The neural computation of internal affective states

Thesis by
Aditya Nair

Caltech

The neural computation of internal affective states

Thesis by

Aditya Nair

In Partial Fulfillment of the Requirements for
the Degree of
Computation and Neural Systems

CALIFORNIA INSTITUTE OF TECHNOLOGY
Pasadena, California

2025
Defended September 12th, 2024

Aditya Nair
ORCID: 0000-0001-5242-5527

Cover illustration by Shaleena Nair.

DEDICATION

To my Acha and Amma.

For creating the river that eventually led me to flow to this point even when the incline all around them was uphill.

ACKNOWLEDGEMENTS

Growing up in Kerala, I had photos of scientific institutions on my desk as a young child that I would gaze at before going to bed. I often wondered what it would be like to be a scientist at one of those places, akin to the ones I read about with fervor. If you had told that younger version of me that I would one day be at Caltech, exploring fascinating questions about the brain, he'd probably have told you to keep dreaming. But that dream became a reality, made possible only due to the immense generosity of many individuals who, to my deep gratitude, saw something in me that I often didn't.

Foremost among those individuals is my advisor, David Anderson, who saw beyond my naivety when I first told him that I wanted to explore a computational angle on understanding social behavior. He challenged my ideas and molded me every step of the way. I will never forget the exhilarating, several-hour-long meetings we had, often multiple times a week, where we pondered what it all means. In David, I witnessed the importance of an unbridled and unrelenting passion for science. His capacity to keep absorbing and learning is awe-inspiring. I hope to one day emulate even a tenth of that as I continue looking to him to shape me into a better scientist.

I am deeply indebted to my computational advisors, Scott Linderman and Ann Kennedy, for guiding me through my graduate school journey. Ann is an amazing teacher who helped me translate ideas into practical methods. The seeds of most ideas in this thesis were born through our conversations, for which I am incredibly grateful. Scott gave me the confidence I needed to make it through my entire PhD. He introduced me to the universe of dynamical systems and always had a solution for every methodological problem we faced. His intuition for identifying the important questions and his lightning-like insight in explaining data that left me scratching my head for days are qualities I hope to possess one day.

My thesis committee Pietro Perona, Ueli Rutishauser and Ralph Adolphs have been a source of immense support. Pietro forced me to think beyond the methods I was familiar with, while Ueli's lessons in the power of single neuron computation resulted in many of the key ideas

in the later portion of this thesis. Ralph grounded this work in every conversation and gave me the confidence to dream about ambitious ideas for the future.

Early on in my graduate journey, I was extremely fortunate to have Mark Schnitzer and Surya Ganguli as mentors. I still remember our earliest meetings, where David would remind Mark "don't go easy on him," and I am exceedingly grateful that they both didn't. That early teaching experience left its mark at every junction of my graduate career. Surya was the first to introduce me to the word "attractor," a concept you will see repeated perhaps too many times in this thesis. His instantaneous insight was like watching a god in action, and I am incredibly grateful to have had that experience.

Nearly everything I've done in science is owed to my scientific fathers: Dr. Lim Kah Leong and Dr. George Augustine. Dr. Lim introduced me to the entire field of neuroscience. Early on, he set the standard for the importance of passion and compassion in science. He believed in me during a period when I certainly did not, and it was his unwavering insistence that led me to apply to Caltech. Words can hardly describe the love and respect I have for him, and this thesis is only possible because of the seeds he planted in me about 10 years ago.

Prof. George introduced me to the world of neural circuits and computational modeling. He treated me like a graduate student and gave me the freedom to explore science in ways that undergraduates are rarely given. And, in doing so, taught me the importance of being ambitious and working unrelentingly toward your goals. His lab was a data heaven, and I reveled in the opportunity to discover new things at every turn. Most of all, he taught me humility, and I will never forget his lessons.

The most enjoyable part of this thesis was that none of this work was done alone—it was the result of the collaborative efforts of my dearest friends in the Anderson lab. Brady was our lab elder, who often graced me with his blunt feedback: "I don't believe any of this," which forced me to explore my work even more deeply until he finally did believe it. Amit pulled me out of my mid-grad school slump with his excitement and sage-like wisdom, becoming my Brady after the original left for MIT. Our paper together was truly an adventure of a

lifetime. I still remember the moment where we witnessed neurons in VMHvl integrate for the first time. It was one of those movie moments of pure magic that reminds you why you decided to go into science. Mengyu showed me what perseverance really looks like. She navigated every punch thrown at our paper with the skill of an expert boxer, delivering a swift finishing blow at the last revision with her heroic experiments. George was a ronin, showing me the importance of being critical of your own data. His ability to step back, not get lost in the excitement of positive results, and deeply question everything to reveal the underlying truth is something I admire immensely. Jineun is the very definition of efficiency. Our short time working together is only juxtaposed with the mountain of skills that I have learnt from her.

Towards the end of this thesis, I finally took an experimental turn thanks to Jineun, Amit and Lindsey who spent so much time teaching me the importance of every step. I would also like to thank entire Anderson lab, especially my fellow students, Kathy, Shuo, Elena, Charles, and Lexi for their company and current and former lab members including Tomomi, Bin, Joe, Stefanos, Emma, Takumi, Kichi, and Vivian for insights at every lab meeting. Liliana and Celine held the fort down in our lab and all the work in this thesis is built upon their kindness and generosity.

One of the most rewarding parts of this journey has been mentoring undergraduate students on a variety of machine learning projects. I want to thank Nestor, Rohan, Jadon, Isa, Dexter, and Angel for trusting me as a mentor to guide them on making all the cool new tools that I hope will be out soon.

The adoption of the ideas in this thesis will depend upon the accessibility of tools and methods discussed. I'm immensely grateful to David, Ralph, Pietro, Mary Sikora, and Helen O'Connor for providing me a platform to enable this through the Chen Data Science and AI for Neuroscience Summer School (ChenDataSAI). Organizing this workshop was a crash course in teaching and pedagogy and I'm excited for its future, one that Justin Bois and I have been envisioning together.

I grew up in a time in India when pursuing a career in science was often seen as throwing your future away. But I was incredibly fortunate to have parents who broke the mold on every cliché. They did everything to ensure my sister and I were happy, often at the cost of their own. They encouraged us to do our best and instilled in us the values of hard work that made this thesis possible. I dedicate this thesis to them.

My sister has been my rock and scaffold that has kept me steady when all else fell apart. Our bond is something beyond words and as I've grown older, I've come to realize how rare and special it truly is. Her resolve and strength have been the inspiration that has enabled me to complete this thesis. My little niece and nephew Aeva and Aru became my world and along with my brother-in-law Vishnu, filled my life with endless joy.

My partner's parents, John and Chris, became my home away from home, always going the extra mile to warmly weave Malayali snacks and culture into their lives. I'm deeply grateful to them for embracing me as their son, celebrating every little milestone with pride, alongside my new brother and sister, Stephen and Elise. They've given me a sense of belonging and love that many can only dream of on their immigrant journey.

This thesis would not exist without my dearest friends, Pradeesh, Shaun, Peter, Sachin and Shinoj, who have been my lifeline when life gets too difficult.

Lastly, all my love and gratefulness go to my partner, Julia. There were many times early on during this thesis when I wondered if I should have ever left Singapore. Meeting her has shown me otherwise and has been the single most important thing that happened during this thesis. Her boundless creativity inspires me, and her love, patience, and care have given my life its deepest meaning. I can't wait for the journeys that lie ahead for us.

Adi, August 6th, 2024.

ABSTRACT

The study of neural computation has long concentrated on our cognitive abilities, with extensive research dissecting the mechanisms of memory, decision-making, and navigation. In contrast, the realm of social innate behavior and emotion has often been treated as a simpler problem, overlooking the immense complexity and biological significance it entails. This thesis aims to bring neural computation into the domain of emotional or affective states, employing data-driven modeling methods that approximate neural activity as dynamical systems. The application of these methods has uncovered brain representations that encode key qualities of persistence and escalation associated with aggressive states, formalized as line attractors. These emergent features of neural circuits arise from the complex interplay of connectivity and network dynamics, challenging long-held notions of subcortical computation. This discovery led us to rigorously test various key properties of line attractor dynamics. Through closed-loop modeling and holographic neural activation, we demonstrate that the line attractor is intrinsic to the mammalian hypothalamus, providing some of the first causal evidence of this property for any continuous attractor. These experiments also suggest that functional connectivity within the hypothalamus underpins the stability of this attractor. Furthermore, using a new cell-type-specific gene-editing system, we show that the implementation of this line attractor depends on neuropeptides, indicating a non-canonical mechanism that contributes to the robustness of this innate attractor. Finally, we reveal that line attractors encode emotional states beyond aggression, including states of sexual receptivity in the female hypothalamus. Longitudinal recordings of neural data across the estrus cycle show that the line attractor disappears during non-estrus states, suggesting long-timescale modulation of attractor dynamics by hormones. Together, these studies present a new paradigm for understanding subcortical computation underlying internal states and suggest a canonical motif that the brain reuses to encode diverse internal affective states.

PUBLISHED CONTENT AND CONTRIBUTIONS

Aditya Nair, Tomomi Karigo, Bin Yang, Surya Ganguli, Mark J. Schnitzer, Scott W. Linderman, David J. Anderson†, and Ann Kennedy†. An approximate line attractor in the hypothalamus encodes an aggressive state. *Cell* 186, no. 1 (2023): 178-193.
DOI: https://doi.org/10.1016/j.cell.2022.11.027

A.N participated in conceptualizing the study, performed all computational analysis and dynamical system modelling, and co-wrote the paper.

Amit Vinograd*, **Aditya Nair***, Joseph Kim, Scott W. Linderman and David J. Anderson†. Causal evidence of a line attractor encoding an affective state. *Nature* (2024).
DOI: https://doi.org/10.1038/s41586-024-07915-x

A.N participated in conceptualizing the study, performed experiments, all computational analysis and dynamical system modelling and co-wrote the paper.

Mengyu Liu*, **Aditya Nair***, Nestor Coria, Scott W. Linderman and David J. Anderson†. Encoding of female mating dynamics by a hypothalamic line attractor. *Nature* (2024).
DOI: https://doi.org/10.1038/s41586-024-07916-w

A.N participated in conceptualizing the study, performed all computational analysis and dynamical system modelling, and co-wrote the paper.

George Mountoufaris, **Aditya Nair**, Bin Yang, Dong-Wook Kim, Samuel Kim, Scott W. Linderman David J. Anderson†. A line attractor encoding a persistent internal state requires neuropeptide signaling. *Cell* (2024).
DOI: https://doi.org/10.1016/j.cell.2024.08.015

A.N participated in conceptualizing the study, performed computational analysis and all dynamical systems modelling and co-wrote the paper.

*co-first author publications

Ambu Hu, David Zoltowski, **Aditya Nair**, David J. Anderson, Lea Ducker and Scott Linderman. Modeling latent neural dynamics with gaussian process switching linear dynamical systems. *Advances in Neural Information Processing Systems 38 (2024).*
DOI: https://doi.org/10.48550/arXiv.2408.03330

A.N contributed data and computational analysis for this study.

Brandon Weissbourd†, Tsuyoshi Momose, **Aditya Nair**, Ann Kennedy, Bridgett Hunt, and David J. Anderson†. A genetically tractable jellyfish model for systems and evolutionary neuroscience. *Cell* 184, no. 24 (2021): 5854-5868.
DOI: https://doi.org/10.1016/j.cell.2021.10.021

A.N performed computational analysis and preliminary electrophysiology experiments.

Ana Badimon*, Hayley J. Strasburger*, Pinar Ayata*, Xinhong Chen, **Aditya Nair**, Ako Ikegami, Philip Hwang et al. Negative feedback control of neuronal activity by microglia. *Nature* 586, no. 7829 (2020): 417-423.
DOI: https://doi.org/10.1038/s41586-020-2777-8

A.N performed computational analysis and network modeling for this study.

*co-first author publications

TABLE OF CONTENTS

PROLOGUE

An organizational note

In his book Vision, David Marr articulates a framework for understanding the brain across multiple levels: the computational, the algorithmic, and the implementational. These concepts have been widely applied to illuminate the mysteries of perception, navigation, and cognition. Yet, they have seldom been turned toward a realm of our daily existence that resonates deeply within us and shapes our very being: our emotions.

In this thesis, I introduce a novel framework for understanding the neural computation of emotional (affective) states through the lens of emergent attractor dynamics. The thesis unfolds across several chapters, each delving into a different facet of this computation:

In "Representation," I explore how data-driven machine learning methods can unearth computations at a representational level in an unsupervised manner. This journey led to the discovery of a line attractor deep within the mammalian hypothalamus, encoding an aggressive state and challenging long-held beliefs about hypothalamic function.

In "Perturbation," I detail efforts to gather causal evidence for this line attractor. Here, I present the first evidence of the existence of an intrinsic line attractor in mammals, utilizing closed-loop modeling of neural data and single-cell holographic activation of single neurons.

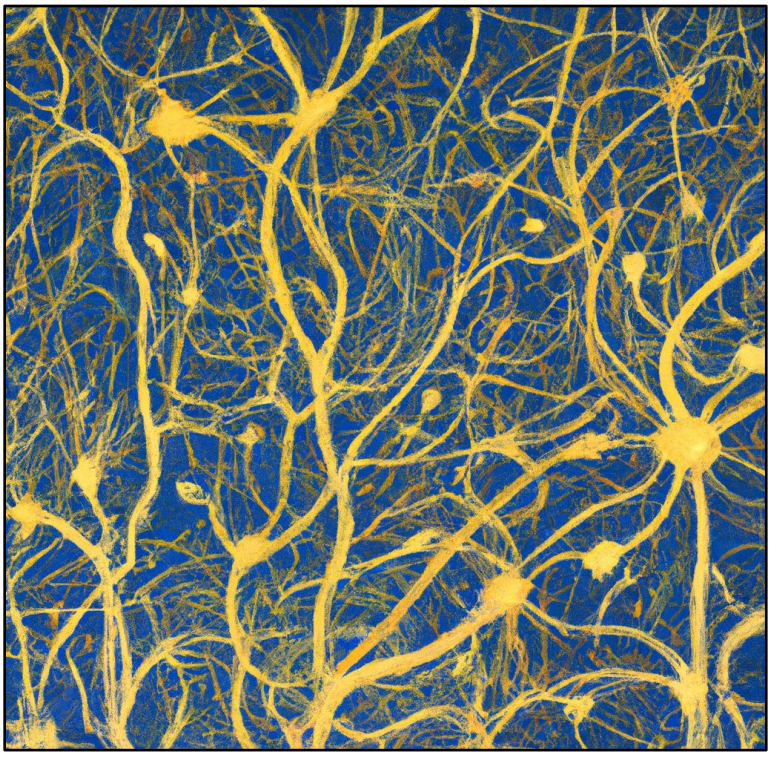
In "Implementation," I reveal how this computation is realized in a non-canonical form, relying on neural communication via neuropeptides to achieve the long timescales that set this attractor apart from others in the cortex.

Finally, in "Generalization," I demonstrate that line attractor dynamics extend beyond encoding aggression, revealing their role in encoding states of sexual arousal with remarkably similar properties.

Together, these studies propose a new paradigm for understanding the computation of our emotions, offering a blueprint that may extend to other affective states as well.

Chapter I

INTRODUCTION



"എത്ര കണ്ടാലും മതി, കാണാതെപോകുന്ന, കാഴ്ചയുടെ,
കാഴ്ചയെക്കുറിച്ചൊരുറ്റുമില്ല."

Vallathol Narayana Menon, Sahityamanjari, 1917

Translation: "No matter how much one sees, it is never enough, for there is always something about the sight that remains unseen."

Chapter 1

The neural coding of affective states:

a dynamical systems perspective

This chapter provides a synthesis of all studies in this thesis and provides a comprehensive overview of concepts in dynamical systems applied to neuroscience.

Summary

Our experience of the world is constantly shaped by the motivational and emotional drives which comprise our internal state. An emerging view suggests that such states are encoded by the coordinated dynamics of populations of neurons deep in the subcortex. In this chapter, I introduce a dynamical system framework to understand internal states and bring together evidence across several studies that key features of internal emotional (affective states) such as aggression and sexual drives are represented by emergent continuous attractor dynamics in the hypothalamus. These dynamics are discovered using data-driven modeling techniques that approximate neural data as dynamical systems and discover computations relevant to internal states in an unsupervised manner. The mechanistic implementation of these dynamics challenges dominant assumptions of subcortical computation and presents a new avenue to study the emergence of slow state-encoding dynamics across scales; from single neurons to network interactions and neuromodulation.

Introduction

A fundamental goal in neuroscience is to understand how animals use sensory cues to produce behavioral outputs. Essential to this process is the concept of the internal state, a low-dimensional representation of the organism's drive, motivation or affect, which can strongly influence how perceived sensory signals are transformed into behavioral decisions[1,2]. For example, internal states can shape an organism's intent and drive to display motivated behaviors: an animal's state of aggression can influence the extent of attack behavior towards a conspecific in a resident-intruder assay[2], or its state of thirst can modulate the extent of licking behavior in animals engaged in operant conditioning tasks[3]. State-dependent behaviors provide adaptability, allowing organisms to flexibly modify their actions based on past experiences and current needs[1]. The past decade has seen a surge in studies that identify neural circuits involved in a variety of innate goal-directed behaviors, including those driven by physiological homeostasis, such as hunger[4,5] and thirst[3,6,7], and affective states such as aggression[8-11], mating[1,2,12], and predator defense[1,13,14]. The time is ripe to begin to investigate whether and how these circuits give rise to emergent neural population dynamics and the role of these dynamics in states and behavior coding.

Determining how and where the internal states underlying these behaviors are encoded in the brain is a long-standing challenge, particularly in the case of affective internal states[15,16]. Targeted neural perturbations have been key to revealing the role of neural circuits in the control of internal states[10,12,17,18]. This perturbation-first strategy has revealed that specific cell types within a distributed hypothalamic network, including the VMHvl[10,11], MeA[19,20], and BNST[9,17,18], are essential for the regulation of an aggressive state. However, such perturbations in different brain regions can yield similar behavioral phenotypes which obscure important differences in neural computations performed by that region or information represented in neural activity. For example, loss-of-function manipulations that block a behavior at a given node in a pathway that transforms a sensory stimulus into a response do not distinguish whether that node plays an instructive or merely a permissive role. Furthermore, such manipulations do not distinguish whether the node encodes stimulus

identity, stimulus associations, motor program or internal state. While gain-of-function manipulations can distinguish whether loss-of-function phenotypes reflect permissive or instructive functions, they do not resolve issues of neural computation and function. For example, the evocation of a given behavior by stimulating a particular node could reflect the activation of a motor program, an internal drive state or the presentation of a fictive sensory stimulus. Solving these challenges require studies that focus on neural representation to disentangle these functions.

However, such studies of representation are hindered by the problem of how internal states can be identified. Many studies determine the presence of an internal state based on observable behavior, often using subjective criteria[1]. For instance, a state of aggressiveness is typically inferred by quantifying the duration and number of attack bouts using a resident--intruder assay[8,9]. While such overt behavior may express an internal state, this behavior-driven approach makes it difficult to distinguish the neural encoding of an internal state from that of the behavior that expresses it.

Observational studies of internal affective states are particularly complicated by the challenge of recording neural activity in freely moving animals as they engage in innate behaviors. Studies focusing on neural representation typically rely on animals that engage in highly trained behaviors where neural data can be collected from hundreds of trials of a stereotyped task[21,22]. Methods developed for analyzing data from highly trained animals do not directly apply to single-trial conditions in naturalistic innate behaviors, where each encounter with a conspecific, for instance, might result in a different distribution of behaviors. Therefore, new methods to gain insights from such observational data are paramount for studying neural representation in internal affective states.

In this perspective, we propose an alternative approach to identify and study affective internal states through unsupervised analysis of neural data from naturalistic behavior, based on dynamical systems analysis. We show how studying the dynamics of neural activity can serve as a "hypothesis generator," uncovering features at the "manifold level" of analysis[23-

[25], such as attractors[26,27], that may represent or encode key internal state properties such as persistence and scalability[16,28]. Recent results from the application of this neural dynamics approach in turn suggest a generalized conceptual framework for understanding the instantiation and implementation of internal affective states across different behaviors. Moreover, this approach emphasizes an integrated computational process that spans multiple levels—from individual neurons to network mechanisms and neuromodulation—providing an opportunity to achieve explanations of brain function that span and link different biological scales and levels of abstraction.

Paradoxes in the neural representation of innate behavior

When incorporating neural activity into the study of naturalistic behavior, bulk measurements such as fiber photometry have been widely utilized to uncover behavior-related activity in specific brain regions[29]. However, this low-dimensional approach can obscure the heterogeneous activity of individual neurons or physiologically distinct neuronal subpopulations, potentially resulting in inaccurate interpretations of a brain region's role in encoding a state. For instance, while bulk measurements[18] in $BNST^{pr}$ indicated an encoding of conspecific sex (male or female) based on the overall level of activity (low=male; high=female), single-cell imaging data[17] have revealed distinct populations of male- vs. female-tuned neurons in a female-biased ratio. Analysis of neural population activity, in conjunction with cell-type specific perturbations, has implicated this region in regulating the transition between appetitive and consummatory behaviors, thereby extending the interpretation beyond the identification of conspecific sex[17].

In addition, single-cell measurements can uncover significant paradoxes that challenge the straightforward relationship between specific regions and their roles in initiating a behavior, as indicated by experiments based on perturbations. For instance, activation of $VMHvl^{esr1}$ neurons can reliably trigger aggressive behavior through optogenetic stimulation[11], yet single-cell imaging consistently shows that only a small number of neurons are specifically tuned and time-locked to attack[10,12]. Most neurons in this nucleus and others such as the BLA

exhibit mixed selectivity, a feature prominently seen in studies of the cortex where neurons encode for multiple variables in a multiplexed fashion. Furthermore, MPOA^{Esr1} neurons which alone are neither necessary nor sufficient to induce aggression, contain a much higher proportion of aggression-tuned neurons than VMHvl[12]. Addressing these paradoxes motivates further exploration of the information processed at the level of population activity and dynamics.

The need to incorporate dynamics in the study of internal states.

Neural circuits can encode information at a population level using various strategies. They may utilize a cell-identity code, where separable populations encode different features, or might use a distributed code, where overlapping neurons encode information in a multiplexed manner. Studies of neural representations in innate behavior-contributing neural circuit nodes have unveiled many examples of cell-identity codes. For instance, in circuits such as VMHvl and BNST, separable populations of neurons are activated in response to the sex of the intruder[17,30]. Similarly, in VMHdm, separable subsets of neurons are activated by threatening stimuli of different modality such as odor versus sound[13]. However, it has been challenging to infer whether these cell-identity codes in state-contributing circuits encode a representation of a stimulus, or of an internal state evoked by the stimuli. For instance, male-activated neurons in VMHvl might encode a representation of intruder sex, or a state of aggression that is associated with the presence of a male-intruder. These alternatives are difficult to disentangle because male mice only exhibit naturalistic aggression towards male conspecifics.

Studies examining the neural representation of internal states have also discovered distributed population codes in different organisms. Research in mice, fish, and worms has used supervised methods such as linear decoders to identify distributed population signals that encode opposing states, such as exploration and exploitation or defensive behaviorr[31-33]. Similar distributed signals have also been uncovered in brain-wide recordings of thirst states[7] and stress states[34,35]. However, these supervised population coding approaches require

defining an axis to classify either different types of states, different intensities of states or use reinforcement-learning approaches to define state variables. This limits their use in affective internal states such as aggression and mating, where it is challenging to define such a state axis during naturalistic behavior, in part due to the trial-trial variability notable in these states.

Methods used to study population coding, which include matrix factorization, linear decoders, choice-probability and generalized linear models (GLMs), can identify neural populations that are "tuned" to particular stimuli or behavioral features. However, by themselves they do not distinguish the encoding of behavior from the encoding of an underlying internal state. The reason is that an essential feature that distinguishes internal states from behavior is their *dynamics*. Behaviors are punctate: they are typically expressed in bouts, with defined onsets and offsets. In contrast, internal motive states are often of longer duration than the behavioral bouts that express them. In other words, they reflect persistent activity that is maintained across overt, observable episodes of behavior[16]. Therefore, if such patterns of persistent activity can be identified in a population of neurons during an otherwise episodic behavior, they are candidates for encoding internal states – they make something visible (patent) that is otherwise invisible (latent) at the level of behavior. In order to discover such features in neural activity, it is necessary to analyze the dynamics of the system. One way to do that is to model the neural data as a dynamical system.

What is a dynamical system, and why is it a useful way to approach the neural coding of internal states? A dynamical system is a system that displays a particular pattern of time-evolving activity, for example in response to a stimulus[36]. This pattern of temporal dynamics reflects the intrinsic properties of the system, which can include those of its components, and their connectivity[36]. A dynamical system can be described by a series of equations, which essentially define how quickly activity changes, and in what direction, at different points in a "state space" a coordinate system that encapsulates the dimensions along which activity evolves over time. Importantly, the exact pattern of activity in a dynamical system can be highly variable and dependent on "initial conditions" (the set of inputs and the state of the system at the beginning of a particular time epoch). Yet the equations for the system will be

able to predict the direction in which activity evolves, for a given set of initial conditions (see Box 1). In other words, it is a way of making something that is complicated – high-dimensional, constantly changing and variable from trial to trial -- simpler to visualize and understand. It is ideally suited for the analysis of naturalistic behaviors, which typically evolve in a unique sequence every time they are performed.

Unsupervised data-driven discovery of population dynamics

Population dynamics can emerge at multiple levels in a neural circuit: at the input to a circuit, through intrinsic dynamics as a feature of the interconnectivity between neurons, or even within single neurons. Complex dynamics may also emerge through meso-scale interactions between nuclei or brain regions[37]. In systems characterized by strong recurrence, emergent intrinsic dynamics resulting from the interconnectivity between regions and between neurons significantly influence neural dynamics. The framework of dynamical systems theory provides a conceptual foundation in engineering and physical sciences for understanding how feedback influences ongoing dynamics in physical systems[36]. This approach has recently been instrumental in elucidating emergent neural dynamics[38,39]. Within this framework, neural populations conceptualized seen as a dynamical system[38,40]. Such systems, through their temporal evolution—integrating a wide range of inputs via intrinsic dynamics—facilitate computations crucial for generating movement, making decisions, and mapping space in the brain[38]. They also afford behavioral flexibility[41].

Dynamical systems are naturally suited to display features such as persistence and escalation or ramping in activity and have been hypothesized as candidates that encode features of internal states[30]. Studying the properties of a dynamical system can uncover important emergent features of neural circuits (refer to Box 1). This is performed by applying dynamical systems analysis to neural population activity. Here, dynamical models of the type illustrated in Equation 1 (Box 1, Figure 1A) are directly fit to the activity of a neuron population. One such emergent feature is attractor dynamics, a component of interconnected neural circuits that allows information to be maintained in persistent neural activity[27,39,42,43].

While theoretical studies have long implicated the role of attractor dynamics in encoding variables such as memory[44,45], spatial locations[46] or eye position[39], recent experimental studies have discovered attractor-like representations through the application of dynamical systems models to neural data[25-27]. For example, a recent influential study by Inagaki et al., has linked observational studies with targeted perturbation to reveal experimental evidence for "point" attractor dynamics, which allows working memory to be held within a circuit[12]. Using neural perturbations, this study demonstrated that activity among neurons contributing to a point attractor encoding short term memory was robust to perturbation and could be used to guide the animal's choice in a memory task.

Attractors can exist in other topologies such as "line" or "continuous" attractors, which allow graded forms of information to be held persistently in circuits[26]. These continuous attractors can function as integrators, allowing a circuit to accumulate and represent a graded, integrated variable[26,43]. Each integrated value can be persistently represented within the circuit, enabling the encoding of continuous variables. The integrated positions function as "fixed points" which constrain activity within the circuit to exist within these points unless perturbed by external inputs to the circuit or noise. Studies applying dynamical systems tools to neural activity have discovered representations of continuous attractors important for encoding head-direction across species[47-49], with the graded form of persistence allowing for different angles of head-direction to be uniquely and persistently represented in activity[47-51]. Similar methods have also revealed continuous attractors encoding evidence in decision-making paradigms[22], for reward history in learning paradigms[52] and for encoding space[50,51].

Recent advances in machine learning have enabled the approximation or 'fitting' of neural data by dynamical systems, allowing for discovery of the governing equations that produce the observed dynamics. This data-driven approach contrasts with traditional modeling in neuroscience, where models are preconceived, intellectual constructs designed to approximate specific aspects of neural data (refer to Box 2). The process of discovering such dynamical properties requires the use of new unsupervised machine-learning enabled methods that fit various forms of dynamical systems directly to neural data [50]. Some

methods, such as LFADs[52,53] (latent factor analysis via dynamical systems), directly fit non-linear dynamical systems such as recurrent neural networks, while others, such as rSLDS[54,55] (recurrent switching linear dynamical systems), approximate non-linear systems as a set of interacting linear systems. A key feature of both model classes is that they use dimensionality reduction to identify functional groups of neurons termed "latent factors" or "dimensions" with shared dynamical properties. Dynamical models then explain the temporal evolution of activity along these latent factors, either using inputs to that circuit or through intrinsic dynamics via intra- or inter-region connectivity. Since these models are unsupervised, they do not assume any form of attractor dynamics or other dynamic mechanisms in the circuit. Furthermore, since they operate at the level of latent factors, they are agnostic to the exact implementation of the discovered dynamics in the neural circuit, instead revealing the computation performed by the circuit and not the mechanism of that computation. Hence, these emerging methods allow for the discovery of distinct dynamics in neural activity and generate hypothesis for how those dynamics operate at a computational level in the circuit.

Box 1

Primer on dynamical systems theory in neuroscience

In its simplest form, a dynamical system tracks the evolution of a variable (x) over time as a function of inputs to the variable (u) and a set of rules that govern its evolution in the absence of inputs (Figure 1A). These rules are represented by a matrix (A) that represents feedback to x from its activity at previous time steps. The variable x could symbolize the activity of a local population of interconnected neurons or that of neurons connected across various brain regions or nuclei. The structure of this matrix, dictating the strength and direction of feedback in the circuit, can lead to new dynamics. Such dynamics can be emergent in that they emerge through the interaction of circuit components (defined by the matrix A) and are not features of the constituent elements or neurons. For instance, consider a simplified neural circuit comprising two cell populations, x_1 and x_2, (Figure 1B, C-inset).

In this system, the matrix *A* is a 2x2 connectivity matrix which describes the pattern of intra-region and inter-region connectivity between these brain regions (Figure 1B, left).

Analyzing the connectivity matrix using dynamical systems methods, such as eigen decomposition, can reveal the emergent properties of the network. The most common approach is an eigen decomposition of this matrix which can give us an eigen spectrum, a plot of the real and imaginary part of the eigenvalue as a two-dimensional plot (Figure 1B, middle). The magnitude of the eigenvalues can reveal the time constant of dynamics of the two cell populations in our example (Figure 1B, right). From dynamical systems theory, an eigenvalue close to one indicates a corresponding dimension with a large time constant[36]. The time constant allows the cell population to remain persistent if it has received an external input.

Let us consider this scenario in more detail, characterized by strong excitatory connectivity within region x_1 but weak connectivity within x_2 and between both populations (See connectivity matrix in Figure 1B, left and Figure 1C, inset) . We can visualize the activity of both regions in a two-dimensional plot of their activity against each other. This set of all possible combinations of x_1 vs. x_2 activity in a neural system is called the neural state space (Figure 1D). The pattern of connectivity creates constraints within neural state space that is visualized using a "flow field": a set of arrows in each point of neural state space that together dictate the direction and speed with which activity can flow in this space. The flow field is inferred by analysis of the dynamics matrix *A*. The flow of activity in neural state space is represented by trajectories referred to as the population activity vector (PAV), since it summarizes net neural activity in this two-dimensional neural system at any given point. In this scenario, we see that arrows point to multiple points of stability (also termed "fixed points") arranged as a line (Figure 1E). The neural state space and flow field can also be visualized as a 3D topographic "energy landscape" as a gully or trough (Figure 1F). Such a feature in neural state-space is termed a line attractor. If repeated input is provided to both x_1 and x_2, activity can move along the line from one point of stability to

another, allowing the system to integrate inputs over time and represented an integrated variable (Figure 1E, F). Since the fixed points are stable, activity can be pushed to any arbitrary point, allowing the integrated value to encode continuous variables. For this reason, a line attractor is often referred to as a type of "continuous" attractor. Other topologies of continuous attractors exist, such as the ring attractor in *Drosophila*.

The pattern of connectivity can also create other strong constraints on where activity can flow in state-space. In Scenario 2, we illustrate the neural state-space and flow field for a variation on this network where both populations possess strong inhibitory intra-region connectivity with asymmetric but weak inter-region excitatory connectivity (Figure 1G). The flow field for this network shows arrows that point to a single position in state space that acts as a point of stability (Fig. 1H, I). The presence of such a single fixed point reveals a "point attractor," which can be visualized in an energy landscape as a pit (Figure 1J). Activity in other regions of the flow field tends to flow into this pit (hence the name "attractor"); once in it the PAV resists other influences that try to push it out of the pit, e.g. from external inputs to the system (Figure 1H-J). In essence the attractor can be thought of as a "sticky" point in neural state-space. Many more types of flow fields can also be obtained depending upon the precise connectivity of the dynamics matrix[27].

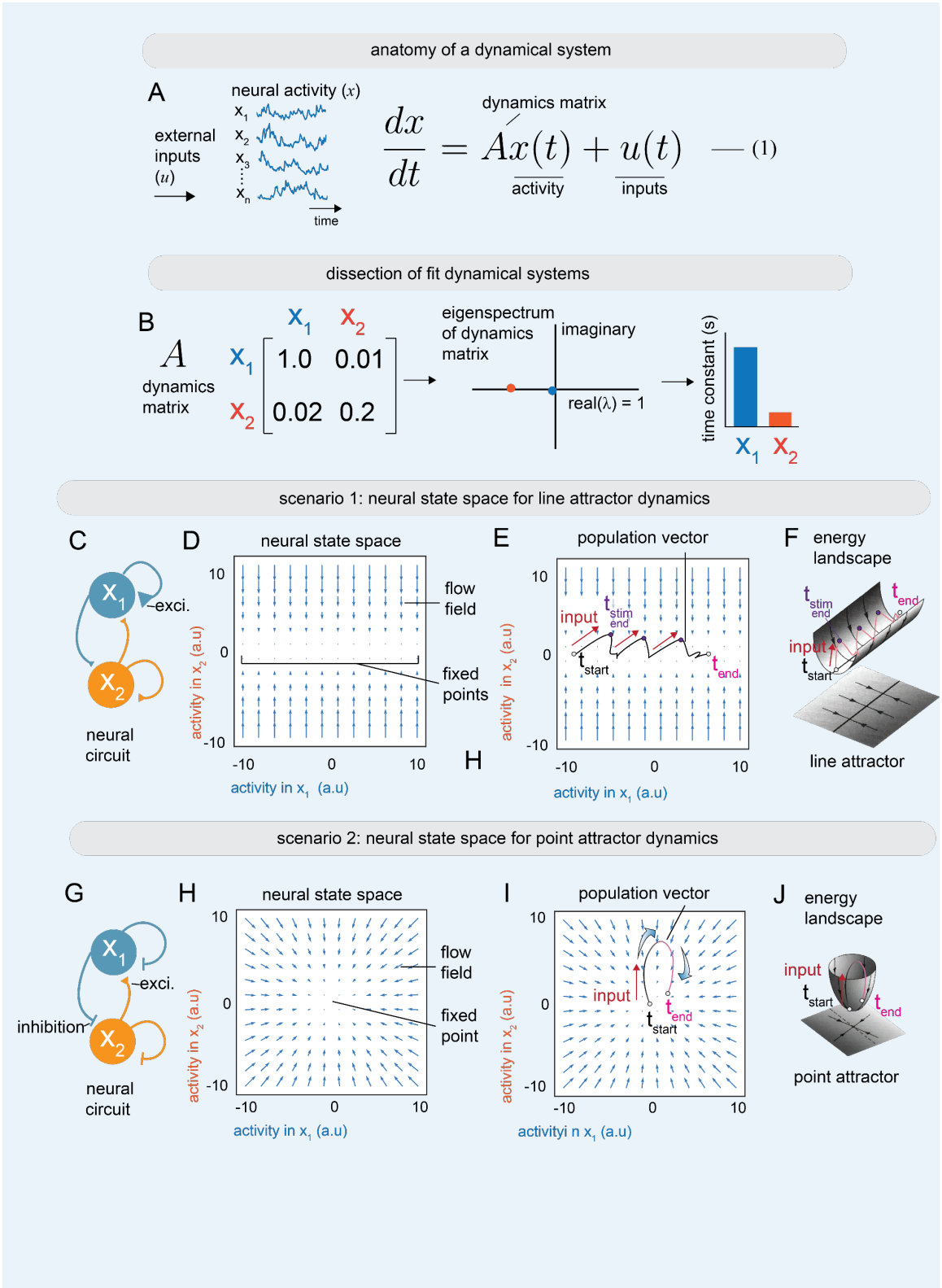

anatomy of a dynamical system
A
neural activity (x)
external inputs (u)
time
dynamics matrix
activity
inputs
— (1)
dissection of fit dynamical systems
B
dynamics matrix
eigenspectrum of dynamics matrix
imaginary
real(λ) = 1
time constant (s)
1.0 0.01
0.02 0.2
scenario 1: neural state space for line attractor dynamics
C
exci.
neural circuit
D
neural state space
flow field
fixed points
activity in x1 (a.u)
activity in x2 (a.u)
E
population vector
input
F
energy landscape
line attractor
scenario 2: neural state space for point attractor dynamics
G
inhibition
H
I
J
point attractor

Figure 1: A: Anatomy of a dynamical system. Left: Cartoon depicting a neural recording with activity (*x)* of several neurons and external inputs (*u*) to the circuit. Right: Equation for a generic linear dynamical system. The dynamical system tracks the evolution of the variable *x*. The dynamics matrix (*A*) accounts for feedback within the network *x* and *u* represents feedforward external inputs to the network. B: Left: A dynamics matrix displaying the functional coupling between x_1 & x_2, interpretable similarly to a 2x2 connectivity matrix, with presynaptic regions on the rows and postsynaptic regions on the columns. Middle: Understanding the dynamics matrix through eigendecomposition, plotted in two dimensions to show the real and imaginary parts of the eigenvalues. Right: The magnitude of the eigenvalues can be used to calculate a time constant which shows the persistence and slow decay along each dimension in the presence of an external input. C: Scenario 1: A cartoon of a coupled neural system consisting of two regions. Region x_1 is marked by strong recurrent excitation. D: The neural state space shows the set of all possible configurations of activity in this system. The flow field, inferred by analysis of the dynamics matrix *A*, shows the constraints that activity must obey in neural state space. The arrows reflect the direction and speed with each activity will move in this system at any given point. The flow field reveals a set of stable "fixed points" which define a line attractor in neural state space oriented parallel to dimension x_1. E: Same as D) but showing the population activity vector (PAV) when the system is provided with a transient input. Each input pushes the system out of the line attractor, following which it returns to the attractor but at a different fixed point from its location when input was initiated. This property allows the neural system to "integrate" (i.e., to accumulate and store information about the history of the inputs) along the line attractor. F: Same as E) but showing the flow field as a 3D topographic energy landscape. The line attractor appears as a trough or funnel, with the base of the trough indicating the stable points of the dynamical system. G: Scenario 2 showing a similar neural system, where both regions show inhibitory recurrent connectivity. H: In this configuration, the flow field reveals a single stable point called a point attractor. I: Same as H) but showing the PAV when transient input is provided to the neural system. J: Same as I) but as a 3D energy landscape. The point attractor appears as a pit in this 3D landscape.

Box 2

Discovering dynamical properties of neural circuits through data-driven modelling

Classical approaches to modeling neural circuits begin with a defined hypothesis about the circuit's properties, such as a specific connectivity pattern or other mechanistic features[13,56]. This ”theory-driven” approach attempts to recreate specific properties of neural activity observed in neural circuits (Figure 2A). For instance, numerous theoretical mechanistic models of the oculomotor integrator have been constructed to explain persistent activity seen in neurons from in-vivo data[43]. In the study of internal states, mechanistic models have been used to explain persistent activity in VMHdm neurons involved in maintaining a defensive state[13]. In each system, different models were constructed that incorporated ad hoc various mechanistic features such as recurrent connectivity, cell-intrinsic persistent activity, or neuromodulator release. These models can then be compared with each other using various metrics of similarity to the data, with the model showing the highest similarity predicted to capture the underlying mechanism. In this approach the observed data are used as a point of reference, rather than serving to directly constrain model construction.

Advances in machine learning over the past several years have enabled a different type of modeling framework that uses machine learning to “learn” the governing equations underlying neural data directly from the data itself[50]. These “data-driven” approaches fit equations of dynamical systems directly to neural data. The fitting process adjusts model parameters through iterative comparisons between a model's predicted neural trajectory and the actual trajectories in the data, using various machine learning algorithms. Some data-driven approaches, such as LFADs (Latent Factor Analysis Via Dynamical Systems)[53] and PINning (Partial In Network Training)[57], attempt to fit nonlinear recurrent network models (RNNs) of observed neural data in a 1-1 manner, providing predictions of

underlying connectivity (Figure 2B). For instance, PINning has been used to create data-driven models of neural activity during decision making tasks in the posterior parietal cortex (PPC)[57]. By analyzing the fit connectivity matrix, the authors discover a new computational mechanism for sequential activity in PPC that uses recurrent connectivity and external inputs. This capacity to discover new computations through analysis of the fit model parameters has evoked considerable interest in this data-driven approach[50].

A distinct class of data-driven approaches fit models at a more algorithmic level to neural data (Figure 2C). These methods, including SLDS (switching linear dynamical systems), use dimensionality reduction to find a set of low-dimensional composite neural signals ("latent variables") which capture most of the variance in the original recordings, and then fit dynamical equations to these low-dimensional variables[55]. In addition to discovering latent population signals using dimensionality reduction, this approach can identify new computations and functions performed by those latent neural dimensions through examination of the fit parameters of the dynamical models. These methods have been recently used to uncover integration signals in the hypothalamus that function as continuous attractors encoding internal states[54,58]. Methods using SLDS models fit fewer parameters than RNNs and are thus well suited for single-trial neural data such as those usually obtained during studies of innate behavior. Moreover, the linear nature of the SLDS model allows for a comprehensive understanding of the fit model using methods from linear dynamical systems theory (see **Box 1**). However, unlike RNNs SLDS models are statistical, not mechanistic, meaning that they do not assign explicit connectivity strengths and dynamics to neurons in a model network.

