_i^k refers to the k-th spike emitted by unit i.

The external current input (I_{ext}) contained task-specific input values along with a constant background current set near the action potential threshold. The output of our spiking model at time t is given by

$$o^{spk}(t) = W_{out}^{spk} \boldsymbol{r}^{spk}(t)$$

where the readout weights (W_{out}^{spk}) are also transferred from the trained rate RNN model.

Other LIF model parameters included the action potential threshold (-40 mV), the reset potential (-65 mV), the absolute refractory period (2 ms), and the constant bias current (-40 pA). Eq. (4.2) was discretized using a first-order Euler method with $\Delta t = 0.05$ ms.

4.3.4 Electrophysiological recordings

Extracellular recordings, previously described in [QMSC11, MQSC11, CQM16], were analyzed to validate our RNN model. The dataset contained spike-train recordings from four rhesus macaque monkeys before and after they learned two DMS tasks. Briefly, for the pre-training condition, the monkeys were rewarded for maintaining fixation on the center of the screen regardless of the visual stimuli shown throughout the trial (Fig. 4.8a). For the post-training condition, the monkeys were trained on two DMS tasks: spatial and feature DMS tasks. For the spatial task (Fig. 4.1b), the monkeys were trained to report if two sequential stimuli matched in their spatial locations. For the feature task, they had to distinguish if two sequential stimuli matched in their shapes. The dataset included spike times from single neurons in the dorsal and ventral PFC, but only the units from the dorsal PFC were analyzed for this study.

4.3.5 Estimation of neuronal timescales

To estimate neuronal timescales, we computed the decay time constant of the spike-count autocorrelation function for each unit during the fixation period [MBF+14]. For each unit, we first binned its spike trains during the fixation period over multiple trials using a non-overlapping 50-ms moving window. Since the fixation duration was 1 s for the experimental data and our model, this resulted in a [Number of Trials $\times$ 20] spike-count matrix for each unit. For the experimental data, the minimum number of trials required for a neuron to be considered for analysis was 11 trials. The average number of trials from all the neurons from the post-training condition was 86.8 $\pm$ 35.1 (mean $\pm$ s.d.) trials. For the pre-training condition, the average number of trials was 95.4 $\pm$ 44.4. For the RNN model, we generated 50 trials for each unit.

Next, Pearson's correlation coefficient (ρ) was computed between two time bins (i.e., two columns in the spike-count matrix) separated by a lag (Δ). The coefficient was calculated for all possible pairs with a maximum lag of 600 ms. The coefficients were averaged for each lag value, and an exponential decay function was fitted across the average coefficient values ($\bar{\rho}$) using the Levenberg-Marquardt nonlinear least-squares method:

$$\bar{\rho}(\Delta) = A\left(\exp\left(-\frac{\Delta}{\tau}\right) + B\right) \tag{4.3}$$

where A and B are the amplitude and the offset of the fit, respectively. The timescale (τ) defines how fast the autocorrelation decays and was used to estimate each neuron's timescale.

The following inclusion criteria (commonly used in previous experimental studies) were applied to the RNN model and the experimental data: (1) minimum average firing rate of 1 Hz during the fixation period for the experimental data and 2.5 Hz for the RNN model, (2) $0 < \tau \leq 500$ ms, (3) $A > 0$, and (4) a first decrease in ρ earlier than $\Delta = 150$ ms. In addition, the fitting was started after the first decrease in autocorrelation. For the experimental dataset, 325 dlPFC units from the post-training condition and 434 units from the pre-training condition

satisfied the above criteria. For the DMS RNN model, 931 units from 40 good performance RNNs and 604 units from 26 poor performance RNNs met the criteria. For the AFC model, 1138 units from 40 RNNs satisfied the criteria.

4.3.6 Cross-temporal decoding analysis

The amount of information encoded by each unit was estimated using cross-temporal decoding analysis [SKS$^+$13, SWFS17, WSB$^+$18]. For both experimental and model data, a Gaussian kernel (s.d. = 50 ms) was first applied to the spike-trains to obtain the firing rate estimates over time. For each cue stimulus identity, each neuron's firing rate timecourses were divided into two splits (even vs. odd trials) and averaged across trials within each split. There were 9 cue conditions (i.e., 9 spatial locations) for the spatial DMS task and 8 cue conditions (i.e., 8 shapes) for the feature DMS task. Within each task, all possible pairwise differences in mean firing rates between any two cue conditions for each neuron in each split were computed. Next, Pearson's correlation coefficient was determined for each pairwise difference condition between the two splits (at each time point across neurons). The correlation coefficients from both tasks (36 pairwise difference conditions for the spatial task and 28 conditions for the feature task) at each time point were averaged after applying the Fisher's z-transformation resulting in a single measure we refer to as a discriminability or decodability score. The within-time discriminability scores were computed from the correlation coefficients at $t_1 = t_2$ where t_1 and t_2 refer to the time points used for the two splits.

Nonparametric cluster-based permutation tests were utilized to account for multiple comparisons and to determine significant discriminability (Fig. 4.3a) and differences in discriminability between short and long τ subgroups (Figs. 4.3 and 4.8) [MO07]. To identify significant clusters in the cross-temporal matrices (Fig. 4.3a and Fig. 4.8c,f), cue stimulus condition labels were randomly shuffled for 1,000 times within each split to construct the null distribution. A point was considered significant if its value exceeded the 95th percentile of the null distribution,

and the largest cluster size (i.e., number of contiguous points that were significant) from the data was compared against the null distribution of the largest cluster size values to correct for multiple comparisons. To determine if within-time decoding timecourses were significantly different between long and short τ groups (Fig. 4.3b and Fig. 4.8c,f), τ group labels were randomly shuffled for 1,000 times within each split and each task. Again, a time point was considered significant if it was greater than the 95th percentile of the null distribution. Similar multiple comparison correction, as described above, was applied.

Cross-temporal decoding matrices and within-time decoding timecourses for the dlPFC data (Figs. 4.3 and 4.8) were smoothed for better visualization, but all statistical tests were performed on unsmoothed data.

4.3.7 Connectivity rewiring method

For Fig. 4.4e, we characterized which connection type contributed the most to the long neuronal timescales observed in the DMS RNN model by randomly shuffling connections belonging to each type ($I \rightarrow I$, $I \rightarrow E$, $E \rightarrow I$, or $E \rightarrow E$) while preserving the original distribution of the connection types. For the $I \rightarrow I$ type, all the outward connections from each inhibitory unit to other inhibitory units were first identified. These connections were then rewired randomly in a manner that preserved their connection identity (i.e., $I \rightarrow I$). This procedure was repeated for the other three synaptic types. For Fig. 4.5, all the synaptic weights corresponding to each connection type were either decreased or increased by 30% without rewiring.

To quantify the amount of cue-specific information maintained during the delay period in each of the four shuffling conditions (Fig. 4.4f), we performed the within-time decoding analysis (see above) for all the units in each RNN per shuffling condition. This resulted in 40 within-time decoding timecourses (one for each RNN) for each rewiring condition.

4.3.8 Cue stimulus selectivity

In order to identify inhibitory units selective for each of the two cue stimuli (-1 or +1), we computed a cue preference index (θ) for each unit using:

$$\theta_i = \frac{r_{i,+1} - r_{i,-1}}{r_{i,+1} + r_{i,-1}}$$

where $r_{i,+1}$ refers to the average firing rate of unit i across positive cue stimulus trials (50 trials) during the cue stimulus window, while $r_{i,-1}$ indicates the average activity across negative cue stimulus trials (50 trials). Thus, $\theta_i > 0$ indicates that unit i prefers the positive cue stimulus over the negative stimulus. Based on this selectivity measure, two subgroups of inhibitory units (one for $\theta > 0$ and the other for $\theta < 0$) were identified for each DMS RNN.

4.3.9 Spike-count Fano factors

The relationship between spike-count variability and neuronal timescales was investigated by computing trial-to-trial spike-count Fano factors during the fixation period (Fig. 4.7). For each unit included in the timescale analysis, the variance of the total number of spikes within the 1-s fixation window across trials was first computed. The Fano factor was then calculated by dividing the variance by the mean spike count. The trials used for computing the Fano factors were identical as those used for estimating the neuronal timescales for both neural and RNN data.

4.3.10 Reconfiguring pre-trained RNNs

In Fig. 4.8g,h, the continuous-variable rate RNNs trained to perform the AFC and DMS tasks were used. For Fig. 4.8g, only the input weights (W_{in}) for the AFC RNNs were re-trained via the same gradient descent algorithm to perform the DMS task. The $I \rightarrow I$ connections were either unaltered (yellow in Fig. 4.8g) or increased by 200% (orange in Fig. 4.8g). In Fig. 4.8h,

only the input weights for the DMS RNNs were reconfigured to perform the AFC task. The maximum number of training trials was set to 6,000 trials for computational efficiency.

4.4 Results

4.4.1 Spiking recurrent neural network model

To study how stable temporal dynamics associated with WM emerge, we trained a spiking RNN model to perform a WM task. The model used in the present study is composed of leaky integrate-and-fire (LIF) units recurrently connected to one another (see Section 4.3).

The WM task we used to train the spiking RNNs was a delayed match-to-sample (DMS) task (Fig. 4.1a; see Section 4.3). The task began with a 1 s long fixation period (i.e., no external input) followed by two sequential input stimuli (each stimulus lasting for 0.25 s) separated by a delay period (0.75 s). The input signal was set to either -1 or +1 during the stimulus window. If the two sequential stimuli had the same sign (-1/-1 or +1/+1), the network was trained to produce an output signal approaching +1 after the offset of the second stimulus. If the stimuli had opposite signs (-1/+1 or +1/-1), the network produced an output signal approaching -1.

Using a method that we had previously developed, we configured the recurrent connections required for the spiking model to perform the task [KLS19]. Briefly, we trained continuous-variable rate RNNs to perform the task using a gradient descent algorithm, and the trained networks were then mapped to LIF networks. In total, we "trained" 40 LIF RNNs of 200 units (80% excitatory and 20% inhibitory units) to perform the task with high accuracy (accuracy $>$ 95%; see Section 4.3).

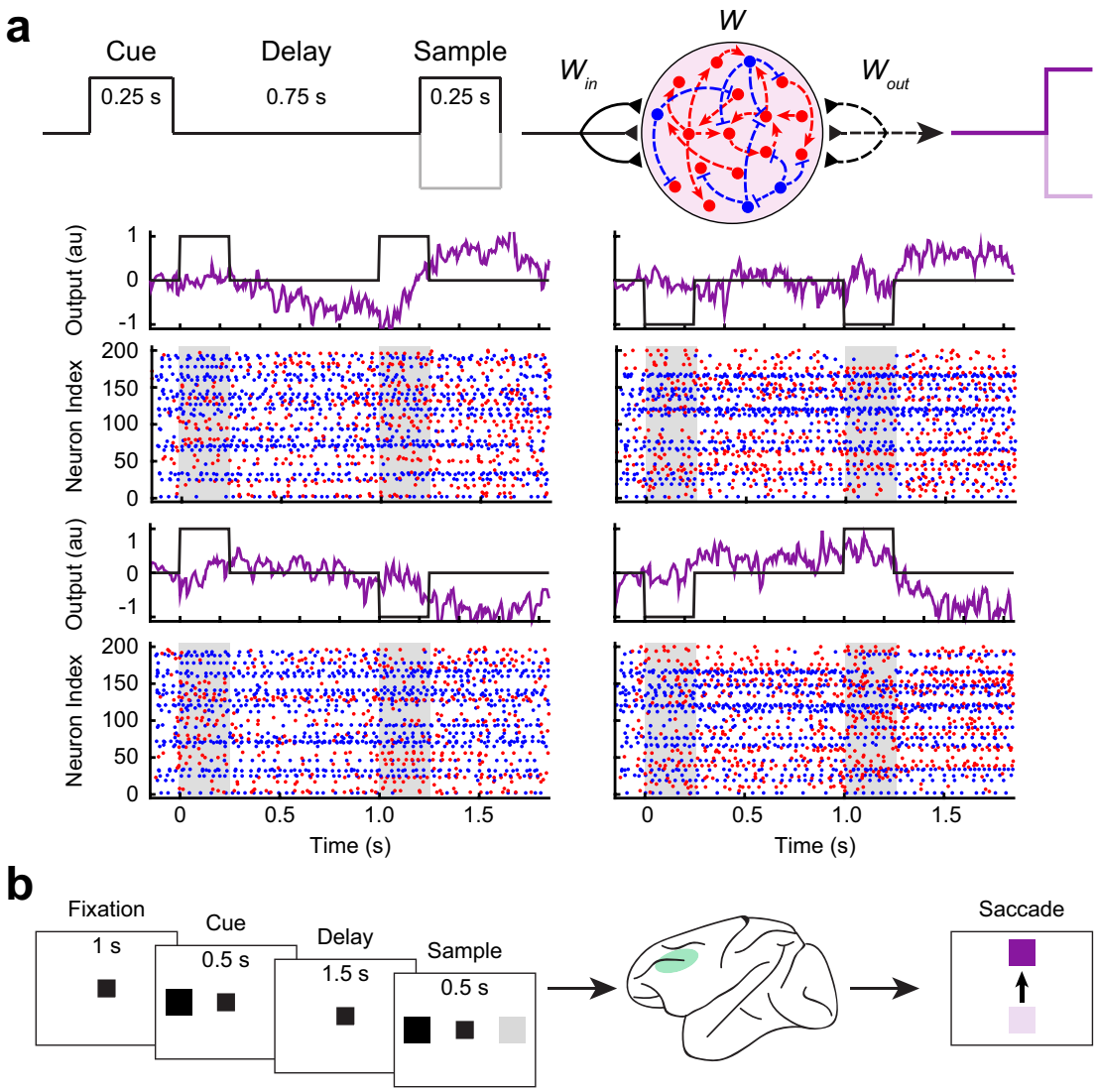


Figure 4.1: **Recurrent neural network model and experimental data. a**, Spiking recurrent neural network (RNN) model contained excitatory (red circles) and inhibitory (blue circles) units recurrently connected to one another. The model was trained to perform a delayed match-to-sample (DMS) task. Each RNN contained 200 units (80% excitatory and 20% inhibitory), and 40 RNNs were trained to perform the DMS task. The dashed lines (recurrent connections and readout weights) were optimized via a supervised learning method. Example output signals along with the corresponding spike raster plots from a trained RNN is shown. Gray shading, stimulus window. **b**, Spatial DMS task paradigm used to train four rhesus monkeys [CQM16]. Extracellular recordings from the dorsolateral prefrontal cortex (green area) were analyzed.

4.4.2 Experimental data

To ensure that our spiking model is a biologically valid one for probing neuronal timescales observed in the cortex, we also analyzed a publicly available dataset containing extracellular

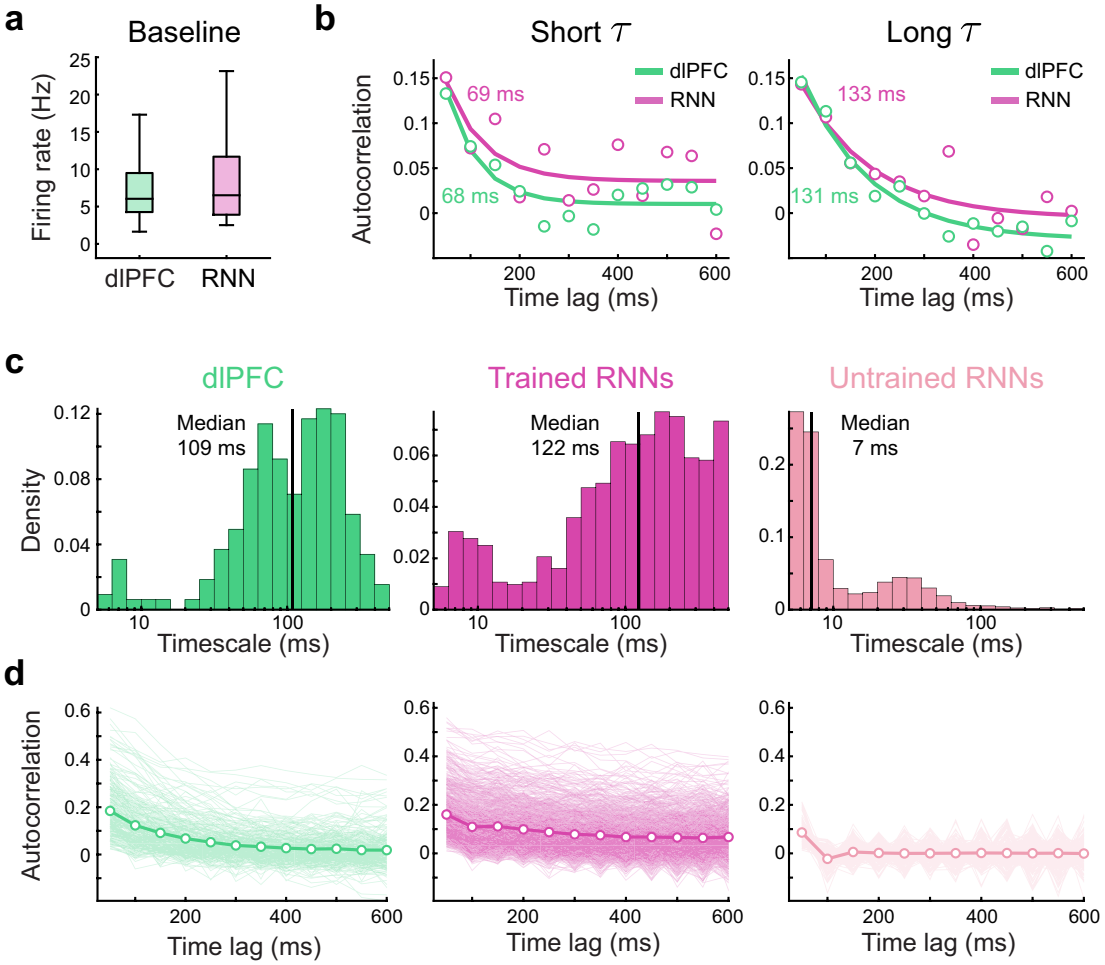


Figure 4.2: **RNN model trained on the DMS task and the dlPFC data contain units with long timescales. a**, Distribution of the firing rates during the fixation period was not significantly different between the experimental data and the RNN model ($P < 0.70$, two-sided Wilcoxon rank-sum test). **b**, Autocorrelation decay curves from example units with short (left) and long (right) timescale values. **c**, Histograms of the distribution of the timescales from the experimental data ($n = 325$; green), trained RNNs ($n = 931$; magenta), and random RNNs ($n = 3963$; light magenta). Solid vertical lines represent median $\log(\tau)$. **d**, Autocorrelation decay curves from single units (light) and the population average autocorrelation (bold) for the dlPFC data, trained RNNs, and random RNNs. For the random RNNs, only 20% of the total single unit traces shown. Boxplot central lines, median; red circles, mean; bottom and top edges, lower and upper quartiles; whiskers, 1.5*interquartile range; outliers not plotted.

spike trains recorded from the dorsolateral prefrontal cortex (dlPFC) of four rhesus monkeys [QMSC11, MQSC11, CQM16]. The monkeys were trained on spatial and feature DMS tasks. A trial for both task types began with a fixation period (1 s in duration) during which the monkeys

were required to maintain their gaze at a fixation target. For a spatial DMS trial, the monkeys were trained to report if two sequential stimuli separated by a delay period (1.5 s) matched in spatial location (Fig. 4.1b). For a feature DMS trial, the monkeys were required to distinguish if two sequential stimuli (in the same spatial location) matched in shape. More details regarding the dataset and the tasks can be found in Section 4.3 and in [QMSC11, MQSC11].

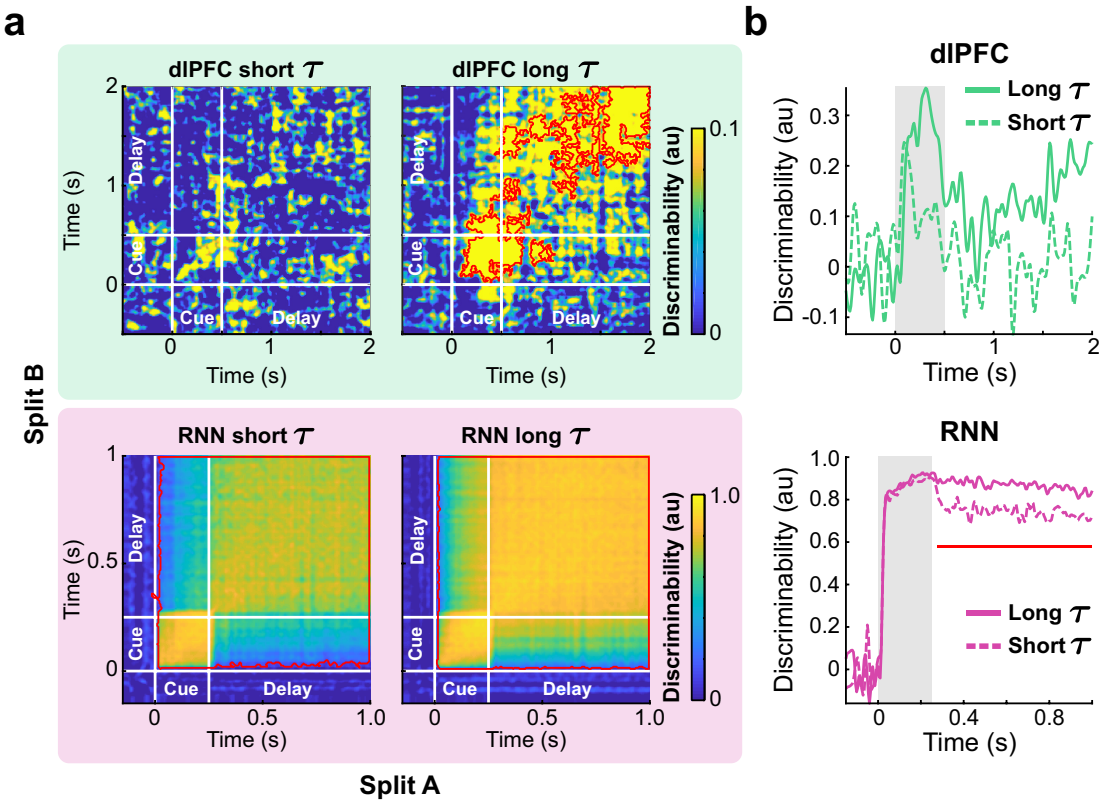


Figure 4.3: **Long τ units maintain cue stimulus information during the delay period robustly. a**, Cross-temporal discriminability matrices for the dlPFC data (top row) and the RNN model (bottom row). Red contours indicate significant decodability (cluster-based permutation test, $P < 0.05$; see Section 4.3). **b**, Within-time discriminability timecourses from the short (dashed) and long (solid) τ groups for the dlPFC data and the RNN model. Gray shading, cue stimulus window. Red lines indicate significant differences in decoding between the short and long τ groups (cluster-based permutation test, $P < 0.05$; see Section 4.3).

4.4.3 Long neuronal timescales in both RNN model and experimental data

Previous studies demonstrated that higher cortical areas consist of neurons with long, heterogeneous timescales using the spike-count autocorrelation decay time constant as a measure of a neuron's timescale [MBF$^+$14, CTW$^+$18, WSB$^+$18]. Here, we sought to confirm that our spiking RNNs trained on the DMS task and the neural data were also composed of units with predominantly long timescales. For each unit from our RNNs and the dlPFC, we computed the autocorrelation decay time constant (τ) of its spike-count during the 1 s fixation period (see Section 4.3) [MBF$^+$14]. The baseline activities (average firing rates during the fixation period) of the units that satisfied the inclusion criteria were comparable between the dlPFC data and our model (Fig. 4.2a; see Section 4.3). Both data contained units with slow temporal dynamics (i.e., long τ values) and short τ units whose autocorrelation function decayed fast (Fig. 4.2b). Furthermore, the distribution of the timescales was heavily left-skewed for both data (Fig. 4.2c,d, left and middle panels) underscoring overall slow temporal properties associated with WM. On the other hand, random RNNs (sparse, random Gaussian connectivity weights) were dominated by units with extremely short timescales (Fig. 4.2c,d, right panels), suggesting that the long τ units observed in the trained RNNs were the result of the supervised training.