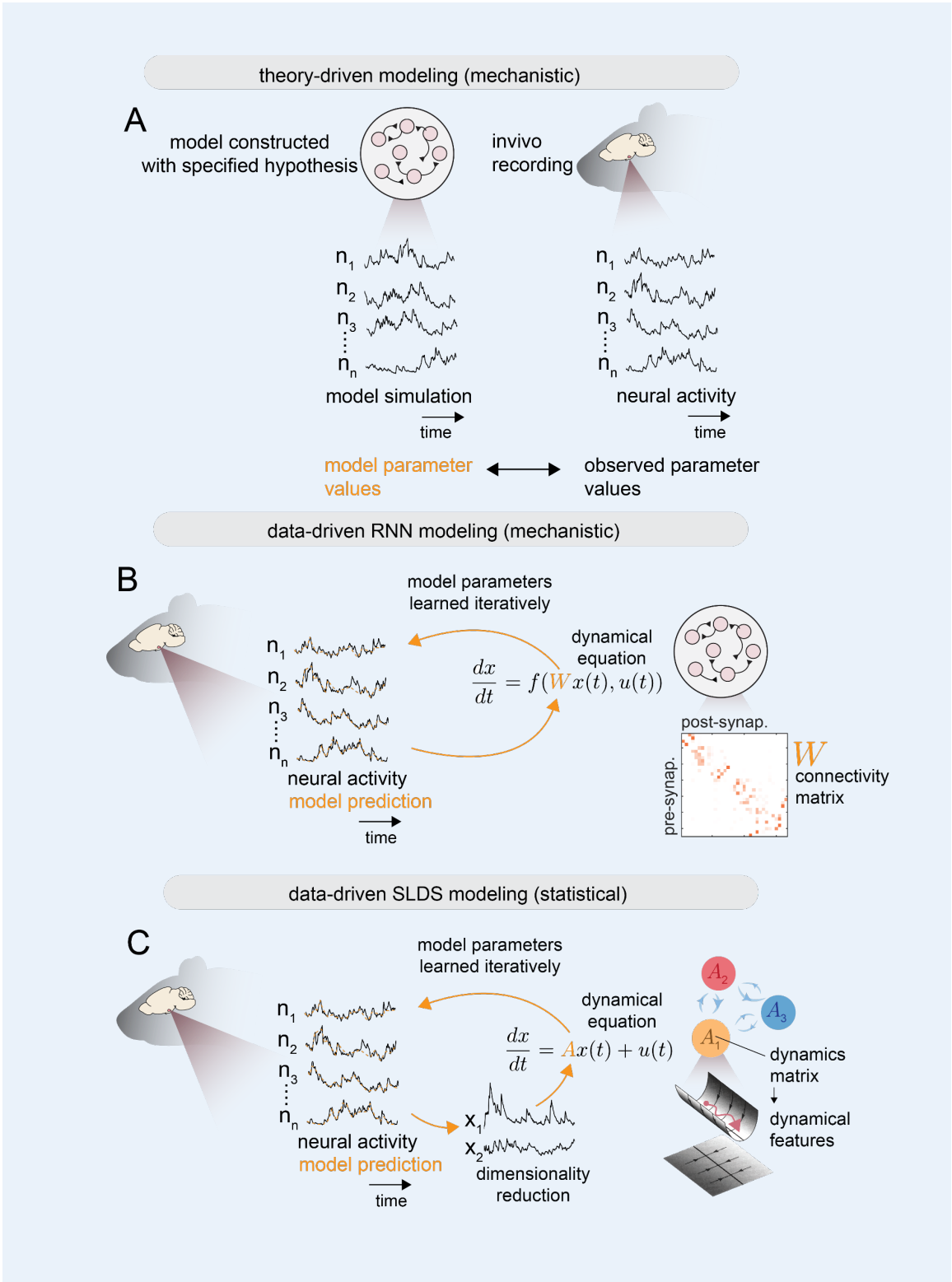

theory-driven modeling (mechanistic)
A
model constructed with specified hypothesis
invivo recording
n1
n2
n3
nn
model simulation
time
neural activity
time
model parameter values
observed parameter values
data-driven RNN modeling (mechanistic)
B
model parameters learned iteratively
dynamical equation
$\frac{dx}{dt} = f(Wx(t), u(t))$
neural activity
model prediction
time
post-synap.
pre-synap.
W
connectivity matrix
data-driven SLDS modeling (statistical)
C
model parameters learned iteratively
dynamical equation
$\frac{dx}{dt} = Ax(t) + u(t)$
neural activity
model prediction
time
x1
x2
dimensionality reduction
A2
A3
A1
dynamics matrix
dynamical features

Figure 2: A: Illustration of the theory-driven modeling approach to understand neural computation. In this approach, models are constructed with specific hypothesis such as detailed connectivity patterns or intrinsic properties and then compared to other models and to neural data. B: Illustration of the data-driven RNN modeling approach to understand neural computation. In this framework, non-linear dynamical equations are fit to neural data and the fit connectivity matrix can studied to reveal neural computations. C: Illustration of the data-driven SLDS modelling approach to understand neural computation. This framework uses dimensionality reduction to obtain latent dimension of neural data and fits dynamical equations to the latent dimensions. Analysis of the fit dynamics matrix can reveal novel computations in the circuit studied.

Dynamical analysis reveals an encoding of internal states as continuous attractors in state-controlling circuit nodes across behaviors.

A set of recent studies have applied unsupervised dynamical systems analysis directly to neural activity from state- contributing circuit nodes (male VMHvl^{Esr1} neurons[54] and female VMHvl$^{Esr1+,Npy2r-}$ neurons[58]). Functional manipulation of these circuit nodes has revealed a role in regulating internal states of aggression and sexual receptivity, but neural representation studies have failed to identify an encoding of behavior at the single-cell level[10,11,58,59]. By incorporating dynamics into the study of neural activity, these studies discover latent factors with properties of ramping and stability that correlate with the computations of scalability and persistence that are essential features of many internal states (Figure 3). Nair et al., discovered a latent factor termed the "integration dimension" whose activity resembles that of a neural integrator that accumulates some input variable during aggression[54] (Figure 3A). Activity along the integration dimension ramps up as an animal starts sniffing a male intruder and peaks as animals engage in attack behavior (Figure 3A). Notably, activity in this dimension remains persistent during the inter-bout intervals of

attack, suggesting that this latent factor resembles a state of aggression rather than a representation of attack motor behavior. Importantly, by dissecting the fit dynamical system through examination of the fit parameters, the authors find that the integration dimension functions as a line or continuous attractor, with the graded nature of the state mapping on to various "fixed points", or points of persistence, along the attractor (Figure 4A).

Remarkably, another study, Liu*, Nair*, et al., also found a qualitatively similar integration dimension during mating behavior that correlates with female sexual receptivity[58] (Figure 3C). This study used similar dynamical models to reveal a continuous attractor that integrates male contact to create a continuous representation of receptivity (Figure 3D). The integrated value along the attractor correlated strongly with the amount of receptive behavior displayed by females in a given mating trial. Surprisingly, the authors found that the line attractor disappears and then reappears in a periodic manner across the estrous cycle, with single neurons contributing to the attractor undergoing a transformation in their dynamics from persistent to transient. Importantly, in both studies, activity along the integration dimension was distinct from that of the population mean of all recorded neurons, underscoring the importance of dynamical methods to reveal neural signals with distinct, behaviorally relevant dynamics.

These studies suggest that continuous attractor dynamics might be a type of canonical computation reused by subcortical nuclei to encode key properties of internal states, namely its persistence and scalability. However, further work remains to elucidate the nature of both the external input variable that is integrated by the attractor and the resulting output state variable from the integration process. For example, while Liu*, Nair* et al., suggest that the receptivity encoding continuous attractor is integrating male contact, the actual dynamics of an input to the circuit that resembles such contact are yet to be uncovered. Furthermore, the state output variable that is integrated by such continuous attractor might reflect qualities of arousal, motivational drive or both. Hence, elucidating the downstream behavioral functions of state-encoding continuous attractors represents an important area of future research.

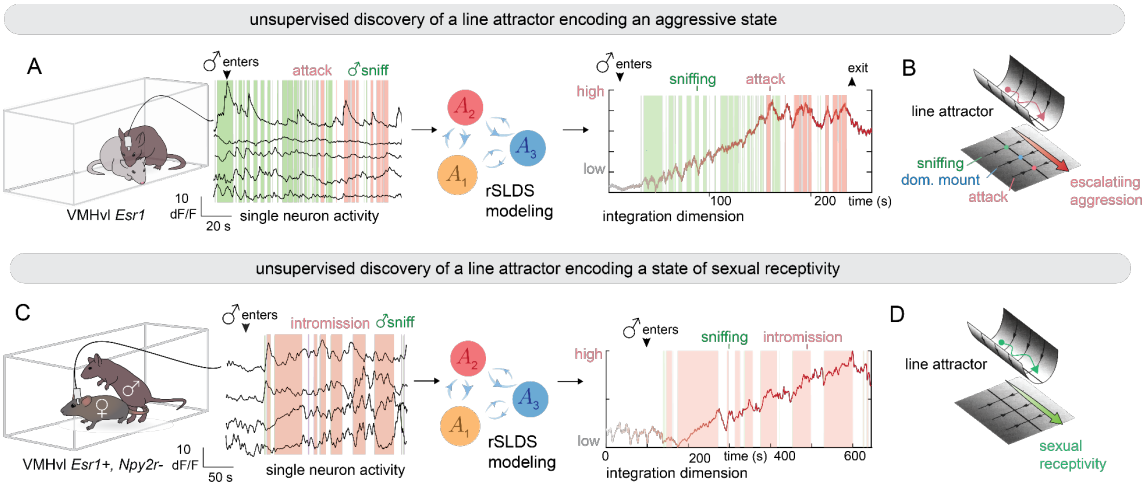


Figure 3: A: Summary of results from Nair et al., 2023. Unsupervised dynamical systems analysis is used to discover a dimension of activity termed the integration dimension in male VMHvl. This dimension possesses properties of ramping activity and persistent. B: This modeling approach discovers a line attractor in male VMHvl that correlates with an escalating aggressive state. C: Summary of results from Liu*, Nair* et al., 2024. Unsupervised dynamical analysis uncovers a similar integration during mating behavior in female mice with similar properties of ramping and persistence. D: Dynamical analysis discovers a line attractor in female VMHvl that correlates with escalating sexual receptivity.

Revealing the causal implementation of state-encoding continuous attractors

Understanding how continuous attractor dynamics are implemented in subcortical nuclei requires us to understand whether the attractor is an inherited feature of inputs to the nucleus or might be intrinsic to the nucleus under study. Given that sources of persistent, time-varying sensory stimuli might be present in the environment, the observed continuous attractor could simply reflect ramping sensory stimuli, and the brain region being studied might not possess true attractor dynamics. Arbitrating between these possibilities requires perturbation experiments to rigorously test attractor dynamics in subcortical regions.

Behavioral perturbations can be informative to reveal the intrinsic nature of persistence in attractor circuits. By examining dynamics along the aggression-encoding continuous attractor in VMHvl^{Esr1} neurons upon removal of a male intruder, Nair et al., observed persistence and slow decay, as predicted by an attractor system, as activity did not collapse upon removal of male-derived sensory stimuli[54] (Figure 4A, B). Liu*, Nair* et al., performed a more sophisticated version of behavioral perturbation for the receptivity-encoding continuous attractor in VMHvl$^{Esr1+,Npy2r-}$ neurons by using wireless optogenetics to induce behavioral pauses in a mating assay. By activating VMHdm^{Sf1} neurons in a male intruder and examining activity in the female continuous attractor in VMHvl$^{Esr1+,Npy2r-}$ neurons, the authors observed slow decay along the attractor, with activity remaining persistent and not collapsing when the male ceased interacting with the female[58]. In some mice, the authors even observed continued ramping activity, suggesting an intrinsic source to ramping in this system.

A rigorous test of the intrinsic nature of attractor circuits, however, requires direct experimental perturbations of neural activity in conjunction with neural recordings. These perturbations may be "on-manifold" if they are targeted to neurons contributing to the attractor and "off-manifold" if they target other attractor-orthogonal neurons in the dimensionally reduced neural state space. While on-manifold perturbations are crucial to reveal conclusive evidence for the intrinsic nature of persistence in attractor networks, off-manifold perturbations are critical to illustrate the attractive nature of fixed points in an attractor. Until recently, on-manifold perturbations were only performed on a head-direction encoding continuous attractor in Drosophila[49], while off-manifold perturbations were only performed in a murine point attractor and was not directed to specific neuron subsets. In a recent study, Vinograd*, Nair*, et al., utilized recent observations that VMHvl^{Esr1} neurons mirror aggression to instantiate the aggression-encoding continuous attractor dynamics in the VMHvl of head-fixed mice[60]. Using holographic on-manifold activation, they reveal the capacity of the hypothalamic continuous attractor to integrate inputs to various fixed points,

and by providing off-manifold activation, they also reveal the attractive nature of those identified fixed points (Figure 4D-F). Neural perturbations have also been used to reveal the intrinsic nature of the female receptivity-encoding continuous attractor[58]. Liu*, Nair*, et al., perform neural perturbations in freely moving animals, by using generic neural inactivation to induce off-manifold perturbations of the female line attractor, revealing the attractive nature of fixed points identified by the data-driven models used in the study (Figure 4G-H).

These studies demonstrate the capacity of hypothalamic circuits to create intrinsic continuous attractor dynamics that encode representations of internal states. However, such intrinsic dynamics may emerge through either local connectivity or long-range connectivity. Given the densely interconnected nature of hypothalamic circuits[61-63], further investigation is needed to understand how attractor dynamics are instantiated across state-contributing nuclei.

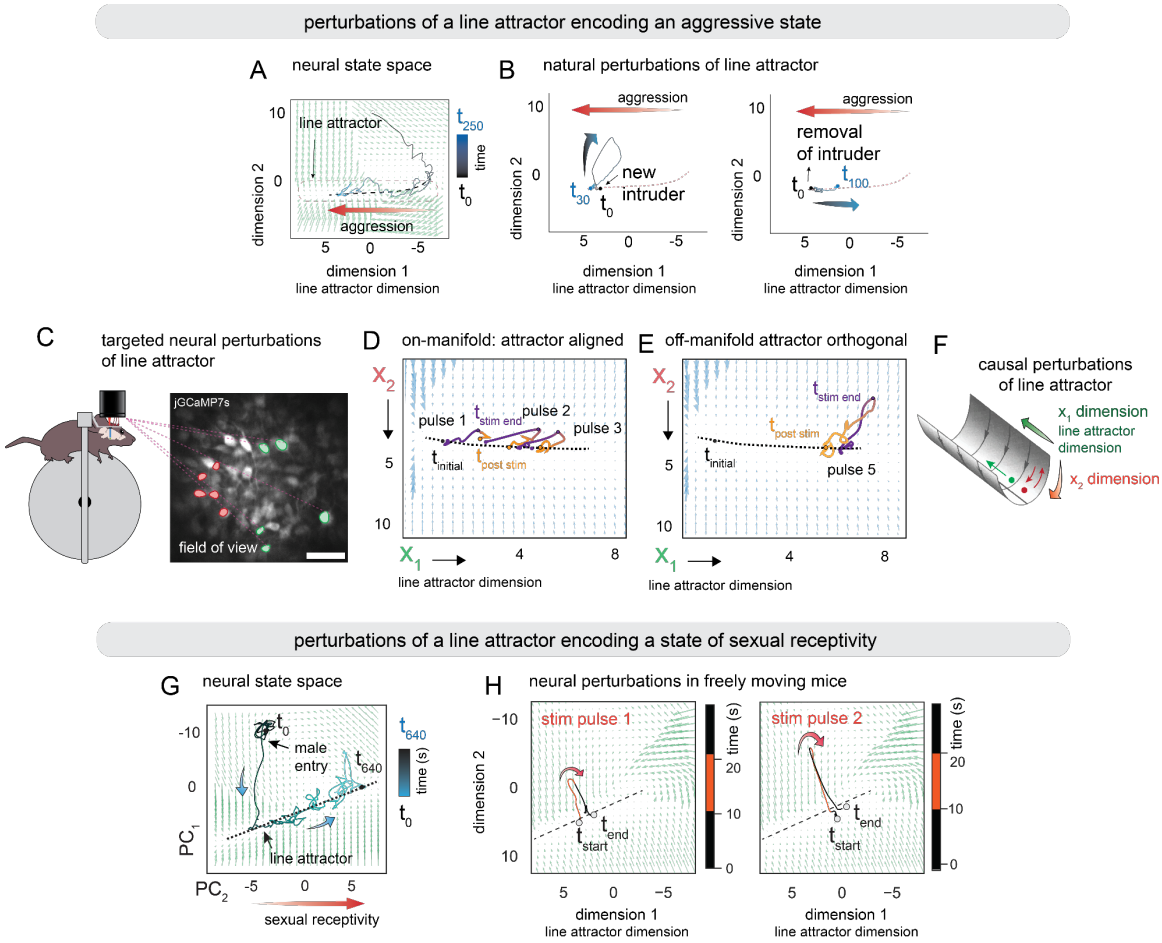


Figure 4: A: Neural state space showing line attractor dynamics in Nair et al., 2023. B: Behavioral perturbations performed in Nair et al., 2023: Left: Introduction of a male intruder results in a transient off-manifold response, with the neural trajectory returning to the line attractor as the resident interacts with the new intruder. Right: Removal of all intruders from the cage results in a slow movement of the neural trajectory along the line attractor, as predicted by the time-constant of the fit dynamical system model. C: Schematic of experiments performed in Vinograd*, Nair*, 2024. Targeted perturbations are performed towards either attractor aligned (x_1) or orthogonal neurons (x_2). D: Activation of attractor aligned (x_1) neurons results in an on-manifold movement of the neural trajectory in neural state space. Each pulse of activation is integrated by the circuit, allowing the trajectory to move to the end of the line. E: Activation of attractor-orthogonal neurons results in an off-manifold response where the neural trajectory is perturbed in a transient manner orthogonal to the

line. Activity returns to the line at the end of the stimulation period. F: Cartoon summarizing results in Vinograd*, Nair*, et al., 2024 showing effects of on- and off-manifold perturbation in neural state space. G: Neural state space showing line attractor dynamics in Liu*, Nair*, et al., 2024. H: Neural perturbations, in conjunction with imaging in freely moving animals results in transient off-manifold movements of the neural trajectory in state space.

Linking internal state encoding continuous attractors to circuit mechanisms

The success of dynamical models in revealing representations of state-encoding continuous attractors has relied on statistical models that are agnostic to implementation-level details of these emergent dynamics. Demonstrating the intrinsic nature of these dynamics is crucial for showing that the discovered signals exhibit attractor dynamics, utilizing some form of intra- or inter-region connectivity, and are not simply graded inputs persistently present in the environment. Furthermore, significant open questions remain regarding the circuit-level implementations of these discovered dynamics. More broadly, the mechanisms of how neural dynamics emerge within the neural circuits that they are studied within is yet to be fully understood, raising important questions about how population-manifold approaches can be linked to a circuit-level understanding[25].

Recent studies on hypothalamic line attractors have begun to address questions of neural implementation. Vinograd*, Nair* et al., provide evidence for functional connectivity among attractor-contributing neurons within the aggression-encoding continuous attractor in $VMHvl^{Esr1}$ neurons. Interestingly, they find that such functional connectivity is specific to the attractor-contributing ensemble and attractor-orthogonal neurons do not possess similar connectivity (Figure 5A-C). By employing computational models, they propose that this connectivity is not implemented through traditional glutamatergic connectivity, as theorized for cortico-hippocampal and oculomotor attractor networks[43], but instead through slower timescale neuropeptide-mediated connections. This study suggests that slow timescale neuromodulators such as neuropeptides, released either locally or through macro-level inter-region connectivity are necessary to explain the slow timescale intrinsic dynamics of the observed continuous attractor[60].

In another study, Mountoufaris, Nair et al., provided evidence for this hypothesis by performing a targeted CRISPR-based knockout of neuropeptide receptor signaling in VMHvl^{Esr1} neurons[64] (Figure 5D). This study finds that the knockout of oxytocin (OXT) and vasopressin (AVP) receptor signaling results in the loss of the aggression-encoding continuous attractor in VMHvl^{Esr1} neurons, providing causal evidence of a need for peptide signaling to implement a hypothalamic line attractor. Dynamical systems analysis on OXT/AVP knockout mice reveal that the resulting dimension no longer possess the property of integration but rather shows activity time-locked to each bout of male-interaction (Figure 5E). As a result, the system possesses a point attractor where the point of stability represents the baseline activity of the nucleus (Figure 5F). While these results point to the importance of peptide signaling towards the hypothalamic line attractor, the relevant peptides are not synthesized by VMHvl^{Esr1} neurons. Thus, further investigation is necessary to reveal the differential contribution of inter-region and intra-region connectivity for peptide-mediated continuous attractor dynamics.

The importance of slow neuromodulator signaling to hypothalamic continuous attractors also appears to generalize to other internal states. By recording neurons across the estrus cycle, Liu*, Nair* et al., show that the female receptivity-encoding continuous attractor in VMHvl is lost during non-proestrus states when levels of cycling hormones of estrogen and progesterone are low, suggesting a strong modulation by steroidal hormones on a timescale of days (Figure 5G-H). Dynamical analysis from mice in the non-proestrus stage of the estrus cycle reveal that the resulting dimensions no longer integrate but possess activity time-locked to periods of interaction with the male (Figure 5H). This suggests that hormones may also play a role in instantiating hypothalamic line attractor dynamics (Figure 5I).

These new studies challenge dominant assumptions in two different domains of neuroscience: 1) They suggest that traditional theoretical frameworks for attractor dynamics[43] need to be expanded to accommodate new classes of mechanisms such as

peptide-mediated attractor dynamics as discovered in hypothalamic continuous attractors. 2) They indicate that the hypothalamus performs a greater degree of computation than previously recognized, challenging the conventional view that equates the region to a simpler network of behavioral action-specific relay stations[65,66]. This repositions the hypothalamus as a crucial platform to understand how attractor dynamics can emerge through multi-level mechanisms, incorporating hormonal and neuropeptide signaling with recurrent connectivity to elucidate the encoding of internal states.

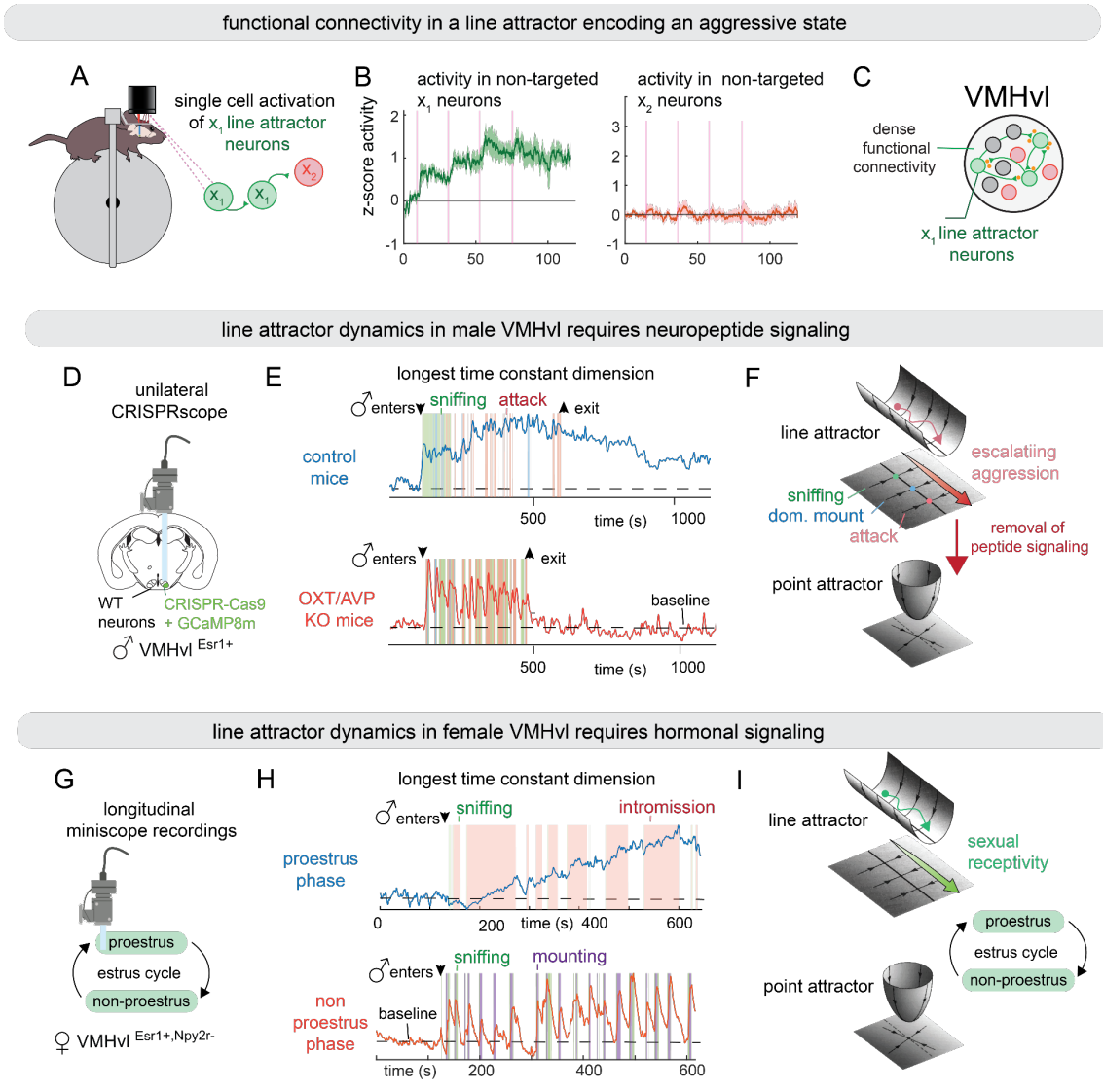


Figure 5: A: Schematic of experiments in Vinograd*, Nair*, et al., 2024: single line attractor contributing neurons (x_1) are activated and untargeted x_1 and attractor-orthogonal (x_2) neurons are examined. B: Left: Activity of untargeted x_1 neurons showing functional connectivity within the line attractor contributing ensemble. Right: Activity of untargeted x_2 neurons showing lack of functional connectivity between x_1 and x_2 neurons. C: Summary of results in Vinograd*, Nair* et al., 2024, showing selective functional connectivity within the line attractor ensemble. D: Schematic of experiments in Mountoufaris, Nair, et al., 2024: miniscope recordings are performed in animals where oxytocin (OXT) and vasopressin (AVP) receptors in VMHvl are knocked out in Esr1 neurons using CRISPR unilaterally. This design allows researchers to study the effect of the knockout on neural activity, independent of its effect on behavior. E: Longest time constant dimension from dynamical system modeling in control (top) and OXT/AVP KO (bottom) mice. This dimension in control mice shows a long time constant with persistent activity that decays slowly in the post-intruder period. In the KO mice however, this dimension possesses a smaller time constant and thus is unable to integrate, resulting in activity time-locked to interaction bouts. F: Summary of Mountoufaris, Nair, et al., showing the loss of line attractor dynamics upon removal of neuropeptide signaling. G: Schematic of experiments in Liu, Nair, et al., 2024: miniscope recordings are performed longitudinally across the estrus cycle in female mice. H: Longest time constant dimension in the same mice during the proestrus phase (top) and non-proestrus phase (bottom). This dimension during the proestrus shows a long time constant with slow ramping and persistent activity. During the non-proestrus phase, this dimension possesses a smaller time constant and thus is unable to integrate, resulting in activity time-locked to interaction bouts. I: F: Summary of Liu*, Nair*, et al., showing the transformation of line attractor dynamics through the estrus cycle.

Testing the limits of the dynamical systems perspective in internal states

The discovery of internal state-encoding continuous attractors introduces a new paradigm to identify and study internal affective states through their unsupervised identification using neural dynamics. However, many critical questions remain about the value of this perspective in the field of affective neuroscience. Demonstrating that movement along hypothalamic continuous attractors causally influence behavior is essential to advance beyond correlation-

based studies of attractor dynamics. However, conducting such experiments requires genetic access to neurons that contribute to continuous attractor dynamics. Since these neurons form a functional subpopulation, activity-based tagging methods may be best suited for accessing these subpopulations[67]. Whether these neuronal dynamics correspond neatly to genetic cell types marked by specific genes remains an open and significant question for enabling targeted perturbation studies of these dynamics.

Population based studies of dynamics often assume that the dynamics are "read-out" by downstream targets of the brain region under study, an assumption that remains to be experimentally validated. Hypothalamic continuous attractors may be "read" by other recurrently connected hypothalamic regions and by pre-motor centers such as the PAG. Understanding these transformations will require simultaneous imaging of hypothalamic regions, along with imaging of activity in the downstream brain region. New imaging techniques that allow for simultaneous two-photon imaging within multiple brains regions will be crucial for determining whether and how continuous attractor dynamics are read-out downstream.

A consistent feature of data-driven studies of hypothalamic continuous attractor dynamics is that key attractor parameters vary across individuals, even when they are genetically identical. For instance, Nair et al., found that the leak-rate of the aggression-encoding continuous attractor is strongly correlated with the fraction of time a mouse spends performing attack behavior, suggesting that experience and/or other features can modify the stability of the inferred continuous attractor[54] (Figure 6A, C). Building on this result, Vinograd*, Nair*, et al., found that the leak-rate of a similar attractor in head-fixed mice is positively correlated with the degree of functional connectivity within the attractor-contributing ensemble of neurons (Figure 6B, C). These results also suggest that attractor properties may be altered by conditions such as disease states. An exciting direction for this perspective is whether disease states predictably modify attractor properties and whether drug-based treatments might be used to restore such changes. Several neuropsychiatric disorders such Major Depression Disorder (MDD) feature persistent changes in behavior[68],

suggesting that impaired attractor-dynamics might be involved in explaining aspects of disease etiology[69-71]. Furthermore, the unsupervised discovery of encoded states directly from neural activity holds significant promise as biomarkers for a broad spectrum of psychiatric disorders with variable and undefined behavioral expressions. This underscores the transformative potential of understanding continuous attractor dynamics, not only for advancing basic neuroscience but also for developing targeted interventions in clinical settings.

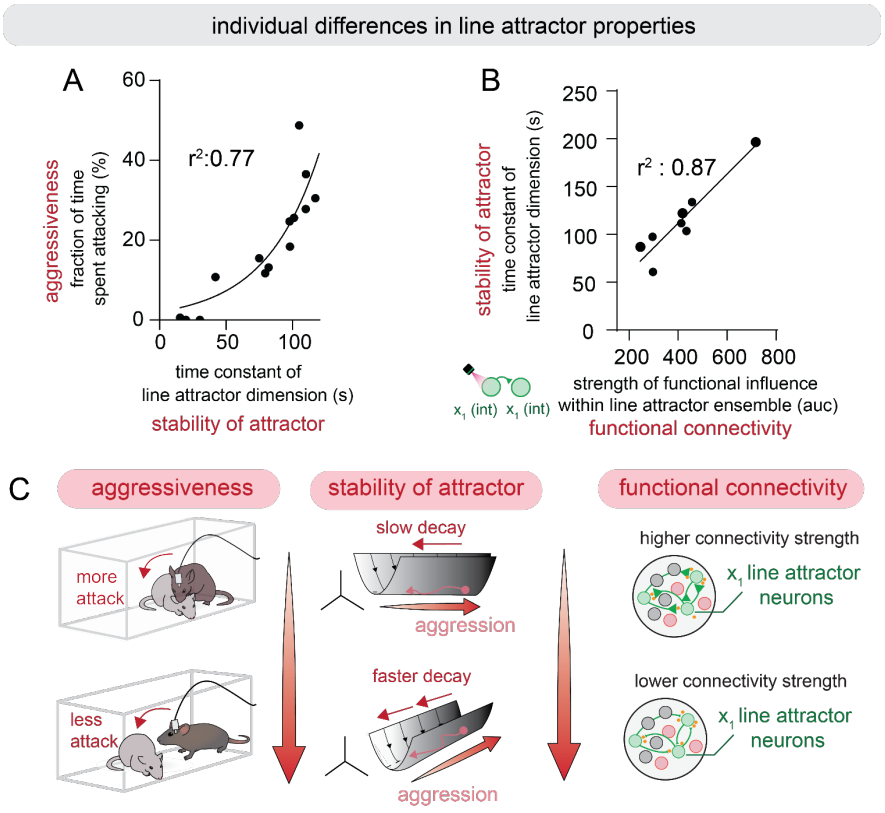


Figure 6: A: Results from Nair et al., showing a correlation between the stability of the line attractor (time constant of the line attractor dimension) and aggressiveness (fraction of time spent attacking) in individual mice. B: Results from Vinograd*, Nair*, et al., showing a correlation between the stability of the line attractor and strength of functional connectivity within the line attractor contributing ensemble (level of activity evoked in untargeted line attractor neurons upon activation of a single neuron) across individual mice. C: Summary of individual differences and line attractor properties in Nair et al., 2023 and Vinograd*, Nair*, et al., 2024.

Conclusions

The advent of machine learning in neuroscience has facilitated data-driven discoveries of neural dynamics across a wide range of paradigms. When applied to neural data acquired during innate behaviors, these tools can uncover an encoding of internal states implemented as continuous attractors, with key features such as persistence and scalability reflected in the emergent properties of attractor dynamics. While future work is necessary to address critical questions of behavioral causality, implementation and read-out, the paradigm holds promise to link the domains of manifold-based and circuit-based neuroscience[25] that are increasingly becoming indispensable to understand neural computations in internal states. We hope and anticipate that the foundational ideas outlined in this perspective will inspire exciting research at the intersection of machine learning, circuit recording and manipulations for the foreseeable future.

References

1. Flavell, S. W., Gogolla, N., Lovett-Barron, M., & Zelikowsky, M. The emergence and influence of internal states. *Neuron* **110**, 2545-2570 (2022). https://doi.org:10.1016/j.neuron.2022.04.030

2 Anderson, D. J. Circuit modules linking internal states and social behaviour in flies and mice. *Nat Rev Neurosci* **17**, 692-704 (2016). https://doi.org:10.1038/nrn.2016.125

3 Allen, W. E. *et al.* Thirst-associated preoptic neurons encode an aversive motivational drive. *Science* **357**, 1149-1155 (2017). https://doi.org:10.1126/science.aan6747

4 Sternson, S. M., Nicholas Betley, J. & Cao, Z. F. Neural circuits and motivational processes for hunger. *Curr Opin Neurobiol* **23**, 353-360 (2013). https://doi.org:10.1016/j.conb.2013.04.006

5 Sutton, A. K. & Krashes, M. J. Integrating hunger with rival motivations. *Trends Endocrinol Metab* **31**, 495-507 (2020). https://doi.org:10.1016/j.tem.2020.04.006

6 Sternson, S. M. Exploring internal state-coding across the rodent brain. *Curr Opin Neurobiol* **65**, 20-26 (2020). https://doi.org:10.1016/j.conb.2020.08.009

7 Allen, W. E. *et al.* Thirst regulates motivated behavior through modulation of brainwide neural population dynamics. *Science* **364**, 253 (2019). https://doi.org:10.1126/science.aav3932

8 Guthman, E. M. & Falkner, A. L. Neural mechanisms of persistent aggression. *Curr Opin Neurobiol* **73**, 102526 (2022). https://doi.org:10.1016/j.conb.2022.102526

9 Lischinsky, J. E. & Lin, D. Neural mechanisms of aggression across species. *Nat Neurosci* **23**, 1317-1328 (2020). https://doi.org:10.1038/s41593-020-00715-2

10 Remedios, R. *et al.* Social behaviour shapes hypothalamic neural ensemble representations of conspecific sex. *Nature* **550**, 388-392 (2017). https://doi.org:10.1038/nature23885

11 Lee, H. *et al.* Scalable control of mounting and attack by Esr1+ neurons in the ventromedial hypothalamus. *Nature* **509**, 627-632 (2014). https://doi.org:10.1038/nature13169

12 Karigo, T. *et al.* Distinct hypothalamic control of same- and opposite-sex mounting behaviour in mice. *Nature* **589**, 258-263 (2021). https://doi.org:10.1038/s41586-020-2995-0

13 Kennedy, A. *et al.* Stimulus-specific hypothalamic encoding of a persistent defensive state. *Nature* **586**, 730-734 (2020). https://doi.org:10.1038/s41586-020-2728-4

14 Kunwar, P. S. *et al.* Ventromedial hypothalamic neurons control a defensive emotion state. *Elife* **4** (2015). https://doi.org:10.7554/eLife.06633

15 LeDoux, J. Rethinking the emotional brain. *Neuron* **73**, 653-676 (2012). https://doi.org:10.1016/j.neuron.2012.02.004

16 Anderson, D. J. & Adolphs, R. A framework for studying emotions across species. *Cell* **157**, 187-200 (2014). https://doi.org:10.1016/j.cell.2014.03.003

17 Yang, B., Karigo, T. & Anderson, D. J. Transformations of neural representations in a social behaviour network. *Nature* **608**, 741-749 (2022). https://doi.org:10.1038/s41586-022-05057-6

18 Bayless, D. W. *et al.* Limbic neurons shape sex recognition and social behavior in sexually naive males. *Cell* **176**, 1190-1205 e1120 (2019). https://doi.org:10.1016/j.cell.2018.12.041

19 Hong, W., Kim, D. W. & Anderson, D. J. Antagonistic control of social versus repetitive self-grooming behaviors by separable amygdala neuronal subsets. *Cell* **158**, 1348-1361 (2014). https://doi.org:10.1016/j.cell.2014.07.049

20 Nordman, J. C. *et al.* Potentiation of divergent medial amygdala pathways drives experience-dependent aggression escalation. *J Neurosci* **40**, 4858-4880 (2020). https://doi.org:10.1523/JNEUROSCI.0370-20.2020

21 Churchland, M. M. *et al.* Neural population dynamics during reaching. *Nature* **487**, 51-56 (2012). https://doi.org:10.1038/nature11129

22 Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78-84 (2013). https://doi.org:10.1038/nature12742

23 Ebitz, R. B. & Hayden, B. Y. The population doctrine in cognitive neuroscience. *Neuron* **109**, 3055-3068 (2021). https://doi.org:10.1016/j.neuron.2021.07.011

24 Jazayeri, M. & Afraz, A. Navigating the neural space in search of the neural code. *Neuron* **93**, 1003-1014 (2017). https://doi.org:10.1016/j.neuron.2017.02.019

25 Langdon, C., Genkin, M. & Engel, T. A. A unifying perspective on neural manifolds and circuits for cognition. *Nat Rev Neurosci* **24**, 363-377 (2023). https://doi.org:10.1038/s41583-023-00693-x

26 Khona, M. & Fiete, I. R. Attractor and integrator networks in the brain. *Nat Rev Neurosci* **23**, 744-766 (2022). https://doi.org:10.1038/s41583-022-00642-0

27 Miller, P. Dynamical systems, attractors, and neural circuits. *F1000Res* **5** (2016). https://doi.org:10.12688/f1000research.7698.1

28 Adolphs, R. & Anderson, D. J. *The neuroscience of emotion: A new synthesis*. (Princeton University Press, 2018).

29 Gunaydin, L. A. *et al.* Natural neural projection dynamics underlying social behavior. *Cell* **157**, 1535-1551 (2014). https://doi.org:10.1016/j.cell.2014.05.017

30 Kennedy, A. *et al.* Internal states and behavioral decision-making: Toward an Integration of Emotion and Cognition. *Cold Spring Harb Symp Quant Biol* **79**, 199-210 (2014). https://doi.org:10.1101/sqb.2014.79.024984

31 Grundemann, J. *et al.* Amygdala ensembles encode behavioral states. *Science* **364** (2019). https://doi.org:10.1126/science.aav8736

32 Ji, N. *et al.* A neural circuit for flexible control of persistent behavioral states. *Elife* **10** (2021). https://doi.org:10.7554/eLife.62889

33 Marques, J. C., Li, M., Schaak, D., Robson, D. N. & Li, J. M. Internal state dynamics shape brainwide activity and foraging behaviour. *Nature* **577**, 239-243 (2020). https://doi.org:10.1038/s41586-019-1858-z

34 Hultman, R. *et al.* Brain-wide electrical spatiotemporal dynamics encode depression vulnerability. *Cell* **173**, 166-180 e114 (2018). https://doi.org:10.1016/j.cell.2018.02.012

35 Walder-Christensen, K. *et al.* Electome network factors: Capturing emotional brain networks related to health and disease. *Cell Rep Methods* **4**, 100691 (2024). https://doi.org:10.1016/j.crmeth.2023.100691

36 Strogatz, S. H. *Nonlinear dynamics and chaos: with applications to physics, biology, chemistry, and engineering*. (CRC Press, 2018).

37 Inagaki, H. K. *et al.* A midbrain-thalamus-cortex circuit reorganizes cortical dynamics to initiate movement. *Cell* **185**, 1065-1081 e1023 (2022). https://doi.org:10.1016/j.cell.2022.02.006

38 Vyas, S., Golub, M. D., Sussillo, D. & Shenoy, K. V. Computation through neural population dynamics. *Annu Rev Neurosci* **43**, 249-275 (2020). https://doi.org:10.1146/annurev-neuro-092619-094115

39 Seung, H. S. How the brain keeps the eyes still. *Proc Natl Acad Sci U S A* **93**, 13339-13344 (1996). https://doi.org:10.1073/pnas.93.23.13339

40 Sussillo, D. Neural circuits as computational dynamical systems. *Curr Opin Neurobiol* **25**, 156-163 (2014). https://doi.org:10.1016/j.conb.2014.01.008

41 Remington, E. D., Egger, S. W., Narain, D., Wang, J. & Jazayeri, M. A Dynamical systems perspective on flexible motor timing. *Trends Cogn Sci* **22**, 938-952 (2018). https://doi.org:10.1016/j.tics.2018.07.010

42 Major, G. & Tank, D. Persistent neural activity: prevalence and mechanisms. *Curr Opin Neurobiol* **14**, 675-684 (2004). https://doi.org:10.1016/j.conb.2004.10.017

43 Wang, X. J. & Goldman, M. S. *Neural Integrator Models*. (Elsevier Ltd, 2009).

44 Pereira, U. & Brunel, N. Attractor dynamics in networks with learning rules inferred from in vivo data. *Neuron* **99**, 227-238 e224 (2018). https://doi.org:10.1016/j.neuron.2018.05.038

45 Hopfield, J. J. Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci U S A* **79**, 2554-2558 (1982). https://doi.org:10.1073/pnas.79.8.2554

46 Tsodyks, M. Attractor neural network models of spatial maps in hippocampus. *Hippocampus* **9**, 481-489 (1999). https://doi.org:10.1002/(SICI)1098-1063(1999)9:4<481::AID-HIPO14>3.0.CO;2-S

47 Chaudhuri, R., Gercek, B., Pandey, B., Peyrache, A. & Fiete, I. The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep. *Nat Neurosci* **22**, 1512-1520 (2019). https://doi.org:10.1038/s41593-019-0460-x

48 Hulse, B. K. & Jayaraman, V. Mechanisms underlying the neural computation of head direction. *Annu Rev Neurosci* **43**, 31-54 (2020). https://doi.org:10.1146/annurev-neuro-072116-031516

49 Kim, S. S., Rouault, H., Druckmann, S. & Jayaraman, V. Ring attractor dynamics in the Drosophila central brain. *Science* **356**, 849-853 (2017). https://doi.org:10.1126/science.aal4835

50 Durstewitz, D., Koppe, G. & Thurm, M. I. Reconstructing computational system dynamics from neural data with recurrent neural networks. *Nat Rev Neurosci* **24**, 693-710 (2023). https://doi.org:10.1038/s41583-023-00740-7

51 Gardner, R. J. *et al.* Toroidal topology of population activity in grid cells. *Nature* **602**, 123-128 (2022). https://doi.org:10.1038/s41586-021-04268-7

52 Sylwestrak, E. L. *et al.* Cell-type-specific population dynamics of diverse reward computations. *Cell* **185**, 3568-3587 e3527 (2022). https://doi.org:10.1016/j.cell.2022.08.019

53 Pandarinath, C. *et al.* Inferring single-trial neural population dynamics using sequential auto-encoders. *Nat Methods* **15**, 805-815 (2018). https://doi.org:10.1038/s41592-018-0109-9

54 Nair, A. *et al.* An approximate line attractor in the hypothalamus encodes an aggressive state. *Cell* **186**, 178-193 e115 (2023). https://doi.org:10.1016/j.cell.2022.11.027

55 Linderman, S. *et al.* in *Artificial intelligence and statistics.* 914-922 (PMLR).

56 Linderman, S. W. & Gershman, S. J. Using computational theory to constrain statistical models of neural data. *Curr Opin Neurobiol* **46**, 14-24 (2017). https://doi.org:10.1016/j.conb.2017.06.004

57 Rajan, K., Harvey, C. D. & Tank, D. W. Recurrent Network Models of Sequence Generation and Memory. *Neuron* **90**, 128-142 (2016). https://doi.org:10.1016/j.neuron.2016.02.009

58 Liu, M., Nair, A., Linderman, S. W. & Anderson, D. J. Periodic hypothalamic attractor-like dynamics during the estrus cycle. *bioRxiv* (2023). https://doi.org:10.1101/2023.05.22.541741

59 Liu, M., Kim, D. W., Zeng, H. & Anderson, D. J. Make war not love: The neural substrate underlying a state-dependent switch in female social behavior. *Neuron* **110**, 841-856 e846 (2022). https://doi.org:10.1016/j.neuron.2021.12.002

60 Vinograd, A., Nair, A., Kim, J., Linderman, S. W. & Anderson, D. J. Intrinsic dynamics and neural implementation of a hypothalamic continuous attractor encoding an internal state. *Nature* (2024).