4.4.4 Long neuronal timescales are essential for stable coding of stimuli

Next, we investigated to see if units with longer τ values were involved with more stable coding compared to short τ units using cross-temporal decoding analysis [SKS$^+$13, SWFS17, WSB$^+$18]. Briefly, for each cue stimulus identity, the trials of each unit were divided into two splits in an interleaved manner (i.e., even vs. odd trials). All possible pairwise differences (in instantaneous firing rates) between cue conditions were computed within each split. Finally, a Fisher-transformed Pearson correlation coefficient was computed between the pairwise differences of the first split at time t_1 and the differences of the second split at time t_2 (see Section 4.3).

Therefore, a high Fisher-transformed correlation value (i.e., high discriminability) represents a reliable cue-specific difference present in the network population.

We performed the above analysis on short and long neuronal timescale subgroups from the neural data and the RNN model. A unit was assigned to the short τ group if its timescale was smaller than the lower quartile value. The upper quartile was used to identify units with large τ values. There were 64 units in each subgroup for the experimental data. For the RNN model, there were 230 units in each subgroup.

The cross-temporal discriminability analysis revealed that stronger cue-specific differences (i.e., higher discriminability) across the delay period were present in the long τ subgroup compared to the short τ subgroup for both data (Fig. 4.3a). The significant decodability during the delay period for the dlPFC dataset mainly stemmed from the spatial task dataset (Fig. 4.9). The within-time discriminability (i.e., taking the diagonal values of the cross-temporal decoding matrices) for the long τ group was significantly higher than the discriminability observed from the short τ group throughout the delay period for the RNN model (Fig. 4.3b). Although significant within-time discriminability was not observed for the dlPFC data (Fig. 4.3b, top), [WSB$^+$18] reported significant within-time decodability during the delay period in the primate lateral prefrontal cortex, consistent with our model findings.

4.4.5 Strong inhibitory connections give rise to task-specific temporal receptive fields

Neuronal timescales extracted from cortical areas have been shown to closely track the anatomical and functional organization of the primate cortex [MBF$^+$14, CKG$^+$15]. For instance, sensory areas important for detecting incoming stimuli house neurons with short timescales. On the other hand, higher cortical areas, including prefrontal areas, may require neurons with stable temporal receptive fields that are capable of encoding and integrating information over a longer timescale.

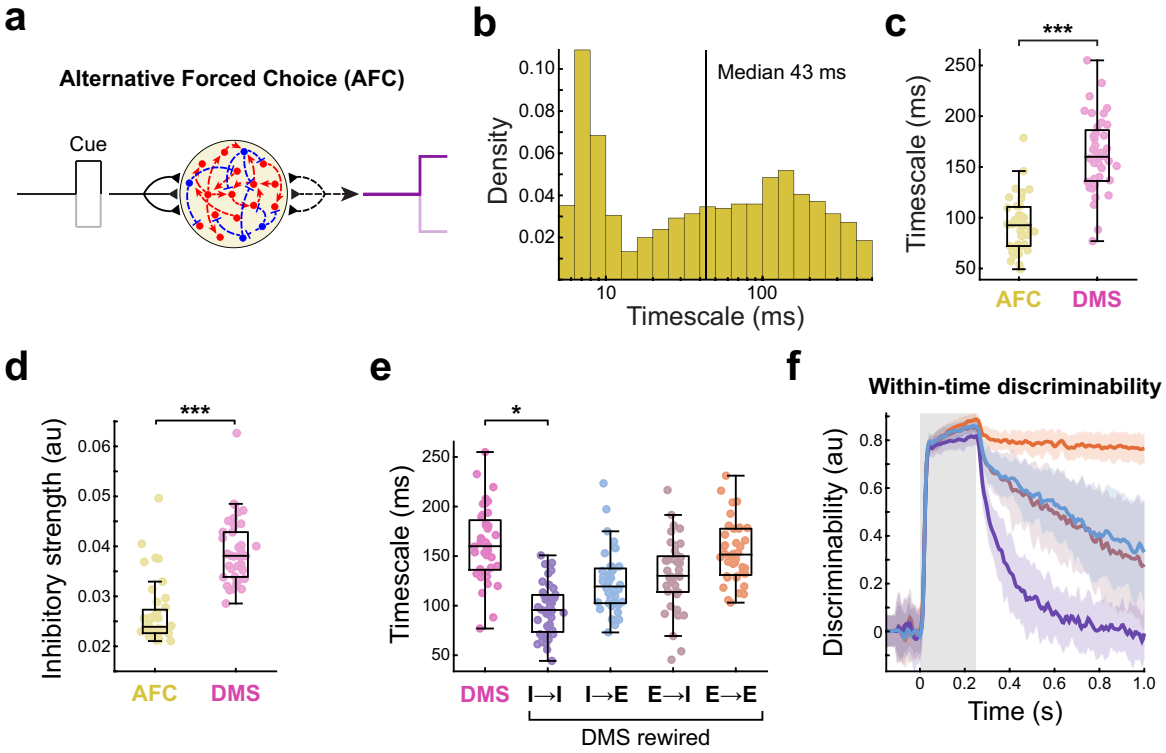


Figure 4.4: **Inhibitory synaptic weights lead to task-specific timescales.** **a**, Task paradigm for the alternative forced choice (AFC) task. **b**, Distribution of the neuronal timescales extracted from 40 RNNs trained on the AFC task. Solid vertical line represents median $\log(\tau)$. **c**, Average timescale values from the AFC and DMS RNNs. Each circle represents the average value from one RNN. **d**, Average recurrent inhibitory synaptic strengths from the AFC and DMS models. **e**, Average timescales from the DMS RNNs with each synaptic type rewired randomly (Friedman test, $F = 74.19$, $P < 0.0001$). **f**, Within-time discriminability timecourses averaged across all the DMS RNNs for each rewiring condition. Same color scheme as **e**. The bold line indicates the mean timecourse averaged across 40 RNNs (and all units). Colored shading, $\pm$ standard deviation (s.d.). Gray shading, cue stimulus window. Boxplot central lines, median; red circles, mean; bottom and top edges, lower and upper quartiles; whiskers, 1.5*interquartile range; outliers not plotted. $*P < 0.01$, $***P < 0.0001$ by Wilcoxon signed-rank test (**c**,**d**) or Dunn's multiple comparisons test (**e**).

To investigate if such functional specialization also emerges in our spiking model, we trained another group of spiking RNNs ($n = 40$ RNNs) on a simpler task that did not require WM. The non-WM task, which we refer to as two-alternative forced choice (AFC) task, required the RNNs to respond immediately after the cue stimulus: output approaching -1 for the "-1" cue and +1 for the "+1" cue (Fig. 4.4a; see Section 4.3). Apart from the task paradigm, all the other model parameters were identical to the parameters used for the DMS RNNs.

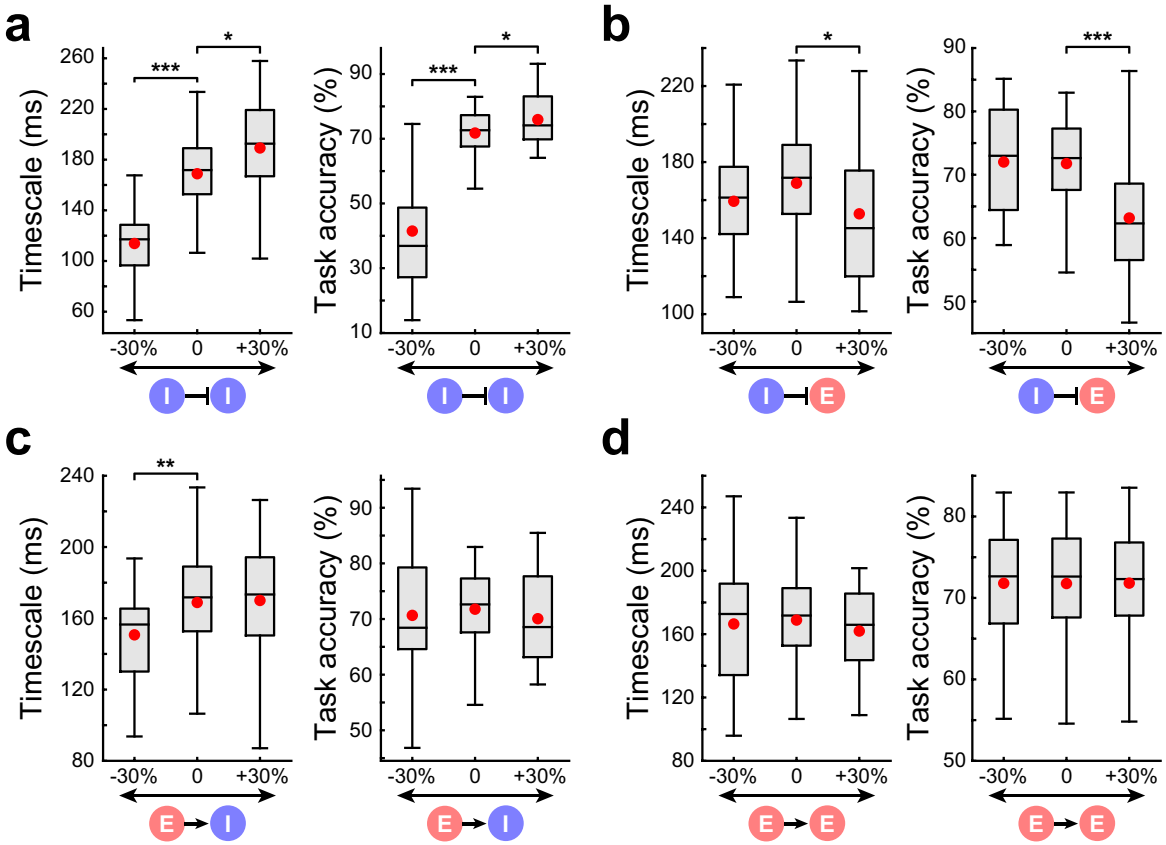


Figure 4.5: ***I* → *I* connectivity strength strongly mediates both neuronal timescales and task performance.** **a**, **b**, **c**, **d**, Timescales and task performance changes when $I \rightarrow I$ (**a**), $I \rightarrow E$ (**b**), $E \rightarrow I$ (**c**), or $E \rightarrow E$ (**d**) connection strength was either decreased or increased by 30%. Boxplot central lines, median; red circles, mean; bottom and top edges, lower and upper quartiles; whiskers, 1.5*interquartile range; outliers not plotted. $^{*}P < 0.05$, $^{**}P < 0.005$, $^{***}P < 0.0001$ by Wilcoxon signed-rank test.

Because the AFC task paradigm did not require the RNNs to store information related to the cue stimulus, we expected that these networks would exhibit faster timescales compared to the DMS RNNs. Consistent with this hypothesis, the AFC RNNs did not contain as many long τ units as the DMS RNNs (Fig. 4.4b), and the timescales averaged by network were also significantly faster for the AFC RNNs (Fig. 4.4c).

To gain insight into the circuit mechanisms underlying the difference in the timescale distributions of the AFC and DMS RNN models, we compared the recurrent connectivity patterns between these two models. The most notable difference was the inhibitory synaptic strength,

which was significantly greater for the DMS RNNs (Fig. 4.4d). In order to confirm if strong inhibitory signaling led to the long timescales observed in the DMS model, we randomly rewired all the connections belonging to each of the four synaptic types ($I \rightarrow I$, $I \rightarrow E$, $E \rightarrow I$, and $E \rightarrow E$) and computed the timescales again (see Section 4.3). Of the four conditions, only rewiring $I \rightarrow I$ synapses resulted in significantly shorter timescales than the timescales from the intact DMS model (Fig. 4.4e), and the distribution of the timescales pooled from all 40 RNNs with $I \rightarrow I$ connections shuffled resembled the distribution obtained from the AFC model (Fig. 4.10). In addition, the amount of cue-specific information maintained during the delay period (as measured by the within-time decoding timecourses) was the lowest for the $I \rightarrow I$ rewired condition (Fig. 4.4f), suggesting that shuffling $I \rightarrow I$ synapses was detrimental to memory maintenance.