61 Lo, L. *et al.* Connectional architecture of a mouse hypothalamic circuit node controlling social behavior. *Proc Natl Acad Sci U S A* **116**, 7503-7512 (2019). https://doi.org:10.1073/pnas.1817503116

62 Saper, C. B. & Lowell, B. B. The hypothalamus. *Curr Biol* **24**, R1111-1116 (2014). https://doi.org:10.1016/j.cub.2014.10.023

63 Hahn, J. D., Sporns, O., Watts, A. G. & Swanson, L. W. Macroscale intrinsic network architecture of the hypothalamus. *Proc Natl Acad Sci U S A* **116**, 8018-8027 (2019). https://doi.org:10.1073/pnas.1819448116

64 Mountoufaris, G., Nair, A., Yang, B., Kim, D. W. & Anderson, D. J. Neuropeptide signaling is required to implement a line attractor encoding a persistent internal behavioral state. *bioRxiv* (2023). https://doi.org:10.1101/2023.11.01.565073

65 Moffitt, J. R. *et al.* Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science* **362** (2018). https://doi.org:10.1126/science.aau5324

66 Ishii, K. K. *et al.* A labeled-line neural circuit for pheromone-mediated sexual behaviors in mice. *Neuron* **95**, 123-137 e128 (2017). https://doi.org:10.1016/j.neuron.2017.05.038

67 DeNardo, L. & Luo, L. Genetic strategies to access activated neurons. *Curr Opin Neurobiol* **45**, 121-129 (2017). https://doi.org:10.1016/j.conb.2017.05.014

68 Czeh, B., Fuchs, E., Wiborg, O. & Simon, M. Animal models of major depression and their clinical implications. *Prog Neuropsychopharmacol Biol Psychiatry* **64**, 293-310 (2016). https://doi.org:10.1016/j.pnpbp.2015.04.004

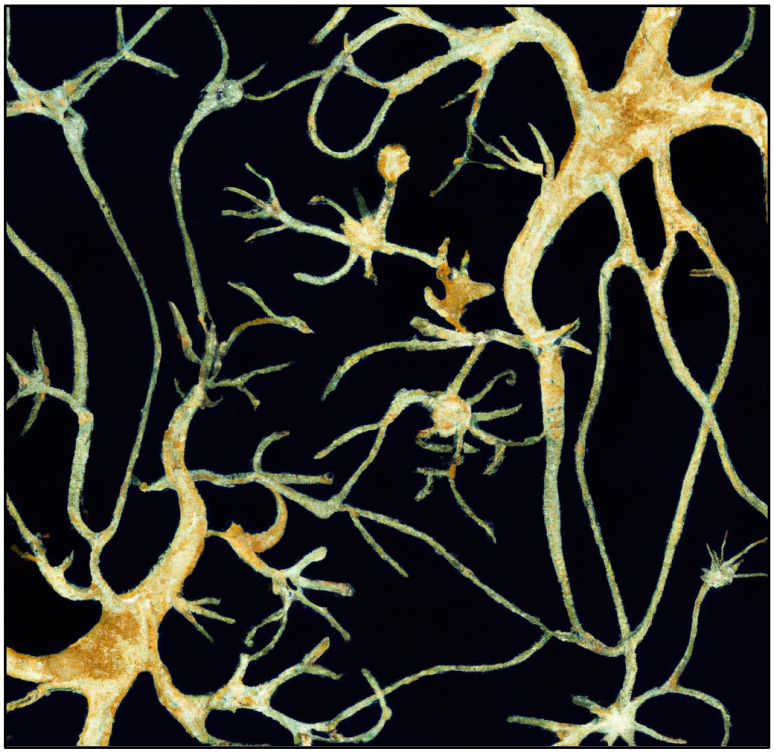
69 LeDuke, D. O., Borio, M., Miranda, R. & Tye, K. M. Anxiety and depression: A top-down, bottom-up model of circuit function. *Ann N Y Acad Sci* **1525**, 70-87 (2023). https://doi.org:10.1111/nyas.14997

70 Rolls, E. T. A non-reward attractor theory of depression. *Neurosci Biobehav Rev* **68**, 47-58 (2016). https://doi.org:10.1016/j.neubiorev.2016.05.007

71 Rolls, E. T., Loh, M. & Deco, G. An attractor hypothesis of obsessive-compulsive disorder. *Eur J Neurosci* **28**, 782-793 (2008). https://doi.org:10.1111/j.1460-9568.2008.06379.x

Chapter II

REPRESENTATION



"ആർക്കും ശ്രദ്ധിക്കാത്ത കാഴ്ചകൾക്കു മുന്നിൽ നിമിഷമെങ്കിലും നിൽക്കുക. അവതന്നെ പറയാൻ എത്രയോ കുറിക്കാത്ത കഥകളുണ്ട്."

M. T. Vasudevan Nair, Naalukettu, 1958

Translation: "Pause for a moment before the sights that no one else notices. They themselves hold countless untold stories."

Chapter II

An approximate line attractor in the hypothalamus encodes an aggressive state

This chapter details the initial study which uncovered a representation of an internal aggressive state as a line attractor using data-driven machine learning.

Published as **Aditya Nair**, Tomomi Karigo, Bin Yang, Surya Ganguli, Mark J. Schnitzer, Scott W. Linderman, David J. Anderson, and Ann Kennedy. An approximate line attractor in the hypothalamus encodes an aggressive state. *Cell* 186, no. 1 (2023): 178-193.

Summary

The hypothalamus regulates innate social behaviors, including mating and aggression. These behaviors can be evoked by optogenetic stimulation of specific neuronal subpopulations within MPOA and VMHvl, respectively. Here we perform dynamical systems modeling of population neuronal activity in these nuclei during social behaviors. In VMHvl, unsupervised analysis identified a dominant dimension of neural activity with a large time constant (>50s), generating an approximate line attractor in neural state space. Progression of the neural trajectory along this attractor was correlated with an escalation of agonistic behavior, suggesting that it may encode a scalable state of aggressiveness. Consistent with this, individual differences in the magnitude of the attractor time constant were strongly correlated with differences in aggressiveness. In contrast, line attractors were not observed in MPOA during mating; instead, neurons with fast dynamics were tuned to specific actions. Thus, different hypothalamic nuclei employ distinct neural population codes to represent similar social behaviors.

Introduction

A fundamental problem in neuroscience is to understand how the brain controls innate behaviors. Many such behaviors are governed by the hypothalamus, a deep subcortical brain region present in all vertebrates[1,2]. Classical brain stimulation and lesion experiments have implicated different hypothalamic regions ("nuclei") in diverse innate behaviors (reviewed in[3-9]). More recently, optogenetic stimulation has identified genetically marked neuronal subpopulations that can evoke such behaviors[10-13] (reviewed in[14-17]). Genetic ablation or reversible silencing has demonstrated that these subpopulations are essential for natural occurrences of these behaviors[10-12,18].

An important open question is how the activity of these neural subpopulations during naturally occurring behavior reflects their "causative" function. Relatively few single unit recordings have been performed in hypothalamic nuclei because of their inaccessibility[13,19-21]. Recordings of bulk calcium signals[22] have confirmed that these neuronal subpopulations are active during the natural behaviors they can artificially evoke[23-25] . However, this averaging method obscures individual cell activity patterns.

Miniature head-mounted microscopes allow calcium imaging with single-cell resolution in freely moving animals[26,27]. Application of this approach to the hypothalamus has identified cells exhibiting stimulus-locked activity during natural behavior[28-30]. For example, imaging of estrogen receptor type 1 (Esr1)-expressing neurons in the medial preoptic area (MPOA), whose optogenetic activation can elicit mounting behavior in male mice[31,32], has revealed cells that respond specifically during spontaneous mounting of females (see also Figure 1E). Such results, together with single-cell transcriptomic analysis, have reinforced the prevailing view that the hypothalamus controls different survival behaviors via genetically determined, functionally specific neuronal subpopulations[33,34].

The case of aggression, however, presents a paradox seemingly at odds with this view. On the one hand, optogenetic stimulation of $Esr1^{+}$ neurons in the ventrolateral subdivision of the ventromedial hypothalamus (VMHvl) neurons triggers attack behavior[12,35-37], identifying these neurons as the likely cellular substrate of electrical brain-stimulated aggression[4,7,38]. Conversely, genetic ablation of VMHvl neurons expressing the progesterone receptor (PR; co-expressed with Esr1) or optogenetic silencing of $VMHvl^{Esr1}$ neurons blocks natural aggression[18],[12] . On the other hand, miniscope imaging of $VMHvl^{Esr1}$ neurons during natural fighting revealed surprisingly few cells that exhibited time-locked, attack-specific activity[29]. Instead most such neurons exhibited "mixed selectivity," responding during different phases of an aggressive interaction. Different subsets of $Esr1^{+}$ neurons responded to male vs. female conspecifics, suggesting an encoding of conspecific sex[29,31,39]. Nevertheless, decoders trained on $VMHvl^{Esr1}$ neural imaging data could accurately distinguish episodes of attack from sniffing[29].

Thus observational vs. perturbational studies of $VMHvl^{Esr1}$ neurons yield seemingly inconsistent views: these neurons causally control aggressive behavior, yet very few of them are specifically "tuned" to attack. There are two possible explanations for this paradox. First, the small fraction of $VMHvl^{Esr1}$ neurons that are more active during attack may be the ones responsible for the specific causative influence of this population. Alternatively, the majority of $VMHvl^{Esr1}$ neurons, despite their mixed behavioral selectivity, may control attack through some type of population code.

In other systems where there is no clear correlation between single-unit spiking patterns and behavior, modeling neural populations as a dynamical system[40-42] (reviewed in[43]) has revealed signals in the dynamics of population activity that can robustly predict motor actions[44,45]. We have therefore carried out similar modeling of $VMHvl^{Esr1}$ neural activity dynamics during naturalistic social behaviors, using legacy data from previous studies[29,31,39]. Our results reveal line attractor dynamics in VMHvl that correlate with escalating levels of aggressive behavior, suggesting they may represent or encode an aggressive internal state.

Strikingly, line attractor dynamics are absent in MPOA activity during both mating and aggression. This analysis therefore reveals fundamental differences in the neural coding of social behaviors by different hypothalamic nuclei.

Results

Cellular tuning analysis confirms behaviorally selective neural populations in MPOA but not in VMHvl

Calcium imaging of MPOA^{Esr1} or VMHvl^{Esr1} neurons revealed distinct patterns of neuronal activation during social interactions[29,31] (Figure 1A, B). To quantify these differences, we re-analyzed calcium imaging data[31] from sexually experienced male C57Bl/6N$^{Esr1\text{-}2A\text{-}Cre/+}$ mice during standard resident-intruder assays, using male or female BalbC intruders (Figure 1C, D). We then computed the mean activity of each neuron during each of 14 different hand-annotated actions and clustered them using a regression model (VMHvl: N= 306 neurons from 3 mice; MPOA: N= 391 neurons from 4 mice, see Methods).

Confirming previous observations[29,31], many MPOA clusters contained neurons only active during specific behavioral actions, such as intromission or mounting towards females (Figure 1E). In contrast, most VMHvl^{Esr1} neurons were activated in response to either males or females, with very few neurons showing behavior-specific activation (Figure 1F).

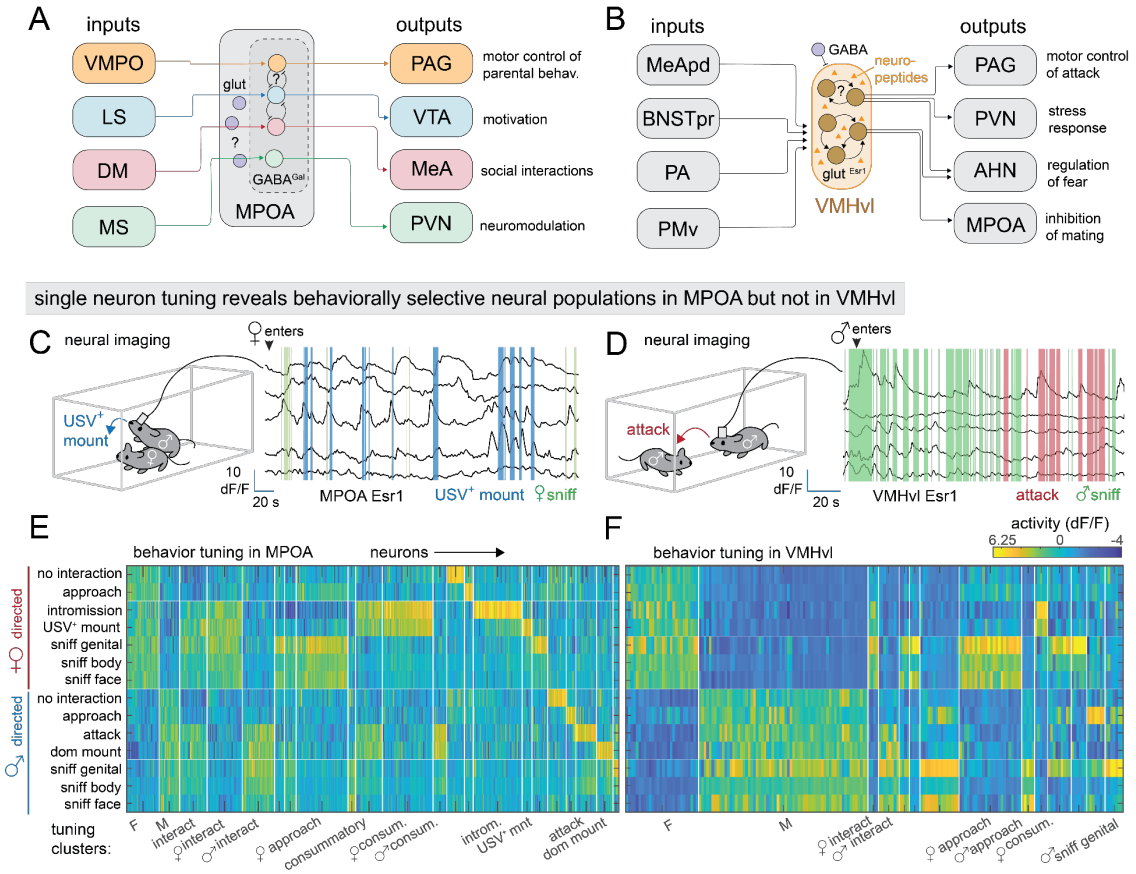


Figure 1: Cytoarchitectures and cellular representations in a neural system regulating social behavior A, B: cytoarchitecture of MPOA (A) and VMHvl (B). C, D: example traces from Esr1+ neurons in MPOA (C) and VMHvl (D). E, F: clustering of recorded Esr1+ neurons in MPOA (E, n =306 neurons from 3 mice) and VMHvl (F, n = 391 neurons from 4 mice) using a regression model. Rows, hand-annotated behaviors; columns, individual neurons.

Unsupervised dynamical systems analysis of neural activity during social behavior

In other systems, population analysis via fit dynamical systems has revealed a neural encoding of behavioral actions that were not apparent in neuron-by-neuron analysis[43,44,46,47]. We therefore investigated whether behavioral representations among VMHvl^{Esr1} neurons might be encoded at a population level, using an unsupervised dynamical systems approach.

To do so, we fit a dynamical model to the population activity of VMHvl^{Esr1} cells from each of multiple mice (n=6), from two different studies[31,39] in which recordings were made throughout male-male or male-female encounters (average duration 5.1± 0.68 min and 11.4± 0.68 min, respectively; mean ± SEM). Specifically, we fit a recurrent switching linear dynamical system (rSLDS) model[48], which approximates a complex non-linear dynamical system as a composite of more easily interpretable linear dynamical systems, or "states" (Supplemental Figure S1A).

rSLDS first reduces neural activity to a set of latent variables (also called "dimensions" or "factors"), defining a low-dimensional "state-space" in which the time-evolving population neural activity vector can be analyzed (Figure 2A①). Population activity in this low-dimensional space is then segmented into a set of discrete states (Figure 2A②), while fitting a linear dynamical system model (Figure 2A③) to neural activity within each state. Each state has a different dynamics matrix, which dictates how neural activity evolves over time from any given point within that state space. Quantitative examination of parameters from this matrix after model fitting can unveil dynamical properties of the neural circuit, such as the time constant of each dimension[42]. Finally, to visualize more easily the dynamical properties of each state, we plotted its "flow field" in 2D using principal component analysis (PCA) (Figure 2A④ right, see Methods).

In fitting the rSLDS model, we chose the minimum number of states and dimensions that could capture 90% of observed variance in neural activity, determined using cross validation in each mouse separately (Supplemental Figure S1B-E; 7-8 dimensions (7.2± 0.1, N=6 mice) and 3-4 states). We evaluated the "goodness of fit" of each model iteration using both the log likelihood of the data[48] and an additional metric that we call the "forward simulation error" (Supplemental Figure S1F, FSE, see Methods). Plotting the FSE over time allows visualization of periods wherein model performance drops (Supplemental Figure S1G). By this metric, our best-fit models captured most of the variance in neural data (model performance (1 – FSE) = 0.72 ± 0.02, N = 6 mice; Supplemental Figure S1H).

The rSLDS framework allows the fit dynamical system models to be either autonomous or to receive external input. Since VMHvl neuron firing rates correlate with the distance to another male or to male mouse urine[49], likely reflecting the concentration of chemosensory cues[50], we used the distance between animals and their facing angle as a proxy for external sensory input strength[49,51] (see Methods).

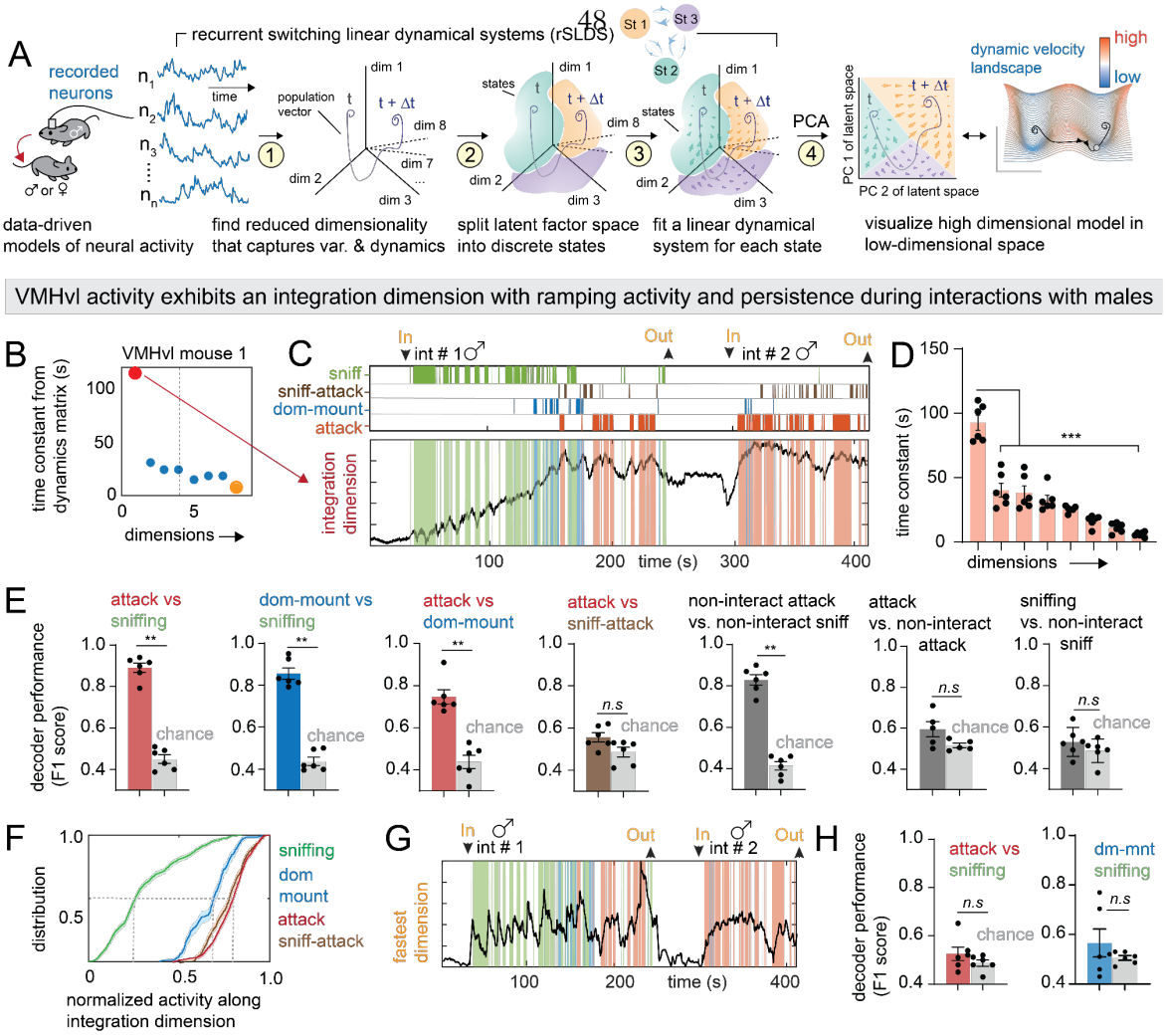


Figure 2: Dynamical analysis of VMHvl neural activity reveals an integrator dimension that correlates with aggressive escalation. A: schematic illustrating rSLDS analysis. B: time constants of rSLDS dimensions (see A①) in attack enriched state from VMHvl mouse 1. Dimensions with longest (red dot) and shortest (yellow dot) time constants are indicated. C: projection onto time axis of integration dimension with overlayed behavior annotations. D: average time constant of all dimensions, arranged in decreasing order. (***$p < 0.001$, n = 6 mice). E: average F1 score of binary decoder of behavior pairs trained on integration dimension activity (**$p < 0.005$, *$p<0.01$, n = 6 mice). F: cumulative distribution of integration dimension value (normalized) for different behaviors. G: projection of fastest dimension in example VMHvl mouse 1. H: performance of binary decoder of behavior pairs trained on fastest dimension activity (n = 6 mice). For additional data see Supplemental Figure S1, S2 and S3.

rSLDS analysis of VMHvl neural activity discovers an integration dimension that correlates with aggressive escalation

Next, we performed retrospective alignment of the unsupervised neural data model with behavioral annotations over time. This comparison revealed that the probability of attack was elevated during a single rSLDS state (state 3, Supplementary Figure S1I-K). Importantly, attacks were not time-locked to the onset/offset of this state; rather epochs of this state outlasted individual attack bouts (state 3 epoch duration: 79.5± 5.5s, attack bout duration: 4.86± 0.44s, N = 6 mice, Supplemental Figure S1I$_5$, J$_3$. K$_3$). This suggests that the state did not simply represent motor activity (Supplemental Figure S1A, cf. Case 2 vs 1).

To understand better the neural population dynamics related to attack behavior, we examined the dynamics matrix for this state, which describes how dimensionally reduced neural activity in that state changes over time. The eigenvalues of this matrix reflect the rate at which activity along each of these dimensions decays to zero following external input, and can be converted to a time constant for each dimension[52,53]. Input to dimensions with short time constants will quickly decay to zero, whereas input to dimensions with long (large) time constants persists and decays slowly. Strikingly, one of the rSLDS dimensions had an estimated time constant of over 100 seconds that was significantly higher than that of all other dimensions (Figure 2B, red dot, 2C, 2D, N =6 mice). Because systems with long time constants approximately integrate their input over time, we refer to the longest time constant dimension as the "integration" dimension[54,55].

The integration dimension accounted for 19.5%± 1.9% of the overall variance in neural activity (N = 6 mice). In contrast a support vector machine (SVM) decoder trained to distinguish attack from sniffing periods explained much less variance (0.3% ± 0.1%, N = 6 mice, $p<0.001$, Supplemental Figure S2B) Examining the activity of individual neurons that were weighted strongly in the integration dimension (Supplemental Figure S2D) revealed that around 20% of neurons per animal contributed to this dimension, with some showing

ramping and persistent activity (Supplemental Figure S2I-J, L, N). Moreover most of these neurons were tuned to male intruders (Supplemental Figure S3A, B). Thus, the integration dimension encapsulates a signal that is present at the level of at least some individual neurons, but is also an emergent property of the population[47].

We next compared the time-varying activity of the integration dimension with the animals' actions during aggressive encounters. In mouse 1, activity along the integration dimension was low during sniffing, ramped up at the onset of dominance mounting (a low-intensity aggressive behavior[31]), and increased further to a stable plateau value as the animal attacked (Figure 2C). A cumulative distribution function (cdf) of the normalized level of activation along the integration dimension during sniffing, dominance mounting, and attack revealed that these three behaviors occurred at low, medium and high values of this dimension, respectively (Figure 2F; distribution means: sniffing: 0.30, dominance mount: 0.66, attack: 0.82, N = 6 mice).

Remarkably, a binary classifier created by thresholding the value of the integration dimension could distinguish periods of sniffing from attack, or from dominance mounting, with a high F1 score (0.89 ± 0.02, N=6 mice, Figure 2E). The same method could also distinguish dominance mounting vs. attack (F1 score 0.74 ± 0.03, N=6 mice, Figure 2E). However such classifiers could not distinguish behaviors occurring close together in time, such as attack and sniff-attack (defined as periods of sniffing that occurred within one second prior to attack, as described recently[56]), perhaps due to the gradual ramping of activity along this dimension. Remarkably, none of the other seven fit dimensions could be used to distinguish aggressive behaviors from sniffing with above chance accuracy (Figure 2G, H; Supplemental Figure S2C).

The foregoing analysis suggested that a low-dimensional signal in VMHvl represents escalating aggressive behaviors. To account for possible spurious behavioral correlations due to the slow decay of activity in this dimension, we devised a version of session permutation

as described recently[57], by cross validating decoder thresholds between animals (see Methods). This more rigorous paradigm could still decode behaviors with high F1 scores (Supplemental Figure S2H).

Sniffing, attack, and dominance mounting are performed in bouts separated by short inter-bout intervals (IBIs). Because of its slow ramping and stable plateau, activity in the integration dimension did not decay during such IBIs and therefore could not distinguish behavioral bouts from adjacent IBIs (Supplemental Figure S1A, Case 1 vs 2). However decoders trained on this activity could distinguish IBIs from sniffing versus attack epochs, which were behaviorally indistinguishable to a human observer, with a high F1 score (0.83 ± 0.02, N=6 mice; Figure 2E, Supplemental Figure S1A, right, Case 2).

Thus, our unsupervised approach uncovered a one-dimensional signal in VMHvl^{Esr1} neural population activity that closely tracks and scales with an animal's escalating level of aggressiveness and is reflected in the activity of approximately 20% of individual VMHvl^{Esr1} neurons. Different aggressive actions are observed as activity along this dimension reaches different thresholds, suggesting an aggression-intensity code in VMHvl^{Esr1} activity. The level of activity along the integration dimension could not be fully predicted from pose features such as the acceleration, facing angle, or velocity of the resident, or from the distance between mice (mean R^2: 0.28 ± 0.04, N= 6 mice, Supplemental Figure S2A). Tracking metric used as inputs to the model were also not predictive of behavior annotations (Supplemental Figure S3F-H). Furthermore, models of VMHvl fit without any tracking inputs also recovered an integration dimension with similar time constants (Supplemental Figure S3D). These results further highlight that the relationship between the integration dimension and escalating aggressive behavior is not due to the incorporation of inputs such as facing angle and distance between mice. Even the incorporation of additional tracking metrics such as speed and area of the ellipse fit to the resident mouse did not improve rSLDS fits, suggesting that VMHvl was likely not integrating features of these sensory related signals (Supplemental Figure 3I).

This relationship between VMHvl^{Esr1} activity and aggression is consistent with our observation that increasing the intensity of optogenetic stimulation of VMHvl^{Esr1} neurons progressively evokes sniffing, dominance mounting and attack[12], actions that can be decoded from the integration dimension as its activity ramps up.

VMHvl contains an approximate line attractor that represents escalating aggressiveness

We examined next how the integration dimension of the fit model influences the overall topology of neural state-space during social behavior (Figure 3A, see Methods). PCA indicated that the first two PCs accounted for 68.5%± 1.2% of the total variance in VMHvl activity (N=6 mice). In all imaged animals, PC1 showed slow ramping dynamics (Figure 3A, Supplemental Figure S4C, PC_1 (behavior-triggered average, N = 6 mice). We confirmed that the rSLDS integration dimension makes the largest contribution to this PC (Supplemental Figure S4A). Activity along PC2 was high when a new intruder was introduced (Figure 3A, Supplemental Figure S4C, PC_2 (behavior-triggered average, N = 6 mice)), but was otherwise low.

To visualize neural state space dynamics, we next generated a 2D flow field in PC space, whose vectors at each point indicate how neural dynamics evolve according to the fit rSLDS model (see Figure 2A4). This revealed a region of low vector flow that forms an approximate line attractor (Figure 3B, *right*, 3C), meaning that the neural population activity vector tends to move towards persistent points along a line[58] (Figure 3D, t_{50}-t_{340}).). To quantitatively delimit this attractor, we calculated the points in the flow field where vector length is at a minimum ("slow points;" see Methods) and linked these points into a dashed line (Figure 3D, dashed black line). Such approximate line attractors were observed in multiple mice (Figure 3E, F and Supplemental Figure 4D, E). Importantly, these line attractors are largely

aligned with the PC1 axis, which principally reflects variance in the slow integration dimension identified by rSLDS (Figure 2B, Supplemental Figure S2A).

To quantitatively test for the existence of a line attractor in each mouse, we devised a "Line Attractor Score" as the base-2 log of the ratio of the largest to the second-largest time constants of the eight rSLDS dimensions (Figure 2C). According to basic concepts in dynamical systems theory[52], this ratio has a relatively high value in systems containing a single integration dimension (forming an approximate line attractor), and is otherwise close to zero. We find that all mice with VMHvl recordings possess a line attractor score greater than zero, indicating the presence of a line attractor (Fig 3G, n = 6 mice).

As population activity progressed along the line attractor from low to high values of PC1, behavior progressed from sniffing to dominance mounting to attack (Figure 3D-F, 3L and Supplementary Video 1). This reflects the "ramping up" of activity seen in the integration dimension as social behavior progresses through these phases (Figure 2B), suggesting an encoding of an underlying continuous variable, as seen in line attractors in other regions[42,46,59-61].

To visualize the dynamical topology of the rSLDS model, we represented the 2D flow field as a 3D landscape, by converting the length of the flow-field vectors at each position in neural state space into the height (z-axis) of the landscape (Figure 3B); the x-y axes are still represented by PC1 and PC2. In this topographic representation, a line attractor appears as a region shaped like a trough or gully, reflecting a slow rate of change (short vectors). A point attractor would appear as a locus of slow rate of change at the base of a cone (Figure 3B[59]). We observed a trough-like structure in the 3D dynamics landscape in each imaged animal (Figure 3H-K, Supplemental Figure S4P), along which neural activity progressed slowly as aggression escalated (Supplementary Video 2). Consistent with the persistent, slow-decaying activity characteristic of "leaky" neural integrators[55,59], VMHvl activity remained high following intruder removal, and slowly decayed along the trough of the attractor over tens of seconds (Supplemental Figure S4H-J).

Although the animals' behavior appears to occur while the system is in the line attractor, it could be that other rSLDS dimensions also show a change in their activity during behavior. To test this possibility, we computed each behavior's "dynamic velocity" by calculating the average vector length across all eight rSLDS dimensions at all time points in which a given behavior occurred. (Figure 3M, see Methods). Time points associated with initial intruder entry had the highest dynamic velocity and were present on the walls of the trough (Figure 3H-K), whereas aggressive behaviors exhibited low dynamic velocities and were distributed along the base of the trough (Figure 3N).

Once the system is in the line attractor, input that is not aligned with the attractor should produce a transient excursion of the population activity vector out of the trough; however once that input decays the vector should move back into the trough close to where it started from[42] (Supplemental Figure S4F). We tested this prediction using a subset of experiments in which one intruder male was removed, and a second male introduced 30-60 seconds later. Strikingly, the introduction of a new intruder male drove a rapid rise in neural firing rates that pushed VMHvl activity away from the trough of the line attractor (Figure 3C-E, intruder #2). However, this signal decayed relatively quickly and the system re-entered the line attractor at nearly the same point (Supplemental Figure S4F-J and Supplemental Video 1). Importantly, the system recovered to the point in the attractor where it had been prior to introduction of the second intruder, regardless of when in the trial the first intruder was removed (Supplemental Figure S4K).

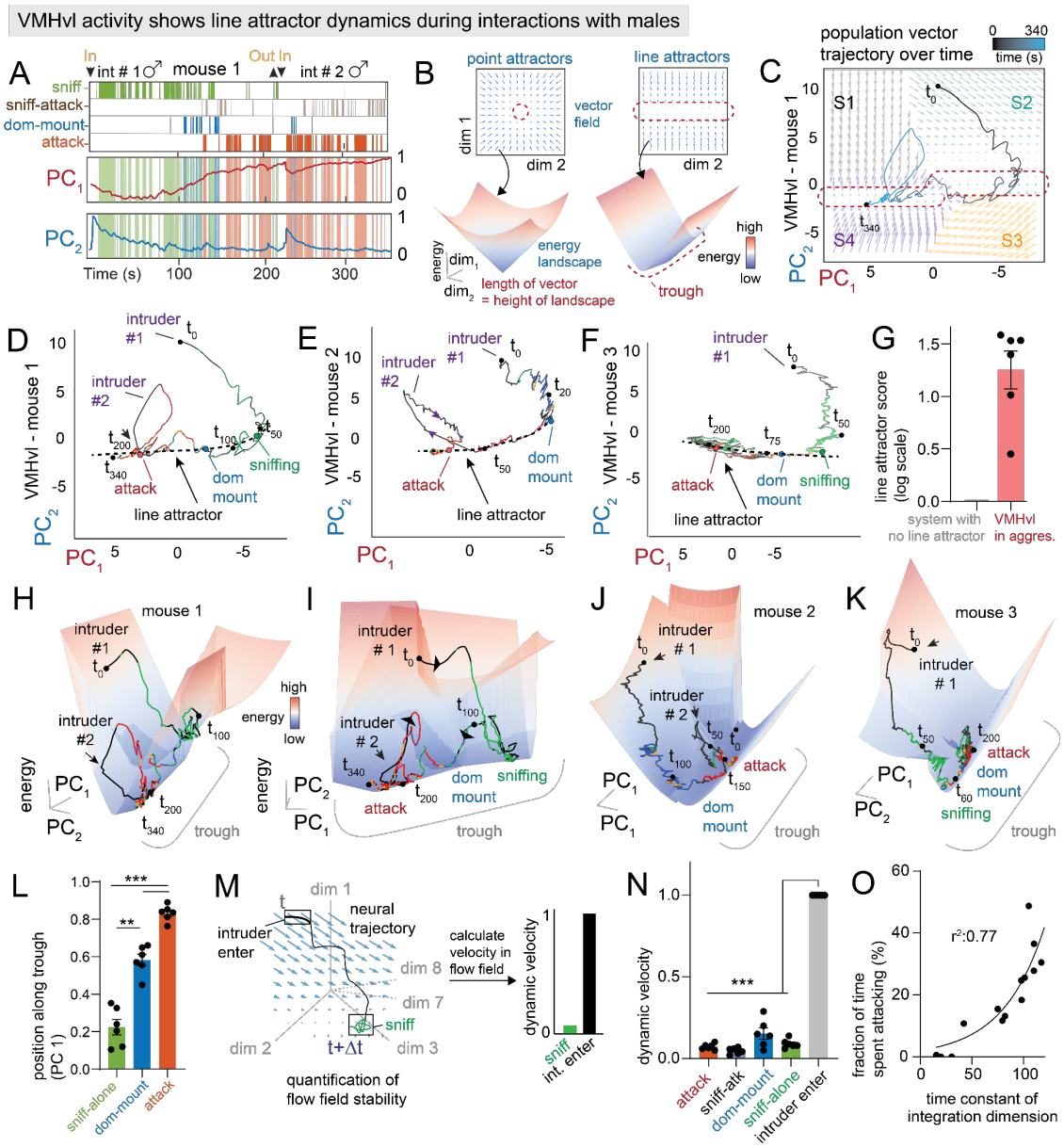


Figure 3: VMHvl contains an approximate line attractor that integrates aggressive escalation. A: behavior rasters shown with first two principal components of dynamical system (see Methods) for example VMHvl mouse 1. B: inferred dynamics shown as a flow field (with attractor highlighted) and 3D landscape for point attractors (left) and line attractors (right). C: neural state space with population trajectories and inferred flow field colored by rSLDS states for VMHvl mouse 1, with line attractor highlighted. D-F: neural state space for VMHvl mouse 1 (D), mouse 2 (E) and mouse 3 (F)

with line attractor highlighted (see Methods). G: line attractor score (see Methods) for VMHvl (red bar, n = 6 mice). H,I: inferred 3D dynamic landscape in VMHvl mouse 1(H,I). J,K: Same as H but for VMHvl mouse 2 (J) and mouse 3 (K) L: position of various behaviors along trough, i.e, PC1 in neural state space (n = 6 mice, **p<0.005, *p<0.01) M: schematic showing quantification of dynamic velocity. N: dynamic velocity for various behaviors in VMHvl (***p<0.001, n = 6 mice) O: relationship between the time spent attacking and the time constant of the integration dimension of individual mice (r2: 0.77, n = 14 animals). For additional data see Supplemental Figure S4.

The time constant of the integration dimension in VMHvl predicts levels of aggressiveness across animals

Although VMHvl^{Esr1} imaging data from different mice always revealed a single integration dimension with a long time constant, the magnitude of this time constant varied across individuals. Unexpectedly, we observed a trend in which animals that displayed more aggressive behavior (calculated as the fraction of time spent attacking) also exhibited an integration dimension with a longer time constant (Figure 3O, $r^2 = 0.77$, n = 14 animals). This relationship held for imaging data from different studies[29,31,39] using different versions of GCaMP (6s vs. 7f; Supplemental Figure S4L-O). This striking correlation of integration time constant with time spent attacking suggests that individual differences in aggressiveness may be reflected in the intrinsic dynamics of VMHvl^{Esr1} neurons.

Mating behaviors are represented using rotational dynamics in the MPOA

Since rSLDS was able to uncover evidence for integration in VMHvl, we next examined whether the same analysis would uncover population dynamics important for mating in MPOA, by fitting models to MPOA^{Esr1} neural data recording during interactions with female intruders (Karigo et al., 2021).

Fit models of MPOA required three rSLDS states in every animal, with mounting and intromission mostly occurring in single but different states (Supplemental Figure S5A-J). Unlike in VMHvl, the bout length of mating behaviors was similar to that of the

corresponding state (Supplementary Figure S5 D, E). Strikingly, the eigenvalues of the dynamics matrix for such states did not include dimensions with long time-constants (Figure 4A). Instead, the first two PCs of the fit model revealed fast dynamics that were highly correlated with specific behaviors (Figure 4B). PC1 peaked at the onset of USV^{+} mounting bouts, while PC2 peaked during intromission (Figure 4B, C, behavior triggered average, N = 3 mice).

The 2D flow-field in PCA space revealed that neural dynamics were dominated by a rotational flow, with activity during mating epochs exhibiting periodic orbits (Figure 4D, F). The phase of the rotations was correlated with progression through sniffing, mounting, and intromission (Figure 4D, F, Supplemental Figure S5K-M), and corresponded to the sequential activation of different neurons during these successive behaviors (Figure 4E, G, Supplemental Figure S5A). Accordingly, the "sequentiality index" of the data[62] was significantly greater than shuffled data or random matrices of similar sizes (seq. index = 0.22 ± 0.01, N = 3 mice, shuffle seq. index = 0.10± 0.002, N = 3 mice, Figure 4H).

We assessed the relationship between the phase of rotational trajectories and behavior by calculating the angle of the population activity vector relative to its value at the start of sniffing (Figure 4I). This revealed that sniffing, mounting and intromission occurred at characteristic angles of the population vector (sniffing: 18.6° ± 6.2 °, mounting: 79.61°±13.6°, intromission: 132.2°± 8.1 °, N = 3 mice; Figure 4J). High dynamic velocities were associated with mounting and intromission, in striking contrast to the low dynamic velocity during attack behavior in VMHvl (Figure 3N, 4K).

To quantitatively assess the presence of line attractor dynamics, we computed the Line Attractor Score for MPOA. These values were close to zero and significantly different from those for VMHvl during aggression (Figure 4L). Thus, unlike the slow ramping and persistent dynamics identified in VMHvl, rSLDS discovered fast, sequential and behaviorally time-locked rotational dynamics in MPOA.

A direct comparison of key quantitative dynamics parameters highlights the key differences between VMHvl and MPOA (Figure 5A-D, G-H). Nevertheless in both regions evolving behavior tracks a single continuous variable: the value of the integration dimension in VMHvl, and the angle of the orbit in MPOA (Figure 5E, F). These variables are instantiated as a line attractor vs. rotational flow, respectively (Figure 5I, J).

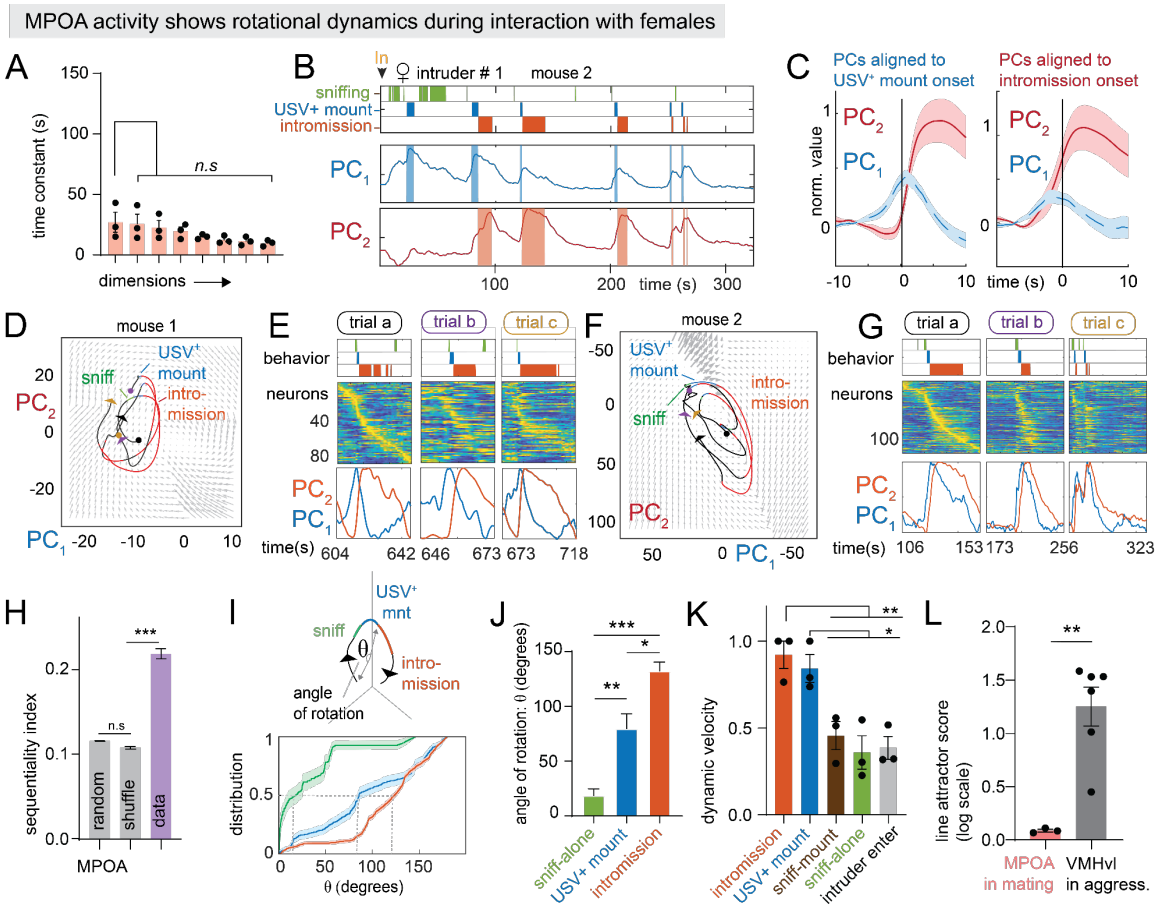


Figure 4: Mating behaviors are represented using rotational dynamics in the MPOA. A: time constants of rSLDS dimensions in mating behavior-enriched state in MPOA (n = 3 mice) B: behavior rasters shown with first two principal components of latent factors for example MPOA mouse 2. C: behavior triggered average of top two principal components aligned to USV+ mount onset (left) and intromission (right) onset (n = 3 mice). D: neural state space with rotational population trajectories from mating episodes shown in E of MPOA mouse 1, colored by behaviors performed by resident mouse. E: sequential activity of MPOA neurons during mating episodes whose rotational population trajectories are shown in D. F,G: same as D,E but for MPOA mouse 2. H: sequential index for MPOA (n = 3 mice, ***p<0.001). I: calculation of angle of rotation (θ) aligned to the start of sniffing during mating episodes (top). Empirical cumulative distribution of θ for various behaviors (n = mice, bottom). J: quantification of θ for various mating behaviors (n = 3 mice, ***p<0.001, **p<0.005, *p<0.01, top). Schematic depicting θ for mating behaviors (bottom). K: dynamic velocity for mating

behavior in MPOA (n = 3 mice).. L: line attractor score for MPOA activity in mating behaviors towards females (left, pink bar, n = 3 mice) and VMHvl activity in aggressive behavior towards males (right, grey bar, n = 6 mice, **p<0.005, data from Fig 3G reproduced for comparative purposes). For additional data see Supplemental Figure S5.

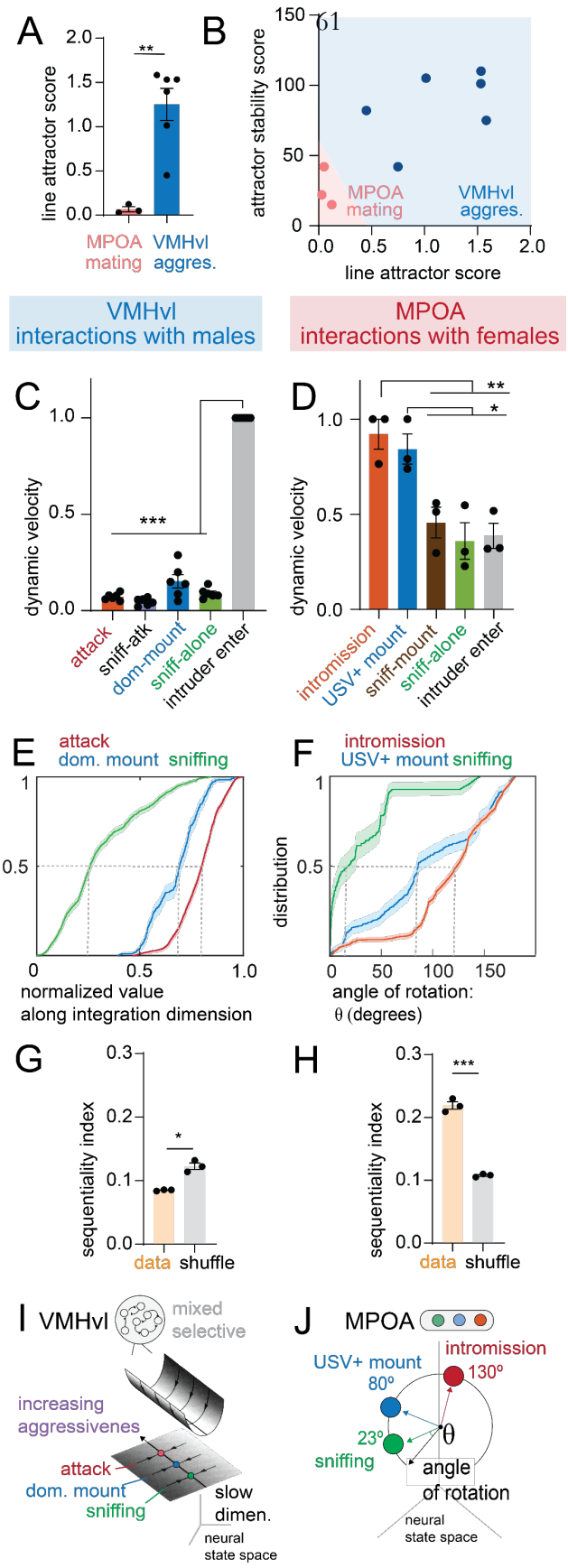

A
line attractor score
2.0
1.5
1.0
0.5
0.0
**
MPOA mating
VMHvl aggres.
B
attractor stability score
150
100
50
0
MPOA mating
VMHvl aggres.
0.0 0.5 1.0 1.5 2.0
line attractor score
VMHvl interactions with males
MPOA interactions with females
C
dynamic velocity
1.0
0.5
0.0
***
attack
sniff-atk
dom-mount
sniff-alone
intruder enter
D
dynamic velocity
**
*
intromission
USV+ mount
sniff-mount
sniff-alone
intruder enter
E
attack
dom. mount
sniffing
distribution
1
0.5
0 0.5 1.0
normalized value
along integration dimension
F
intromission
USV+ mount
sniffing
distribution
0 50 100 150
angle of rotation:
θ (degrees)
G
sequentiality index
0.3
0.2
0.1
0.0
*
data
shuffle
H
sequentiality index
***
data
shuffle
I
VMHvl
mixed selective
increasing aggressivenes
attack
dom. mount
sniffing
slow dimen.
neural state space
J
MPOA
intromission
USV+ mount
80°
130°
23°
sniffing
θ
angle of rotation
neural state space

Figure 5: Distinct neural coding schemes for similar behavior in VMHvl vs MPOA. A: line attractor score for mating behavior in MPOA and aggressive behavior in VMHvl (n = 3 mice for MPOA, n = 6 mice for VMHvl), reproduced from Figure 4L. B: scatter plot for line attractor score versus attractor stability score (magnitude of largest time constant) separates VMHvl and MPOA. C,D: dynamic velocity score in VMHvl during aggression (C) and MPOA during mating (D), reproduced from Figure 3N and Figure 4K respectively. E: empirical cumulative distribution of value of integration dimension (normalized) in VMHvl for various aggressive behaviors, reproduced from Figure 2F. F: empirical cumulative distribution of angle of rotation (normalized) in MPOA for various mating behaviors , reproduced from Figure 4I. G,H: Sequentiality index in MPOA (n = 3 mice), reproduced from Figure 4E, and in VMHvl (H) in aggression (n = 3 mice). I: summary of line attractor dynamics in VMHvl. J: summary of rotational dynamics in MPOA.

VMHvl exhibits an approximate line attractor encoding reproductive behavior

The foregoing findings raised the question of whether the contrasting dynamics in VMHvl vs. MPOA reflect differences specific to aggression vs. mating, or rather generic differences in behavioral coding between these nuclei. To address this, we fit rSLDS models to $VMHvl^{Esr1}$ and $MPOA^{Esr1}$ neuronal activity during mating vs. aggression, respectively.

Models fit to VMHvl activity during male-female encounters yielded a single integration dimension with a long time constant, created by neurons that displayed ramping and persistent activity (Figure 6A red dot, 6B, 6D, Supplemental Figure S6L). In addition, the duration of the rSLDS-discovered mating states in VMHvl tended to outlast individual bouts of mating actions (Supplemental Figure S6D, H), similar to the case of aggression in VMHvl (Supplementary Fig. 1I-K).

The cumulative distribution of the value of the integration dimension during various behaviors revealed that sniffing occurred at the lowest values, USV^{+} mounting at intermediate values, and intromission at the highest values of this dimension (Supplemental Figure S6J). Strikingly, pairwise decoders trained on this dimension performed with high

accuracy (intromission vs sniffing: F1 = 0.92± 0.01 N = 4 animals; mounting vs sniffing: F1=0.81± 0.02 N = 6 animals Figure 6C). Such decoders could also distinguish periods of non-interaction between mounting bouts from those between sniffing bouts (Supplementary Figure S6I). Thus, VMHvl^{Esr1} neuronal dynamics during mating resembled those exhibited during aggression. However, the integration dimension seen during mating was biased towards neurons tuned to female intruders[29], while male-tuned neurons primarily contributed to this dimension during aggression (7.73% ± 0.8% overlap, n = 6 mice, Supplemental Figure S6M, Supplemental Figure S3A,B).

As for aggression, a single dimension of the rSLDS model for mating exhibited a long time constant, yielding a high Line Attractor Score (Figure 6D, H). The first two PCs of the fit model were similar to those seen during aggression, with PC1 exhibiting ramping during the progression from sniffing to mounting to intromission (Figure 6E). Examination of the underlying 2D vector flow field revealed an approximate line attractor (Figure 6F) and a corresponding trough shape in the 3D dynamic velocity landscape, with neural activity moving along the trough as the animal progressed from the appetitive to consummatory phases of mating (Supplemental Figure 6K). Transient movements out of the line attractor occurred only during the introduction of a new intruder and were aligned with PC2 (Figure 6F, intruder #2). Accordingly, periods during intruder entrance had high dynamic velocities, while mating behaviors had low dynamic velocities (Figure 6G).

Thus rSLDS modeling of VMHvl^{Esr1} neuronal activity during mating revealed an approximate line attractor, with many features similar to those observed during aggression. However, the mating and aggression line attractors incorporate primarily female- vs. male-selective neurons, respectively (Supplemental Figure 6M). These data suggest that line-attractor dynamics are a general feature of social behavior coding in VMHvl, rather than a unique signature of aggression *per se*.

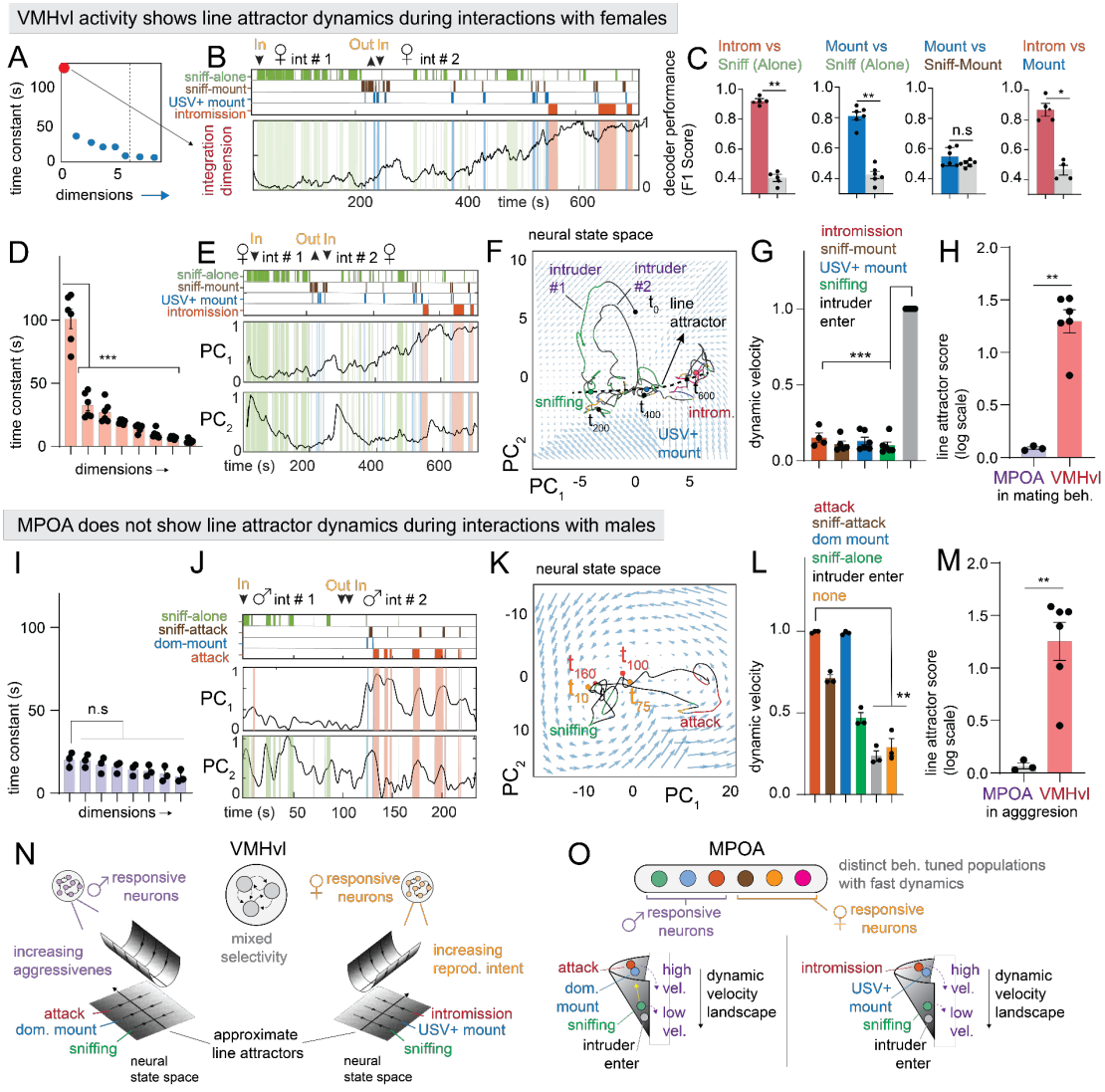


Figure 6: Distinct coding schemes of VMHvl and MPOA are region-specific, not intruder specific. A: left: time constants of rSLDS dimensions of mating enriched state from example VMHvl mouse 1. The red dot highlights the integration dimension. B: projection of integration dimension with overlayed behavior annotations. C: F1 score for decoding behavior pairs from integration dimension (**$p < 0.005$, *$p<0.01$, $n = 4$ mice for comparisons involving intromission as only 4/6 mice showed this behavior. $n = 6$ mice for all other comparisons). D: time constant arranged in decreasing order. ($p < 0.001$, $n = 6$ mice). E: behavior rasters shown with PCs of dynamical system for example VMHvl mouse 1. F: neural state space with population trajectories for VMHvl mouse 1 colored by behavior annotations and flow field showing a line attractor. G: quantification of dynamic velocity during

mating behavior in VMHvl ($p<0.001$, n = 6 mice). H: line attractor score for MPO (n = 3 mice) and VMHvl (n = 6 mice) during mating behavior with females (**$p<0.005$). I: time constants of rSLDS dimensions from MPOA during aggression. J: behavior rasters shown with PCs of dynamical system for example MPOA mouse 1. K: neural state space with population trajectories for MPOA mouse 1 colored by behavior annotations and flow field. L: dynamic velocity during aggressive behavior in MPOA (**$p<0.005$, n = 3 mice). M: line attractor score for MPO (n = 3 mice) and VMHvl (n = 6 mice, reproduced from Fig 3G) during aggressive behavior (**$p<0.005$). N: Schematic illustrating two line attractors discovered in VMHvl encoding aggressiveness and mating intent. 0: Schematic illustrating dynamics seen in MPOA showing similarity in stability of behaviors during interactions with males and females. For additional data see Supplemental Figure S5.

MPOA does not exhibit line attractor dynamics during aggression

Finally, we fit rSLDS models to MPOA^{Esr1} neuronal dynamics during male-male encounters. Analogous to the case of mating behaviors, we found a state (state 3) that is closely aligned to the onset and offset of attack behavior (Supplementary Figure S6N). No dominant "slow" dimension was apparent in the time constants of the rSLDS dimensions (Figure 6I, M). Reflecting this, PC1 of rSLDS state space exhibited a fast increase in activity at the onset and offset of attack (Figure 6J, Supplemental Figure S6O, blue trace), in contrast to the slow attack-related dynamics in VMHvl (Supplemental Figure S6O, red trace).

Visualizing the 2D MPOA flow field in PC space revealed little change in the population trajectory during investigation (Figure 6K). During attack bouts, activity showed excursions into a separate region of state space, but quickly returned to the "sniffing" region after fighting (Figure 6K), reflecting the activation of different neuronal subsets (Figure 1E). Accordingly, attack and dominance-mounting had high dynamic velocities in MPOA, rather than the low dynamic velocities in VMHvl (Figure 6L, Supplemental Figure S6P). Lastly, the Line Attractor Score in MPOA during aggression had a value close to zero and was significantly different from that of VMHvl (Figure 6M, n = 3 for MPOA, n = 6 mice for VMHvl), confirming the absence of line attractor dynamics.

In MPOA, therefore, we find a representation of male-male encounters that alternates between investigatory and aggressive states, with the latter largely time-locked to the onset and offset of attack bouts. Strikingly, MPOA activity during aggression lacks the persistence, ramping and line attractor dynamics seen in VMHvl. Together with our analysis of VMHvl activity during mating, these results support the conclusion that MPOA and VMHvl exhibit fundamentally different coding of the same social behaviors.

Discussion

MPOA and VMHvl control social behaviors using different population codes

Here we report that MPOA^{Esr1} and VMHvl^{Esr1} neurons utilize very different schemes for the neural coding of mating and aggression, despite the fact that optogenetic perturbation specifically elicits mating in MPOA^{Esr1} neurons and attack in VMHvl^{Esr1} neurons. GCaMP imaging of Esr1^{+} neurons in MPOA indicates that specific actions can be decoded according to which cells are active[31], consistent with transcriptomic studies[33]. In contrast, most VMHvl^{Esr1} neurons exhibit mixed behavioral selectivity in both imaging and transcriptomic studies[29,63]. Thus, MPOA represents behavior via a cell identity code, while VMHvl apparently does so via population coding.

Our studies suggest a possible mechanism underlying this population code. rSLDS analysis of VMHvl neural activity during male-male social interactions revealed one dimension of neural activity with a long time-constant that exhibits progressively increasing activity during escalating aggressive encounters. In a topological representation, these dynamics can be visualized as a progression along a stable "trough" or gully, which has the characteristics of an approximate line attractor[59]. In contrast, rSLDS analysis of MPOA revealed rotational dynamics, generated by the sequential activity of behavior-specific cell types during each bout of mating. Put simply, VMHvl coding of behavior appears to be analog, while MPOA coding of behavior appears more digital.

In other neural systems, line attractors often encode a continuous, low-dimensional variable[42,46]. Here, this variable may correspond to the intensity of an aggressive internal state. VMHvl neurons have previously been implicated in the motivation to engage in fighting, as operationalized using instrumental conditioning assays[23]. However, such assays cannot measure aggressive motivation during attack itself, for technical reasons. The escalating (scalable) nature of aggression has ethological relevance as a means of

establishing dominance while minimizing the risk of injury[64]. Unexpectedly, in comparing data across multiple animals we discovered a strong positive correlation between each mouse's level of aggressiveness and the magnitude of the time constant of its integration dimension. This result reveals a neural correlate of individual differences in aggressiveness within VMHvl.

The different neural codes for social behavior we have uncovered in VMHvl and MPOA may reflect their distinct neurochemical and cytoarchitectonic features. VMH neurons are primarily glutamatergic. Recurrent connectivity among excitatory neurons is often invoked as a mechanism to achieve persistent activity[54,55,59]. Indeed, there is evidence that glutamatergic neurons in VMHdm that encode persistent defensive behaviors exhibit local connectivity[65]. However, slow dynamics can also be achieved using neuromodulatory signaling, and there is indirect evidence for peptidergic transmission in VMHvl[63,66]

By contrast, MPOA neurons are 85% GABAergic; to our knowledge there is no way to achieve similar graded and persistent signals within a population of inhibitory neurons. However, GABAergic neurons could provide a substrate for reciprocal inhibitory connections between action-specific subpopulations. Such connectivity could produce winner-take-all dynamics or feed-forward dis-inhibitory circuits that control transitions between sequential action phases of mating, e.g., from sniffing to mounting[39], giving rise to the rotational dynamics observed in neural data. The existence of such circuits in MPOA can be investigated using slice physiology or in vivo imaging once genetic access to the appropriate cell types is achieved.

Why should MPOA and VMHvl utilize such different strategies for the coding of closely related social behaviors? It is tempting to attribute this difference in population dynamics to distinct features of reproductive vs aggressive behavior. For example, aggressive encounters can dynamically escalate or de-escalate to avoid serious injury or death to the combatants, whereas male mating must proceed to completion (ejaculation) to be reproductively

beneficial. These differences are well-suited to control by ramping and rotational neuronal dynamics, respectively. In this view, the different properties and coding strategies of VMHvl and MPOA may have evolved to be optimally adaptive for fighting and mating, respectively.

However, our analysis also revealed approximate line attractor dynamics in a subset of $VMHvl^{Esr1}$ neurons that is female-tuned and active during mating[29]. This suggests that line attractor-like dynamics are a general property of behavioral coding by VMHvl, not an aggression-specific feature. Conversely, MPOA contains specific $Esr1^{+}$ neurons highly tuned to attack which do not exhibit line attractor dynamics (although there is no evidence that these neurons play a causative role in aggression). These data suggest that MPOA and VMHvl more likely encode different features of a given social behavior, such as action selection vs. motive state intensity, respectively. If so, then by extension the hypothalamus may contain GABAergic populations that control action-selection during aggression. Indeed, the anterior hypothalamic nucleus (AHN), which has a similar neurochemical and cytoarchitectonic structure as MPOA, can promote defensive attack[67,68]; it will be interesting to see whether rotational dynamics are observed in this structure. By the same token, PMv which controls aggression and is also primarily glutamatergic[50,69,70], may utilize population coding like VMHvl.

Potential functions of the VMHvl line attractor

Line attractors have been identified in cortical and hippocampal regions involved in cognitive functions, such as decision-making, spatial mapping and sensory discrimination[42,46]. It is unexpected to find such neural dynamics in the hypothalamus, which is widely viewed as controlling innate behaviors via action-specific cell types (as observed in MPOA[33]). What function(s) might such attractor dynamics serve, in the context of innate behaviors? Two explanations are possible, which are not mutually exclusive.

As mentioned earlier, progression along the line attractor may encode the intensity of an internal motive state of aggressiveness. This is supported by our finding that the integration dimension that contributes to this attractor can distinguish periods of non-social interaction during high- vs. low-intensity phases of aggressive escalation (Figure 2D). In this view, the line attractor serves to maintain the system in a stable internal motive state that persists continuously during stochastic expressions of observable attack.

Previous studies have indicated that the greatest source of variance in VMHvl^{Esr1} neural activity is intruder sex[29]. Whether VMHvl encodes intruder sex *per se*, or an internal motive state tightly correlated with intruder sex, has been difficult to distinguish because males only attack other males and not females. In female mice, however, lactating mothers attack intruders of both sexes. Recently, we identified a subset of VMHvl^{Esr1} neurons in females that express the GPCR gene *Npy2r*, called ß cells, which are both necessary for maternal aggression and sufficient to promote attack in non-aggressive virgins [37]. Bulk calcium measurements revealed that ß cells are strongly active during maternal aggression towards both male and female intruders. However, these cells display low activity in individual females that are non-aggressive [37]. Thus in females the encoding of aggressive state by VMHvl^{Esr1} neurons can be decoupled from the encoding of intruder sex. These data reinforce the idea that in males, the VMHvl^{Esr1} line attractor (which reflects a dimension weighted primarily by male-selective neurons) encodes aggressiveness, rather than simply intruder sex.

An alternative function for the line attractor is that it may serve as an integrator that accumulates "evidence" used to make behavioral decisions, such as the decision to switch from sniff to dominance mount, or from dominance mount to attack. Such a function would require that different behaviors be triggered at different threshold values of the integrator. This type of ramp-to-threshold mechanism has been suggested to control sequential actions during male courtship behavior in *Drosophila* [71] and predator escape in mice [72]. These two

hypotheses are not incompatible: the attractor could encode both the intensity of an internal state, and (indirectly) the selection of actions at different state intensities.

Line attractor dynamics could also serve useful functions in the context of behavioral plasticity and individual variation. For example, VMHvl^{Esr1} neurons show increased selective tuning for male vs. female intruders as a function of social experience [29], and exhibit a form of long-term potentiation that underlies the increase in aggressiveness that occurs when mice win a series of fights [73]. It will be interesting to determine whether changes in flow field dynamics or attractor properties are associated with these forms of experience-dependent plasticity. Finally, we note that differences in line attractor properties were observed among mice which exhibited different and characteristic levels of aggressiveness (Figure 3O). It is possible that individual differences in aggressiveness may reflect, or be caused by, individual constraints on population dynamics in VMHvl.

Testable predictions of the line-attractor model

Our rSLDS model of VMHvl dynamics makes several testable predictions and raises several interesting questions for future investigation. First, it predicts that once in the attractor, the system will return quickly to it following perturbations that move it out of this stable trough. This behavior is suggested by the brief excursion out of the attractor that occurs when a first intruder is removed and replaced by a second one. However, it would be ideal to demonstrate this directly by transiently activating neurons that contribute to the attractor, and determining whether the system rapidly returns to it following stimulus offset, as has been demonstrated for point attractors underlying working memory in ALM [74]. Another prediction is that selectively inactivating the VMHvl^{Esr1} neurons that exhibit slow dynamics should eliminate activity along the line attractor. Such experiments will require combined optogenetic perturbations and calcium imaging in this deep subcortical structure. Such experiments will also be critical to confirm whether line attractor properties indeed play a causative role in controlling levels of aggressiveness.

The results herein show that about 20% of VMHvl^{Esr1} neurons exhibit persistent activity and ramping dynamics, raising the question of whether these cells constitute a genetically determined subpopulation. Single-cell RNAseq experiments have shown that the Esr1^{+} population in VMHvl can be subdivided into 6-7 distinct transcriptomic subtypes [63]. Whether any of these subtypes selectively contributes to attractor dynamics can be addressed once genetic drivers specific for these subtypes are available. An additional question is whether the slow dynamics observed for some VMHvl^{Esr1} neurons reflects recurrent connectivity between them, as has been demonstrated for fear-encoding neurons in VMHdm [65], or the release of slow-acting neuromodulators such as neuropeptides. Recurrent connectivity in VMHvl can be investigated by slice electrophysiology[66] and ultimately by EM connectomics. VMHvl^{Esr1} neurons are known to express multiple neuropeptides, as well as receptors for neuropeptides and other neuromodulators. New sensors for detecting neuromodulator release [75,76], as well as methods for dynamically perturbing neuromodulator function *in vivo*, should help to address these questions in the future.

Limitations of the study

Our discovery of line attractor dynamics in VMHvl derives from quantitative analysis of a dynamical system model fit to neural data. While this analysis has revealed several conditions required for line attractor dynamics, such as persistence in the absence of input and robustness to behavioral perturbation, a definitive test requires experimental perturbation of neural activity[58]. Perturbations are also required to determine the contributions to line attractor dynamics of region-intrinsic vs extrinsic (i.e, via other nuclei) recurrent dynamics and feedback, as well as whether the attractor is truly "autonomous" and not input-driven. The biological line attractor in VMHvl is a 'leaky' approximation of a mathematically defined line attractor, exhibiting slow decay over time scales similar to line attractors discovered in other neural systems[46,59]. Further knowledge of the underlying neural

mechanisms is required to understand the extent to which the region of stability identified here approximates a true line attractor.

Experimental model and subject details

Neural imaging data (Karigo et al., 2021, Remedios et al., 2017, Yang and Anderson, 2022)

We analyzed data from three sets of previous experiments [29,31,39] All experiments were approved by the Institute Animal Care and Use Committee (IACUC) and the Institute Biosafety Committee (IBC) at the California Institute of Technology (Caltech). All experiments utilized heterozygous *Esr1*$^{cre/+}$ knock-in mice on a C457BL6/N background (B6N.129S6(Cg)-*Esr1*$^{tm1.1(cre)And}$/J, JAX strain #017911). Expression of GCaMP6s (Remedios et al., 2017, Karigo et al., 2021) or GCaMP7f (Yang et al., 2022) was achieved by stereotaxic injection of a Cre-dependent GCaMP-expressing adeno-associated viruses (AAVs). Briefly, for data obtained from Karigo et al., 2021, mice expressing GCaMP6s selectively in *Esr1* neurons in either the medial preoptic area (MPOA) or the ventrolateral subdivision of the ventromedial hypothalamus (VMHvl), were allowed to interact with BALB/c male and female intruders in a standard resident intruder assay (Karigo et al., 2021). Male or female intruders were introduced into the home cage in a random order, with a 5-10 min interval between intruder session. Each session typically lasted 10-20 minutes. Behavior videos of interacting animals were annotated using a custom MATLAB-based interface. A total of 7 behaviors including sniffing, dominance-mount, attack, mount, intromission, interact (periods where animals were close to each other but other behaviors were absent) were annotated with male and female intruders. A head-mounted micro-endoscope (Inscopix, Inc.) was used to acquire Ca^{2+} imaging data at 15Hz from either MPOA^{Esr1}

neurons (total of 583 neurons from 3 mice) or VMHvl^{Esr1} neurons (total of 421 neurons from 3 mice) for neural data analysis described in sections below.

For data obtained from Yang et al., 2022, Esr1-Cre mice in which GCaMP7f was expressed selectively in *Esr1* neurons in VMHvl, were allowed to interact with BALB/c male intruders in a standard resident intruder assay. In addition to the behaviors annotated for above, male intruders were also "dangled", where the ano-genital region of the dangled intruder is held next to the resident mouse. A head-mounted micro-endoscope was used to acquire Ca^{2+} imaging data at 30Hz from VMHvl^{Esr1} neurons (386 neurons from 3 mice) for neural data analysis described in sections below.

For data obtained from Remedios et al, 2017, Esr1-Cre mice in which GCaMP6s was expressed selectively in *Esr1* neurons in VMHvl were allowed to interact with BALB/c male intruders in a standard resident intruder assay. A head-mounted micro-endoscope was used to acquire Ca^{2+} imaging data at 30Hz from VMHvl^{Esr1} neurons (358 neurons from 3 mice) for neural data analysis described in sections below. rSLDS models were fit to data from n=14 mice to extract the time constant of the integration dimension used for correlation with individual differences in aggressiveness in Figure 3O. However 8 of those mice were excluded from decoder analysis of sniffing, mounting and attack, either because they were highly aggressive and attacked without any prior sniffing or dominance mounting (5 mice), or because they were non-aggressive and failed to attack (3 mice). Typically 20-25% of male mice from the C57BL6 background fail to show aggression in resident-intruder assays (Stagkourakis et al., 2020).

Method Details

Tuning rasters for single neurons

We examined the tuning properties of single neurons in VMHvl^{Esr1} or MPOA^{Esr1} by creating behavior tuning rasters (Figure 1 C, D). We first computed the mean activity of each neuron for each of the 14 manually annotated behavioral actions. To group neurons, we created a set of 40 regressors representing combinations of behavioral actions, and grouped neurons by which single regressor captured the most variance in each cell's activity. In addition to regressors for individual behaviors, example regressors include signals such as all male-directed actions, all female-directed actions, all male-directed/female-directed/sex-invariant investigative behaviors, and all male-directed/female-directed/sex-invariant consummatory behaviors. Neurons for which no single regressor captured at least 50% of variance in behavior-averaged activity were omitted from the visualization (approximately 5% of cells.)

Computation of pose features for input to dynamical model

As external input to the dynamical model (see next section), we selected two features of animal pose estimates produced by the Mouse Action Recognition System (MARS, [51] The first of these is the distance between animals, computed as the distance between centroids of ellipses fit to the poses of the two mice. The second is the facing angle of the resident towards intruder mouse, defined as the angle between a vector connecting the centroids of the two mice and a vector from the centroid to the nose of the resident mouse. In addition we also fit dynamical models with either no input or with additional inputs in the form of the speed of the resident (computed as the mean change in position of centroids of the head and hips, computed across two consecutive frames) and area of ellipse fit to the resident mouse's pose.

Dynamical system models of neural data

We model neural activity using a recurrent switching linear dynamical systems (rSLDS) according to previous methods[48,77]. Briefly, rSLDS is a generative model that breaks down non-linear time series data into sequences of linear dynamical modes. The model relates three sets of variables: a set of discrete states (z), a set of continuous latent factors (x) that captures the low-dimensional nature of neural activity, and the activity of recorded neurons (y). The model also allows for external inputs (u) which consists of extracted pose features including the distance between animals and the facing angle between the resident and intruder mouse. The model is formulated as follows: At each time t = 1,2,…T_n, there is a discrete state $z_t \in \{1,2,\dots,K\}$.. In a standard SLDS, these states follow Markovian dynamics, however rSLDS allows for the transitions between states to depend recurrently on the continuous latent factors (x) and external inputs (u) as follows:

$$p(z_{t+1} = k, z_t = j, x_t) \propto exp\{Rx_t + Wu_t + r\} \quad (1)$$

where R , W and r parameterizes a map from the previous discrete state, continuous state and external inputs using a softmax link function to a distribution over the next discrete states. The discrete state z_t determines the linear dynamical system used to generate the continuous latent factors at any time t:

$$x_t = A_{z_t}x_{t-1} + V_{z_t}u_t + b_{z_t} \quad (2)$$

where $A_k \in \mathbb{R}^{d\times d}$ is a dynamics matrix, $V_{z_t} \in \mathbb{R}^{d\times m}$ is a matrix that describes the contribution of external inputs (u_t) to each dimension of the latent space and $b_k \in \mathbb{R}^d$ is a bias vector, where d is the dimensionality of the latent space and m is the dimensionality of the external inputs. Thus, the discrete state specifies a set of linear dynamical system parameters and specify which dynamics to use when updating the continuous latent factors.

Lastly, we can recover the activity of recorded neurons by modelling activity as a linear noisy Gaussian observation $y_t \in \mathbb{R}^N$ where N is the number of recorded neurons:

$$y_t = Cx_t + d \quad (3)$$

For $C \in \mathbb{R}^{N \times D}$ and $d \sim N(0, S)$, a gaussian random variable. Overall, the system parameters that rSLDS needs to learn consists of the state transition dynamics, library of linear dynamical system matrices and neuron-specific emission parameters, which we write as:

$$\theta = \{A_k, V_k, b_k, C, d, R, W, r\}$$

These parameters are estimated using maximum likelihood using approximate variational inference methods as described in detail in [48,77].

Model performance is reported as the *evidence lower bound (ELBO)* which is equivalent to the Kullback-Leibler divergence between the approximate and true posterior, $KL(q(x, z; \varphi) \,||\, p(x, z \,|y; \theta))$ using 5-fold cross validation.

Since the ELBO is sensitive to the inclusion of regularizers and the amount of data used during fitting, we also provide an additional "forward simulation error (FSE)" model evaluation metric calculated as follows: given observed neural activity in state space at time *t*, we predict the trajectory of the population activity vector over an ensuing small time interval Δt using the model, then compute the mean squared error (MSE) between that trajectory and the observed data at time $t + \Delta t$ (Supplemental Figure S1F). This MSE is computed across all dimensions of the latent space and repeated for all times *t*. This error metric is normalized to a 0-1 range in each animal across the whole recording to obtain a bounded measure of model performance (Supplemental Figure S1F) . This metric is

computed across cross-validation folds and can provide intuition about time segments where model performance drops

Code used to fit rSLDS on neural data is available in the SSM package: (https://github.com/lindermanlab/ssm)

Code to generate flow fields and energy landscapes from fit dynamical systems is available in (https://github.com/DJALab/VMHvl_MPOA_dynamics)

Estimation of time constants

We estimated the time constant of each mode of linear dynamical systems using eigenvalues λ_a of the dynamics matrix of that system, derived by [53] as:

$$\tau_a = \left| \frac{1}{\log(|\lambda_a|)} \right|$$

Calculation of line attractor score

To provide a quantitative measure of the presence of line attractor dynamics, we devised a line attractor score defined as:

$$line\ attractor\ score = \log_2 \frac{t_n}{t_{n-1}}$$

where t_n is the largest time constant of the dynamics matrix of a dynamical system and t_{n-1} is the second largest time constant. This measure would be zero in a system without line attractor dynamics due to the similar magnitudes of the first two largest time constants and would be greater than one for systems that possess a line attractor.

Decoding behavior from integration dimension

We trained a frame-wise decoder to discriminate pairs of behavior (such as sniffing vs attack) from the activity of the integration dimension on individual frames of a behavior (sampled at 15Hz) as described previously (Karigo et al., 2021). We first created 'trials' from bouts of social behavior by merging all bouts that were separated by less than five seconds. We then trained a linear support vector machine (SVM) to identify a decoding threshold that maximally separates the values of our normalized "integration dimension" signal on frames during which behavior A occurred from values on frames during which behavior B occurred, for the pair-wise behavioral comparison. 'Shuffled' decoder data was generated by setting the decoding threshold on the same "trial", but with the behavior annotations randomly assigned to each behavior bout. We repeated shuffling 20 times for each intruder and each imaged mouse. We report performances of actual and shuffled 1D-threshold "decoders" as the average F1 score of the fit decoder, on data from all other "trials" for each mouse. For significance testing, the mean accuracy of the decoder trained on shuffled data was computed across mice, with shuffling repeated 1000 times for each mouse. Significance is determined by bootstrapping; we considered observed F1 scores significant if they fell above the 97.5th percentile of the distribution of chance F1 scores as done previously [29].

As a stringent test for spurious correlations due to the slow decay seen in the integration, we performed a variation of session permutation [57] as follows. Consider a neural signal that displays a slow ramp in activity, which can be used to decode attack from sniffing Supplemental Figure S2E). If this correlation was spurious and occurred due to slow drift in activity, that decoding threshold would perform poorly if used on the integration dimension from another mouse (Supplemental Figure S2F). On the other hand, that same threshold would produce a high F1 score if the correlation was not spurious as shown in Supplemental Figure S2G. To implement this paradigm, we used the decoding threshold obtained in a given mouse on the integration dimension from all other mice and averaged the final performance.

Low dimensional (PCA) representation of dynamical system

Since the latent states are invariant to linear transformations, it is possible to apply a suitable transformation to obtain an equivalent model using rSLDS. We use PCA for this transformation as it allows us to describe our high dimensional rSLDS latent space in a concise manner with few dimensions while capturing the overall dynamics. To perform this the following steps are applied:

1. Given latent factors: $x_1, x_2, \ldots, x_t$ of the raw neural data y_t
2. Compute a whitening transformation W such that Wx is the identity
3. Compute the transformed linear dynamical system $x_t' = Wx_t$ with new emission matrix $C' = CW^{-1}$.
4. Compute the singular value decomposition (SVD) of the new emission matrix $C' = USV^T$. Let $P = SV^T$, such that $P^{-1} = \frac{V}{S}$
5. Compute the final transformed latent states (i.e principal components) $x_t'' = P^{-1}x_t{}' = P^{-1}Wx_t$

In this final transformation, since the singular values are ordered, the first two components of x_t'' accounts for the most variance in the raw neural data y_t. This method of applying PCA also accounts for the emission matrix C of the fit dynamical system.