4.4.6 Inhibitory-to-inhibitory connections regulate both neuronal timescales and task performance

Given our findings that $I \rightarrow I$ connections are important for long neuronal timescales and information encoding, we next investigated if $I \rightarrow I$ synapses could be manipulated to provide more stable temporal receptive fields and to improve WM maintenance.

Recent studies revealed that optogenetically stimulating SST or VIP interneurons that specifically inhibit PV interneurons could improve memory retrieval [KD17, XLT$^+$19, CC19]. Based on these experimental observations, we expected that strengthening $I \rightarrow I$ synapses would increase neuronal timescales and task performance of the DMS RNNs. To test this hypothesis, we first generated another group of RNNs with poor DMS task performance (26 RNNs; mean accuracy $\pm$ s.e.m., 71.77 $\pm$ 1.43 %). Next, we modeled the effects of optogenetic manipulation of VIP/SST neurons by either decreasing or increasing $I \rightarrow I$ synaptic strength ($W_{I \rightarrow I}$) in each network by 30% (see Section 4.3). Decreasing the connection strength led to significantly shorter timescales compared to the RNNs without any modification (Fig. 4.5a, left). Strengthening

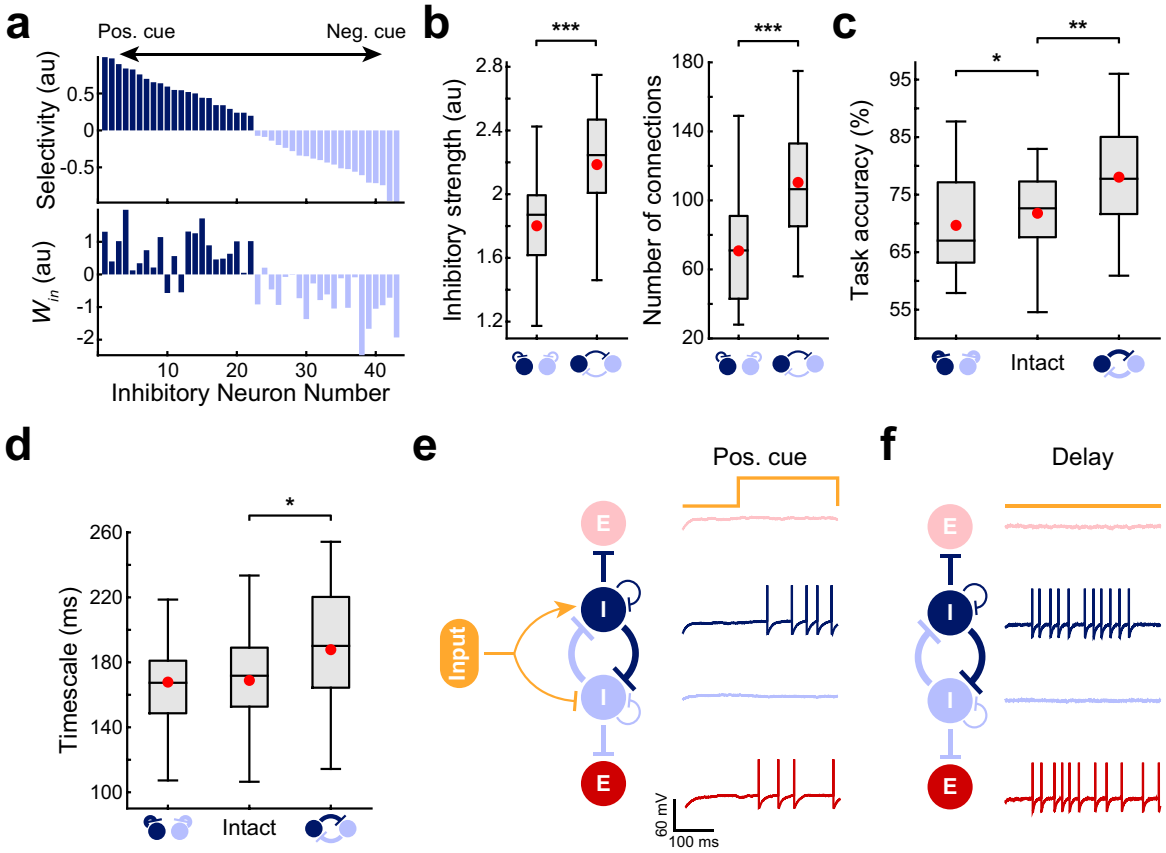
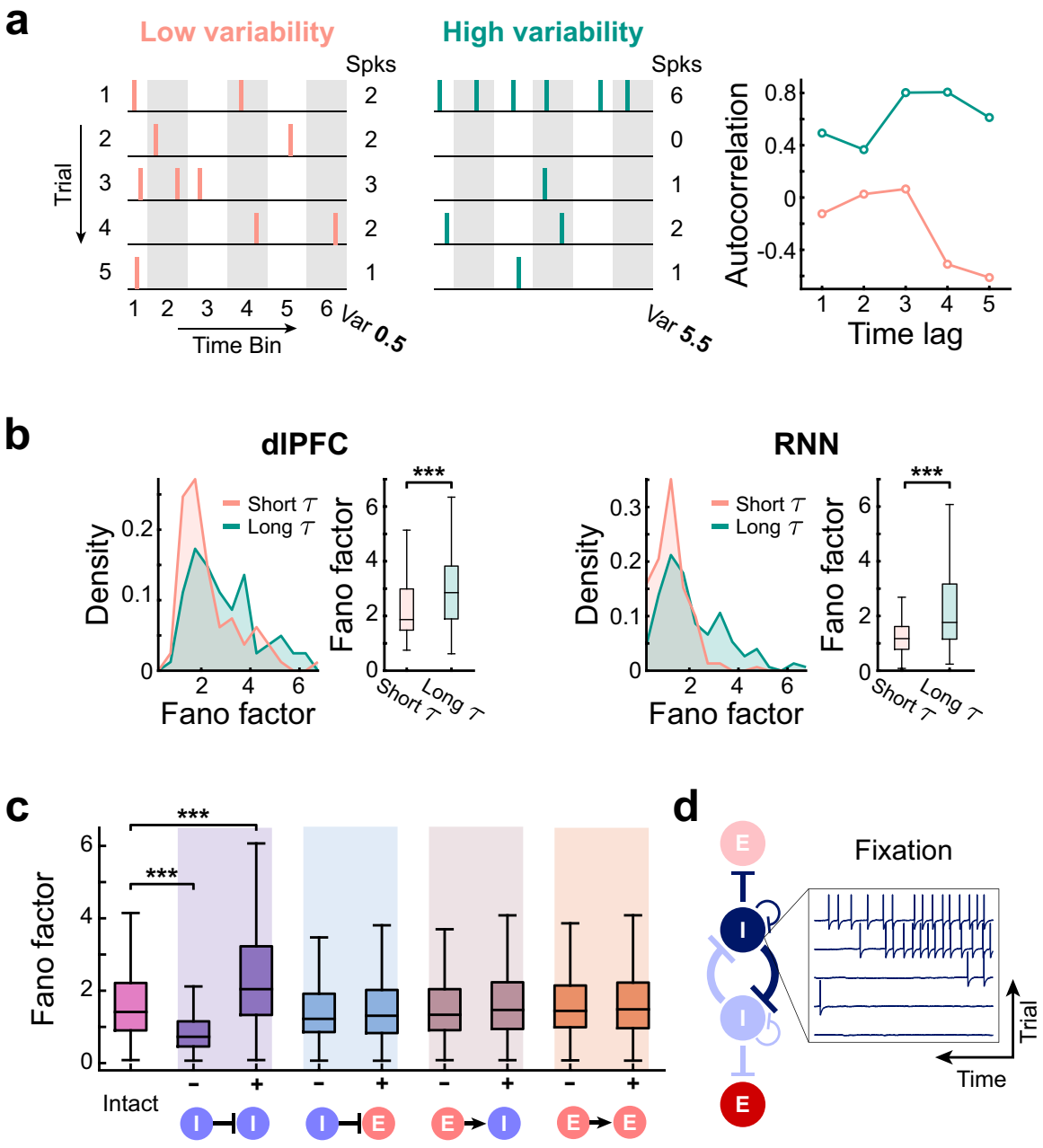


Figure 4.6: **Two oppositely tuned inhibitory subgroups mutually inhibit each other for WM maintenance. a**, Cue preference selectivity (top) and input weights (W_{in}; bottom) from inhibitory units of an example DMS RNN. The selectivity index values are sorted in descending order. **b**, Average inhibitory strengths (left) and number of inhibitory connections (right) within and across two oppositely tuned inhibitory subgroups from all 40 DMS RNNs. **c**, Average task performance of the DMS RNN model when the within-group or across-group inhibition was increased by 30%. **d**, Average neuronal timescales of the DMS RNNs when the within-group or across-group inhibition was increased by 30%. **e**, **f**, Schematic illustration of the circuit mechanism employed by the DMS RNN model during the cue stimulus window (**e**) and delay period (**f**). The positive cue stimulus was used as an example, and membrane voltage tracings from example units are shown. Dark blue and dark red units indicate units that prefer the positive cue stimulus, while the light blue and light red units favor the negative cue. For simplicity, only recurrent inhibitory connections are shown. Boxplot central lines, median; red circles, mean; bottom and top edges, lower and upper quartiles; whiskers, 1.5*interquartile range; outliers not plotted. $*P < 0.05$, $**P < 0.005$, $***P < 0.0001$ by two-sided Wilcoxon rank-sum test (**b**) or Wilcoxon signed-rank test (**c**, **d**).

$W_{I\to I}$ resulted in a moderate but significant increase in neuronal timescale (Fig. 4.5a, left). The task performance of the RNNs followed the same pattern: decreasing $W_{I\to I}$ severely impaired

WM maintenance while increasing $W_{I\to I}$ significantly improved task performance (Fig. 4.5a, right). Increasing $W_{I\to I}$ further did not correspond to a significant increase in timescale and task performance (Fig. 4.11). For $I \to E$ connections, only enhancing $W_{I\to E}$ resulted in significant changes in both timescale and task performance (Fig. 4.5b). Manipulating $E \to I$ synapses did not affect the task performance, but decreasing $W_{E\to I}$ significantly shortened the timescales (Fig. 4.5c). Altering the excitatory-to-excitatory connections did not produce any significant changes (Fig. 4.5d). Overall, these findings suggest that $I \to I$ synapses tightly mediate both temporal stability and WM maintenance. The findings also indicate that the main downstream effect of $I \to I$ connections is to disinhibit excitatory units.

Figure 4.7: **High trial-to-trial spike-count variability during the fixation corresponds to long neuronal timescale.** **a**, Schematic illustrating how high spike-count variability across multiple trials can result in slow decay of the autocorrelation function. **b**, Comparison of the spike-count Fano factors from the short and long τ groups in the neural data (left) and the DMS RNN model (right). **c**, Average Fano factors from the DMS model with each of the synaptic type either decreased ("-") or increased ("+") by 30% (Kruskal-Wallis test, $H = 665.2$, $P < 0.0001$). **d**, Spiking activity of an example inhibitory unit during the fixation period across 5 trials. The trials were sorted by the number of spikes. Units that were strongly modulated by the disinhibitory circuit mechanism showed highly dynamic baseline firing patterns across trials. Boxplot central lines, median; red circles, mean; bottom and top edges, lower and upper quartiles; whiskers, 1.5*interquartile range; outliers not plotted. $***P < 0.0001$ by two-sided Wilcoxon rank-sum test (**b**) or Dunn's multiple comparisons test (**c**).


a
Low variability
High variability
Spks
Trial
Time Bin
Var 0.5
Var 5.5
Autocorrelation
Time lag
b
dlPFC
RNN
Short τ
Long τ
Density
Fano factor
***
c
Intact
d
Fixation
Trial
Time

4.4.7 Unique inhibitory-to-inhibitory circuitry for WM maintenance

So far, our results indicate that (1) microcircuitry involving specific $I \rightarrow I$ connectivity patterns is important for WM (Fig. 4.4e) and (2) $I \rightarrow I$ can be strengthened to enhance both neuronal timescales and task performance (Fig. 4.5a). Here, we dissect the DMS RNN model to elucidate how specific and strong $I \rightarrow I$ connections lead to stable memory retention.