Dynamic velocity as a measure of stability in a dynamical system and visualization as 3D landscape

We devised a metric termed the "dynamic velocity" to quantify the average intrinsically generated rate of change of the fit dynamical system during a given behavior of interest. We first calculated the average norm of $A_{z_t}x_t$ for every value of x_t associated with a given behavior, for a given state *z*. We then averaged this value across states, giving a definition of $V_b = \frac{1}{n(Z)}\sum_{z\in Z}\left(\frac{1}{n(T_b)}\sum_{t\in T_b}\left\|A_{z_t}x_t\right\|\right)$, where Z is the set of states, T_b is the set of all timepoints during which behavior *b* occurred, $\|\cdot\|$ is the Euclidean norm, and n(·) is the number of elements in a set. Finally, to facilitate comparison across animals, we normalized

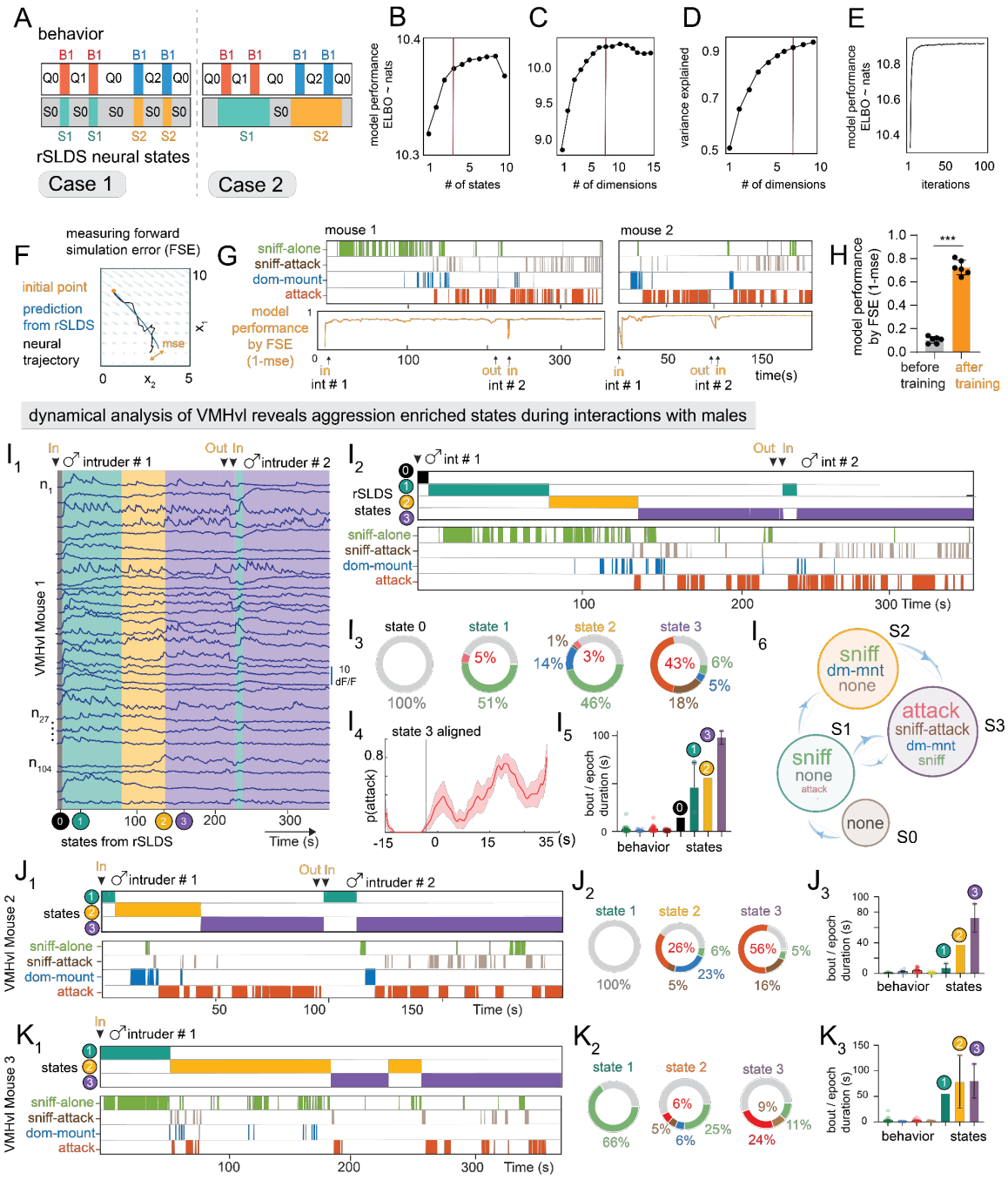
this value to a 0-1 range, with respect to its maximum across behaviors in each animal. Low values of this measure close to zero indicate regions with high stability while large values indicate unstable regions of neural state space.

We also converted the flow-fields obtained from rSLDS into a 3D landscape for visualization by calculating the dynamic velocity at each point in neural state space and using it as the height of a 3D landscape.

Quantification and statistical analysis

Data were processed and analyzed using Python, MATLAB, and GraphPad (GraphPad PRISM 9). All data were analyzed using two-tailed non-parametric tests. Mann-Whitney test were used for binary paired samples. Friedman test was used for non-binary paired samples. Kolmogorov-Smirnov test was used for non-paired samples. Multiple comparisons were corrected with Dunn's multiple comparisons correction. Not significant (NS), $P > 0.01$; *$P < 0.01$; **$P < 0.005$; ***$P < 0.001$; ****$P < 0.0001$.

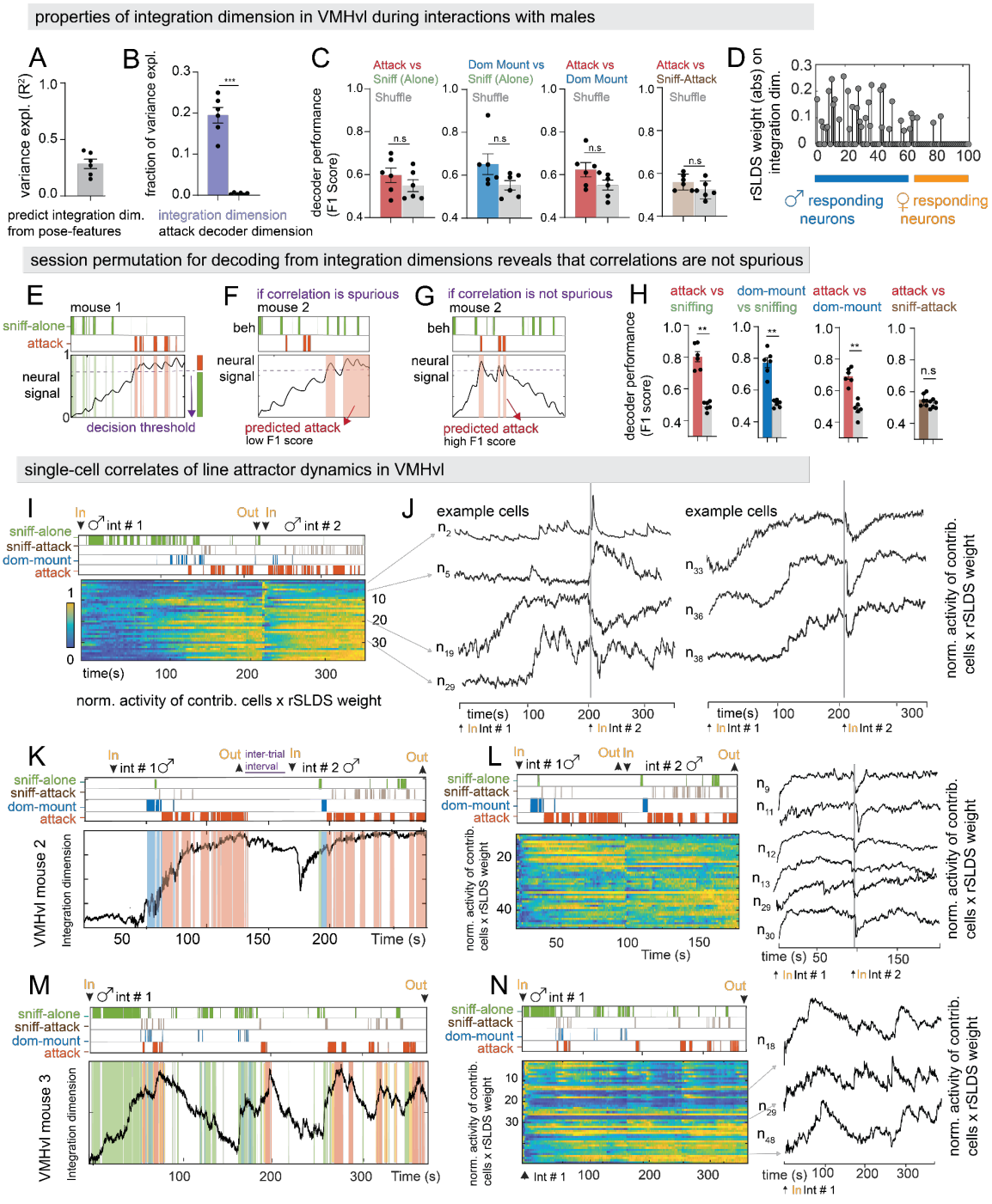
Supplemental Figure 1



Supplementary Figure 1: Unsupervised discovery of aggression-enriched states in VMHvl

Related to Figure 2. A: types of neural states identified by rSLDS. B1, B2: behaviors; Q0, Q1: periods of quiescence between behavior bouts; S0,S1,S2: rSLDS states. Case 1: rSLDS states cannot distinguish behavior vs internal states. Case 2: rSLDS reflects internal state-encoding due to persistence during behavioral quiescence. B: optimization of number of rSLDS states in example VMHvl mouse 1. Model performance is measured as ELBO (see methods). C: same as B, but for dimensionality. D: variance explained by dimension chosen in C. E: convergence of model performance. F: creation of a bounded model performance metric (forward simulation error, FSE, see methods). G: FSE for VMHvl mouse 1 & 2. H: average model performance (FSE) before and after training (n = 6 mice,***$p<0.001$).: I1: rSLDS states in VMHvl mouse 1. I2: comparison of rSLDS states with behaviors. I3: behavioral composition of rSLDS states. State 3 possesses the highest amount of attack behavior across mice (see panel J, K). I4 : probability of attack aligned to the onset of state 3 (n = 6 mice). I5: timescale of behavior bouts and discovered states epochs. I6: state transition diagram from empirical transition probabilities. J: Same as F2, F3, F5 but for VMHvl mouse 2. K: Same as F2, F3, F5 but for VMHvl mouse 3.

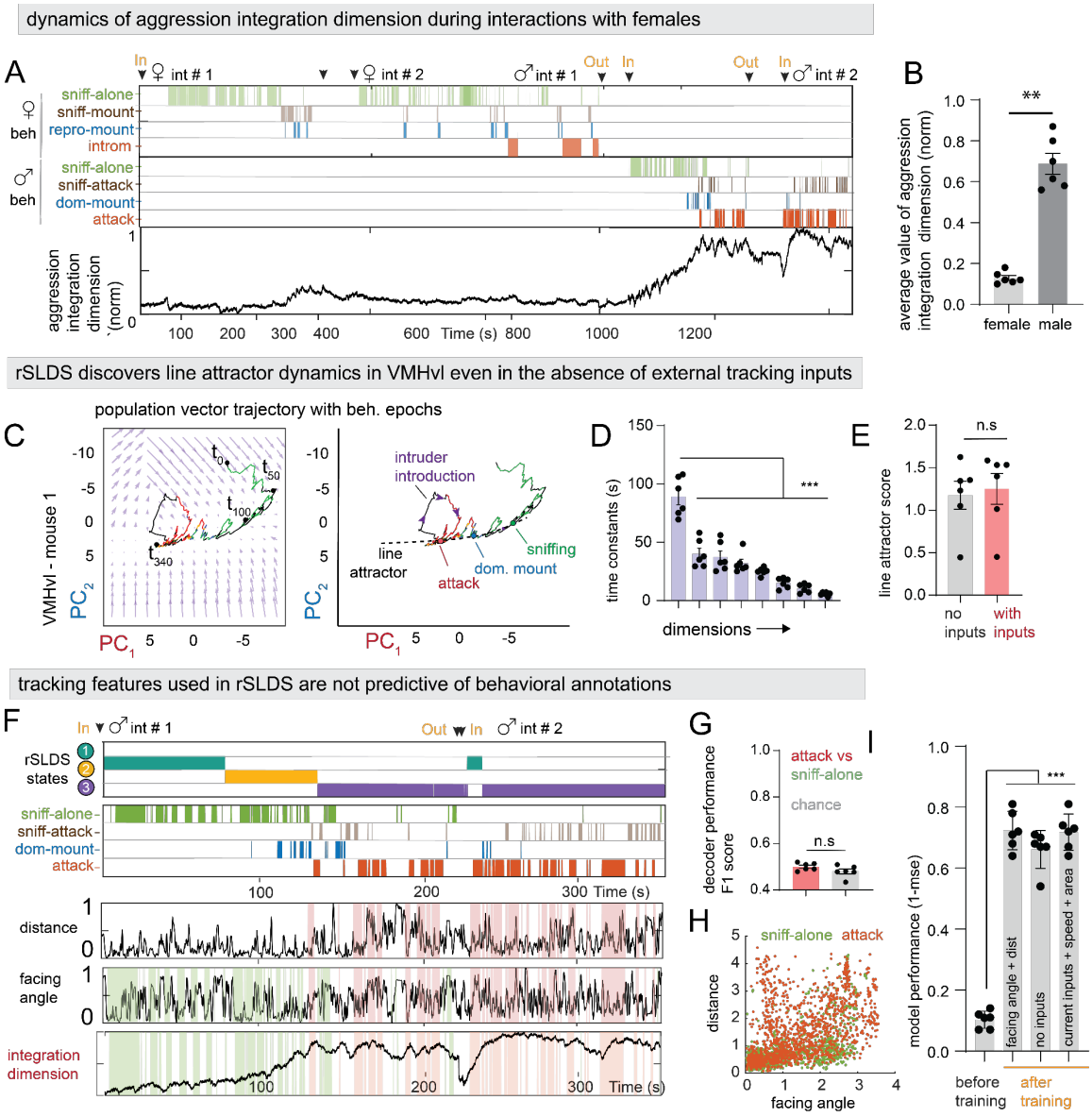
Supplemental Figure 2



Supplementary Figure 2: Characterization of aggression-integration dimension

Related to Figure 2. A: variance explained by a generalized linear model trained to predict integration dimension from pose-features including distance between mice, facing angle, speed, acceleration, and velocity of resident mouse (mean: 0.28 ± 0.04 R2, n = 6 mice). B: fraction of overall variance explained by integration dimension (purple) compared to variance explained by decoder dimension trained to distinguish attack from sniff bouts (integration dimension mean: 19.5%± 1.9%, attack decoder mean: 0.3% ± 0.1%, n= 6 mice, ***p<0.001). C: decoding behaviors from non-integration dimensions (average across dimensions, n = 6 mice). D: absolute rSLDS weight on integration dimension of VMHvl mouse 1 (cell number on x-axis), sorted by choice probability values for male vs female intruder encounter. E-G: paradigm to account for spurious correlations: decoding threshold obtained using integration dimension of mouse 1 (E, purple line) is used on integration dimension from mouse 2 (F). Spurious correlations lead to low F1 scores (F) while true correlations retain high F1 scores (G). H: decoding behaviors using paradigm described above (**p < 0.005, n = 6 mice) I: normalized activity of neurons times rSLDS weight for cells with significant weights for integration dimension of VMHvl mouse 1. J: example cells from I. K: integration dimension in VMHvl mouse 2. L: same as I for VMHvl mouse 2. M,N: Same as K,L for VMHvl mouse 3.

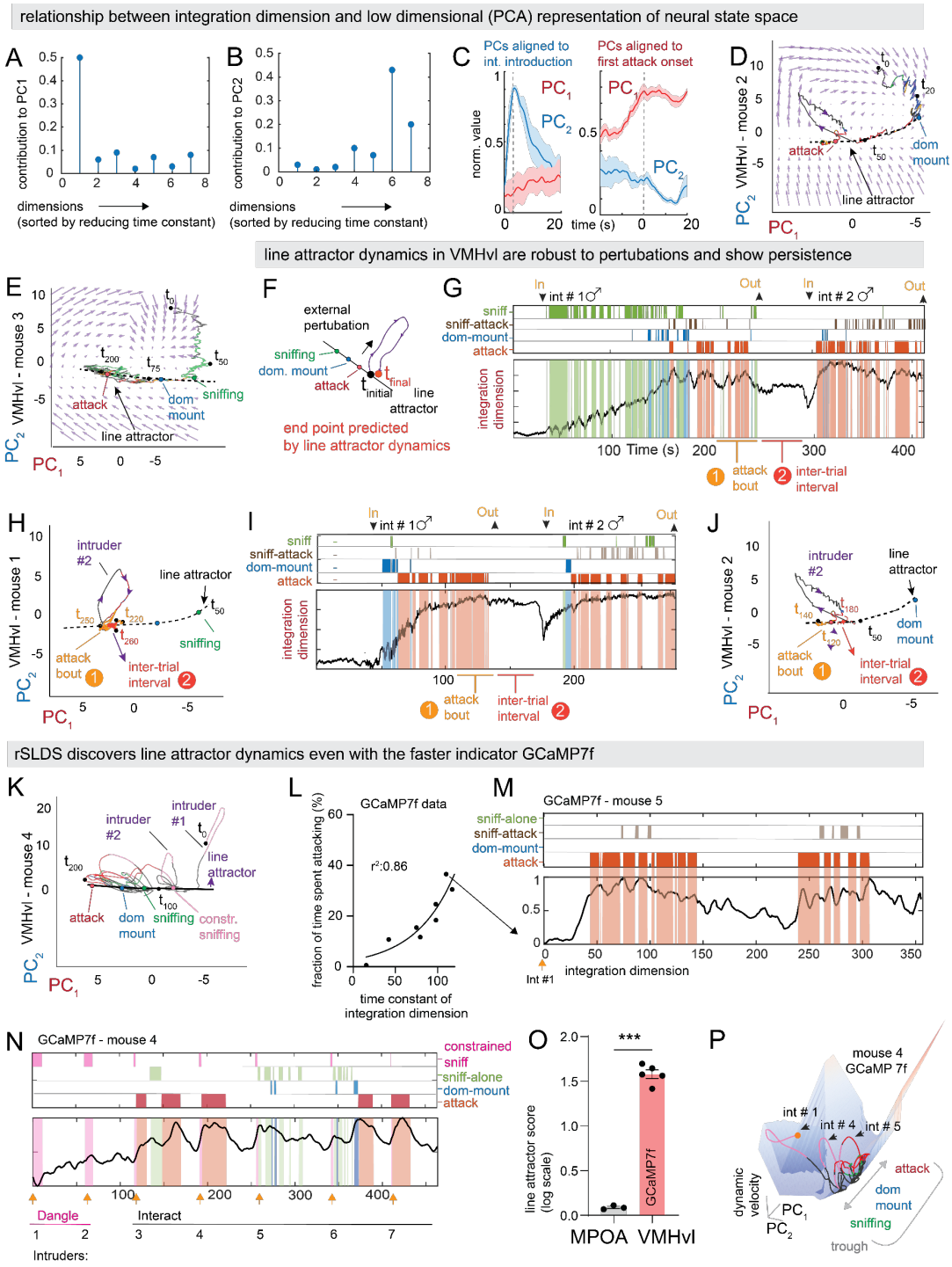
Supplemental Figure 3



Supplementary Figure 3: Characterization of aggression-integration dimension and dependence on tracking feature based external inputs.

Related to Figure 2. A: aggression integration dimension in female and male trials in VMHvl Mouse 1. B: mean projection of neural activity from female vs male trials onto the aggression integration dimension (n = 6 mice, **p<0.005). C: low dimensional dynamics and flow field from model with no behavioral inputs included with line attractor highlighted. D: time constants from the fit dynamical system (n = 6 mice). E: line attractor score for VMHvl models without input. F: tracking features used in rSLDS shown alongside discovered states and integration dimension in VMHvl mouse 1. G: performance of decoder used to separate attack frames from sniff-alone frames using the distance between mice and facing angle of the resident. H: scatter plot of distance between mice and facing angle of resident. I: model performance (1-FSE) for different types of external inputs (n = 6 mice); current inputs = distance between animals, facing angle of resident. (***p<0.001).

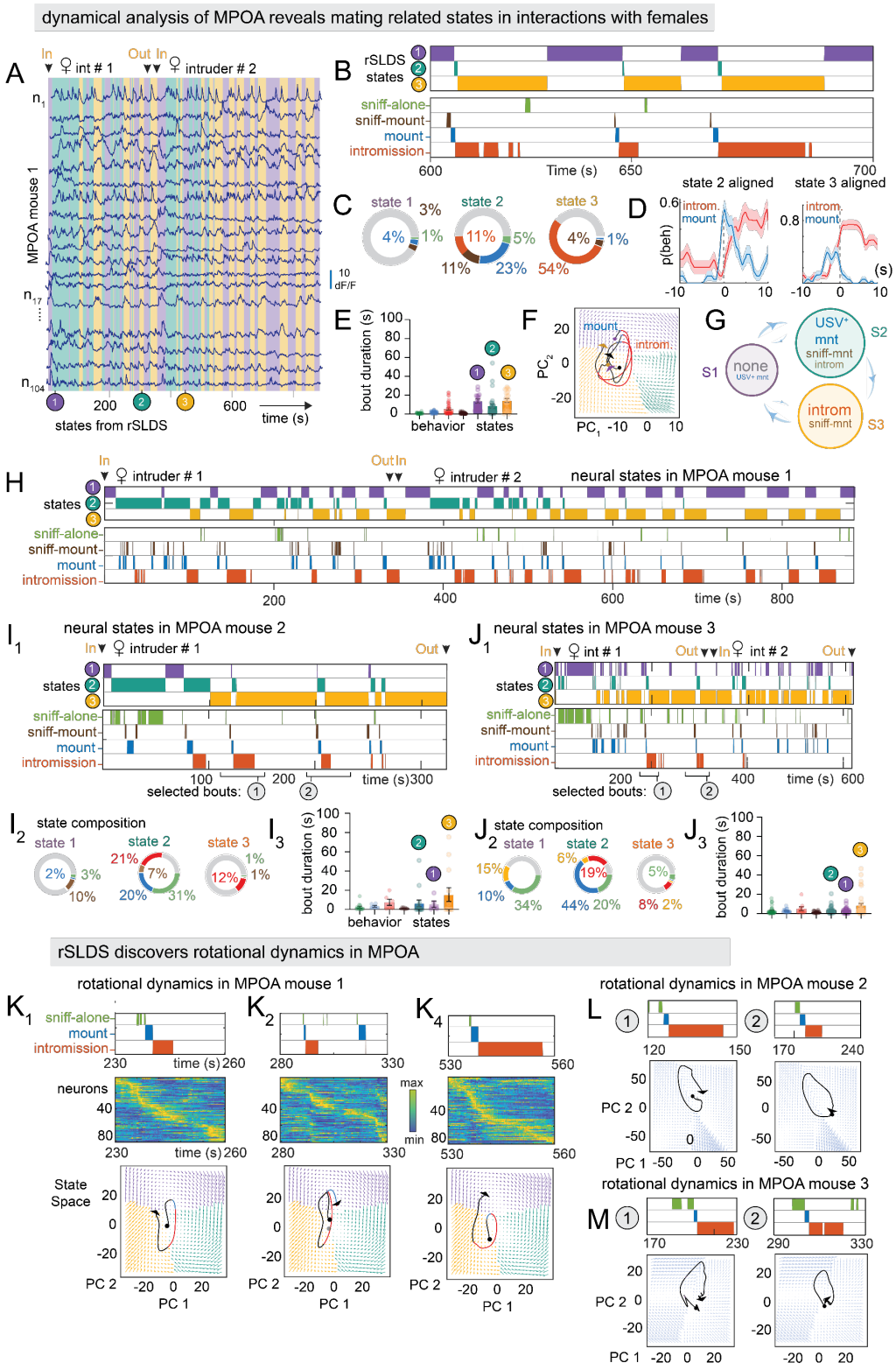
Supplemental Figure 4



Supplementary Figure 4: Properties of line attractor dynamics in VMHvl.

Related to Figure 3. A,B: absolute PCA weights of PC1(A) and PC2(B) on dimensions of dynamical system sorted by decreasing time constant in VMHvl mouse 1. C: behavior triggered average of top two principal components aligned to introduction of first intruder or first attack onset (n = 6 mice). D,E: low dimensional dynamics and flow field showing line attractor dynamics for VMHvl mouse 2 and mouse 3 with line attractor highlighted. F: schematic showing how perturbations orthogonal to a line attractor do not alter the position of the system. G: integration dimension in VMHvl mouse 1 (reproduced from Fig 2B) with attack bout (1) and inter-trial interval (2) highlighted. H: neural state space with line attractor highlighted in VMHvl mouse 1, showing the persistence of activity during the inter-trial interval shown in G. The introduction of intruder #2 acts as an orthogonal perturbation and activity returns to the same point along the attractor. I,J: Same as G,H for VMHvl mouse 2. K: neural state space with line attractor highlighted in VMHvl mouse 4. The introduction of intruder #2 occurs earlier in the trial when the animal displays sniffing behavior but results in a similar perturbation as above. L: relationship between fraction of time spent attack vs time constant of integration for animals with GCaMP7f recordings (n= 8 mice). M: integration dimension in VMHvl mouse 5 (GCaMP 7f) shows the same persistence and slow decay of activity. N: same as M for VMHvl mouse 4 (GCaMP 7f). O: line attractor score for mice with GCaMP7f recordings (***$p<0.001$). P: dynamics landscape for VMHvl mouse 4 (GCaMP 7f) showing a trough shaped landscape.

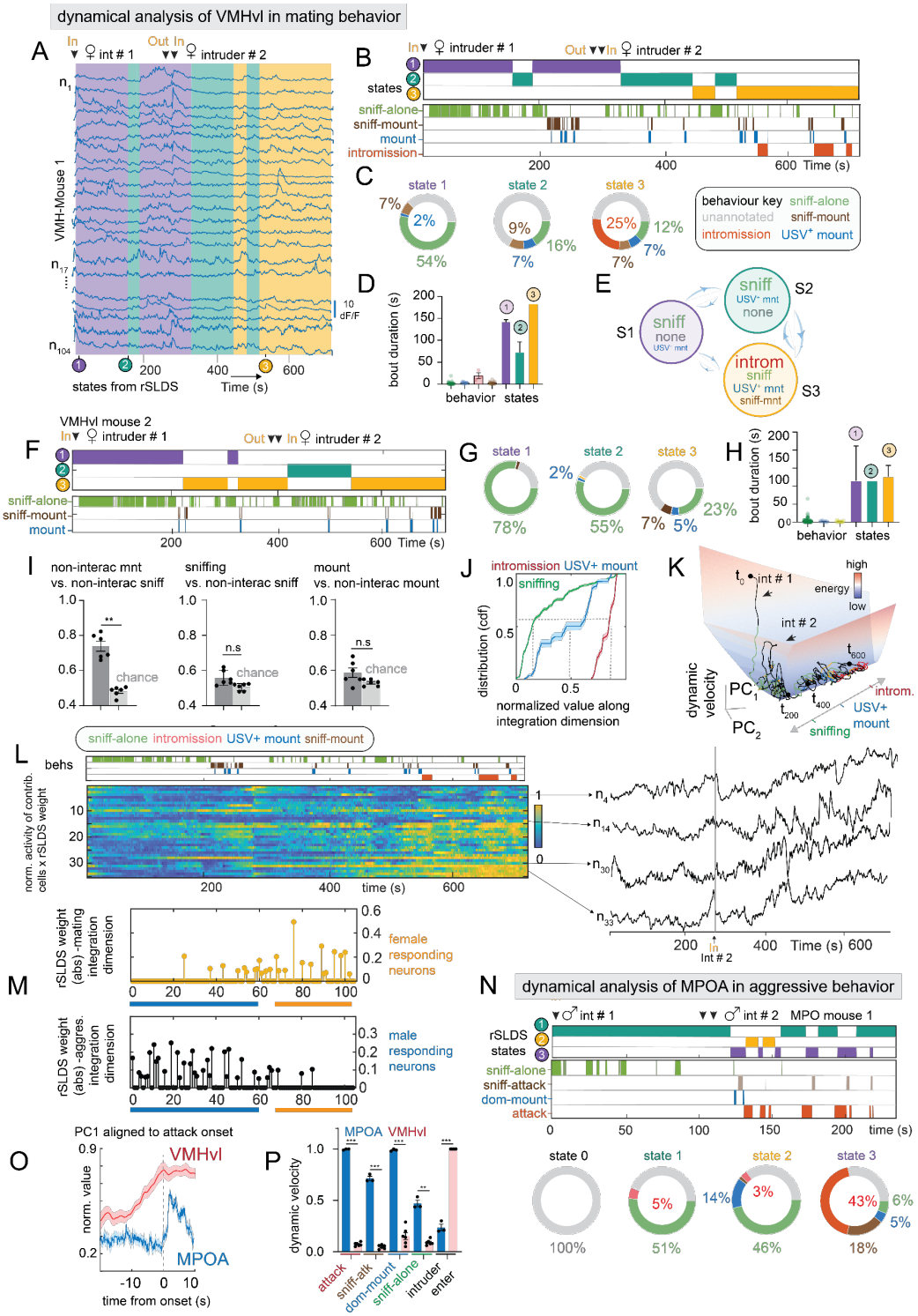
Supplemental Figure 5



Supplementary Figure 5: Mating enriched states and rotational dynamics in MPOA.

Related to Figure 4. A: rSLDS states in MPOA mouse 1. B: comparison of rSLDS states with behavior in MPOA mouse 1 for period from t = 600s to t = 700s. C: behavioral composition of rSLDS states. D: probability of intromission and USV+ mounting aligned to the onset of state 2 and state 3 (also see panel I, J, n = 3 mice). E: timescale of behavioral bouts and states epochs. F: Reproduced from Figure 4D but with state-specific inferred flow-field colors. G: state transition diagram from empirically calculated transition probabilities. H: state and behavior raster for MPOA mouse 1 for entire recording. I1: same as H for MPOA mouse 2, selected mating bouts highlighted. I2: behavioral composition of rSLDS states (bottom). I3: timescale of behavioral bouts and states epochs. J1-3: same as I1-3 for MPOA mouse 3, selected mating bouts highlighted. K: rotational trajectories for 3 mating episodes in MPOA mouse 1. L: same as K, for mating bouts highlighted in highlighted in I1 for MPOA mouse 2. M: same as K, for mating bouts highlighted in highlighted in J1 for MPOA mouse 3.

Supplemental Figure 6



Supplementary Figure 6: Dynamical analysis of VMHvl activity in mating behavior and MPOA activity in aggression.

Related to Figure 6. A: rSLDS states in VMHvl mouse 1 during interactions with female intruders. B: comparison of rSLDS states with behaviors. C: behavioral composition of rSLDS states. State 3 possesses the highest amount of mating behavior across mice (see panel H). D: timescale of behavior bouts and state epochs. E: state transition diagram from empirical transition probabilities. F-H: Same as B-D for VMHvl mouse 2. This mouse did not achieve intromission. I: decoding behaviors from integration dimension (**$p<0.005$).). J: empirical cumulative distribution of value of integration dimension (normalized) for various behaviors. K: dynamics velocity landscape showing a progression of mating behavior along the trough for VMHvl mouse 1. L: normalized activity times rSLDS weight for cells contributing significantly to integration dimension of VMHvl mouse 1. M: absolute rSLDS weight on integration dimension of VMHvl mouse 1 during mating behavior (top, yellow dots) and aggression (bottom, black dots) sorted by choice probability values for male vs female intruder encounter. N: top: state and behavior raster for MPOA mouse 1 during aggressive behavior. State 3 is aligned closely to the onset of attack bouts, bottom: behavioral composition of discovered states. O: behavior triggered average of principal component 1 in VMHvl (red line) and MPO (blue line) ($n = 3$ mice for MPOA, $n = 6$ mice for VMHvl). P: comparison of dynamic velocity for similar behavior between VMHvl and MPOA (reproduced from Figure 6F, 6K) (**$p<0.005$,***$p<0.001$) ($n = 3$ mice for MPOA, $n = 6$ mice for VMHvl).

References

1. Hahn, J.D., Sporns, O., Watts, A.G., and Swanson, L.W. (2019). Macroscale intrinsic network architecture of the hypothalamus. *Proc Natl Acad Sci U S A 116*, 8018-8027. 10.1073/pnas.1819448116.

2. Saper, C.B., and Lowell, B.B. (2014). The hypothalamus. *Curr Biol 24*, R1111-1116. 10.1016/j.cub.2014.10.023.

3. Paredes, R.G., and Baum, M.J. (1997). Role of the medial preoptic area/anterior hypothalamus in the control of masculine sexual behavior. *Annu Rev Sex Res 8*, 68-101.

4. Siegel, A., Roeling, T.A., Gregg, T.R., and Kruk, M.R. (1999). Neuropharmacology of brain-stimulation-evoked aggression. *Neurosci Biobehav Rev 23*, 359-389. 10.1016/s0149-7634(98)00040-2.

5. Canteras, N.S. (2002). The medial hypothalamic defensive system: hodological organization and functional implications. *Pharmacology Biochemistry and Behavior 71*, 481-491.

6. King, B.M. (2006). The rise, fall, and resurrection of the ventromedial hypothalamus in the regulation of feeding behavior and body weight. *Physiol Behav 87*, 221-244. 10.1016/j.physbeh.2005.10.007.

7. Kruk, M.R. (2014). Hypothalamic attack: A wonderful artifact or a useful perspective on escalation and pathology in aggression? A viewpoint. *Curr Top Behav Neurosci 17*, 143-188. 10.1007/7854_2014_313.

8. Swanson, L.W. (2005). Anatomy of the soul as reflected in the cerebral hemispheres: neural circuits underlying voluntary control of basic motivated behaviors. *J Comp Neurol 493*, 122-131. 10.1002/cne.20733.

9. Simerly, R.B. (2002). Wired for reproduction: Organization and development of sexually dimorphic circuits in the mammalian forebrain. *Annu Rev Neurosci 25*, 507-536. 10.1146/annurev.neuro.25.112701.142745.

10. Wu, Z., Autry, A.E., Bergan, J.F., Watabe-Uchida, M., and Dulac, C.G. (2014). Galanin neurons in the medial preoptic area govern parental behaviour. *Nature 509*, 325-330. 10.1038/nature13307.

11. Atasoy, D., Betley, J.N., Su, H.H., and Sternson, S.M. (2012). Deconstruction of a neural circuit for hunger. *Nature 488*, 172-177. 10.1038/nature11270.

12. Lee, H., Kim, D.W., Remedios, R., Anthony, T.E., Chang, A., Madisen, L., Zeng, H., and Anderson, D.J. (2014). Scalable control of mounting and attack by Esr1+ neurons in the ventromedial hypothalamus. *Nature 509*, 627-632. 10.1038/nature13169.

13. Lin, D., Boyle, M.P., Dollar, P., Lee, H., Lein, E.S., Perona, P., and Anderson, D.J. (2011). Functional identification of an aggression locus in the mouse hypothalamus. *Nature 470*, 221-226. 10.1038/nature09736.

14. Yamaguchi, T. (2022). Neural circuit mechanisms of sex and fighting in male mice. *Neurosci Res 174*, 1-8. 10.1016/j.neures.2021.06.005.

15. Zha, X., and Xu, X.H. (2021). Neural circuit mechanisms that govern inter-male attack in mice. *Cell Mol Life Sci 78*, 7289-7307. 10.1007/s00018-021-03956-x.

16. Augustine, V., Lee, S., and Oka, Y. (2020). Neural control and modulation of thirst, sodium appetite, and hunger. *Cell 180*, 25-32. 10.1016/j.cell.2019.11.040.

17. Sternson, S.M. (2013). Hypothalamic survival circuits: blueprints for purposive behaviors. *Neuron 77*, 810-824. 10.1016/j.neuron.2013.02.018.

18. Yang, C.F., Chiang, M.C., Gray, D.C., Prabhakaran, M., Alvarado, M., Juntti, S.A., Unger, E.K., Wells, J.A., and Shah, N.M. (2013). Sexually dimorphic neurons in the ventromedial hypothalamus govern mating in both sexes and aggression in males. *Cell 153*, 896-909. 10.1016/j.cell.2013.04.017.

19. Esteban Masferrer, M., Silva, B.A., Nomoto, K., Lima, S.Q., and Gross, C.T. (2020). Differential encoding of predator fear in the ventromedial hypothalamus and periaqueductal grey. *J Neurosci 40*, 9283-9292. 10.1523/JNEUROSCI.0761-18.2020.

20. Kato, A., and Sakuma, Y. (2000). Neuronal activity in female rat preoptic area associated with sexually motivated behavior. *Brain Res 862*, 90-102. 10.1016/s0006-8993(00)02076-x.

21. Mandelblat-Cerf, Y., Ramesh, R.N., Burgess, C.R., Patella, P., Yang, Z., Lowell, B.B., and Andermann, M.L. (2015). Arcuate hypothalamic AgRP and putative POMC neurons show opposite changes in spiking across multiple timescales. *Elife 4*. 10.7554/eLife.07122.

22. Gunaydin, L.A., Grosenick, L., Finkelstein, J.C., Kauvar, I.V., Fenno, L.E., Adhikari, A., Lammel, S., Mirzabekov, J.J., Airan, R.D., Zalocusky, K.A., et al. (2014). Natural neural projection dynamics underlying social behavior. *Cell 157*, 1535-1551. 10.1016/j.cell.2014.05.017.

23. Falkner, A.L., Grosenick, L., Davidson, T.J., Deisseroth, K., and Lin, D. (2016). Hypothalamic control of male aggression-seeking behavior. *Nature Neuroscience 19*, 596-604.

24. Zhao, Z.D., Yang, W.Z., Gao, C., Fu, X., Zhang, W., Zhou, Q., Chen, W., Ni, X., Lin, J.K., Yang, J., et al. (2017). A hypothalamic circuit that controls body temperature. *Proc Natl Acad Sci U S A 114*, 2042-2047. 10.1073/pnas.1616255114.

25. Li, Y., Zeng, J., Zhang, J., Yue, C., Zhong, W., Liu, Z., Feng, Q., and Luo, M. (2018). Hypothalamic Circuits for Predation and Evasion. *Neuron 97*, 911-924 e915. 10.1016/j.neuron.2018.01.005.

26. Ghosh, K.K., Burns, L.D., Cocker, E.D., Nimmerjahn, A., Ziv, Y., El Gamal, A., and Schnitzer, M.J. (2011). Miniaturized integration of a fluorescence microscope. *Nature Methods 8*, 871-878.

27. Ziv, Y., Burns, L.D., Cocker, E.D., Hamel, E.O., Ghosh, K.K., Kitch, L.J., El Gamal, A., and Schnitzer, M.J. (2013). Long-term dynamics of CA1 hippocampal place codes. *Nature Neuroscience 16*, 264-266.

28. Jennings, J.H., Ung, R.L., Resendez, S.L., Stamatakis, A.M., Taylor, J.G., Huang, J., Veleta, K., Kantak, P.A., Aita, M., Shilling-Scrivo, K., et al. (2015). Visualizing hypothalamic network dynamics for appetitive and consummatory behaviors. *Cell 160*, 516-527. 10.1016/j.cell.2014.12.026.

29. Remedios, R., Kennedy, A., Zelikowsky, M., Grewe, B.F., Schnitzer, M.J., and Anderson, D.J. (2017). Social behaviour shapes hypothalamic neural ensemble representations of conspecific sex. *Nature 550*, 388-392.

30. Krzywkowski, P., Penna, B., and Gross, C.T. (2020). Dynamic encoding of social threat and spatial context in the hypothalamus. *Elife 9*. 10.7554/eLife.57148.

31. Karigo, T., Kennedy, A., Yang, B., Liu, M., Tai, D., Wahle, I.A., and Anderson, D.J. (2021). Distinct hypothalamic control of same-and opposite-sex mounting behaviour in mice. *Nature 589*, 258-263.

32. Wei, Y.-C., Wang, S.-R., Jiao, Z.-L., Zhang, W., Lin, J.-K., Li, X.-Y., Li, S.-S., Zhang, X., and Xu, X.-H. (2018). Medial preoptic area in mice is capable of mediating sexually dimorphic behaviors regardless of gender. *Nature Communications 9*, 1-15.

33. Moffitt, J.R., Bambah-Mukku, D., Eichhorn, S.W., Vaughn, E., Shekhar, K., Perez, J.D., Rubinstein, N.D., Hao, J., Regev, A., and Dulac, C. (2018). Molecular, spatial, and functional single-cell profiling of the hypothalamic preoptic region. *Science 362*.

34. Ishii, K.K., Osakada, T., Mori, H., Miyasaka, N., Yoshihara, Y., Miyamichi, K., and Touhara, K. (2017). A labeled-line neural circuit for pheromone-mediated sexual behaviors in mice. *Neuron 95*, 123-137. e128.

35. Hashikawa, K., Hashikawa, Y., Tremblay, R., Zhang, J., Feng, J.E., Sabol, A., Piper, W.T., Lee, H., Rudy, B., and Lin, D. (2017). Esr1+ cells in the ventromedial hypothalamus control female aggression. *Nature Neuroscience 20*, 1580-1590.

36. Yang, T., Yang, C.F., Chizari, M.D., Maheswaranathan, N., Burke, K.J., Jr., Borius, M., Inoue, S., Chiang, M.C., Bender, K.J., Ganguli, S., and Shah, N.M. (2017). Social Control of Hypothalamus-Mediated Male Aggression. *Neuron 95*, 955-970 e954. 10.1016/j.neuron.2017.06.046.

37. Liu, M., Kim, D.W., Zeng, H., and Anderson, D.J. (2022). Make war not love: The neural substrate underlying a state-dependent switch in female social behavior. *Neuron*. 10.1016/j.neuron.2021.12.002.

38. Hess, W.R., and Brügger, M. (1943). Das subkortikale Zentrum der affektiven Abwehrreaktion. *Helvetica Physiologica et Pharmacologica Acta.*

39. Yang, B., Karigo, T., and Anderson, D.J. (2022). Transformations of neural representations in a social behaviour network. *Nature 608*, 741-749. 10.1038/s41586-022-05057-6.

40. Miller, P. (2016). Dynamical systems, attractors, and neural circuits. *F1000Res 5*. 10.12688/f1000research.7698.1.

41. Sussillo, D. (2014). Neural circuits as computational dynamical systems. *Curr Opin Neurobiol 25*, 156-163. 10.1016/j.conb.2014.01.008.

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

43. Vyas, S., Golub, M.D., Sussillo, D., and Shenoy, K.V. (2020). Computation through neural population dynamics. *Annual Review of Neuroscience 43*, 249-275.

44. Shenoy, K.V., Sahani, M., and Churchland, M.M. (2013). Cortical control of arm movements: a dynamical systems perspective. *Annual Review of Neuroscience 36*, 337-359.

45. Churchland, M.M., Cunningham, J.P., Kaufman, M.T., Ryu, S.I., and Shenoy, K.V. (2010). Cortical preparatory activity: representation of movement or first cog in a dynamical machine? *Neuron 68*, 387-400. 10.1016/j.neuron.2010.09.015.

46. Hulse, B.K., and Jayaraman, V. (2020). Mechanisms underlying the neural computation of head direction. *Annual Review of Neuroscience 43*, 31-54.

47. Ebitz, R.B., and Hayden, B.Y. (2021). The population doctrine in cognitive neuroscience. *Neuron 109*, 3055-3068. 10.1016/j.neuron.2021.07.011.

48. Linderman, S., Johnson, M., Miller, A., Adams, R., Blei, D., and Paninski, L. (2017). Bayesian learning and inference in recurrent switching linear dynamical systems. (*PMLR*), pp. 914-922.

49. Falkner, A.L., Dollar, P., Perona, P., Anderson, D.J., and Lin, D. (2014). Decoding ventromedial hypothalamic neural activity during male mouse aggression. *J Neurosci 34*, 5971-5984. 10.1523/JNEUROSCI.5109-13.2014.

50. Itakura, T., Murata, K., Miyamichi, K., Ishii, K.K., Yoshihara, Y., and Touhara, K. (2022). A single vomeronasal receptor promotes intermale aggression through dedicated hypothalamic neurons. *Neuron 110*, 2455-2469 e2458. 10.1016/j.neuron.2022.05.002.

51. Segalin, C., Williams, J., Karigo, T., Hui, M., Zelikowsky, M., Sun, J.J., Perona, P., Anderson, D.J., and Kennedy, A. (2021). The Mouse Action Recognition System (MARS) software pipeline for automated analysis of social behaviors in mice. *Elife 10*. 10.7554/eLife.63720.