Focusing on inhibitory units only, we first characterized the cue stimulus selectivity from each inhibitory unit in an example DMS network (see Section 4.3). Analyzing the selectivity index values revealed two distinct subgroups of inhibitory units in the network: one group of units favoring the positive cue stimulus and the other group selective for the negative stimulus (Fig. 4.6a, top). The input weights (W_{in}) that project to these units closely followed the selectivity pattern (Fig. 4.6a, bottom).

Given these two subgroups with distinct selectivity patterns, we next hypothesized that mutual inhibition between these two groups (across-group inhibition) was stronger than within-group inhibition. Indeed, inhibition between the oppositely tuned inhibitory populations was significantly greater (both in synaptic strength and number of connections) than inhibition within each subgroup across all RNNs (Fig. 4.6b). To confirm that the behavioral improvement we observed with $I \rightarrow I$ enhancement in Fig. 4.5a was largely due to the strengthened across-group inhibition, we increased across-group and within-group $I \rightarrow I$ connections separately (see Section 4.3). The DMS RNN performance improved following enhancement of the across-group inhibition, while increasing the within-group inhibition impaired performance (Fig. 4.6c). In addition, across-group $I \rightarrow I$ enhancement resulted in a significant increase in neuronal timescale (Fig. 4.6d).

In summary, these findings imply that robust inhibition of oppositely tuned inhibitory subpopulations is critical for memory maintenance in our RNN model. For example, a positive cue stimulus activates the inhibitory subgroup selective for that stimulus and deactivates the negative stimulus subgroup (Fig. 4.6e). Through disinhibition, a group of excitatory units that

favor the positive cue stimulus also emerges. During the delay period, the inhibition strength between these two inhibitory subgroups dictates the stability of the cue-specific activity patterns generated during the stimulus window (Fig. 4.6f).

4.4.8 Circuit mechanism for WM generates units with long neuronal timescales

The circuit mechanism (Fig. 4.6e,f) explains why enhancing $I \rightarrow I$ connections results in improved WM performance, but it is still not clear how this same mechanism also produces units with long timescales.

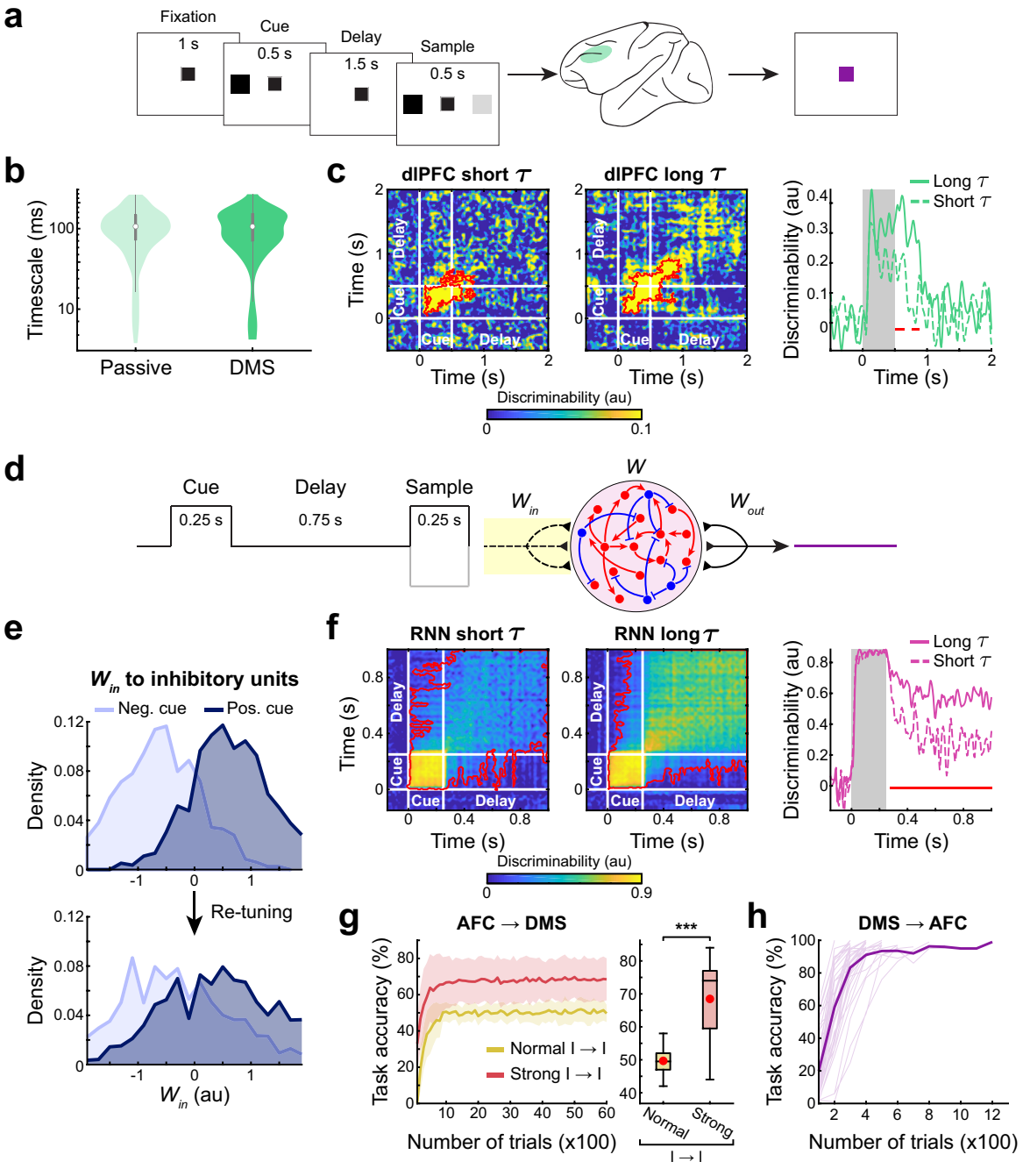
Here, we first demonstrate that a high trial-to-trial spike-count variability during the fixation period could give rise to slow decay of the spike-count autocorrelation function. If a neuron exhibits highly variable activity patterns across trials such that it is highly active (i.e., persistent firing) in some trials and relatively silent in other trials, the Pearson correlation between any two time bins within the fixation window could be large (Fig. 4.7a). On the other hand, firing activities with a low trial-to-trial variability could result in a weak correlation between two time bins. To directly test this positive relationship between trial-to-trial variability and neuronal timescales, we computed spike-count Fano factors (spike-count variance divided by spike-count mean across trials; see Section 4.3) for the short and long τ subgroups in both neural and model data. The Fano factor values for the short timescale subgroup were significantly smaller than the values obtained from the long τ group for both data (Fig. 4.7b). There was also a significant positive correlation between the spike-count Fano factors and neuronal timescales across all the units in both data (Spearman rank correlation, $r = 0.25, P < 0.0001$ for dlPFC; $r = 0.28, P < 0.0001$ for RNN; Fig. 4.12).

Manipulating each of the four synaptic types (decreasing or increasing synaptic strength by 30%) in our DMS RNN model revealed that $I \rightarrow I$ connections strongly modulated the spike-count Fano factors (Fig. 4.7c). Enhancing $I \rightarrow I$ synaptic strength led to units with more variable

spiking patterns across trials, whereas reducing the strength resulted in smaller Fano factors (example shown in Fig. 4.13).

In our RNN model, strong $I \rightarrow I$ synapses give rise to both excitatory and inhibitory units behaving in a highly variable manner during the fixation period (Fig. 4.7d). For instance, an inhibitory unit selective for the positive stimulus could be partially activated in some trials by chance (i.e., via random noise during the fixation period), and this, in turn, could silence a portion of the negative stimulus inhibitory population (light blue circle in Fig. 4.7d). This leads to variable firing activities across trials in inhibitory units. Furthermore, the dynamic activity of the inhibitory population could be transferred to the excitatory population via disinhibition. Therefore, $I \rightarrow I$ connections play a central role in conferring the network with highly dynamic baseline firing patterns, which then translate to high τ values.

Figure 4.8: **Strong $I \rightarrow I$ connections intrinsic to prefrontal cortex. a**, Passive task paradigm used by [CQM16] to train the same four monkeys before they learned the DMS tasks (Fig. 4.1b). **b**, Distribution of the neuronal timescales from the monkeys before (i.e., passive) and after they learned the DMS tasks. **c**, Cross-temporal decoding matrices and within-time decoding timecourses from the short and long τ subgroups. **d**, Passive task paradigm used to re-train our DMS RNNs. Only the input weights (dashed lines with yellow shading) were trained. **e**, Distribution of the input weights projecting to the two inhibitory subgroups tuned to the two cue stimuli from all 40 DMS RNNs before (top) and after (bottom) re-training. **f**, Cross-temporal decoding matrices and within-time decoding timecourses from the short and long τ subgroups for the re-trained DMS RNNs. **g**, Task performance during re-training of the AFC rate RNNs to perform the DMS task (left) and average performance at the end of training (right). The task performance significantly increased when $I \rightarrow I$ connections were strengthened (orange; see Section 4.3). Shaded area, $\pm$ s.d. **h**, Task performance during re-training of the DMS rate RNNs to perform the AFC task. Individual networks (light) and mean across 40 DMS RNNs (bold). Boxplot central lines, median; red circles, mean; bottom and top edges, lower and upper quartiles; whiskers, 1.5*interquartile range; outliers not plotted. Red contours indicate significant discriminability (cluster-based permutation test, $P < 0.05$; see Section 4.3). Red lines indicate significant differences in decoding between the short and long τ groups (cluster-based permutation test, $P < 0.05$; see Section 4.3). $^{***}P < 0.0001$ by Wilcoxon signed-rank test.


a
Fixation
1 s
Cue
0.5 s
Delay
1.5 s
Sample
0.5 s
b
Timescale (ms)
100
10
Passive
DMS
c
dlPFC short τ
dlPFC long τ
Time (s)
Delay
Cue
Discriminability (au)
Long τ
Short τ
d
Cue
Delay
Sample
0.25 s
0.75 s
0.25 s
W
W_in
W_out
e
W_in to inhibitory units
Neg. cue
Pos. cue
Density
Re-tuning
W_in (au)
f
RNN short τ
RNN long τ
g
AFC → DMS
Normal I → I
Strong I → I
***
Task accuracy (%)
Number of trials (x100)
h
DMS → AFC

4.4.9 Strong $I \rightarrow I$ is an intrinsic property of prefrontal cortex

Cognitive flexibility is one of the hallmarks of the prefrontal cortex [MC01, GR11]. If higher-order areas are indeed wired with specific and robust $I \rightarrow I$ synapses that give rise to stable temporal receptive fields, then what would happen to these connections during learning? Would learning a new task disrupt the existing $I \rightarrow I$ connectivity structure, thereby abolishing the previously established timescale distribution? To answer these questions, we analyzed neuronal timescales from the same monkeys before they learned the DMS task. For the pre-training condition, the monkeys were trained on a passive task (Fig. 4.8a): they were trained to maintain their gaze at a central fixation point throughout the trial regardless of the stimuli presented around the fixation point [MQC07].

Surprisingly, the timescales from the spike-train data from the dlPFC of the same four monkeys that learned the passive task were similar to the timescales obtained after the monkeys learned the DMS task (Fig. 4.8b). In addition, the cue-specific information maintenance during the delay period by long τ units was largely abolished, and the within-time decoding was similar between long τ and short τ neurons (Fig. 4.8c). These findings suggest that the primate dlPFC was already equipped with stable temporal receptive fields and that learning the DMS task resulted in long τ neurons carrying more information during the delay period while preserving the network temporal dynamic architecture.