52. Strogatz, S.H. (2018). Nonlinear dynamics and chaos with student solutions manual: With applications to physics, biology, chemistry, and engineering (*CRC Press*).

53. Maheswaranathan, N., Williams, A.H., Golub, M.D., Ganguli, S., and Sussillo, D. (2019). Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics. *Advances in Neural Information Processing Systems 32*, 15696.

54. Major, G., and Tank, D. (2004). Persistent neural activity: prevalence and mechanisms. *Curr Opin Neurobiol 14*, 675-684. 10.1016/j.conb.2004.10.017.

55. Goldman, M.S., Compte, A., and Wang, X.J. (2009). Neural Integrator Models. In Encyclopedia of Neuroscience, L.R. Squire, ed. (Academic Press), pp. 165-178. https://doi.org/10.1016/B978-008045046-9.01434-0.

56. Zhu, Z., Ma, Q., Miao, L., Yang, H., Pan, L., Li, K., Zeng, L.H., Zhang, X., Wu, J., Hao, S., et al. (2021). A substantia innominata-midbrain circuit controls a general aggressive response. *Neuron 109*, 1540-1553 e1549. 10.1016/j.neuron.2021.03.002.

57. Harris, K.D. (2021). Nonsense correlations in neuroscience. *bioRxiv*, 2020.2011.2029.402719. 10.1101/2020.11.29.402719.

58. Khona, M., and Fiete, I.R. (2022). Attractor and integrator networks in the brain. *Nat Rev Neurosci.* 10.1038/s41583-022-00642-0.

59. Seung, H.S. (1996). How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences 93*, 13339-13344.

60. Ganguli, S., Bisley, J.W., Roitman, J.D., Shadlen, M.N., Goldberg, M.E., and Miller, K.D. (2008). One-dimensional dynamics of attention and decision making in LIP. *Neuron 58*, 15-25.

61. Sylwestrak, E.L., Jo, Y., Vesuna, S., Wang, X., Holcomb, B., Tien, R.H., Kim, D.K., Fenno, L., Ramakrishnan, C., Allen, W.E., et al. (2022). Cell-type-specific population dynamics of diverse reward computations. *Cell 185*, 3568-3587 e3527. 10.1016/j.cell.2022.08.019.

62. Zhou, S., Masmanidis, S.C., and Buonomano, D.V. (2020). Neural sequences as an optimal dynamical regime for the readout of time. *Neuron 108*, 651-658 e655. 10.1016/j.neuron.2020.08.020.

63. Kim, D.-W., Yao, Z., Graybuck, L.T., Kim, T.K., Nguyen, T.N., Smith, K.A., Fong, O., Yi, L., Koulena, N., and Pierson, N. (2019). Multimodal analysis of cell types in a hypothalamic node controlling social behavior. *Cell 179*, 713-728. e717.

64. Kennedy, A. (2022). The what, how, and why of naturalistic behavior. *Curr Opin Neurobiol 74*, 102549. 10.1016/j.conb.2022.102549.

65. Kennedy, A., Kunwar, P.S., Li, L.Y., Stagkourakis, S., Wagenaar, D.A., and Anderson, D.J. (2020). Stimulus-specific hypothalamic encoding of a persistent defensive state. *Nature 586*, 730-734. 10.1038/s41586-020-2728-4.

66. Shao, Y.Q., Fan, L., Wu, W.Y., Zhu, Y.J., and Xu, H.T. (2022). A developmental switch between electrical and neuropeptide communication in the ventromedial hypothalamus. *Curr Biol 32*, 3137-3145 e3133. 10.1016/j.cub.2022.05.029.

67. Xie, Z., Gu, H., Huang, M., Cheng, X., Shang, C., Tao, T., Li, D., Xie, Y., Zhao, J., Lu, W., et al. (2022). Mechanically evoked defensive attack is controlled by GABAergic neurons in the anterior hypothalamic nucleus. *Nat Neurosci 25*, 72-85. 10.1038/s41593-021-00985-4.

68. Nelson, R.J., and Trainor, B.C. (2007). Neural mechanisms of aggression. *Nat Rev Neurosci 8*, 536-546. 10.1038/nrn2174.

69. Stagkourakis, S., Spigolon, G., Williams, P., Protzmann, J., Fisone, G., and Broberger, C. (2018). A neural network for intermale aggression to establish social hierarchy. *Nat Neurosci 21*, 834-842. 10.1038/s41593-018-0153-x.

70. Soden, M.E., Miller, S.M., Burgeno, L.M., Phillips, P.E.M., Hnasko, T.S., and Zweifel, L.S. (2016). Genetic Isolation of Hypothalamic Neurons that Regulate Context-Specific Male Social Behavior. *Cell Rep 16*, 304-313. 10.1016/j.celrep.2016.05.067.

71. McKellar, C.E., Lillvis, J.L., Bath, D.E., Fitzgerald, J.E., Cannon, J.G., Simpson, J.H., and Dickson, B.J. (2019). Threshold-based ordering of sequential actions during *Drosophila* courtship. *Curr Biol 29*, 426-434 e426. 10.1016/j.cub.2018.12.019.

72. Evans, D.A., Stempel, A.V., Vale, R., Ruehle, S., Lefler, Y., and Branco, T. (2018). A synaptic threshold mechanism for computing escape decisions. *Nature 558*, 590-594. 10.1038/s41586-018-0244-6.

73. Stagkourakis, S., Spigolon, G., Liu, G., and Anderson, D.J. (2020). Experience-dependent plasticity in an innate social behavior is mediated by hypothalamic LTP. *Proc Natl Acad Sci U S A 117*, 25789-25799. 10.1073/pnas.2011782117.

74. Inagaki, H.K., Fontolan, L., Romani, S., and Svoboda, K. (2019). Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature 566*, 212-217. 10.1038/s41586-019-0919-7.

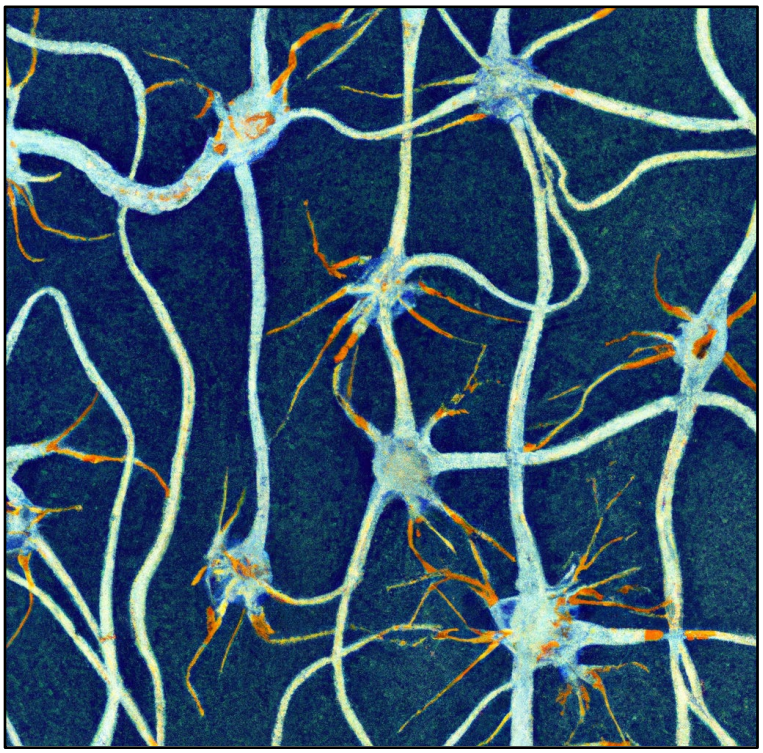
75. Sabatini, B.L., and Tian, L. (2020). Imaging neurotransmitter and neuromodulator dynamics in vivo with genetically encoded indicators. *Neuron 108*, 17-32. 10.1016/j.neuron.2020.09.036.

76. Sun, F., Zeng, J., Jing, M., Zhou, J., Feng, J., Owen, S.F., Luo, Y., Li, F., Wang, H., Yamaguchi, T., et al. (2018). A Genetically Encoded Fluorescent Sensor Enables Rapid and Specific Detection of Dopamine in Flies, Fish, and Mice. *Cell 174*, 481-496 e419. 10.1016/j.cell.2018.06.042.

77. Linderman, S., Nichols, A., Blei, D., Zimmer, M., and Paninski, L. (2019). Hierarchical recurrent state space models reveal discrete and continuous dynamics of neural activity in C. elegans. *BioRxiv*, 621540.

Chapter III

PERTURBATION



"അവൾ ഒരു കല്ലുംതള്ളി, സത്യം പൊട്ടിച്ചിതറിച്ച,

കടലിനെ ഒരു തുള്ളി അകറ്റി നോക്കി."

Kamala Surayya, Ente Katha, 1973

Translation: "She pushed a stone, scattering the truth, and looked at the sea, one drop at a time."

Chapter III

Causal evidence of a line attractor encoding an affective state

This chapter details first-in-class efforts to examine and test line dynamics important for internal states using closed-loop perturbations and modelling.

Published as Amit Vinograd*, **Aditya Nair***, Joseph Kim, Scott Linderman, David J. Anderson. Causal evidence of a line attractor encoding an affective state. *Nature (*2024).

Summary

Line attractors are emergent population dynamics hypothesized to encode continuous variables such as head direction and internal states[1-4]. In mammals, direct evidence of neural implementation of a line attractor has been hindered by the challenge of targeting perturbations to specific neurons within contributing ensembles[2,3]. Linear dynamical systems modeling has revealed that neurons in the hypothalamus exhibit approximate line attractor dynamics in male mice during aggressive encounters[5]. We have previously hypothesized that these dynamics may encode the variable intensity of an aggressive internal motive state. Here, we report that these neurons also showed line attractor dynamics in head-fixed mice observing aggression[6]. We identified and perturbed line attractor-contributing neurons using 2-photon calcium imaging and holographic optogenetic perturbations. On-manifold perturbations yielded integration and persistent activity that drove the system along the line attractor, while transient off-manifold perturbations were followed by rapid relaxation back into the attractor. Furthermore, single-cell stimulation and imaging revealed selective functional connectivity among attractor-contributing neurons. Intriguingly, individual differences among mice in line attractor stability were correlated with the degree of functional connectivity among attractor neurons. Mechanistic RNN modelling indicated that dense subnetwork connectivity and slow neurotransmission[7] best recapitulate our empirical findings. Our work bridges circuit and manifold levels[3], providing causal evidence of continuous attractor dynamics encoding an affective internal state in the mammalian hypothalamus.

Introduction

Neural circuit function has been studied from two vantage points. One focuses on understanding behaviorally specialized neuron types and their functional connectivity[8-10], while the other investigates emergent properties of neural networks, such as attractors[1,3,11]. Attractors of different topologies are theorized to encode a variety of continuous variables, ranging from head direction[12], location in space[2] and internal states[5]. Recent data-driven methodologies have allowed for the discovery of such attractor mediated computations directly in neural data[5,13-16]. Consequently, attractor dynamics have received increasing attention as a major type of neural coding mechanism[2,4,12 3,13].

Despite this progress, establishing that these attractors arise from the dynamics of the observed network remains a formidable challenge[2,3 4]. This calls for combining large-scale recordings with perturbations of neuronal activity *in vivo*. While this has been accomplished for a point attractor that controls motor planning in cortical area ALM[17,18], spatial ensembles that regulate short term memory[19,20], and for a ring attractor in *Drosophila*[21,22], there is no study reporting such perturbations for a continuous attractor in any mammalian system. While theoretical work on continuous attractors in mammals is well-developed[2], the lack of direct, neural perturbation-based experimental evidence of such attractor dynamics has hindered progress towards a mechanistic circuit-level understanding of such emergent manifold-level network features[3].

Estrogen receptor type 1 (*Esr1*)-expressing neurons in the ventrolateral subdivision of the ventromedial hypothalamus (VMHvl^{Esr1}) comprise a key node in the social behavior network and have been causally implicated in aggression[23 24]. Calcium imaging of these neurons in freely behaving animals has revealed mixed selectivity and variable dynamics, with time-locked attack signals sparsely represented at the single-neuron level[25,26]. Application of dynamical system modeling[27] has revealed an approximate line attractor in VMHvl that correlates with the intensity of agonistic behavior, suggesting a population-level encoding of

a continuously varying aggressive internal state[5]. This raises the question of whether the observation of a line attractor in a dynamical systems model fit to VMHvl^{Esr1} neuronal activity reflects inherited dynamics or is instantiated locally.

This question can be addressed, in principle, using all-optical methods to observe and perturb line attractor-relevant neural activity[3,28-30]. A challenge in applying these methods during aggression is that current technology requires head-fixed preparations, and head-fixed mice do not fight. To overcome this challenge, we exploited a recent observation that VMHvl^{PR} neurons (which encompass the *Esr1*$^{+}$ subset) [31,32] [33] mirror inter-individual aggression[6], to instantiate the line attractor in head-fixed subjects. Using this preparation, we performed model-guided, closed-loop on- and off-manifold perturbations[34] of VMHvl^{Esr1} activity. These experiments demonstrate that the VMHvl line attractor indeed reflects causal neural dynamics in this nucleus. They also identified selective functional connectivity within attractor-weighted ensembles, suggesting a local circuit implementation of attractor dynamics. Modeling suggests that this implementation is likely mediated by slow neurotransmission. Collectively, our findings elucidate a circuit-level foundation for a continuous attractor in the mammalian brain.

Results

A line attractor for observing aggression

Recent studies have demonstrated that VMHvl contains neurons that are active during passive observation of, as well as active participation in, aggression and that re-activating the former can evoke aggressive behavior[6]. However, those findings were based on a relatively small sample of VMHvl neurons, which might comprise a specific subset distinct from those contributing to the line attractor (the latter represent ~20-25% of *Esr1*$^{+}$ neurons[5]). To assess whether these “mirror-like” responses can be observed in *Esr1*$^{+}$ neurons that contribute to line attractor dynamics, we performed microendoscopic imaging[35] of VMHvl^{Esr1} neurons

expressing jGCaMP7s in the same freely behaving (FB) animals during engagement in followed by observation of aggression (Extended Data Fig. 1a-e). Analysis using recurrent switching linear dynamical systems (rSLDS)[27] to fit a model to each dataset (Extended Data Fig. 1f) revealed an approximate line attractor under both conditions, exhibiting ramping and persistent activity aligned and maintained across both performed and observed attack sessions (Extended Data Fig. 1g-q, Extended Data Fig. 2 and Extended Data Fig. 3a-f). Activity in the integration dimension ("x_1") aligned with the line attractor during observation of aggression could be reliably used to decode from held-out data instances of both observation of and engagement in attack, suggesting that this dimension encodes a similar internal state variable under both conditions (Extended Data Fig. 3g-h). In addition, the integration dimension was weighted by a consistent and aligned set of neurons under both conditions, suggesting that a highly overlapping set of neurons (70%) contributes to line attractor dynamics during observing or engaging in attack (Extended Data Fig. 4a-d).

The dynamical systems analysis also revealed a dimension orthogonal to the integration dimension ("x_2") that displayed faster dynamics time locked to the entry of the intruder(s) in both conditions (Extended Data Fig. 1g-l). To examine whether the neurons contributing to the two dimensions (x_1 and x_2 neurons) can be separated based on biophysical properties, we examined their baseline activity when solitary animals were exploring their home cage before any interaction. We did not detect a difference in amplitude or decay constant (tau) between x_1 and x_2 neurons (Extended Data Fig. 4e-i). However, we did see a slightly but significantly higher frequency of spontaneous calcium transients in x_2 neurons (Extended Data Fig. 4f, g), suggesting that x_2 neurons are more "spontaneously" active than x_1 neurons when no interaction is taking place.

While these observed attractor dynamics could be generated in VMHvl, they might also arise from unmeasured ramping sensory input or dynamics inherited from an input brain region[36]. Although behavioral perturbations in prior studies have hinted at the intrinsic nature of VMHvl line attractor dynamics[5], a rigorous test requires direct neuronal perturbations[37 34]

targeted to cells that contribute to the attractor. Direct on-manifold perturbation of a continuous attractor has previously been performed only in the *Drosophila* head direction system[12,21]. In mammals, although a point attractor has been perturbed off-manifold using optogenetic manipulation[17,18,28], direct single-cell perturbations of neurons contributing to a continuous attractor *in vivo* has not been reported.

To do this, we employed 2-photon (2P) imaging in head-fixed mice of VMHvl^{Esr1} neurons expressing jGCaMP7s[38] following observation of aggression and removal of the demonstrator mice (Figure 1a-c). As described above, during observation of aggression by the head-fixed mice, rSLDS analysis identified an integration dimension with slow dynamics (x_1) aligned to an approximate line attractor, and an orthogonal dimension with faster dynamics (x_2) (Figure 1d-h, k). We used the mapping between neural activity and the underlying state space to directly identify neurons contributing to each dimension (Figure 1i, j). Neurons contributing to the integration dimension displayed more persistence than those aligned with the faster dimension (Figure 1g, l, m). Importantly, only a small fraction of the neural activity could be explained by movements of the observer mouse (Extended Data Fig. 5a-e). Thus, a line attractor can be recapitulated in head-fixed mice observing aggression, opening the way to 2-photon-based perturbation experiments.

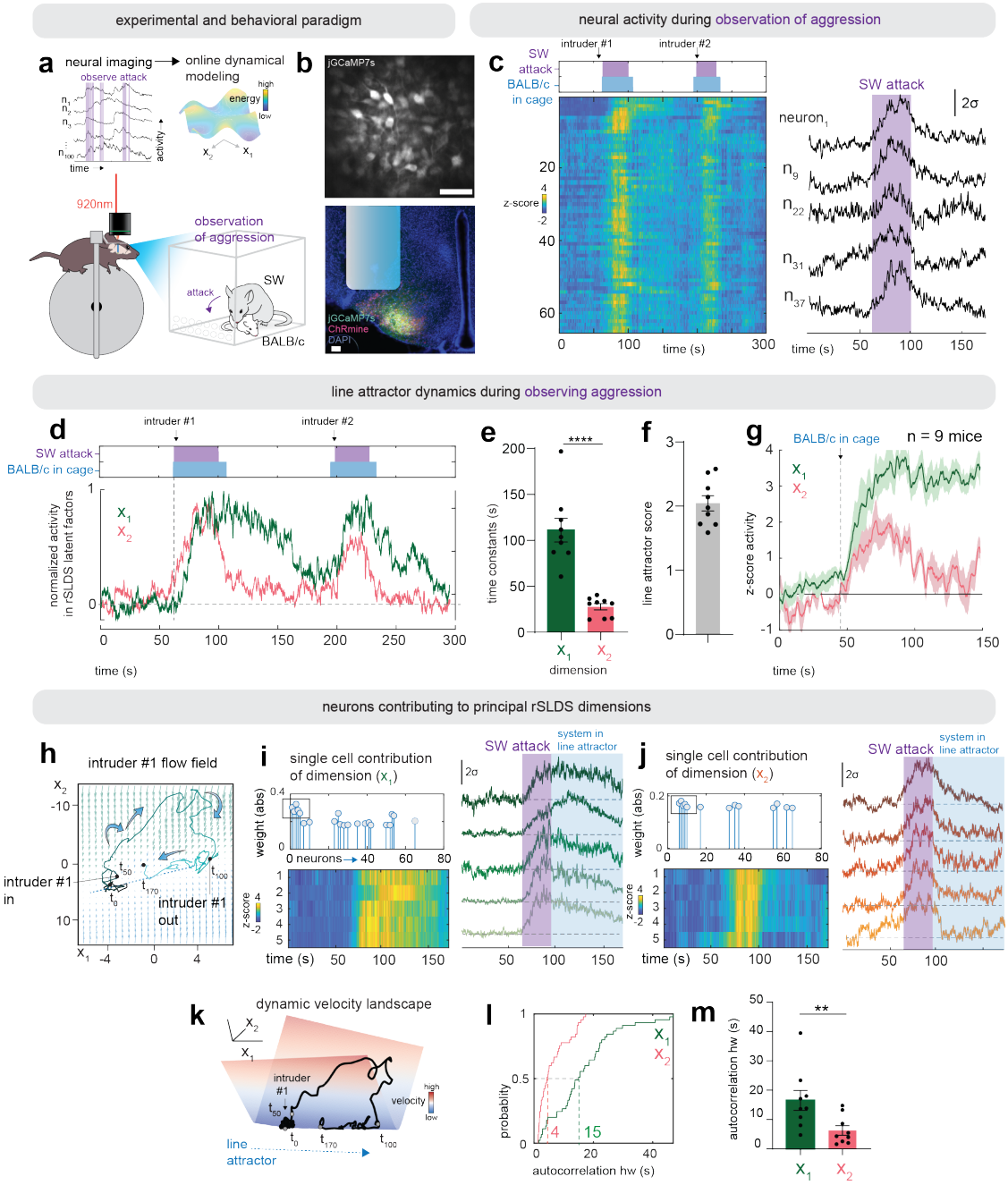

experimental and behavioral paradigm
neural activity during observation of aggression
line attractor dynamics during observing aggression
neurons contributing to principal rSLDS dimensions
a
neural imaging
online dynamical modeling
observe attack
energy
high
low
activity
time
920nm
observation of aggression
SW
attack
BALB/c
b
jGCaMP7s
jGCaMP7s
ChRmine
DAPI
c
intruder #1
intruder #2
SW attack
BALB/c in cage
z-score
time (s)
neuron
2σ
d
normalized activity in rSLDS latent factors
time (s)
e
time constants (s)
****
dimension
f
line attractor score
g
BALB/c in cage
n = 9 mice
z-score activity
time (s)
h
intruder #1 flow field
intruder #1 in
intruder #1 out
i
single cell contribution of dimension (x1)
weight (abs)
0 neurons
time (s)
SW attack
system in line attractor
j
single cell contribution of dimension (x2)
k
dynamic velocity landscape
intruder #1
velocity
line attractor
l
probability
autocorrelation hw (s)
m
autocorrelation hw (s)
**

Figure 1 | Attractor dynamics in head-fixed mice observing aggression.

a. Experimental paradigm for 2-photon imaging in head-fixed mice observing aggression. b. Representative field of view through a GRIN lens in 2-photon setup (top). Fluorescence image of a coronal slice showing expression of jGCaMP7s and ChRmine (bottom). Scale bars – 100µm. c. Left: Neural and behavioral raster from example mouse observing aggression in the 2-photon setup. Arrows indicate insertion of submissive BALB/c intruder to the observation chamber for interaction with an aggressive Swiss Webster mouse (SW). Right: Example neurons from left. d. Neural activity projected onto rSLDS dimensions obtained from models fit to 2-photon imaging data in one example mouse. e. rSLDS time constants across mice (n = 9 mice, ****p<0.0001, Two-tailed Mann Whitney U-test, error bars - sem). f. Line attractor score (see methods) across mice (n = 9 mice, error bars - sem). g. Behavior triggered average of x1 and x2 dimensions, aligned to introduction of BALB/c into resident's cage (n = 9 mice). Dark line – mean activity, shaded surrounding – sem. h. Flow fields from 2P imaging data during observation of aggression from one example mouse. Blue arrows indicate the direction flow of time. i. Top: Identification of neurons contributing to x1 dimension from rSLDS model. Neuron's weight is shown as absolute value. Bottom: Activity heatmap of five neurons contributing most strongly to x1 dimension. Right: Neural traces of the same neurons and an indication of when the system enters the line attractor. j. Same as 1i but for x2 dimension. k. Dynamic velocity landscape from 2P imaging data during observation of aggression from one example mouse. Blue – stable area in the landscape, red – unstable. Black line – trajectory of neuronal activity. l. Cumulative distributions of autocorrelation half width of neurons contributing to x1 (green) and x2 (red) dimensions (n = 9 mice, 45 neurons each for x1 and x2 distributions). m.Mean autocorrelation half width across mice for neurons contributing to x1 and x2 dimensions (n = 9 mice, **p = 0.0078, Two-tailed Mann Whitney U-test, error bars - sem).

Holographic activation reveals integration in VMHvl

Next, to determine whether VMHvl^{Esr1} line attractor dynamics are intrinsic to the nucleus, after removing the demonstrator mice we performed holographic activation of a subset of neurons contributing to the integration dimension (x_1) using soma tagged ChRmine[39], which was co-expressed with jGCaMP7s (Figure 1b, lower). These neurons were identified in real-time using rSLDS fitting of data recorded during observation of aggression (in a manual closed-loop), followed by 2-photon single cell-targeted optogenetic reactivation of those neurons (Figure 2a). In each field of view (FOV), we concurrently targeted five neurons, chosen by the criteria that they 1) contributed most strongly to a given dimension (x_1 or x_2); and 2) could be reliably re-activated by photostimulation (Figure 2a). Repeated pulses of optogenetic stimulation (2 sec, 20 Hz, 5 mW) were delivered with a 20s inter-stimulus interval (ISI) (Figure 2b-d). Under these conditions we observed minimal off-target effects (Extended Data Fig. 6a-h) and did not observe spatial clustering of x_1 or x_2 neurons (Extended Data Fig. 6i-k, Extended Data Fig. 7a-b, and see Methods).

In this paradigm, optogenetically induced activity along the x_1 (but not the x_2) dimension is predicted to exhibit integration across successive photostimulation pulses, based on the time constants of these dimensions extracted from the fit rSLDS model (Figure 1e). Consistent with this expectation, optogenetic re-activation of cohorts of ~5 individual x_1 neurons yielded robust integration along the x_1 dimension, as evidenced by progressively increasing activity during the ISI following each consecutive pulse (Figure 2c-d; n=8 mice). Activated x_1 neurons exhibited activity levels comparable to their response during observation of aggression (Extended Data Fig. 7d-f). Similar results were obtained using an 8s ISI (Extended Data Fig. 8a-b). Providing the same (digital optogenetic) input to the fit rSLDS model also resulted in integration by the model along the x_1 dimension, similar to that observed in the data (Extended Data Fig. 8c). This activity also scaled with different laser powers (Extended Data Fig. 8e-f). Importantly, x_1 stimulation did not evoke appreciable activity in x_2 dimension neurons (Extended Data Fig. 8g-i).

To visualize in neural state space the effect of re-activating x_1 neurons in the absence of demonstrator mice, we projected the data into a 2D flow-field based on the dynamics matrix fit to data acquired during the observation of aggression. Activation pulses transiently moved the population activity vector (PAV) "up" the line attractor, followed by relaxation back down the attractor to a point that was higher than the initial position of the system (Figure 2e-f). To quantify this effect, we calculated the Euclidean distance in state space between the initial time point during the baseline period (denoted $t_{initial}$), to the time point at the end of stimulation or at the end of the ISI following each pulse (denoted $t_{stim\ end}$ and $t_{post\ stim}$ respectively) (Figure 2e-h). This revealed that the x_1 perturbations resulted in progressive, stable on-manifold movement along the attractor with each consecutive stimulation, as measured by the increase in both metrics (Figure 2g, h). However, we found that integration of optogenetic stimulation pulses saturated in the x_1 dimension after the third pulse, suggesting that the line attractor occupies a finite portion of the neural state space (Extended Data Fig. 9a-d).

Importantly, activation of x_2 neurons did not lead to integration (Figure 2i-k) as predicted by the time constant derived from the fit rSLDS model (Figure 1e, red bar). Instead, following each pulse we observed stimulus-locked transient activity in the x_2 dimension followed by a decay back to baseline during the ISI period, across stimulation paradigms (Figure 2k, Extended Data Fig. 8b), with little to no effect on x_1 neurons (Extended Data Fig. 8j-l). In 2D neural state space, we observed that x_2 neuron activation caused transient off-manifold movements of the PAV orthogonal to the attractor axis during each pulse (Figure 2l-o). Following each stimulus, the PAV relaxed back into the attractor, near the initial location it occupied before the stimulus. The small Euclidean distance between $t_{initial}$ and $t_{post\ stim}$ reflected the attractor's stability (Figure 2o).

To examine further the stability of different points along the line attractor, we performed photostimulation of x_2 neurons after first moving activity in neural state space further along the attractor using photostimulation of x_1 neurons (Extended Data Fig. 9e-f). This x_2

perturbation also resulted in transient off-manifold movements of the PAV orthogonal to the line attractor, followed by relaxation to the position occupied after the previous x_1 stimulation (but prior to the x_2 stimulation) – rather than simply relaxing back to baseline (Extended Data Fig. 9g-i). This experiment confirms the attractive nature of different points along the line. Lastly, activation of randomly selected neurons not weighted by either dimension did not produce activity along either the x_1 or x_2 dimension, emphasizing the specificity of our on- and off-manifold holographic activation (Extended Data Fig. 9j-n). Activation of either ensemble did not result in overt changes in behavior of the head-fixed mouse (Extended Data Fig. 5f-j). Together, these findings demonstrate that a subset of VMHvl^{Esr1} neurons (x_1) can integrate direct optogenetic stimulation, moving the PAV along the line attractor, while a different subset (x_2) pushes the PAV out of the attractor.

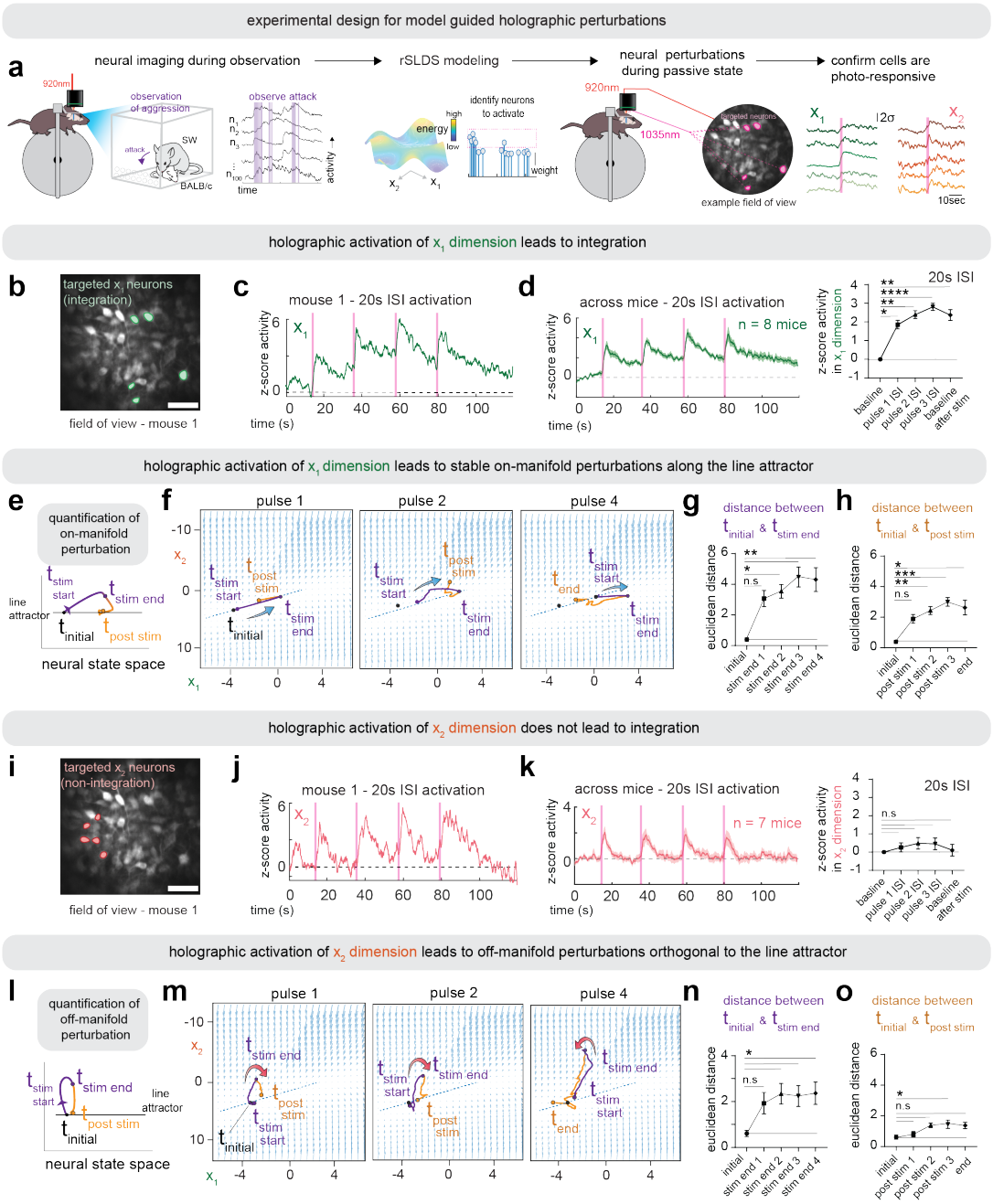


Figure 2 | Holographic perturbations reveal integration dynamics in VMHvl.

a. Experimental paradigm for 2-photon perturbation in head-fixed mice. b. Field of view of five x1 neurons selected for 2-photon activation in example mouse 1. c. Neural activity projected onto x1 dimension after holographic activation of five x1 neurons in example mouse 1. Pink vertical lines – time of activation. d. Left: average activity projected onto x1 dimension from activation of x1 neurons

(average in dark green ± sem in shaded green area, n = 8 mice). Right: Average z-scored activity of projected x1 dimension during baseline or inter stimulus intervals (n = 8 mice, *p = 0.0363, **p = 0.0013, ****p<0.0001, **p = 0.0067. Kruskal-Wallis test with Dunn's correction, error bars - sem). e. Cartoon showing quantification of perturbation along line attractor in neural state space. f. Flow fields from example mouse 1, showing perturbations along line attractor upon activation of x1 neurons. Blue arrows indicate the direction flow of time. g. Euclidian distance between time points tinitial and tstim end across mice (n = 8 mice, n.s p = 0.061, *p = 0.029, **p = 0.0018, **p = 0.0059, Kruskal-Wallis test with Dunn's correction, error bars - sem). h.Same as 2g but for time points tinitial and tpost stim (n = 8 mice, n.s p = 0.1965, **p = 0.0082, ***p = 0.0004, *p = 0.016, Kruskal-Wallis test with Dunn's correction, error bars - sem). i. Field of view of five x2 neurons selected for activation in example mouse 1. j. Neural activity projected onto x2 dimension after holographic activation of x2 neurons in example mouse 1. k. Left: average activity projected onto x2 dimension from activation of x2 neurons (average in dark red ± sem in shaded red area, n = 7 mice). Right: Average z-scored activity of projected x2 dimension during baseline or inter stimulus intervals (n = 7 mice, p>0.99, Kruskal-Wallis test with Dunn's correction, error bars - sem). l. Same as 2e but for x2 activation. m. Flow fields from example mouse 1, showing x2 activation. Red arrows indicate the direction flow of time. n.Same as 2g but for x2 activation (n = 7 mice, n.s p = 0.1554, *p = 0.042, *p = 0.029, *p = 0.029, Kruskal-Wallis test with Dunn's correction, error bars - sem). o.Same as 2h but for x2 activation (n = 7 mice, p>0.05, n.s p = 0.508, *p = 0.0383, n.s p = 0.0508, Kruskal-Wallis test with Dunn's correction, error bars - sem).

Line attractor neurons form ensembles

The integration observed in the foregoing experiments could reflect a cell-intrinsic mechanism, or it could emerge from recurrent interactions within a network[40]. To determine whether the latter mechanism contributes to the line attractor, we first examined whether putative x_1 follower cells (i.e., non-targeted neurons that were photoactivated by stimulation of targeted x_1 neurons) exhibited integration. Indeed, even after excluding the targeted x_1 neurons themselves as well as potentially off-target neurons located within a 50 μm radius of the targeted cell (Extended Data Fig. 6a-h and 10j-n), we observed integration in the remaining x_1 neurons (Extended Data Figure 10a-c). In addition, optogenetically evoked integrated activity in targeted x_1 neurons could be reliably decoded from the activity of their follower x_1 neurons (Extended Data Fig. 10d-f). This decoding was significantly better than that obtained using the activity of non-targeted x_2 neurons; furthermore, the x_2 activity-based decoder performance was slightly worse than decoders trained on neurons chosen randomly (Extended Data Fig. 10g). These analyses suggest that selective functional connectivity between integration dimension-weighted x_1 neurons contributes to line attractor dynamics in VMHvl.

To assess more precisely the extent of functional connectivity among VMHvl^{Esr1} neurons, we activated solitary x_1 or x_2 neurons and performed imaging of non-targeted neurons (Figure 3a). These experiments revealed a slowly decaying elevation of activity during the ISI period in non-targeted x_1 neurons following each pulse of activation (Figure 3b, d) which was mostly positive (Extended Data Fig. 10h, i). Interestingly, the strength of functional connectivity was not positively correlated with distance from the targeted photostimulated cell (Extended Data Fig. 10j-n) and was still observed even after excluding neurons in a 50 μm zone surrounding the targeted neuron to eliminate potential off-target effects due to "spillover" photo-stimulation (Extended Data Fig. 10o, p). Comparing the activity of non-targeted photoactivated x_1 neurons during solitary x_1 neuron photoactivation vs. during targeted 5 x_1 neuron cohort activation revealed that the response strength of the non-targeted

x_1 neurons scaled with the number of targeted x_1 neurons (Extended Data Fig. 10q-r). Importantly, the observed functional coupling between x_1 neurons could not be explained by local clustering of non-targeted x_1 neurons near the targeted cell (Extended Data Fig. 6i-k and 10k-l).

In contrast to the observed x_1-to-x_1 functional connectivity, we observed little activity in non-targeted x_2 neurons following activation of solitary x_1 or x_2 neurons (Figure 3c, e), suggesting that functional x_1-x_1 connectivity is selective. While there was some gradual increase in activity in non-targeted x_1 neurons upon activation of solitary x_2 neurons (Figure 3f-h), that increase was not statistically significant (Figure 3i, j).

The functional connectivity we observed could arise either from a population of sparsely but strongly inter-connected neurons, or from a population with denser connections of intermediate strength[41] (Figure 4a, left). To assess this, we calculated the distribution of pairwise influence scores in our solitary neuron stimulation experiments, defined as the average evoked z-scored activity in each non-targeted photoactivated x_1 neuron following photostimulation of a single targeted cell. To estimate the amount of functional coupling within the x_1 network, we considered the percentage of $x_1 \rightarrow x_1$ pairs that had influence scores higher than the highest $x_1 \rightarrow x_2$ pair, which had a z-score of ~0.6 (Figure 4a, right, vertical line). The fraction of $x_1 \rightarrow x_1$ pairs above this threshold was ~36% (Figure 4a, right). These data suggest that VMHvl^{Esr1} neurons that contribute to the line attractor form relatively dense functional ensembles, confirming theory-based predictions[40].

We next used computational approaches to investigate the kinetics of the observed functional connectivity within x_1 ensembles. Such connectivity could reflect either fast, glutamatergic synapses, as typically assumed for most attractor networks[40]; or they could be slow neuromodulator-based connections that use GPCR-mediated second messenger pathways to sustain long time-scale changes in synaptic conductance. To investigate systematically the density and synaptic kinetics of networks capable of generating line attractors with the

measured integration-dimension (x_I) network time constants, we turned to mechanistic modelling using an excitatory integrate and fire network[7] (Figure 4b). Because VMHvl is >80% glutamatergic[42], we used excitatory networks and analytically calculated the network time constant using an eigen-decomposition of the connectivity matrix[40] (Extended Data Fig. 11a). By varying the synaptic conductance time constant (τ_s) and the density of the integration subnetwork connectivity, we found that only artificial networks based on relatively sparse connectivity (~8-12%) and slow synaptic time constants (~20s) could yield network time constants (τ_n) in the experimentally observed range (~50-200s; Figure 4c, e; red shading). In contrast, networks with fast glutamatergic connectivity failed to do so over the same range of connection densities (Figure 4d, f).

In these purely excitatory network models, the density of connections that yielded network time constants in the observed range was much lower than the experimentally measured value (36%). To match more accurately the empirically observed connection density, we incorporated excitation-recruited fast-feedback inhibition into our integrate-and-fire network[7], since VMHvl is known to receive dense GABAergic innervation from surrounding areas[43,44]. The addition of global strong feedback inhibition allowed networks to match the observed connection density (36%), but importantly, maintained the slow nature of the functional connectivity (20s; Figure 4g; h, left). Indeed, networks simulated with a long τ_s (20s) and dense σ (36%) could integrate digital optogenetic stimulation in a manner like that observed experimentally (Figure 4i-j). In contrast, purely glutamatergic networks (τ_s=100 msec) were unable to integrate at the observed timescales given the measured connectivity density (Figure 4h, right; k-l). Together, these results suggest an implementation of the VMHvl^{Esr1} line attractor that combines slow neurotransmission and relatively dense[41] subnetwork interconnectivity within an attractor creating ensemble.