Based on these findings, we reasoned that prefrontal cortical areas and other higher cognitive areas are endowed with strong $I \rightarrow I$ connections whose connectivity patterns do not undergo significant plastic changes during learning. Instead, learning-related changes occur to the connections stemming from upstream networks that project to these areas. To test this, we asked if we could only optimize the upstream connections (i.e., input weights; W_{in}) of the DMS RNNs to perform a passive version of the DMS task (Fig. 4.8d; see Section 4.3). By freezing the recurrent connections (W), we ensured the previously observed distribution of the timescales (Fig. 4.2b) was preserved. As expected, the distinct distribution of the input weights projecting to the two

inhibitory subpopulations that we observed in Fig. 4.6a was "flattened" after re-training the DMS RNNs to perform the passive task (Fig. 4.8e). Repeating the cross-temporal discriminability analysis on the re-trained RNNs showed that the cue stimulus information during the delay period was not maintained as robustly by long τ units (Fig. 4.8f). However, the long τ units still carried significantly higher information than the short τ units throughout the delay window. Re-tuning the recurrent connections instead of the input weights for the passive task disrupted the existing timescale structure and resulted in significantly faster timescales (Fig. 4.14).

The above results from the experimental data and our model strongly suggest that higher cortical areas might have intrinsically diverse and robust inhibitory signaling. This innate property, in turn, would give rise to long neuronal timescales, and the incoming connections to these areas could undergo plastic changes to support various higher cognitive functions that require integration of information on a slower timescale. Along this line of thought, we hypothesized that the AFC RNNs, which do not have strong inhibitory-to-inhibitory signaling, are not capable of performing WM tasks by simply re-tuning the input weights only. With the recurrent architecture (W) fixed, we attempted to re-train the input weights of the 40 AFC RNNs to perform the DMS task, but none of the networks could be trained successfully (yellow line in Fig. 4.8g). When we repeated the re-training procedure with the $I \to I$ recurrent connections strengthened (see Section 4.3), the performance of the AFC RNNs significantly improved (magenta line in Fig. 4.8g). On the other hand, the input weights of the DMS RNNs could be successfully tuned to perform the AFC task (Fig. 4.8h), further confirming the hierarchical organization of these two RNN models.

4.5 Discussion

In this study, we provide a computational model that gives rise to task-specific spontaneous temporal dynamics, reminiscent of the hierarchy of neuronal timescales observed across primate cortical areas [MBF$^+$14]. When trained on a WM task, our RNN model was composed of

units with long timescales whose distribution was surprisingly similar to the one obtained from the primate dlPFC. In addition, the long-timescale units encoded and maintained WM-related information more robustly than the short-timescale units during the delay period. By analyzing the connectivity structure of the model, we showed that a unique circuit motif that incorporates strong $I \to I$ synapses is an integral component for WM computations and slow baseline temporal properties. Interestingly, $I \to I$ synaptic weights could be manipulated to control both task performance and neuronal timescales tightly. Our work also provides mechanistic insight into how $I \to I$ connectivity supports memory storage and dynamic baseline activity patterns crucial for long neuronal timescales. Lastly, we propose that the microcircuitry we identified is intrinsic to higher-order cortical areas enabling them to perform cognitive tasks that require steady integration of information.

Relating specific baseline spiking activities to the underlying circuit mechanisms has been challenging partly due to the lack of computational models capable of both performing cognitive tasks and capturing temporal dynamics derived from experiments. [BB19] employed Poisson spiking neurons randomly wired to present a flexible WM model, whereas [MRL18] used LIF RNNs constrained by experimental measurements to underscore the importance of inhibitory connectivity in WM. These studies provide biologically plausible models that can explain several experimental and behavioral aspects of WM, but it is unclear if units with stable baseline temporal dynamics are recruited for performing WM maintenance in these models. It is also possible to study neuronal timescales using continuous rate (i.e., non-spiking) RNNs which have been widely used to uncover neural mechanisms behind cognitive processes [MSSN13, SYW16, Mic17, OM19, YJS$^+$19]. Although spontaneous firing rate estimates could be used in place of spike counts to compute the autocorrelation decay time constants, our spiking RNN model allowed us to (1) use the same experimental procedures previously used to estimate neuronal timescales, (2) easily interpret and compare our model results with experimental findings, and (3) uncover spiking statistics (spike-count Fano factors) associated with long neuronal timescales.

Our work revealed that strong $I \rightarrow I$ connections are critical for long neuronal timescales, and we investigated the functional implication of such connections in WM-related behavior. Despite the fact that excitatory pyramidal cells make up the majority of neurons in cortical areas, inhibitory interneurons have been shown to exert greater influence at the local network level [KF14, BBVF$^+$17]. Furthermore, different subtypes of interneurons play functionally distinct roles in cortical computations [PXH$^+$13, KJL$^+$16]. In agreement with these observations, recent studies uncovered the importance of disinhibitory gating imposed by VIP interneurons [PHK$^+$13, KJA$^+$16, KD17, KPdA$^+$19]. Through inhibition of SST and PV neurons, VIP interneurons have a unique ability to disinhibit pyramidal cells and create "holes" in a dense "blanket of inhibition" [KJA$^+$16]. Surprisingly, optogenetically activating VIP neurons in the PFC of mice trained to perform a WM task significantly enhanced their task performance highlighting that disinhibitory signaling is vital for memory formation and recall [KD17]. Similar to VIP neurons, SST interneurons have also been shown to disinhibit excitatory cells for fear memory [CC19, XLT$^+$19]. Intriguingly, the connectivity structures of the RNNs we trained on a WM task using supervised learning also centered around disinhibitory circuitry with strong $I \rightarrow I$ synapses (Fig. 4.6). The strength of the $I \rightarrow I$ connections was tightly coupled to the task performance of the RNNs. Thus, our work suggests that microcircuitry specializing in disinhibition could be a common substrate in higher-order cortical areas that require short-term memory maintenance.

Most notably, our results shed light on exactly how robust $I \rightarrow I$ connections maintain stable memory storage and long neuronal timescales. By dissecting our WM RNN model, we found that strong mutual inhibition between two oppositely-tuned inhibitory subgroups was necessary for maintaining stimulus-specific information during the delay period (Fig. 4.6). We also illustrated that our model units that were strongly modulated by $I \rightarrow I$ synapses displayed highly dynamic baseline activities leading to both large trial-to-trial Fano factors and long neuronal timescales (Fig. 4.7). Although we only considered two cue stimulus types (positive and negative stimuli) for simplicity, our circuit model could be generalized to store more stimulus types. For

example, another group of inhibitory units tuned to a third stimulus type could be added to our circuit design, forming three mutually inhibiting groups. Interestingly, such a circuit mechanism has been recently identified to generate categorical responses in barn owls [MM20]. Our findings also suggest that baseline trial-to-trial spike-count variability and neuronal timescales are reliable indicators of the underlying circuit mechanisms: neurons with asynchronously occurring synchronous firing patterns (i.e., high variability) could make up WM-related microcircuits. Furthermore, we propose that these signatures are area-specific and do not undergo significant changes during learning.

Although our model can capture several experimental findings, a few interesting questions remain for future studies. For example, our spiking RNN model utilizes connectivity patterns derived from a gradient-descent approach, which is not biologically plausible. It will be important to characterize if more biologically valid learning mechanisms, such as reinforcement learning or Hebbian learning, also generate spiking networks with heterogeneous neuronal timescales. Another unexplored aspect is nonlinear dendritic computations. SST interneurons are known for targeting dendrites of pyramidal cells, and such dendritic inhibition has been associated with gating information [YMW16]. Incorporating dendritic processes into our model could elucidate the computational benefits of dendritic inhibition over perisomatic inhibition during WM. In summary, we have explored a neural circuit mechanism that performs logical computations over time with stable temporal receptive fields.

Chapter 4, in full, is a reprint of the material as it has been written in a manuscript that has been submitted for publication: Robert Kim and Terrence J. Sejnowski. Strong inhibitory signaling underlies stable temporal dynamics and working memory in spiking neural networks. Preprint at `https://www.biorxiv.org/content/10.1101/2020.02.11.944751v1` (2020). The dissertation author was the primary investigator and author of this paper.

4.6 Appendix

4.6.1 Code availability

The code for the analyses performed in this work will be made available at https://github.com/rkim35/wmRNN.

4.6.2 Data availability

The trained RNN models used in the present study will be deposited as MATLAB-formatted data in Open Science Framework, https://osf.io/md4wg.

4.6.3 Supplementary figures

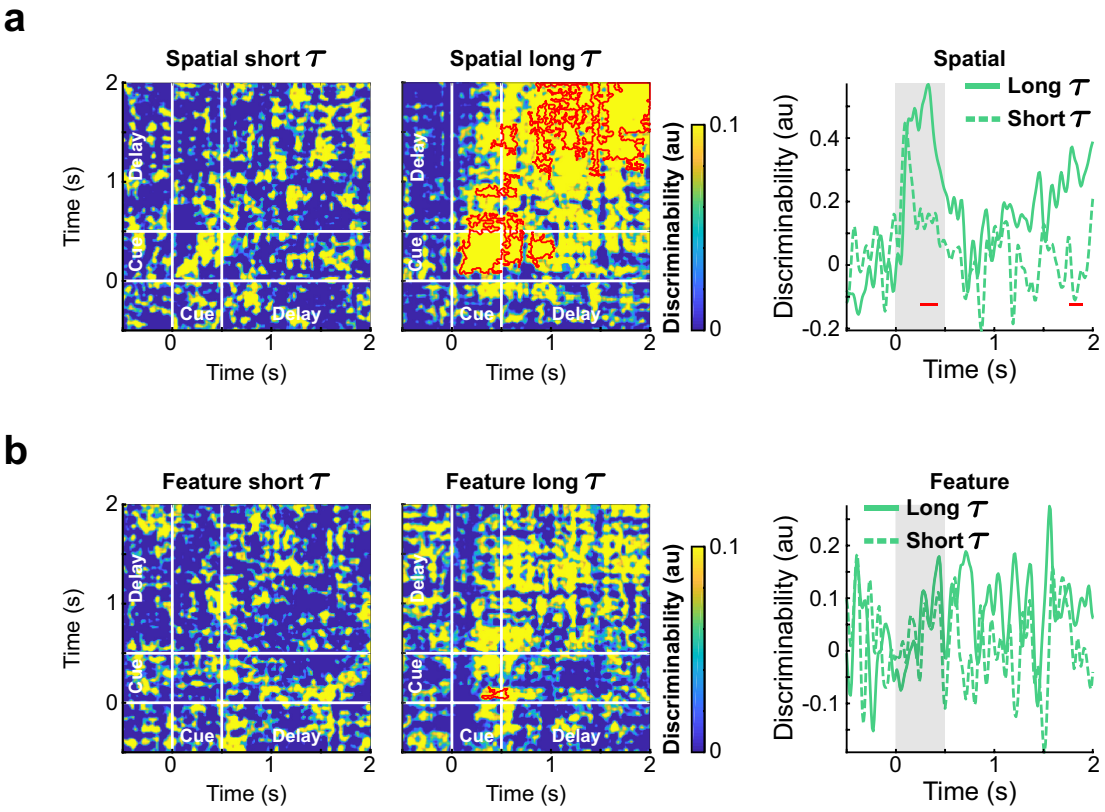


Figure 4.9: **Long τ units maintain cue stimulus information during the delay period of the spatial DMS task. a**, Cross-temporal discriminability matrices and the within-time decoding timecourses from the short and long τ groups of the dlPFC data limited to the spatial DMS task. **b**, Cross-temporal discriminability matrices and the within-time decoding timecourses from the short and long τ groups of the dlPFC data limited to the feature DMS task. Gray shading, cue stimulus window. Red contours indicate significant decodability (see Section 4.3; cluster-based permutation test, $P < 0.05$). Red lines indicate significant differences in decoding between the short and long τ groups (see Section 4.3; cluster-based permutation test, $P < 0.05$).