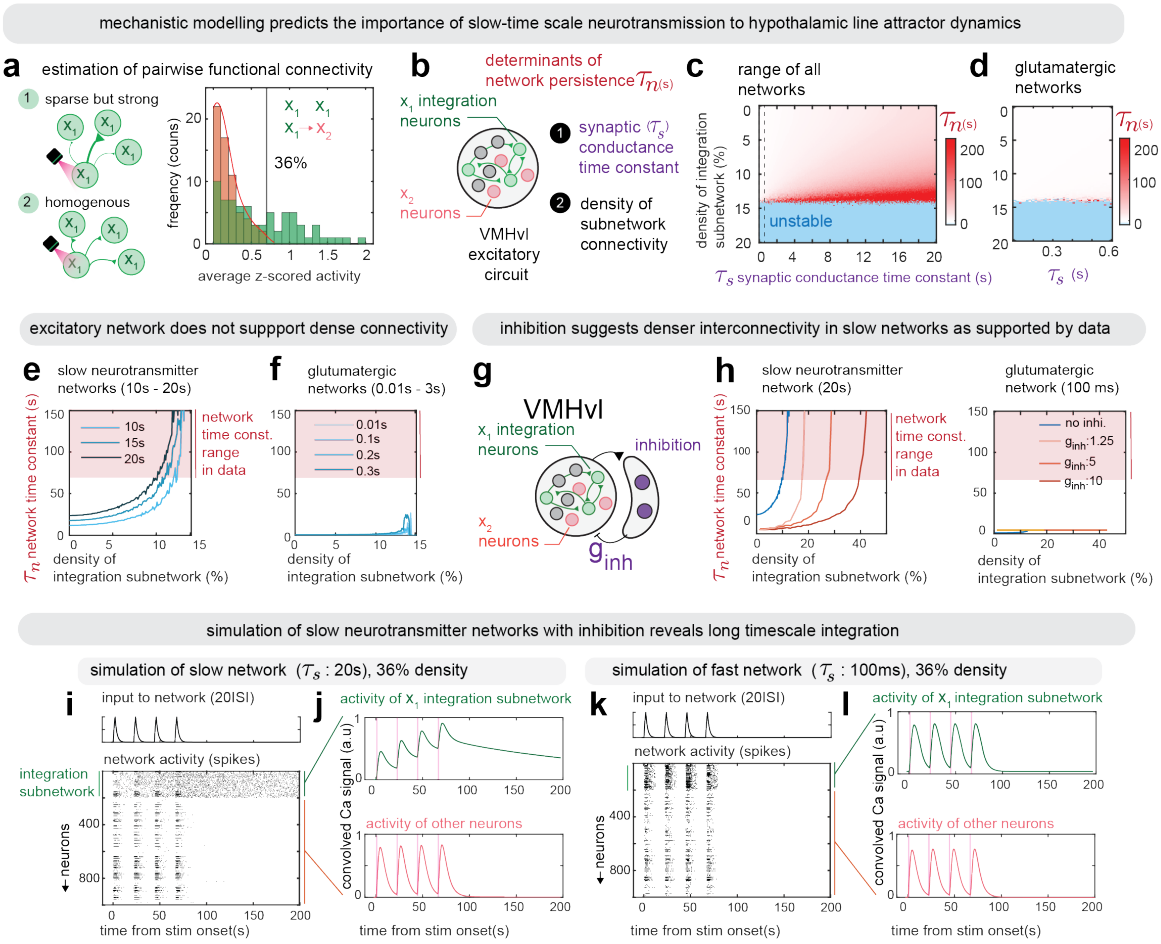


Figure 4 | Mechanistic modelling suggests slow neurotransmission and feedback inhibition.

a. Left: Cartoon illustrating strong but sparse connectivity among x1 neurons (1), or dense interconnectivity within subnetwork (2). Right: Empirical distribution of strength of pairwise functional connectivity between x1 neurons (green) and from x1 to x2 neurons (red) (n = 99 pairs, n = 7 mice). b. Cartoon illustrating different elements of an excitatory network that can determine network level persistent activity. c. Model simulation result showing network time constant (τn) by varying subnetwork connectivity (σ) in range 0-20% density values and τs in range 0-20s. Blue portions - configurations that result in unstable networks with runaway excitation. d. Zoomed in version of 4c (region left of dashed line) showing glutamatergic networks with synaptic conductance time constant (τs) in range 0.01-0.6s. e. Network time constant (τn) against density of integration subnetwork for slow neurotransmitter (τs:10,15,20s). τn varies monotonically with density for large

values of τs. f. Same as 4e but for glutamatergic networks (τs:0.01,0.1,0.2,0.3s). g. Cartoon showing modified VMHvl circuit with fast feedback inhibition incorporated. h. Left: Plot of network time constant (τn) against density of integration subnetwork for a slow neurotransmitter network with τs = 20s, for different values of strength of inhibition (inhibitory gain, ginh: 1.25,5,10). Right: Same as left but for a glutamatergic network with ts= 0.1s. i. Model simulation of a slow neurotransmitter network with fast feedback inhibition (ts:20s, 36% density of subnetwork connectivity). Top: Input (20s ISI) provided to model, Bottom: Spiking activity in network. First 200 neurons (20%) comprise the interconnected integration subnetwork. j. Ca2+ activity convolved from firing rate (see Methods) of integration subnetwork (top) and remaining neurons (bottom). k. Same as 4i but for a fast transmitter network (ts:0.1s, 36% density of subnetwork connectivity). l. Same as 4j but for a fast transmitter network (ts:0.1s, 36% density of subnetwork connectivity).

Attractor stability ties to connectivity

The observed dynamics along the integration dimension exhibits two important characteristics that can reflect the stability of the line attractor: ramping activity up; and slow decay down, the integrator, respectively (Figure 5a). Both of these characteristics might either be intrinsic or be driven by external inputs to the line attractor[5,40]. Previously, we discovered that individual differences in aggressiveness among mice were positively correlated with the stability and decay of the VMHvl line attractor during aggression[5]. We therefore investigated whether individual differences in line attractor ramping or rate of decay might also be correlated with the strength of functional connectivity within the x_1 ensemble (Figure 5b, c). We plotted either the x_1 decay time constants, or the rate of ramp up along the x_1 dimension (obtained from rSLDS models fit to each mouse using data recorded during attack observation), against different quantitative metrics of functional connectivity between targeted x_1 or x_2 neurons and their non-targeted putative follower cells (obtained from the same animals following removal of the demonstrator intruder mice) (Figure 5d, e, Extended Data Fig. 12a).

Strikingly, there was a strong correlation across mice between the time constant of the line attractor measured during the observation of aggression, and the strength of functional connectivity among integration-dimension (x_1) neurons measured by post-observation optogenetic stimulation (Extended Data Fig. 12c, d). The strength of this correlation was higher after the third (r^2=0.87) than the first (r^2=0.59) stimulus (Figure 5g, Extended Data Fig. 12b), indicating that individual differences in integration dynamics become more apparent once the system has already integrated several inputs. In contrast, was no correlation between functional connectivity and the rate of ramp-up, suggesting that the latter might be driven by extrinsic inputs to VMHvl (Figure 5f, Extended Data Fig. 12b-d). Importantly, the correlation between attractor stability and functional connectivity was specific to neurons in the integration (x_1) subnetwork, and did not hold when rSLDS time constants were compared with the influence strength of targeted x_1 neurons on x_2 cells (Extended Data Fig. 12e-h).

Thus, individual differences among mice in the stability of the line attractor during the observation of aggression are correlated with individual differences in the functional connection strength among attractor-contributing neurons.

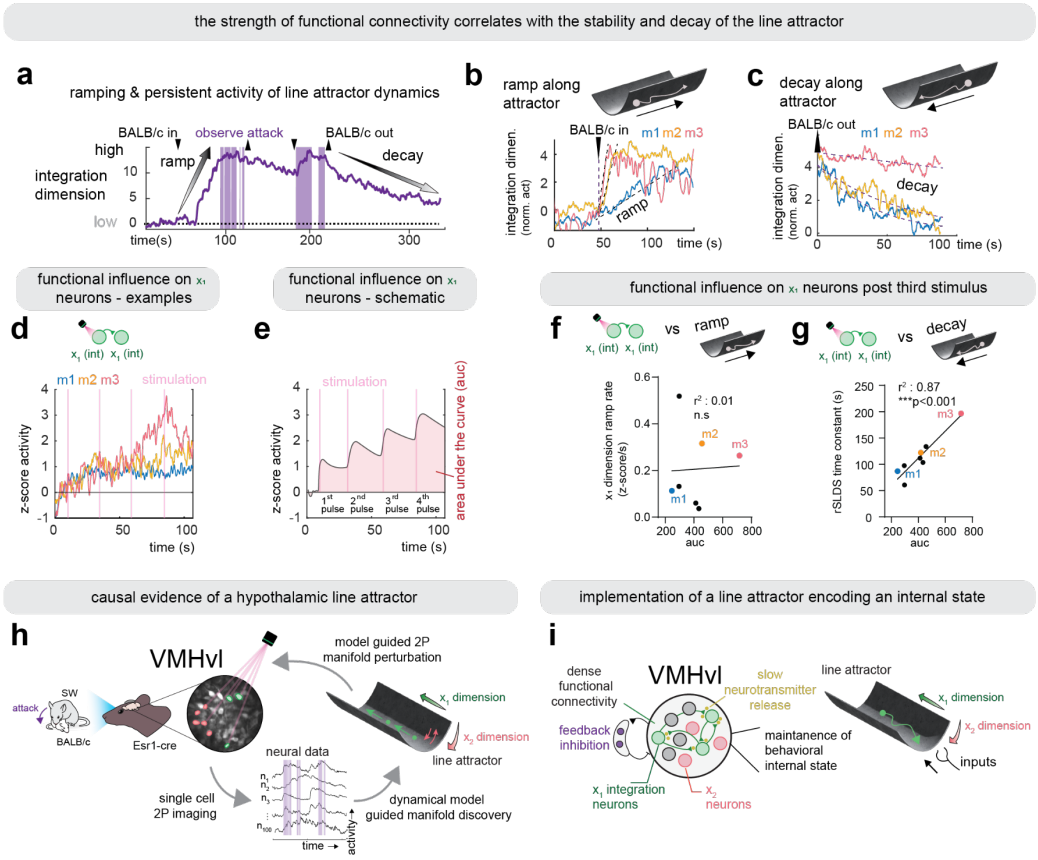


Figure 5 | Strength of functional connectivity reflects line attractor stability.

a. Example neural activity projected onto x1 (integration) dimension of one mouse observing aggression demonstrating the ramp when Balb/c intruder enters the cage (i.e movement up the line attractor) and decay following removal of a Balb/c intruder from cage (i.e movement down the line attractor). b. Dynamics of integration dimension aligned to entry of Balb/c intruder for three example mice. Note the different rates of ramping in different mice. c. Same as 5b, aligned to removal of Balb/c intruder showcasing different rates of decay. d .Z-score activity of non-targeted x1 neurons upon activation of single x1 neurons in the same mice from 5b-c. Pink vertical lines - photostimulation pulses. e. Illustration of different quantification approaches to the change in activity of non-targeted x1 neurons from 5d as either the average z-score activity following first stimulus, or the area under

the curve (auc). Pink vertical lines - photostimulation pulses. f. Correlation between the rate of ramping of the integration dimension obtained from observation of aggression and auc of non-targeted x1 neurons measured using area under the curve (auc) post third stimulus (r2: 0.01, n.s, n = 8 mice). g. Correlation between rSLDS time constant obtained from observation of aggression and auc of non-targeted x1 neurons measured post third stimulus (r2: 0.87, ***p<0.001, n = 8 mice). h. Cartoon depicting summary of results illustrating causal evidence of a hypothalamic line attractor. i. Cartoon depicting implementation of a hypothalamic line attractor encoding a behavioral internal state.

Discussion

Using model-guided closed-loop all-optical experiments, we have provided causal evidence of line attractor dynamics in a mammalian system (Figure 5h, i). Our data and modeling also provided insight into the implementation of the line attractor. We found evidence of relatively dense, selective connectivity among a physiologically distinct subset of *Esr1*$^{+}$ neurons. Whether this subset corresponds to one of the transcriptomically distinct subtypes of *Esr1*$^{+}$ neurons remain to be determined[31]. Our models confirm the importance of rapid feedback inhibition[7], as indicated in invertebrate ring attractor studies[21,45]. However, they differ from conventional continuous attractor models[3,40] by invoking slow neuromodulatory transmission rather than fast glutamatergic excitation. Numerous theoretical studies have posited that continuous attractors relying on recurrent glutamatergic connectivity require precise tuning of synaptic weights to sustain stable attractor dynamics[40,46,47]. The slow neurotransmission predicted by our model may have evolved both to ensure attractor robustness, as well as to implement the relatively long-time scales of internal affective or motive states. These slow dynamics could be implemented by GPCR-mediated signaling triggered by biogenic amines or neuropeptides[48]. Consistent with this prediction, we have recently found that VMHvl line attractor dynamics and aggression are dependent on signaling through oxytocin and/or vasopressin neuropeptide receptors expressed in *Esr1*$^{+}$ neurons[49]. However, that does not exclude a contribution from recurrent glutamatergic excitation in VMH, as in line attractors that mediate cognitive functions on time scales[14,50].

Lastly, our observations indicate a pronounced correlation between individual differences in the functional strength of integration subnetwork connectivity and differences in the measured stability of the line attractor, perhaps reflecting a leaky integrator. Previously we found that in freely behaving animals, individual differences in attractor stability were correlated with individual differences in aggressiveness[5]. By transitivity, this suggests that differences in the strength of functional connectivity within the attractor network might

underlie differences in aggressiveness. Because these differences are observed among genetically identical inbred mice, these observations suggest that attributes of the attractor, such as its connectivity density or strength, may be modifiable (either by genetics and/or experience[26]). Deciphering the underlying mechanisms that grant this attractor its apparent flexibility while maintaining its robustness represents a promising avenue for future research.

Methods

Mice

All experimental procedures involving the use of live mice, or their tissues were carried out in accordance with NIH guidelines and were approved by the Institute Animal Care and Use Committee and the Institute Biosafety Committee at the California Institute of Technology (Caltech). All C57BL/6N (Bl/6N) mice used in this study, including wild-type and transgenic mice, were bred at Caltech. Swiss Webster (SW) male Residents and BALB/c male intruder mice were bred at Caltech. Experimental Bl/6N mice and resident SW mice were used at the age of 8–20 weeks. Intruder BALB/c mice were used at the age of 6–12 weeks and were maintained with three to five cage mates to reduce their aggression. *Esr1* Cre/+ knock-in mice (Jackson Laboratory, stock no. 017911) were back-crossed into the Bl/6N background (>N10) and bred at Caltech. Heterozygous *Esr1*Cre/+ mice were used for cell-specific targeting experiments and were genotyped by PCR analysis using genomic DNA from ear tissue. All mice were housed in ventilated micro-isolator cages in a temperature-controlled environment (median temperature 23 °C, humidity 60%), under a reversed 11/13-h dark/light cycle, with ad libitum access to food and water. Mouse cages were changed weekly.

Viruses

The following adeno-associated viruses (AAVs), along with the supplier, injection titers (in viral genome copies ml–1 (vg ml–1) and injection volumes (in nanoliters), were used in this

study: **AAV1-syn-FLEX-jGCaMP7s-WPRE** (Addgene, no. 104492, roughly 2×10^{12} vg ml–1, 200 nl per injection), **AAVdj-Ef1a-DIO-ChRmine-mScarlet-Kv2.1-WPRE** (Janelia Vector Core, around 2×10^{12} vg ml–1, 200 nl per injection).

Histology

Following completion of 2-photon\miniscope experiments, histological verification of virus expression and implant placement were performed on all mice. Mice lacking virus expression or correct implant placement were excluded from the analysis. Mice were perfused transcardially with 0.9% saline at room temperature, followed by 4% paraformaldehyde (PFA) in 1× PBS. Brains were extracted and post-fixed in 4% PFA overnight at 4 °C, followed by 24 h in 30% sucrose/PBS at 4 °C. Brains were embedded in OCT mounting medium, frozen on dry ice and stored at −80 °C for subsequent sectioning. Brains were sectioned in 80-μm thickness on a cryostat (Leica Biosystems). Sections were washed with 1× PBS and mounted on Superfrost slides, then incubated for 30 min at room temperature in DAPI/PBS (0.5 μg/ml) for counterstaining, washed again and coverslipped. Sections were imaged with epifluorescent microscope (Olympus VS120).

Stereotaxic Surgeries

Surgeries were performed on sexually experienced adult male *Esr1*Cre/+mice aged 6–12 weeks. Virus injection and implantation were performed as described previously[25,51]. Briefly, animals were anaesthetized with isoflurane (5%for induction and 1.5% for maintenance) and placed on a stereotaxic frame (David Kopf Instruments). Virus was injected into the target area using a pulled-glass capillary (World Precision Instruments) and a pressure injector (Micro4 controller, World Precision Instruments), at a flow rate of 50 nl min-1. The glass capillary was left in place for 5 min following injection before withdrawal. Stereotaxic injection coordinates were based on the Paxinos and Franklin atlas[52]. Virus injection: VMHvl, AP: −1.5, ML: ±0.75, DV: −5.75. For 2-photon experiments GRIN lenses (0.6 ×

7.3 mm, Inscopix) were slowly lowered into the brain and fixed to the skull with dental cement (Metabond, Parkell). Coordinates for GRIN lens implantation: VMHvl: AP: −1.5, ML: −0.75, DV: −5.55). A permanent head-bar was attached to the skull with Secure Resin cement (parkell). For micro-endoscope experiments an additional baseplate was attached to the skull (Inscopix).

Housing conditions for behavioral experiments

All male Bl/6N mice used in this study were socially and sexually experienced. Mice aged 8–12 weeks were initially co-housed with a female Bl/6N female mouse for 1 day and were then screened for attack behaviors. Mice that showed attack towards males during a 10 min resident intruder assay were selected for surgery and subsequent behavior experiments. From this point forward, these male mice were always co-housed with a female.

Behavior annotations

Behavior videos were manually annotated using a custom MATLAB-based behavior annotation interface[53,54]. A 'baseline' period of 5 min when the animal was alone in its home cage was recorded at the start of every recording session. Two behaviors during the resident intruder assays were annotated: sniff (face, body, genital-directed sniffing) towards male intruders, and attack (bite, lunge).

Behavioral assays

An observation arena was built from a transparent acrylic (18× 12.5× 18 cm, LxWxH), and a perforated part was put in front of the mice observing aggression. Perforations were 1.27 cm diameter and spread evenly throughout the bottom third of the panel. Before initiation of the assay, the observation arena was scattered with soiled bedding from the cage of the aggressive SW demonstrator. For observation of aggression in freely behaving animals

(miniscope experiments) an observer was first habituates for 15 minutes. Then, a singly housed SW male demonstrator was introduced into the observation arena, followed 1 min later with the insertion of a socially housed stimulus male (BALB/c) in the same compartment. The observation of aggressive encounters persisted for ~1 min, then after 2 minutes a different intruder was introduced for another minute. Observation assays were conducted under white light illumination. For experiments in engaging aggression, the resident mouse was first habituated 15 minutes then a BALB/c intruder mouse was introduced twice for 1-2 minutes. For the experiments comparing neural activity of mice observing aggression and mice engaging aggression, we randomly changed the order of sessions. For mice observing aggression in the 2P setup similar the approach was similar except that the observer mouse was head-fixed and on a treadmill instead of freely behaving in his home cage.

Micro-endoscopic imaging

On the day of imaging, mice were habituated for at least 15 min after installation of the miniscope in their home cage before the start of the behavior tests. Imaging data were acquired at 30 Hz with 2× spatial downsampling; light-emitting diode power (0.1–0.5) and gain (1–7×) were adjusted depending on the brightness of GCaMP expression as determined by the image histogram according to the user manual. A transistor–transistor logic (TTL) pulse from the Sync port of the data acquisition box (DAQ, Inscopix) was used for synchronous triggering of StreamPix7 (Norpix) for video recording.

2-photon imaging and holographic optogenetics

Two to three weeks after surgery mice were habituated to the experimenter's hand by handling 15 minutes a day for three consecutive days. Once animals where habituated to the experimenter's hand, they were manually scooped and gently placed on the treadmill. Mice were head-fixed 3 consecutive days for habituation. Head-fixation was achieved by securing

the head bar into a metal clamp attached to a custom head-stage. During habituation, mice were placed underneath the objective for 15 minutes and given access to random presentations of chocolate milk. Following habituation, combined two-photon imaging and behavior sessions were conducted. jGCaMP7s imaging was acquired via an Ultima 2P Plus and the Prairie View Software (Bruker Fluorescence Microscopy, USA). Individual frames were acquired at 10Hz using a galvo-resonant scanner with a resolution of 1024px x 1024px. We used a long working distance 20x air objective designed for infrared wavelengths (Olympus, LCPLN20XIR, 0.45 numerical aperture (NA), 8.3mm working distance) combined with femtosecond-pulsed laser beam (Chameleon Discovery, Coherent). GCaMP was excited using a 920nm wavelength. For targeted photostimulation, the same microscope and acquisition system (Bruker) was used with a second laser path consisting of a 1035nm high power femtosecond pulsed laser (Monaco 1035-40-40, Coherent), spatial light modulator (512×512-pixel density) to generate multi- point stimulation montages (NeuraLight 3D, Bruker). During photostimulation the mice are head-fixed in complete darkness on a rotating cylinder that enables them to run. Neurons were selected for targeted photostimulation based on two criteria: 1. Their weights from the rSLDS model. 2. If they responded to photostimulation. In case a neuron did not show a significant increase in activity in response to photostimulation, a new neuron was chosen until a total of five photo-sensitive neurons were targeted for each grouped stimulation experiment (Figure 2). During the photostimulation session, a 128-frame average image was generated in order to clearly highlight all neurons. To reduce off-target effect during photo-stimulation, we used small targets (10μm diameter) which were manually restricted to GCaMP expressing neurons. In addition, laser power was adjusted to be a maximum of 5mW per target. We used prairie software to elicit holographic photostimulation (10hz, 2s, 10ms pulse width). Photostimulations were done between frames to avoid laser artefacts. Importantly, to reduce cross activation of the ChRmine from the 920nm laser we kept laser power for imaging to be less than 30mW.

To extract regions of interest, data from mice observing aggression was uploaded to ImageJ. Then, videos were motion corrected using the moco plugin[55]. Motion corrected videos were averaged, and additional contrast and brightness adjustments were made to clearly highlight all neurons in the field of view. Then cells were manually extracted and an rSLDS model was used to identify x_1 and x_2 dimension neurons. Neurons were then identified on the field of view using the prairie view software and were targeted for photo-stimulation. While rSLDS models was running (15-20 minutes, see below), control experiments were conducted.

Micro-endoscopic data extraction

Preprocessing

Miniscope data were acquired using the Inscopix Data Acquisition Software as 2× downsampled .isxd files. Preprocessing and motion correction were performed using Inscopix Data Processing Software. Briefly, raw imaging data were cropped, 2× downsampled, median filtered and motion corrected. A spatial band-pass filter was then applied to remove out-of-focus background. Filtered imaging data were temporally downsampled to 10 Hz and exported as a .tiff image stack.

Calcium data extraction

After preprocessing, calcium traces were extracted and deconvolved using the CNMF-E[56] large data pipeline with the following parameters: patch_dims = [4], gSig = 3, gSiz = 13, ring_radius = 17, min_corr = 0.7, min_pnr = 8. The spatial and temporal components of every extracted unit were carefully inspected manually (SNR, PNR, size, motion artefacts, decay kinetics and so on) and outliers (obvious deviations from the normal distribution) were discarded.

Terminology

We use the following terminology to refer to the design and results of our experiments:

1. "x_1" or "x_2" neurons:" cells that were identified by rSLDS modeling as contributing to dimensions x_1 or x_2, respectively during observation of aggression.

2. "targeted neurons:" rSLDS-identified cells that were purposely photostimulated.

3. "photoactivated" neurons: cells that were empirically found to increase their ΔF/F in response to photostimulation of 1 or more targeted neurons, i.e., photoresponsive neurons. This category includes both the purposely stimulated (targeted) and not purposely stimulated neurons. The latter may include both "off-target" neurons and putative "follower cells."

4. "off-target" neurons: photoactivated neurons that were not purposely photostimulated, but which responded to photostimulation of a selected target cell(s) with an increased ΔF/F because they were close enough to be inadvertently activated by light spillover from the targeted neuron (15μm; see Extended Data Figure 6a-h).

5. "putative follower cells:" neurons that responded to photostimulation and which were outside a 50μm radius around the targeted cell (to conservatively exclude off-target neurons; see Extended Data Figure 6h, 10k-n); they are putative targets (direct or indirect) of the targeted cell.

Dynamical system models of neural data

Recurrent-switching linear dynamical system (rSLDS) models[16,29] are fit to neural data as previously described[15]. Briefly, rSLDS is a generative state-space model that decomposes non-linear time series data into a set of discrete states, each with simple linear dynamics. The model describes three sets of variables: a set of discrete states (z), a set of latent factors (x) that captures the low-dimensional nature of neural activity, and the activity of recorded neurons (y). While the model can also allow for the incorporation of external inputs based on behavior features, such external inputs were not included in our first analysis.

The model is formulated as follows: At each timepoint, there is a discrete state $z_t \in \{1, \ldots, K\}$ that depends recurrently on the continuous latent factors (x) as follows:

$$p(z_{t+1} \mid z_t = k, x_t) = \text{softmax}\{R_k x_t + r_k\} \quad (1)$$

where $R_k \in \mathbb{R}^{K \times K}$ and $r_k \in \mathbb{R}^K$ parameterizes a map from the previous discrete state and continuous state to a distribution over the next discrete states using a softmax link function. The discrete state z_t determines the linear dynamical system used to generate the latent factors at any time t:

$$x_t = A_{z_t} x_{t-1} + b_{z_t} + \epsilon_t \quad (2)$$

where $A_k \in \mathbb{R}^{d \times d}$ is a dynamics matrix and $b_k \in \mathbb{R}^D$ is a bias vector, where D is the dimensionality of the latent space and $\epsilon_t \sim N(0, Q_{z_t})$ is a Gaussian-distributed noise (aka innovation) term.

Lastly, we can recover the activity of recorded neurons by modelling activity as a linear noisy Gaussian observation $y_t \in \mathbb{R}^N$ where *N* is the number of recorded neurons:

$$y_t = C x_t + d + \delta_t \quad (3)$$

For $C \in \mathbb{R}^{N \times D}$ and $\delta_t \sim N(0, S)$, a Gaussian noise term. Overall, the system parameters that rSLDS needs to learn consists of the state transition dynamics, library of linear dynamical system matrices and neuron-specific emission parameters, which we write as:

$$\theta = \{\{A_k b_k, Q_k, R_k, r_k\}_{k=1}^K, C, d, S\} \quad (4)$$

We evaluate model performance using both the evidence lower bound (ELBO) and the forward simulation accuracy (FSA) (Fig. 3a) described in Nair et al., 2023[15] as well as by calculating the variance explained by the model on data.

We employed two-dimensional models, selecting the optimal number of states through 5-fold cross-validation. To ascertain which neurons contributed to each of the two model dimensions (x_1 and x_2), we initially confirmed the orthogonality of these dimensions by

computing the subspace angle between them, (88.1 ± 0. 87º, n = 9 mice). Given this near orthogonality, we then utilized the columns of the emission matrix *C* to identify neurons that contributed to the two separate dimensions of the model.

The contribution of neurons to each latent dimension are defined based on their weights from the emission matrix C, which is initialized by factor analysis and then optimized by rSLDS. In the matrix C, the rows define the weights that create the latent dimensions and columns defined the different latent dimensions (x_1 and x_2) in the model. Model performance is reported both as the evidence lower bound (ELBO) which is equivalent to the Kullback-Leibler divergence between the approximate and true posterior as well as the variance (cvR^2) explained. We cross validated the model using 5-fold cross validation where we trained the data on four arbitrary portions of the data and tested on a left out fifth portion. In all experiments, the model must achieve at least 70% cvR2 before they are used for downstream analysis such as identification of x_1 and x_2 neurons. Models fit to miniscope data during engagement of aggression obtained a cvR2 = 84.7 ± 0.03%, while the same model explains 67.2 ± 0.02% of variance in data obtained from observation of aggression. Flow fields obtained from head-fixed animals observing aggression where fit with input terms representing the presence of the BALB/c intruder.

Estimation of time constants

We estimated the time constant of each dimension of linear dynamical systems using eigenvalues λ_a of the dynamics matrix of that system, derived previously as[57]:

$$\tau_a = \left| \frac{1}{\log(|\lambda_a|)} \right| \quad (5)$$

The intrinsic leak rate is defined based on the time constant of the integration dimension across the whole session. The activity observed by the model takes into account both decays (i.e., the decays after the first and second time the intruder is removed), and therefore gives high prediction to the holographic perturbation experiments (cVR ~85% figures 2f, 2p). Note

also that the dynamics captured by the perturbation experiments more closely resembles the 2nd intruder interaction rather than the first. Furthermore, the SW mouse is still in the observation chamber between BALB/c intruders, but is removed after the 2nd intruder. For this reason, the observed dynamics is mostly consistent and across mice the 2nd decay seems faster.

Calculation of line attractor score

To provide a quantitative measure of the presence of line attractor dynamics, we devised a line attractor score as defined in Nair et al., 2023 as:

$$line\ attractor\ score = \log_2 \frac{t_n}{t_{n-1}} \tag{6}$$

where t_n is the largest time constant of the dynamics matrix of a dynamical system and t_{n-1} is the second largest time constant.

Calculation of auto-correlation half-width

We computed autocorrelation halfwidths by calculating the autocorrelation function for each neuron timeseries data (y_t) for a set of lags as described previously[12]. Briefly, for a time series (y_t), the autocorrelation for lag k is:

$$r_k = \frac{c_k}{c_0} \tag{7}$$

where c_k is defined as:

$$c_k = \frac{1}{T}\sum_{t=1}^{T-k}(y_t - \bar{y})(y_{t+k} - \bar{y}) \tag{8}$$

and c_0 is the sample variance of the data.

Mechanistic modelling

We constructed a model population of N = 1,000 standard current-based leaky integrate-and-fire neurons as previously performed[7]. We first modelled a purely excitatory spiking network in which each neuron has membrane potential x_i characterized by dynamics:

$$\tau_m \frac{dx_i}{dt} = -x_i(t) + g\sum_{j=1}^{N} W p_i(t) + w_i s(t) \quad (9)$$

where $\tau_m = 20ms$ is the membrane time constant, W is the synaptic weight matrix, s is an input term representing external inputs and p represents recurrent inputs. To model spiking, we set a threshold ($\theta = 0.1$), such that when the membrane potential $x_i(t) > \theta$, $x_i(t)$ is set to zero and the instantaneous spiking rate $r_i(t)$ is set to 1.

Spiking-evoked input was modelled as a synaptic current with dynamics:

$$\tau_s \frac{dp_i}{dt} = -p_i(t) + r_i(t), \quad (10)$$

where τ_s is the synaptic conductance time constant. In excitatory networks, the network time constant τ_n was derived as $\frac{\tau_s}{|1-\lambda_{max}|}$, where λ_{max} is the largest eigenvalue of the synaptic weight matrix W[40].

We designed the synaptic connectivity matrix to include a subnetwork of 200 neurons (20% of the network), designated the integration subnetwork as suggested by empirical measurements, with varying densities of random connectivity as highlighted in Fig 3.

Weights of the overall network were sampled from a uniform distribution: $W_{ij} \sim U(0,1/\sqrt{N})$, while weights of the subnetwork were sampled as $W_{ij} \sim U(0,1/\sqrt{N_p})$, where $N_p = 200$.

External input was provided to the network as a smoothened step function consisting of four pulses at 20 ISI as provided in vivo. This stimulus drove a random 25% of neurons in the network.

To account for finite size effects and runaway excitation in networks, we also simulated models with fast feedback inhibition. This was modelled as recurrent inhibition from a single graded input I_{inh} representing an inhibitory population that receives equal input from and provides equal input to, all excitatory units. The dynamics of I_{inh} evolves as:

$$\tau_I \frac{dI_{inh}}{dt} = -I_{inh}(t) + \frac{1}{N}\sum_{n=1}^{N} r_N(t), \quad (11)$$

where $\tau_I = 50ms$ is the decay time constant for inhibitory currents. In this model, outside spiking events, the membrane potential evolved as:

$$\tau_m \frac{dx_i}{dt} = -x_i(t) + g(\sum_{j=1}^{N} Wp_i(t) - g_{inh}I_{inh}(t)) + w_i s(t) \quad (12)$$

Model dynamics were simulated in discrete time using Euler's method with a timestep of 1ms and a small gaussian noise term $\eta_i \sim N(0,1)/5$ was added at each time step. We used g = 1 and varied $g_{inh} = 1,5,10$ as suggested by measurements of inhibitory input to VMHvl[43].

Spatial cluster decoder

To examine whether x_1 and x_2 neurons are spatially clustered in a field of view (FOV), we used a linear support vector machine (SVM) decoder trained to separate cell positions of x_1 and x_2 neurons on each FOV. 'Shuffled' decoder data was generated by randomly assigning neuronal identity. Shuffling was repeated 20 times for each FOV and performance is reported as the average accuracy of each fit decoder.

Decoding behavior from integration dimension.

We trained a frame-wise decoder to discriminate bouts of attack during engaging in aggression from integration dimension activity during observation of attack. We first created

'trials' from bouts of attack during observation and engaging in aggression by merging all bouts that were separated by less than five seconds and balancing the data. We then trained a SVM to identify a decoding threshold that maximally separates the values of our normalized "integration dimension" signal on frames during observation of aggression versus all other frames and tested the accuracy of the trained decoder on held-out frames. 'Shuffled' decoder data was generated by setting the decoding threshold on the same "trial", but with the behavior annotations randomly assigned to each behavior bout. We repeated shuffling 20 times. We then tested the decoder trained on data from observation, on frames during attack while the animals were engaging in aggression. We report performances of actual and shuffled 1D-threshold "decoders" as the average accuracy score of the fit decoder, on data from all other "trials" for each mouse. For significance testing, the mean accuracy of the decoder trained on shuffled data was computed across mice, with shuffling repeated 1000 times for each mouse.

Examining the effect of motion on neural encoding during observation of aggression in head-fixed mice.

We used an analysis designed to detect motion from video recordings of head-fixed mice[58]. To detect motion this method uses singular value decomposition (SVD) to extract groups of pixels showing high differences in luminance or contrast between consecutive frames. We extracted 500 SVDs from our video recordings that reflect different sources of motion including movements of the limbs, whiskers, nose, ears and more. To predict neural activity from behavior, we trained generalized linear models to predict the activity of each neuron k, as a weighted linear combination of the first 10 PCs of the 500 SVDs (reflecting over 90% of the SVDs variance) as follows:

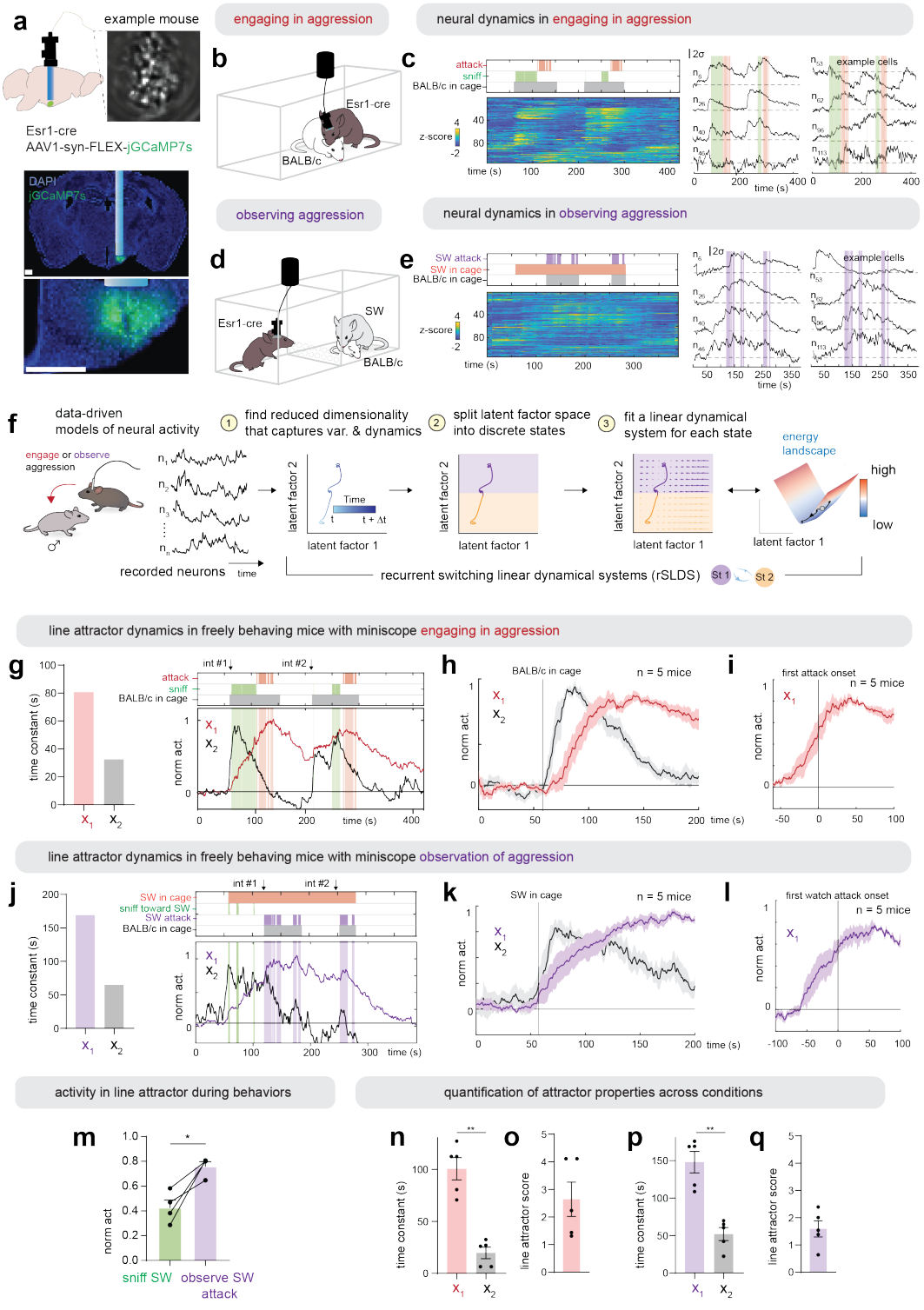
$$y_k(t) = \vec{x}(t)\vec{\beta} + \varphi \quad (13)$$

Here, $y_k(t)$ is the calcium activity of neuron k at time t, $\vec{x}(t)$ is a feature vector of 10 binary reduced SVD dimensions at time lags ranging from t-D to t where D = 10s. $\vec{\beta}$ is a behavior-filter which described how a neuron integrates stimulus over a 10s period (example filters are shown in Extended Data Fig.5c). φ is an error term. The model was fit using 10-fold cross validation with ridge regularization and model performance is reported as cross-validated R^2 (cvR^2).

Statistical analysis

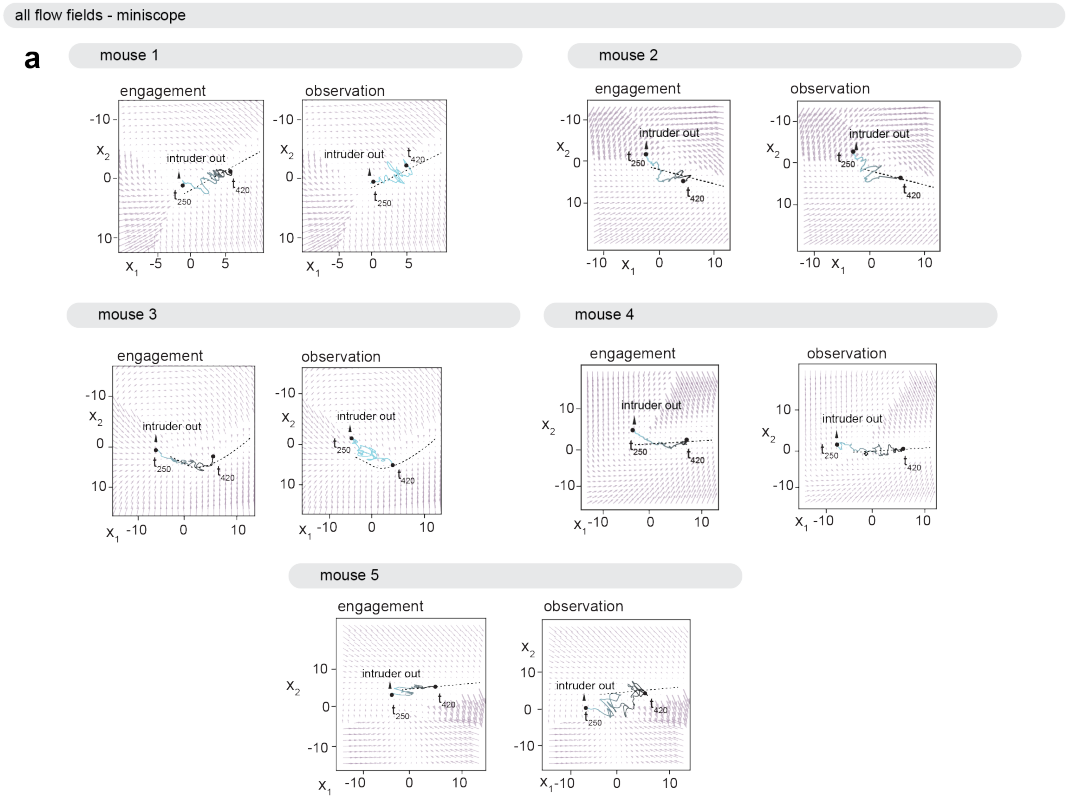
Data were processed and analyzed using Python, MATLAB, and GraphPad (GraphPad PRISM 9). All data were analyzed using two-tailed non-parametric tests. Mann-Whitney *U*-test were used for binary paired samples. Friedman test was used for non-binary paired samples. Kolmogorov-Smirnov test was used for non-paired samples. Multiple comparisons were corrected with Dunn's multiple comparisons correction. Not significant (n.s), $p > 0.05$; *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$; ****$p < 0.0001$.

Extended Data (Supplemental) Figure 1



Extended Data Fig. 1 | Shared line attractor dynamics in engaging and observing aggression. a. Implantation of miniscope, field of view (top) and fluorescence image showing histology (bottom) with jGCaMP7s expression in VMHvl. N = 5 mice. b. Experimental paradigm to record VMHvlEsr1 activity in mice engaging in aggression. c. Left: neural & behavioral raster of example mouse 1 when engaging in aggression. Right: example neurons. d. Experimental paradigm to record VMHvlEsr1 activity in same mice in Ex. Data 1c during observation of aggression. e. Left: neural & behavioral raster of example mouse 1 during observation of aggression. Right: example neurons. f. Overview of rSLDS analysis. g. Left: rSLDS time constants in example mouse 1. Right: Neural activity projected onto two dimensions (x1 & x2) of dynamical system. h. Behavior triggered average of x1 and x2 dimensions, aligned to introduction of male intruder (n = 5 mice, average trace in dark red and black ± sem in shaded area). i. Behavior triggered average of x1 dimensions, aligned to first attack onset (n = 5 mice, average trace in dark red ± sem in shaded red area). j. Left: rSLDS time constants in example mouse 1 during observation of aggression. Right: Neural activity projected onto two dimensions (x1 & x2) of dynamical system. k. Behavior-triggered average of x1 and x2 dimensions from observation of aggression, aligned to introduction of BALB/c into resident's cage (n = 5 mice, average trace in dark purple and black ± sem in shaded area). l. Behavior triggered average of x1 dimensions from observation of aggression, aligned to first bout of observing attack (n = 5 mice, average trace in dark purple ± sem in shaded purple area). m.Average activity in the x1 dimension during sniffing of the SW mouse, vs observing the SW mouse a BALB\c intruder (n = 4 mice, *p = 0.0286, Two-tailed Mann Whitney U-test, error bars - sem). n.rSLDS time constants across mice engaging in aggression (n = 5 mice, *p = 0.0079, Two-tailed Mann Whitney U-test, error bars - sem). o. Line attractor score across mice engaging in aggression (n = 5 mice, error bars - sem). p. rSLDS time constants across mice during observation of aggression (n = 5 mice, *p = 0.0079, Two-tailed Mann Whitney U-test, error bars - sem). q. Line attractor score across mice during observation of aggression (n = 5 mice, error bars - sem).