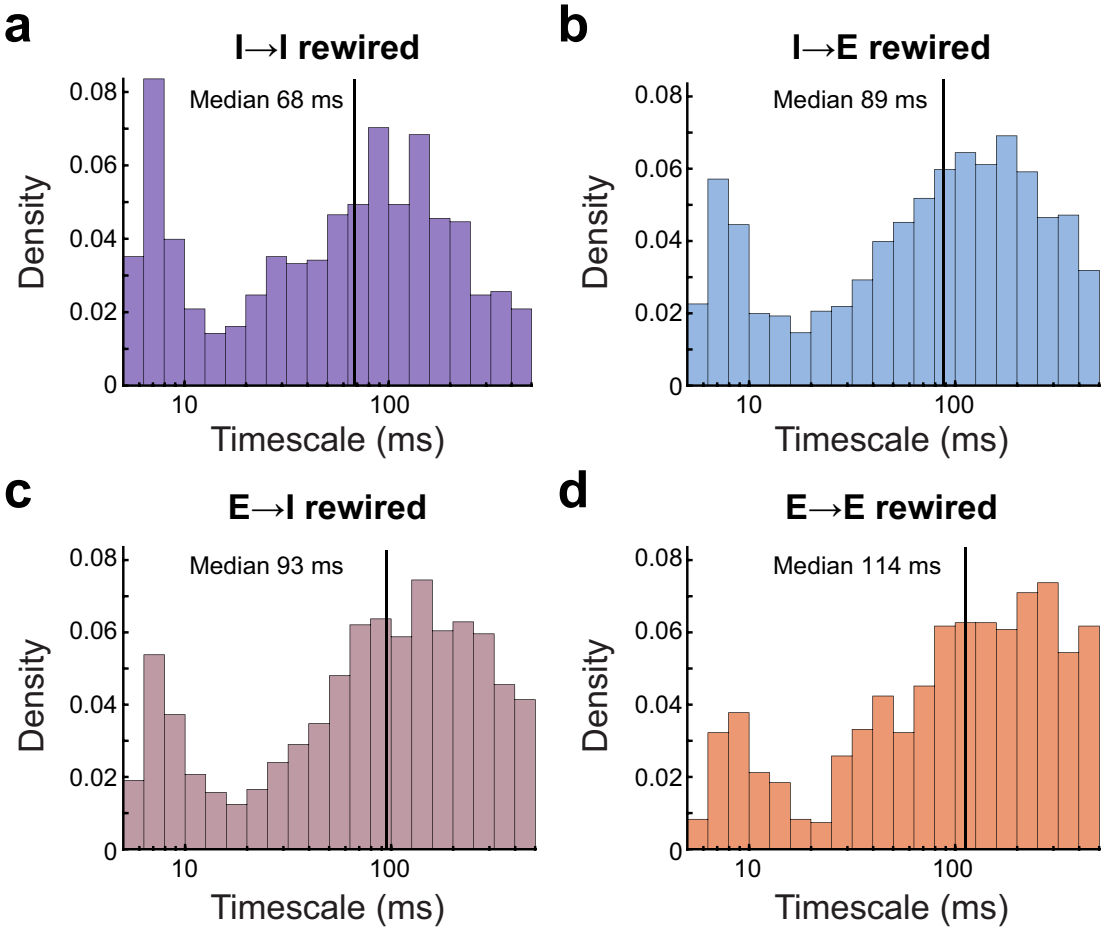


Figure 4.10: **Distribution of the timescales extracted from the DMS RNNs with rewired synaptic connections.** **a-c**, Rewiring $I \to I$ ($n = 824$; **a**), $I \to E$ ($n = 1243$; **b**), or $E \to I$ ($n = 1015$; **c**) connections shortened the neuronal timescales. **d**, Shuffling $E \to E$ ($n = 891$) connections did not alter the distribution significantly. Solid vertical lines represent median $\log(\tau)$.

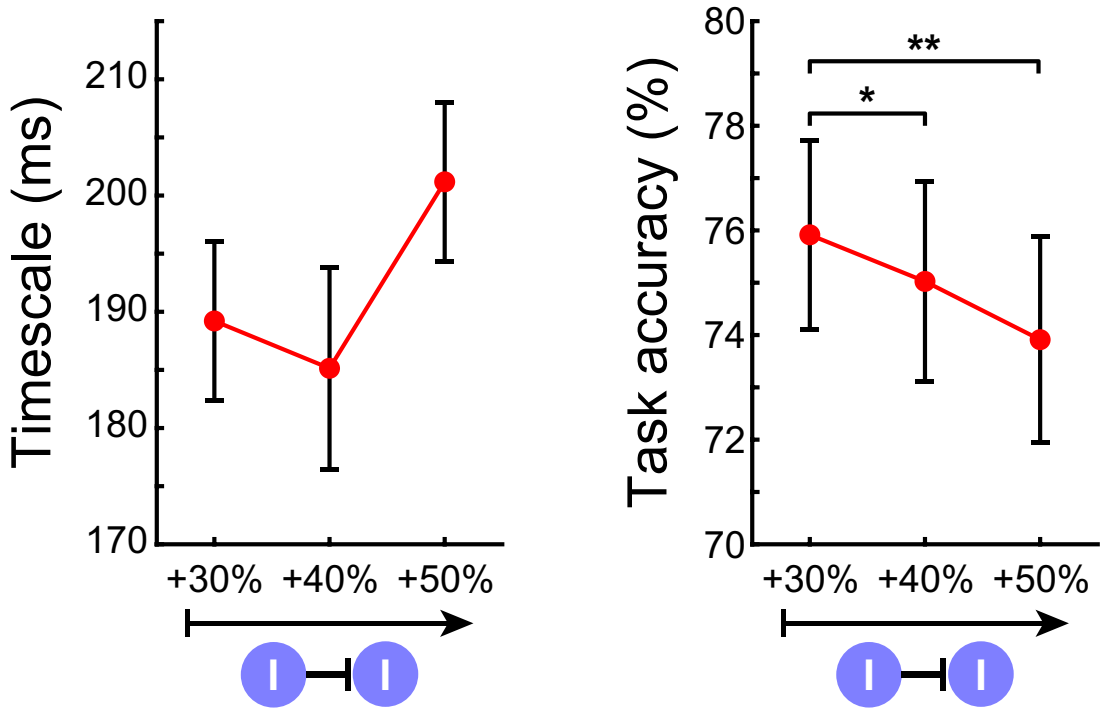


Figure 4.11: **Increasing $I \rightarrow I$ connections does not always lead to increased timescales and task performance.** Strengthening $I \rightarrow I$ connections by more than 30% did not result in significant changes in neuronal timescales (left), but significantly impaired task performance (right). Error bars, $\pm$ s.e.m. $^{*}P < 0.05$, $^{**}P < 0.005$ by two-sided Wilcoxon rank-sum test.

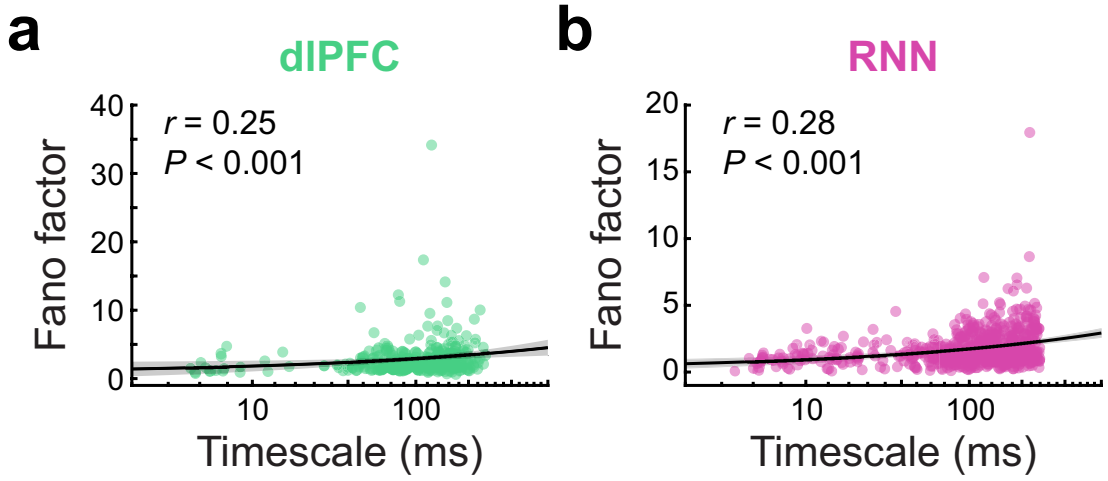


Figure 4.12: **Relationship between trial-to-trial Fano factors and neuronal timescales. a**,**b**, For both neural (**a**) and RNN data (**b**), trial-to-trial spike-count Fano factors were strongly correlated with neuronal timescales. Spearman rank coefficients are shown. Each dot represents a cell or unit. Black solid lines, linear fits to the log-transformed τ; gray shading, 95% confidence interval of the linear fits.

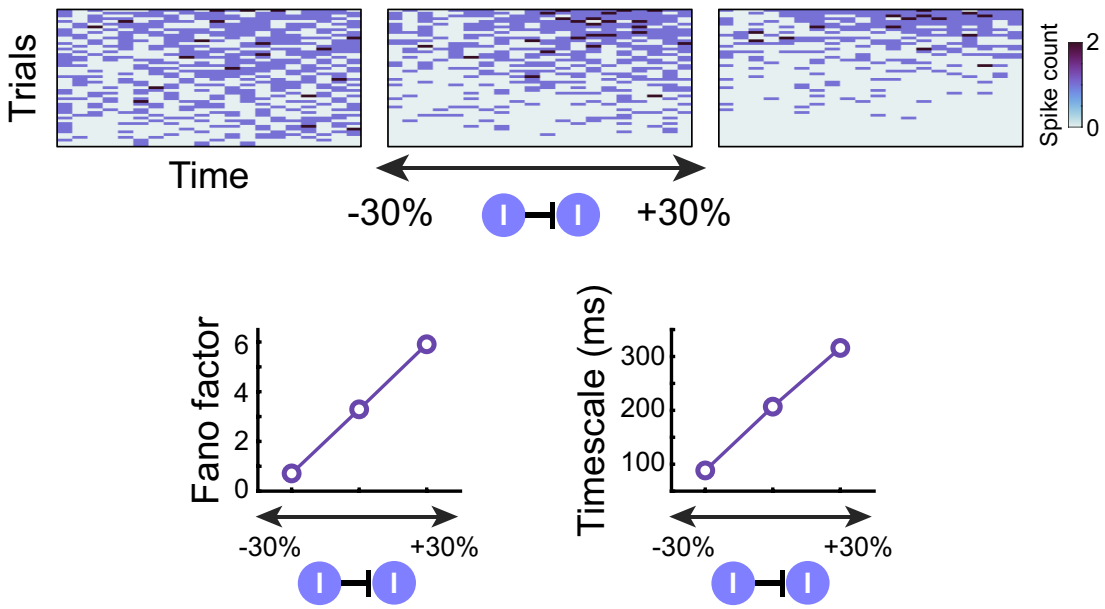


Figure 4.13: **Example unit whose spontaneous firing activity and timescale were strongly modulated by $I \rightarrow I$ strength.** An inhibitory unit from an example DMS RNN displayed highly dynamic baseline firing activity as $I \rightarrow I$ strength increased (top). The unit's Fano factor and timescale increased linearly with increasing $I \rightarrow I$ strength (bottom).

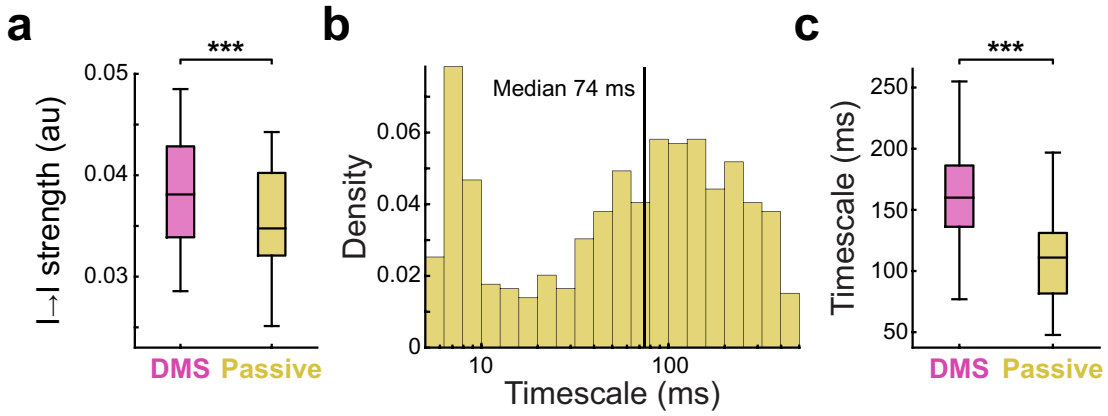


Figure 4.14: **Training DMS RNNs to perform the passive DMS task by re-tuning recurrent connections (W).** **a**, Re-tuning the recurrent connections resulted in weakened $I \rightarrow I$ connection strength. **b**, Distribution of the timescales ($n = 598$) from the re-trained DMS RNNs. Solid vertical line represents median $\log(\tau)$. **c**, In line with the decreased $I \rightarrow I$ strength, the timescales from the re-trained RNNs were significantly shorter than those extracted from the DMS RNNs. Boxplot central lines, median; red circles, mean; bottom and top edges, lower and upper quartiles; whiskers, 1.5*interquartile range; outliers not plotted. $***P < 0.0001$ by Wilcoxon signed-rank test.

Bibliography

[BB19] Flora Bouchacourt and Timothy J. Buschman. A flexible model of working memory. *Neuron*, 103(1):147–160, Jul 2019.

[BBVF+17] Renata Batista-Brito, Martin Vinck, Katie A. Ferguson, Jeremy T. Chang, David Laubender, Gyorgy Lur, James M. Mossner, Victoria G. Hernandez, Charu Ramakrishnan, Karl Deisseroth, Michael J. Higley, and Jessica A. Cardin. Developmental dysfunction of VIP interneurons impairs cortical circuits. *Neuron*, 95(4):884 – 895.e9, 2017.

[CC19] Kirstie A. Cummings and Roger L. Clem. Prefrontal somatostatin interneurons encode fear memory. *Nature Neuroscience*, 23(1):61–74, 2019.

[CKG+15] Rishidev Chaudhuri, Kenneth Knoblauch, Marie-Alice Gariel, Henry Kennedy, and Xiao-Jing Wang. A large-scale circuit mechanism for hierarchical dynamical processing in the primate cortex. *Neuron*, 88(2):419 – 431, 2015.

[CQM16] Christos Constantinidis, Xue-Lian Qi, and Travis Meyer. Single-neuron spike train recordings from macaque prefrontal cortex during a visual working memory task before and after training. *CRCNS.org*, 2016.

[CTW+18] Sean E. Cavanagh, John P. Towers, Joni D. Wallis, Laurence T. Hunt, and Steven W. Kennerley. Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. *Nature Communications*, 9(1), Aug 2018.

[CWKH16] Sean E Cavanagh, Joni D Wallis, Steven W Kennerley, and Laurence T Hunt. Autocorrelation structure at rest predicts value correlates of single neurons during reward-guided choice. *eLife*, 5:e18937, oct 2016.

[FA71] Joaquin M. Fuster and Garrett E. Alexander. Neuron activity related to short-term memory. *Science*, 173(3997):652–654, 1971.

[FTMG17] Valeria Fascianelli, Satoshi Tsujimoto, Encarni Marcos, and Aldo Genovesio. Autocorrelation Structure in the Macaque Dorsolateral, But not Orbital or Polar, Prefrontal Cortex Predicts Response-Coding Strength in a Visually Cued Strategy Task. *Cerebral Cortex*, 29(1):230–241, 2017.

[GR11] Patricia S. Goldman-Rakic. *Circuitry of Primate Prefrontal Cortex and Regulation of Behavior by Representational Memory*, pages 373–417. American Cancer Society, Atlanta, 2011.

[KD17] Tsukasa Kamigaki and Yang Dan. Delay activity of specific prefrontal interneuron subtypes modulates memory-guided behavior. *Nature Neuroscience*, 20(6):854–863, Apr 2017.

[KF14] Adam Kepecs and Gordon Fishell. Interneuron cell types are fit to function. *Nature*, 505(7483):318–326, Jan 2014.

[KJA+16] Mahesh M. Karnani, Jesse Jackson, Inbal Ayzenshtat, Azadeh Hamzehei Sichani, Kasra Manoocheri, Samuel Kim, and Rafael Yuste. Opening holes in the blanket of inhibition: Localized lateral disinhibition by VIP interneurons. *Journal of Neuroscience*, 36(12):3471–3480, 2016.

[KJL+16] Dohoung Kim, Huijeong Jeong, Juhyeong Lee, Jeong-Wook Ghim, Eun Sil Her, Seung-Hee Lee, and Min Whan Jung. Distinct roles of parvalbumin- and somatostatin-expressing interneurons in working memory. *Neuron*, 92(4):902 – 915, 2016.

[KLS19] Robert Kim, Yinghao Li, and Terrence J. Sejnowski. Simple framework for constructing functional spiking recurrent neural networks. *Proceedings of the National Academy of Sciences*, 116(45):22811–22820, 2019.

[KPdA+19] Sabine Krabbe, Enrica Paradiso, Simon d' Aquin, Yael Bitterman, Julien Courtin, Chun Xu, Keisuke Yonehara, Milica Markovic, Christian Müller, Tobias Eichlisberger, Jan Gründemann, Francesco Ferraguti, and Andreas Lüthi. Adaptive disinhibitory gating by VIP interneurons permits associative learning. *Nature Neuroscience*, 22(11):1834–1843, Oct 2019.

[Man03] Dara S. Manoach. Prefrontal cortex dysfunction during working memory performance in schizophrenia: reconciling discrepant findings. *Schizophrenia Research*, 60(2):285 – 298, 2003.

[MBF+14] John D. Murray, Alberto Bernacchia, David J. Freedman, Ranulfo Romo, Jonathan D. Wallis, Xinying Cai, Camillo Padoa-Schioppa, Tatiana Pasternak, Hyojung Seo, Daeyeol Lee, and Xiao-Jing Wang. A hierarchy of intrinsic timescales across primate cortex. *Nature Neuroscience*, 17(12):1661–1663, Nov 2014.

[MC01] Earl K. Miller and Jonathan D. Cohen. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24(1):167–202, 2001.

[MED96] Earl K. Miller, Cynthia A. Erickson, and Robert Desimone. Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience*, 16(16):5154–5167, 1996.

[MGWL17] Maria Medalla, Joshua P. Gilman, Jing-Yi Wang, and Jennifer I. Luebke. Strength and diversity of inhibitory signaling differentiates primate anterior cingulate from lateral prefrontal cortex. *Journal of Neuroscience*, 37(18):4717–4734, 2017.

[Mic17] Thomas Miconi. Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *eLife*, 6:e20899, feb 2017.

[MM20] Nagaraj R. Mahajan and Shreesh P. Mysore. Neural circuit mechanism for generating categorical representations. Preprint at https://www.biorxiv.org/content/10.1101/2019.12.24.887810v2. 2020.

[MO07] Eric Maris and Robert Oostenveld. Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164(1):177 – 190, 2007.

[MQC07] Travis Meyer, Xue-Lian Qi, and Christos Constantinidis. Persistent discharges in the prefrontal cortex of monkeys naïve to working memory tasks. *Cerebral Cortex*, 17:i70–i76, 2007.

[MQSC11] Travis Meyer, Xue-Lian Qi, Terrence R. Stanford, and Christos Constantinidis. Stimulus selectivity in dorsal and ventral prefrontal cortex after training in working memory tasks. *Journal of Neuroscience*, 31(17):6266–6276, 2011.

[MRL18] Gianluigi Mongillo, Simon Rumpel, and Yonatan Loewenstein. Inhibitory connectivity defines the realm of excitatory plasticity. *Nature Neuroscience*, 21(10):1463–1470, Sep 2018.

[MSSN13] Valerio Mante, David Sussillo, Krishna V. Shenoy, and William T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, Nov 2013.

[OM19] A. Emin Orhan and Wei Ji Ma. A diverse range of factors affect the nature of neural representations underlying short-term memory. *Nature Neuroscience*, 22(2):275–283, Jan 2019.

[PHK+13] Hyun-Jae Pi, Balázs Hangya, Duda Kvitsiani, Joshua I. Sanders, Z. Josh Huang, and Adam Kepecs. Cortical interneurons that specialize in disinhibitory control. *Nature*, 503(7477):521–524, Oct 2013.

[PXH+13] Carsten K. Pfeffer, Mingshan Xue, Miao He, Z Josh Huang, and Massimo Scanziani. Inhibition of inhibition in visual cortex: the logic of connections between molecularly distinct interneurons. *Nature Neuroscience*, 16(8):1068–1076, Jun 2013.

[QMSC11] Xue-Lian Qi, Travis Meyer, Terrence R. Stanford, and Christos Constantinidis. Changes in prefrontal neuronal activity after learning to perform a spatial working memory task. *Cerebral Cortex*, 21(12):2722–2732, 04 2011.

[SKS+13] Mark G. Stokes, Makoto Kusunoki, Natasha Sigala, Hamed Nili, David Gaffan, and John Duncan. Dynamic coding for cognitive control in prefrontal cortex. *Neuron*, 78(2):364–375, Apr 2013.

[SWFS17] Eelke Spaak, Kei Watanabe, Shintaro Funahashi, and Mark G. Stokes. Stable and dynamic coding for working memory in primate prefrontal cortex. *Journal of Neuroscience*, 37(27):6503–6516, 2017.

[SYW16] H. Francis Song, Guangyu R. Yang, and Xiao-Jing Wang. Training excitatory-inhibitory recurrent neural networks for cognitive tasks: A simple and flexible framework. *PLOS Computational Biology*, 12(2):e1004792, 2016.

[TLR16] Robin Tremblay, Soohyun Lee, and Bernardo Rudy. GABAergic interneurons in the neocortex: From cellular properties to circuits. *Neuron*, 91(2):260–292, Jul 2016.

[VGH+16] Jared X. Van Snellenberg, Ragy R. Girgis, Guillermo Horga, Elsmarieke van de Giessen, Mark Slifstein, Najate Ojeil, Jodi J. Weinstein, Holly Moore, Jeffrey A. Lieberman, Daphna Shohamy, Edward E. Smith, and Anissa Abi-Dargham. Mechanisms of working memory impairment in schizophrenia. *Biological Psychiatry*, 80(8):617 – 626, 2016. Schizophrenia Biomarkers.

[WSB+18] D. F. Wasmuht, E. Spaak, T. J. Buschman, E. K. Miller, and M. G. Stokes. Intrinsic neuronal dynamics predict distinct functional roles during working memory. *Nature Communications*, 9:3499, Aug 2018.

[XLT+19] Haifeng Xu, Ling Liu, Yuanyuan Tian, Jun Wang, Jie Li, Junqiang Zheng, Hongfei Zhao, Miao He, Tian-Le Xu, Shumin Duan, and Han Xu. A disinhibitory microcircuit mediates conditioned social fear in the prefrontal cortex. *Neuron*, 102(3):668–682, 2019.

[YJS+19] Guangyu Robert Yang, Madhura R. Joglekar, H. Francis Song, William T. Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, 22(2):297–306, Jan 2019.

[YMW16] Guangyu Robert Yang, John D. Murray, and Xiao-Jing Wang. A dendritic disinhibitory circuit mechanism for pathway-specific gating. *Nature Communications*, 7(1), Sep 2016.

Chapter 5

Conclusions

In this dissertation, I presented two methods for investigating local and large-scale dynamics linked to higher cognitive functions. The first method, which is built on delay differential analysis (DDA), was used to characterize and identify brain signals with similar nonlinear large-scale dynamics (Chapter 2). Applying the method to a large dataset consisting of brain signals from both non-psychiatric comparison and schizophrenia participants revealed discrete subgroups whose nonlinear dynamics were indicative of information processing and cognitive functioning. The second method revolves around constructing spiking recurrent neural networks designed to perform cognitive tasks commonly studied in neuroscience (Chapter 3). The utility of the method was explored extensively in Chapter 4, where I showed how trained networks employed inhibitory-to-inhibitory signaling to maintain information transiently. These findings, along with previous experimental findings, suggest that (1) inhibitory interneurons from diverse classes play an integral role in working memory maintenance, and (2) dysfunction of such inhibitory neurons could lead to working memory impairment and other cognitive deficits seen in schizophrenia. Bridging and establishing a relationship between micro- and macro-scale brain dynamics is a natural next step to understand how multiple interacting local circuits translate to distinct large-scale dynamics.