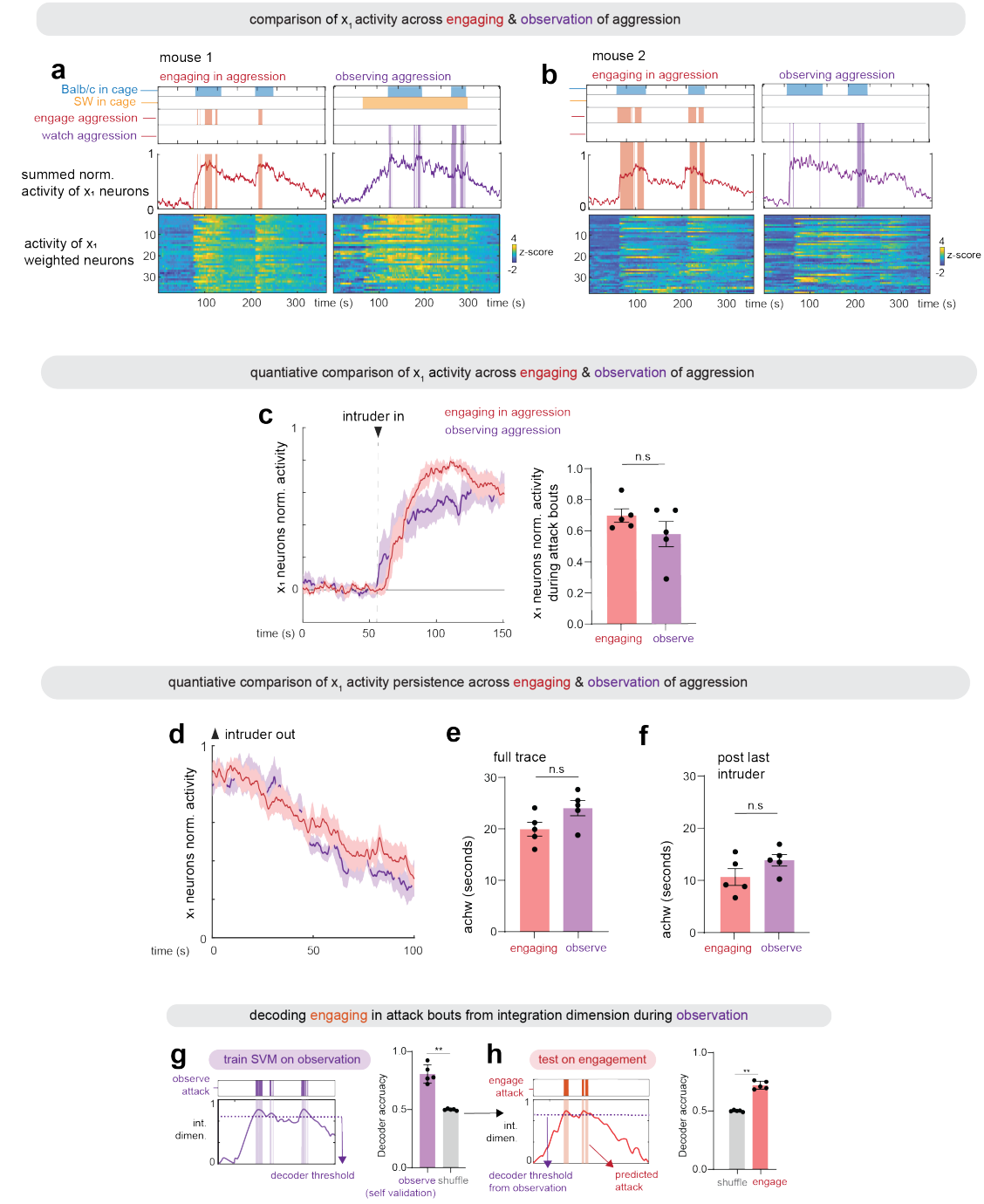
Extended Data Figure 2



Extended Data Fig. 2 | Flow fields from miniscope experiments during engagement and observation of aggression.

Flow fields from all mice showing neural trajectories aligned to removal of the intruder or demonstrator mouse in either observation or engagement of aggression. Dashed lines highlight region of slow points (line attractor).

Extended Data Figure 3

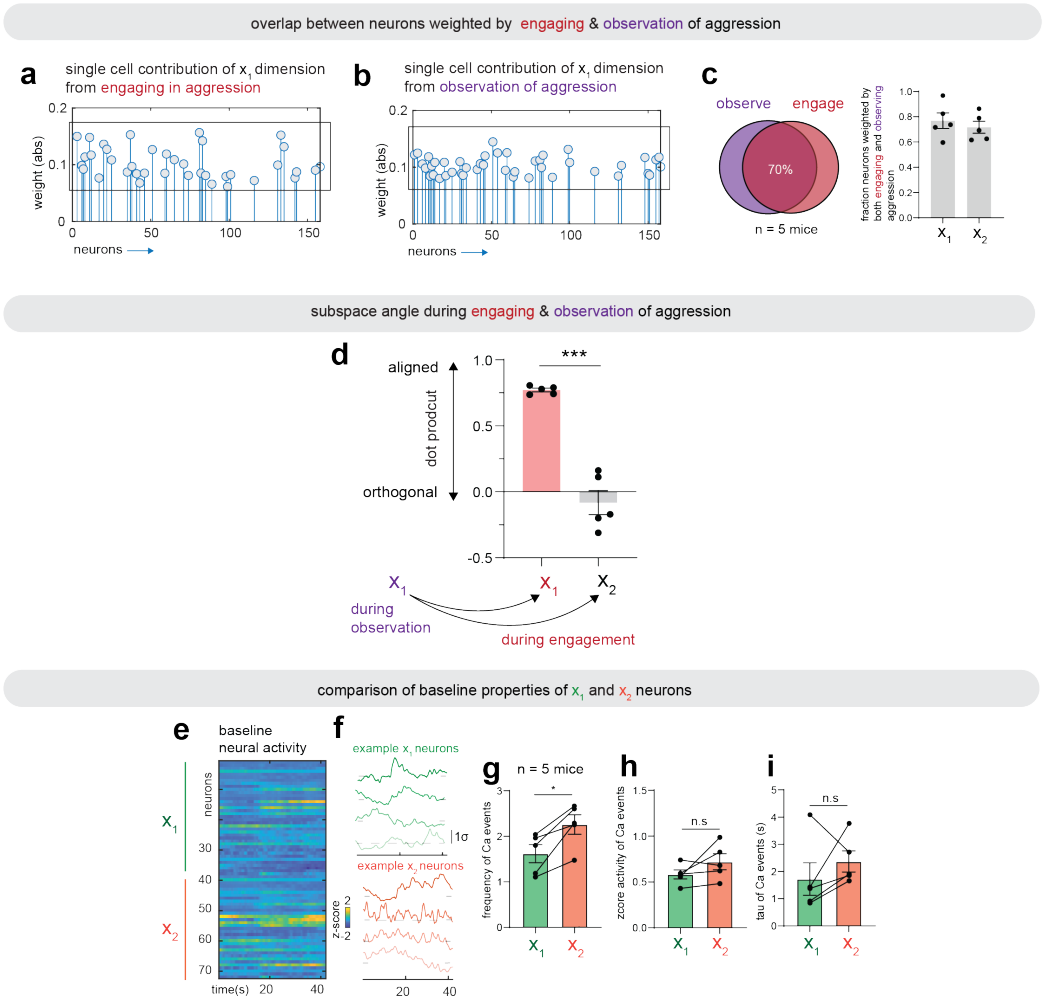


Extended Data Fig. 3 | Comparing neuronal activity of x1 neurons during engaging vs. observing aggression.

a. Normalized neuronal activity of all x1 neurons from example mouse 5 when engaging in aggression (left) and observing aggression (right). Bottom: Raster plots of the activity of all neurons from x1 dimension in mouse 5. b. Same as Extended Data 2c but for example mouse 2.

c. comparing the activity of x1 neurons between observing and engaging in aggression. Left: Average activity across mice (n=5 mice, shaded area is sem). Right: comparison of the activity during observing attack bouts and engaging in attack (n=5 mice, p = 0.42, Two-tailed Mann-Whitney U-test, error bars - sem). d. Activity of x1 neurons aligned to removal of last intruder during observation and engaging in aggression (n=5 mice, shaded area is sem). e. Quantification of autocorrelation half-width for x1 neurons in both conditions during the full interaction (mean achw during observation: 25 ± 0.8s, mean achw during engagement: 20 ± 1.7s, n=5 mice, p=0.125, Two-tailed Mann-Whitney U-test, error bars - sem). f. Quantification of achw for x1 neurons in both conditions aligned to removal of last intruder (mean achw during observation: 14 ± 1s, mean achw during engagement: 11 ± 1.6s, n= 5 mice, p=0.187, Two-tailed Mann-Whitney U-test, error bars - sem). g. Decoding bouts of attack during engaging in aggression from integration dimension activity during observation of attack. Left: Decoder strategy. A SVM decoder was trained on data from integration dimension activity to separate bouts of observing attack from non attack bouts. Right: Quantification of the decoder accuracy performance (n= 5 mice, p = 0.0079, Two-tailed Mann-Whitney U-test, error bars - sem). h. Left: Strategy for testing the decoder. The SVM decoder that was trained on observation of attack is tested with data from engaging in attack. Right: Quantification of the performance of the decoder on engaging vs shuffled data (n= 5 mice, p = 0.0079, Two-tailed Mann-Whitney U-test, error bars - sem).

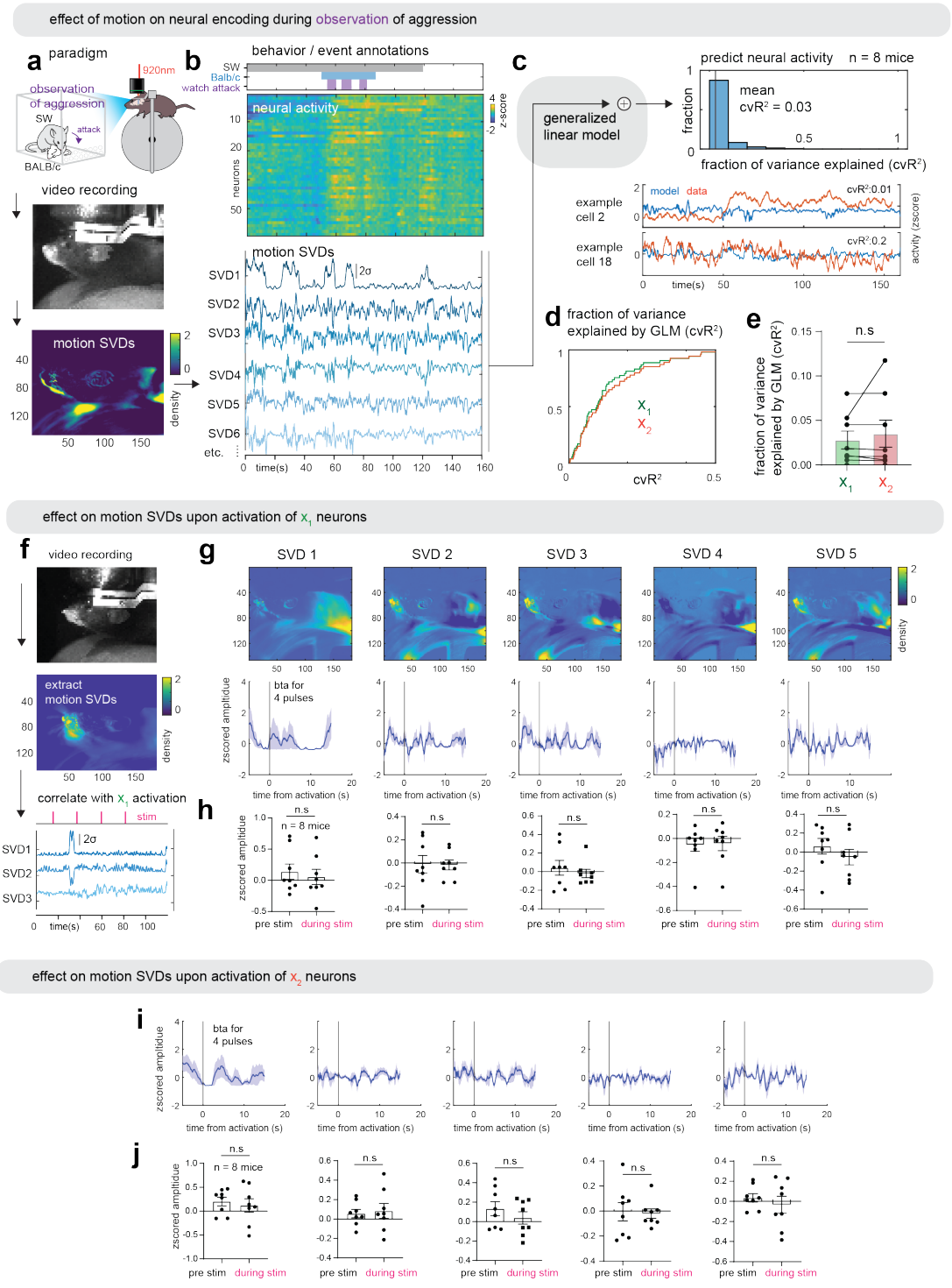
Extended Data Figure 4



Extended Data Fig. 4 | Single cell comparison of integration neurons across conditions.

a. Single cell contribution of x1 dimension (rSLDS weights) from engagement of aggression in example mouse. b. Single cell contribution of x1 dimension (rSLDS weights) from observation of aggression in example mouse. c. Overlap in neurons contributing to line attractor x1 & x2 dimension from rSLDS performing independently in engaging versus observing aggression. Left: Example mouse, Right: Average across 5 mice, error bars - sem. d. Dot product of x1 neural weight vectors during observation vs. engagement in aggression. rSLDS weights of the x1 dimension during observation were compared to model weights of the x1 and x2 dimensions during engagement using a dot product of the two weight vectors. Example raster of baseline activity from one mouse freely behaving while solitary in the cage (n= 5 mice, *p = 0.0079, Two-tailed Mann-Whitney U-test, error bars - sem). e. Example raster of baseline activity from one mouse freely behaving while solitary in its home cage. f. Example single-cell traces from raster in Ex. Data Fig.e. Top - x1 neurons, bottom - x2 neurons. g. Comparison of frequency of Ca+2 transients (above 1.5σ in z-score activity) during baseline recordings across mice (mean frequency x1: 1.6 ± 0.2 events, mean frequency x2: 2.3 ± 0.2 events, n= 5 mice, *p = 0.012, Two-tailed Mann-Whitney U-test, error bars - sem). h. Comparison of the mean amplitude of Ca+2 transients in x1 vs. x2 neurons during baseline recordings, averaged across mice (mean amplitude x1: 0.58 ± 0.04 z-score, mean amplitude x2: 0.71 ± 0.08 z-score, n= 5 mice, p = 0.188, Two-tailed Mann-Whitney U-test, error bars - sem). i. Comparison of the decay time of Ca+2 events during baseline recordings across mice (mean tau x1: 1.7 ± 0.6s, mean tau x2: 2.3 ± 0.4s, n= 5 mice, p = 0.34, Two-tailed Mann-Whitney U-test, error bars - sem).

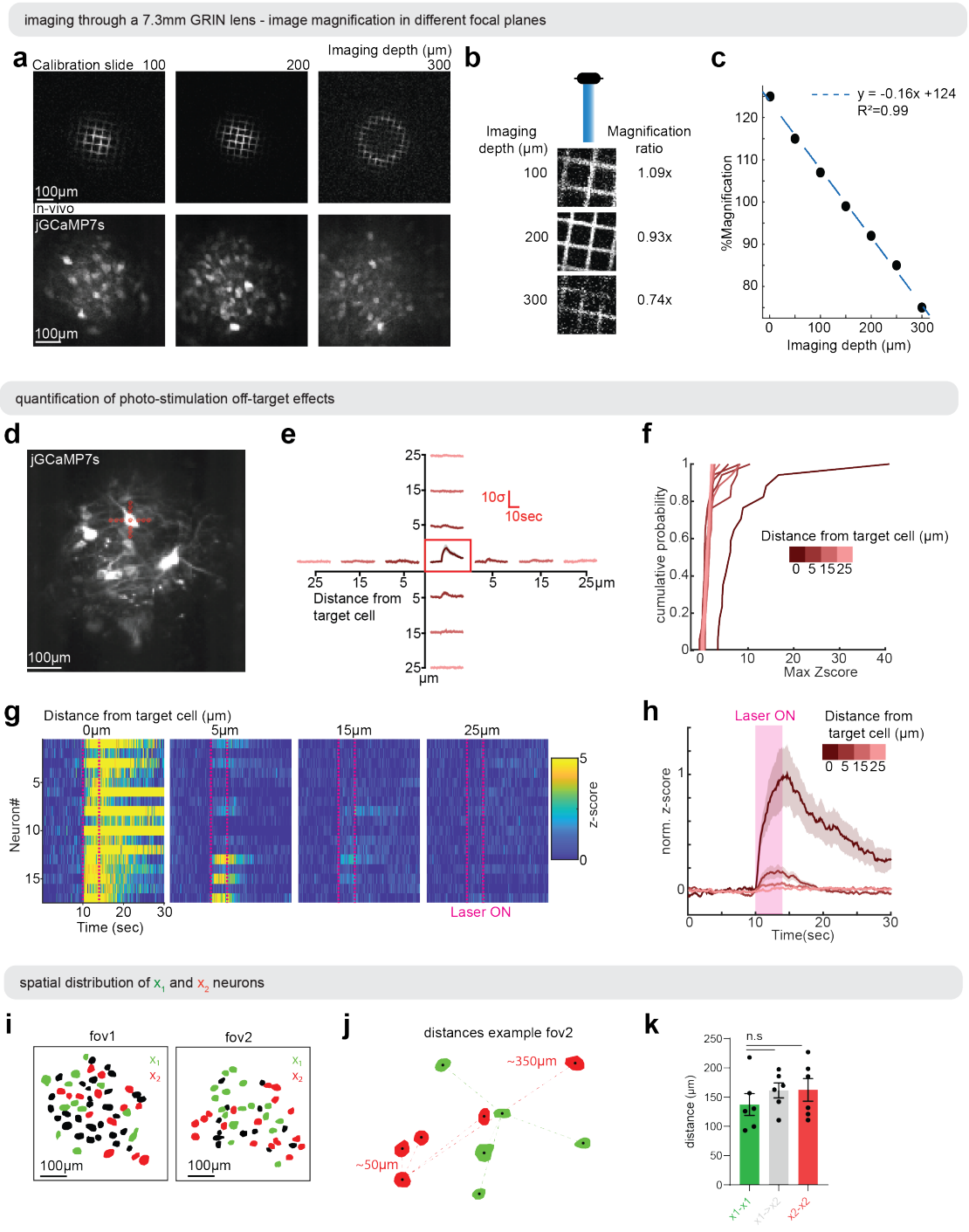
Extended Data Figure 5



Extended Data Fig. 5 | Readouts of behavior and motion in head-fixed mice.

a. Top: Experimental paradigm for 2-photon imaging in head-fixed mice observing aggression. A 920nm 2-photon laser was used to monitor activity of Esr1+ neurons in VMHvl. Middle: One frame from a video recorded during observation of aggression. Bottom: An example of one motion SVD. b. Top: Neural activity raster during observation of aggression. Bottom: examples of SVD outputs over time during observation of aggression. c.Top: Predicted neuronal activity of single neurons and their variance explained by a generalized linear model (GLM) from SVDs readout over time. Bottom: Two example cells with different levels of variance explained. d. Estimated cumulative distribution of variance explained by the GLM of either x1 or x2 neurons across all mice. e.Statistical comparison of variance explained by GLM of x1 activity or x2 activity neurons per mouse (n = 7 mice, p = 0.8125, Two-tailed Mann-Whitney U-test, error bars - sem). f. Top: One frame from a video recorded during group photo-activation of x1 neurons. Middle: An example of one motion SVD. Bottom: Time-evolving activity of top 3 SVDs aligned to x1 activation (vertical red bars = photoactivation pulses). g.Projection of top 5 motion SVDs and stimulus triggered averaged of each SVD aligned to the start of x1 activation. h. Average response in top 5 SVDs during pre-stimulus and stimulus periods (n = 8 mice, p >0.05, Two-tailed Mann-Whitney U-test, error bars - sem). i. Same as g, but for activation of x2 neurons. j. Same as h, but for activation of x2 neurons.

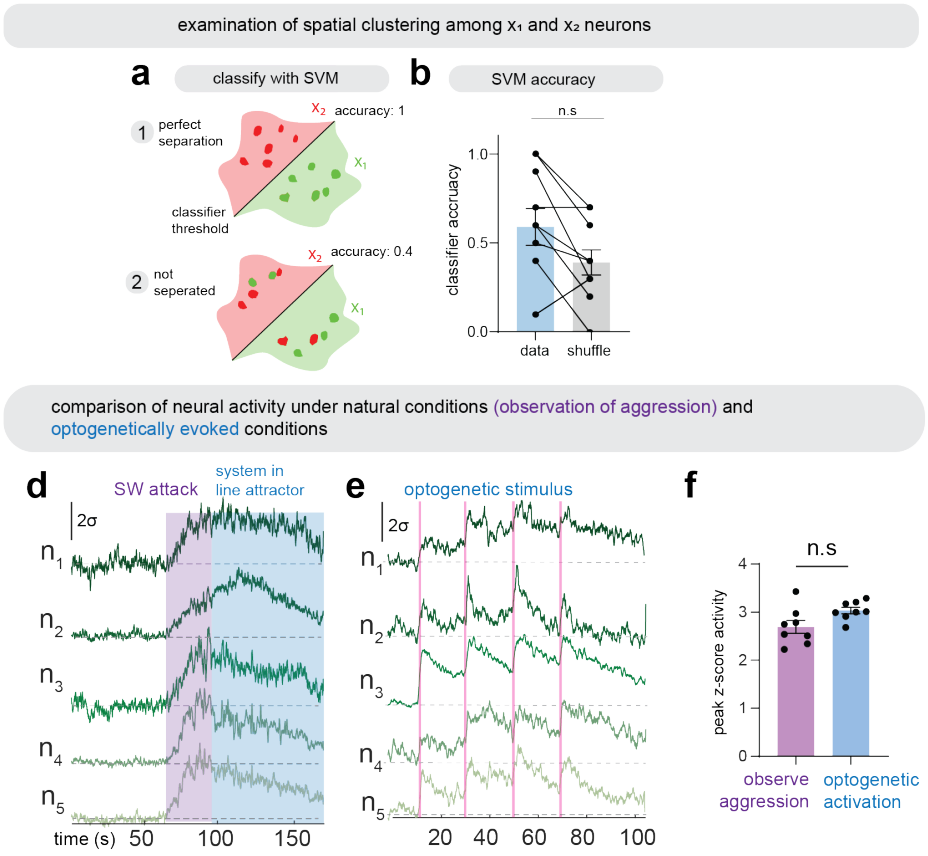
Extended Data Figure 6



Extended Data Fig. 6 | Controls for off-target effects of 2P photoactivation

a. GRIN lens changes the spatial resolution based on the axial depth. Left: imaging a calibration slide with 40μm fluorescent squares at different axial distances below the GRIN lens. Right: imaging in-vivo jGCaMP7s expressing Esr1+ neurons in the VMHvl at different axial distances below the GRIN lens. N = 9 mice. b. Magnification ratio at different imaging depths calculated from the fluorescent calibration slide. c. Quantification of the relationship between imaging depth and magnification error. Linear regression is used to estimate the degree of aberration caused by the GRIN lens. d. Example field of view illustrating the experimental procedure for mapping the spatial resolution of 2P targeted photo-stimulation through the GRIN lens. Reference neurons were targeted first with a spot centered on their somata, and then again stepwise at different distances from the soma center along each of the four cardinal directions, using 10μm diameter stimulation spirals. N = 17 cells. e. Average response of all tested neurons to stimulation at each location from the soma. The red-boxed trace indicates the response observed when the stimulation spot is centered on the reference cell (0 μm).f. Estimated cumulative distribution of the reference cell responses at different distances from soma. Lighter shades of red represent responses at distances progressively further from the soma. n=17 neurons, average trace in dark ± sem in shaded area. g. Average neural activity of all 17 reference neurons tested using the procedure in Ex. Data Fig. e. Shaded area represents standard error of the mean. Note that at 15μm the average response in the reference cells is close to zero. h. Normalized average activity of all neurons at different distances from soma. Each row is a different experiment on a different reference cell. i. Representative examples of field of views from two mice. Green - all x1 neurons, Red – all x2 neurons, black - non x1 or x2 neurons. Fov - field of view. j. Example illustrating how distances are calculated for estimating the spatial clustering of x1 and x2 neurons. k. Quantification of average distance within x1 and x2 neurons and between x1 and x2 neurons, across mice (n=8 mice, $p>0.05$: Kruskal-Wallis test with Dunn's correction for multiple comparison, error bars - sem).

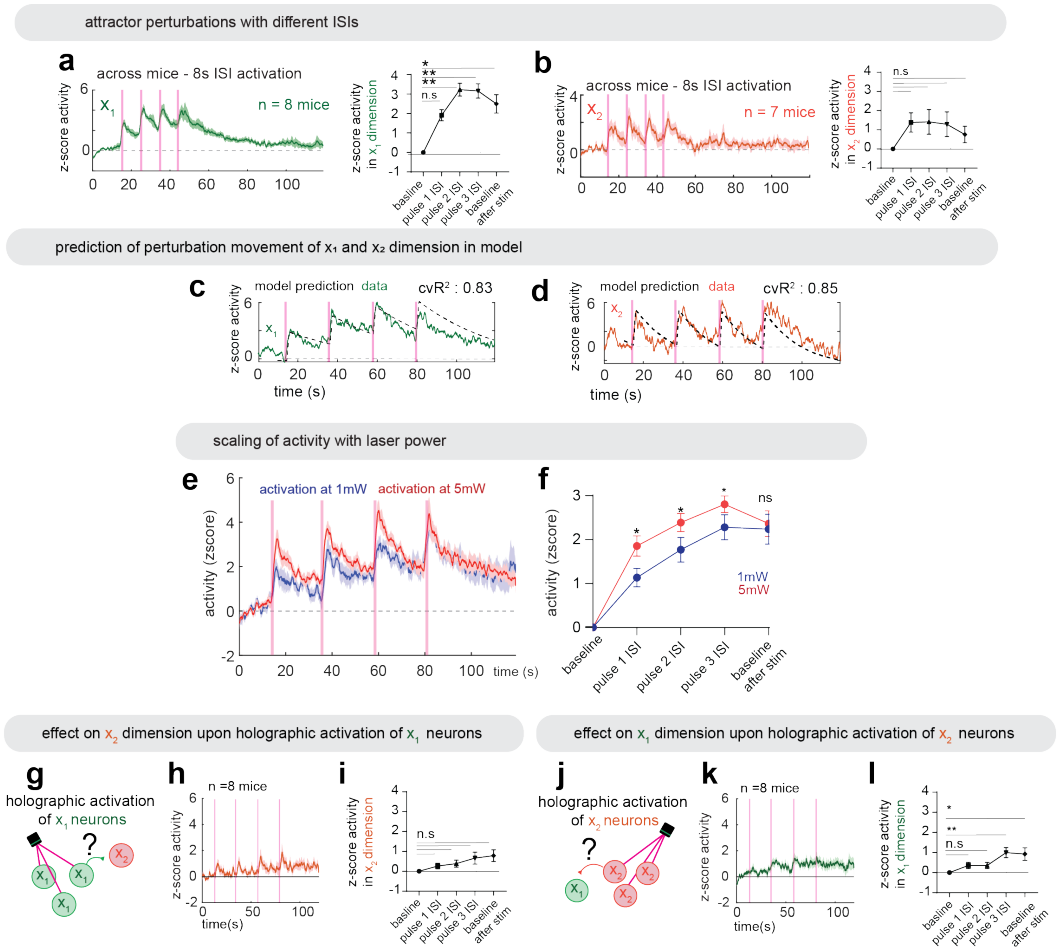
Extended Data Figure 7



Extended Data Fig. 7 | Spatial clustering of neurons and activity comparison.

a. Support vector machine decoder trained to separate cell positions of x1 versus x2 neurons. Scenario 1 shows a cartoon where cells are perfectly separated by the SVM decoder and scenario 2 shows a cartoon where cells are inseparable based on their spatial location and shows low classifier accuracy. b. Accuracy of SVM decoder trained on data versus shuffled control (n=10 mice, p=0.156: Two-tailed Mann-Whitney U-test, error bars - sem). c. Classification width of SVM decoder trained on data versus shuffled control (n=10 mice, p=0.578: Two-tailed Mann-Whitney U-test). d. Neural activity of five x1 neurons selected for grouped optogenetic targeting during observation of aggression. e. Neural activity of same five x1 neurons in Ex. Data 3d during grouped optogenetic activation. f. Comparison of peak z-score of x1 neurons selected for grouped optogenetic activation during observation of aggression and during optogenetic activation (n = 8 mice, p>0.05: Two-tailed Mann-Whitney U-test, error bars - sem).

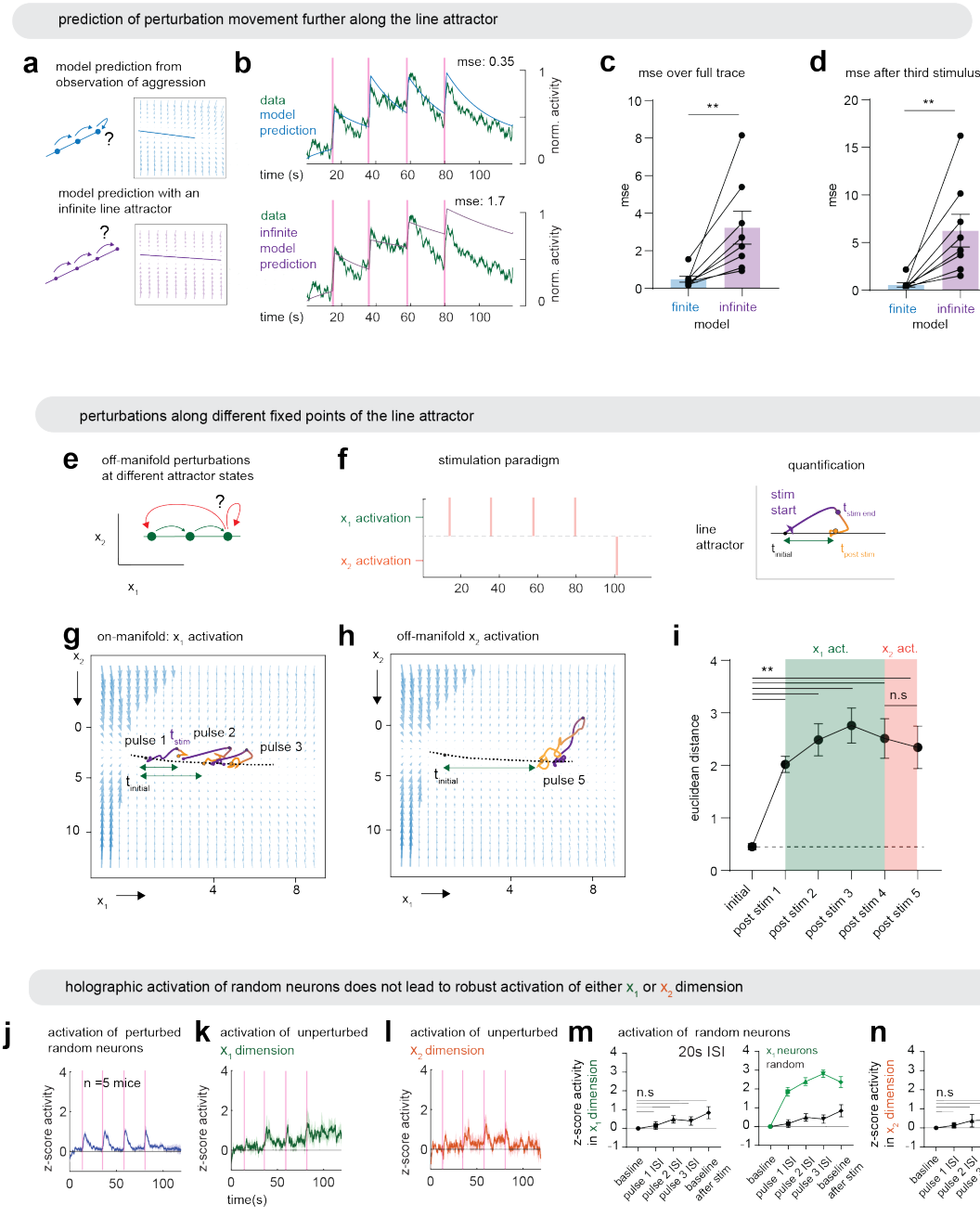
Extended Data Figure 8



Extended Data Fig. 8 | Characterization of line attractor properties.

a. Average activity projected onto x1 dimension from activation of x1 neurons across mice using 8s inter stimulus interval (n = 7 mice). Shaded area – sem. Right: Quantification of average z-scored activity of projected x1 dimension during baseline or inter stimulus intervals (n = 7 mice, n.s p = 0.3, **p = 0.0012, **p = 0.0012, *p = 0.0192 Kruskal-Wallis test with Dunn's correction for multiple comparison, error bars - sem). b. Average activity projected onto x2 dimension from activation of x2 neurons across mice using 8s inter stimulus interval (n = 7 mice). Shaded area – sem. Right: Quantification of average z-scored activity of projected x2 dimension during baseline or inter stimulus intervals (n.s p>0.05, n = 7 mice, Kruskal-Wallis test with Dunn's correction for multiple comparison, error bars - sem). c. Data and model prediction of applying stimulation paradigm in Figure 2c to rSLDS model trained on observing aggression. d. Data and model prediction of applying stimulation paradigm in Figure 2j to rSLDS model trained on observing aggression. e. x1 integration dimension activity with 1mW per neurons (blue) and 5mW per neuron (red). Shaded area – sem. n = 8 mice. f. Quantification of average z-scored activity of projected x1 dimension neurons in 1mW and 5mW per neuron during baseline or various inter stimulus intervals (n = 8 mice, *p = 0.0295, *p = 0.0186, *p = 0.045, n.s p = 0.7, Two-tailed Mann-Whitney U-test, error bars - sem). g. Paradigm for examining activity in x2 dimension upon grouped holographic activation of x1 neurons. h. Average z-score activity of neural activity projected onto x2 dimension across mice (n = 8 mice). Shaded area – sem. i. Quantification of activity in non-targeted x2 dimension upon grouped holographic activation of x1 neurons (n.s, n = 8 mice, Kruskal-Wallis test with Dunn's correction for multiple comparison, error bars - sem). j. Paradigm for examining activity in x1 dimension upon grouped holographic activation of x2 neurons. k.Average z-score activity of neural activity projected onto x1 dimension across mice (n = 8 mice, Shaded area – sem). l. Quantification of activity in non-targeted x1 dimension upon grouped holographic activation of x2 neurons (n.s p = 0.276, n.s p = 0.276, **p = 0.0072, *p = 0.03, n = 8 mice, Kruskal-Wallis test with Dunn's correction for multiple comparison, error bars - sem).

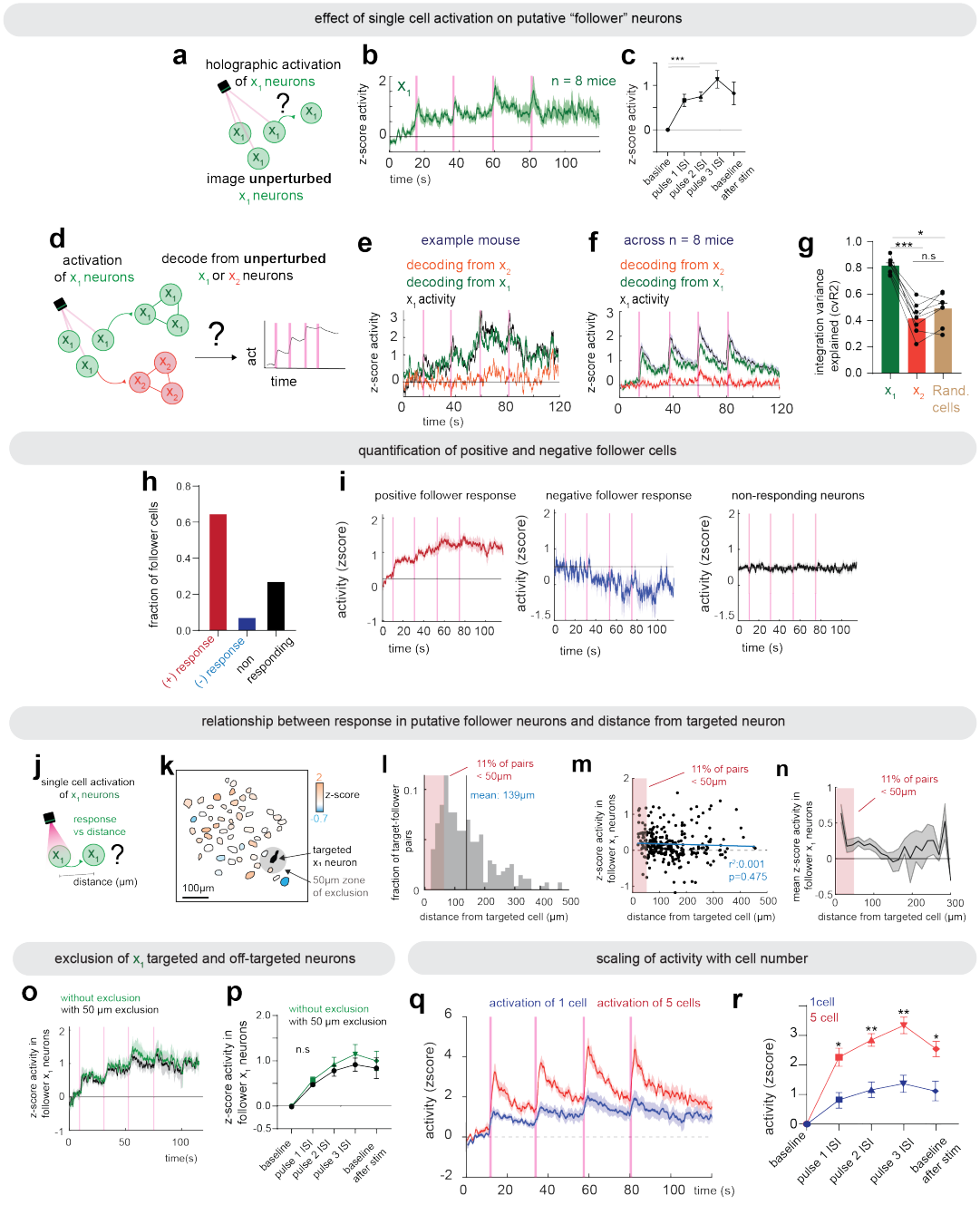
Extended Data Figure 9



Extended Data Fig. 9 | Examination of finite nature and stability of line attractor.

a. Top: model prediction, assuming there is a finite length of the attractor, after the system reaches a certain point along the attractor, further pulses of activity will not cause a further ramp. Bottom: If the line attractor is infinite, then each activation should push the system further along the attractor. b. Example from one mouse comparing the prediction of finite (top) and infinite (bottom) model of the line attractor. Pink lines represent time of photoactivation. Mse - mean square error between model and the data. c. Comparison of the mse of the whole trace between the data and either the finite or infinite models (n = 8 mice, **p<0.001, Two-tailed Mann-Whitney U-test, error bars - sem). d. Same as Ex. Data Fig.9c but comparing only after the third pulse. Note that the scale of the y axis in Ex. Data Fig.9d is twice as big as in Ex. Data Fig.9c (n = 8 mice, **p<0.00, Two-tailed Mann-Whitney U-test, error bars - sem). e. Testing off-manifold perturbations further along the attractor. Experimental design: first we ramp the activity mid-way along the line attractor using activation of x1 neurons, then test the population vector trajectory after targeting of x2 neurons. f. Left: stimulation paradigm. Right: Scheme of the quantification approach for the effect of off manifold targeting further along the attractor. g. State space and the activity ramp following x1 photo-activation (showing only three pulses to avoid clutter). h. Same as Ex. Data Fig. 9g but for x2 photo-activation. i. Quantification of the activity distance from baseline after each photostimulation (n= 8 mice, Kruskal-Wallis test with Dunn's correction for multiple comparison, **p = 0.0025, n.s p>0.05, error bars - sem). j. Effect of grouped holographic activation of randomly selected neurons on activated neurons. Shaded area – sem, n = 5 mice. k. Average z-score activity of non-targeted x1 dimension upon activation of random neurons. Shaded area – sem n = 5 mice. l. Average z-score activity of non-targeted x2 dimension upon activation of random neurons. Shaded area – sem, n = 5 mice. m. Left: Quantification of activity in non-targeted x1 dimension upon grouped holographic activation of random neurons (n.s, p>0.05, Kruskal-Wallis test with Dunn's correction for multiple comparison, n = 5 mice, error bars - sem). Right: Comparison of grouped activation of x1 neurons (green, reproduced from Fig. 2c, right) and grouped activation of random neurons on activity of x1 dimension (black, reproduced from Ex. Data 3m, left, error bars - sem). n. Quantification of activity in non-targeted x2 dimension upon grouped holographic activation of random neurons (n.s, p>0.05, Kruskal-Wallis test with Dunn's correction for multiple comparison, n = 5 mice, error bars - sem).

Extended Data Figure 10

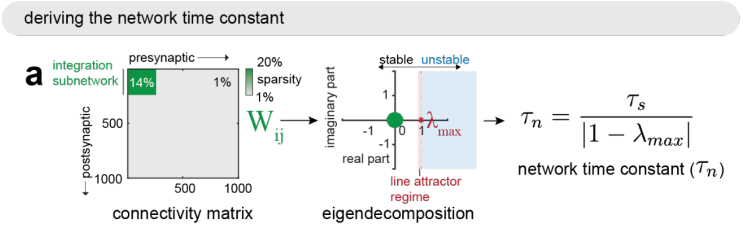


Extended Data Fig. 10 | Impact of functional connectivity measurements on non-targeted neurons a. Experimental design. We grouped activated five x1 neurons (three are shown for illustrative purposes) and examined the activity of non-targeted photoactivated x1 neurons following exclusion of off-target neurons. b. Z-score activity of x1 dimension photoactivated neurons not targeted for photo-stimulation. N = 8 mice. Shaded area – sem. c. Quantification of average z-scored activity of a weighted average of non-targeted x1 dimension neurons during baseline or various inter- photo-stimulation intervals. (n = 8 mice, error bars - sem).d. Experimental design for decoding analysis. We examined whether the activity of non-targeted but photoactivated x1 or x2 dimension neurons can be used to decode integration of direct photo-stimulation by groups of five targeted x1 neurons (three are shown for simplicity), using a support vector machine (SVM) decoder. e. One example mouse showing the activity of targeted x1 dimension neurons (black), activity decoded from non-targeted x1 neurons (green), and activity decoded from x2 non-targeted neurons (orange).f. Same as Ex. Data Fig. 9e but averaged over 8 mice. Shaded area – sem. g. Decoding from non-targeted x1 neurons can explain significantly more variance (80% versus 40%) than non-targeted x2 or randomly selected neurons (n = 8 mice, *p = 0.01, ***p = 0.0003, n.s. p >0.05, Kruskal Wallis test with Dunn's correction for multiple comparisons, error bars - sem). h. Fraction of non-targeted neurons with either positive or negative response (defined by whether their mean response post photostimulation of targeted x1 neuron is 1.5 std above or below baseline activity).i. Averaged activity of non-targeted neurons with either a positive (left), negative (middle) or no significant response (right). Shaded area – sem. N = 8 mice. j. Cartoon illustrating how the relationship between spatial distance and response 4in putative "follower" x1 neurons is assessed. k. Example field of view showing z-score response in all neurons in a field of view. The filled-in black cell is the targeted x1 neuron and the shaded region around it shows a 50µm stringent zone of exclusion. Putative follower cells are shaded according to their z-score response (see color scale). Note that some of the most strongly activated cells are located >100µm from the targeted cell. l. Histogram of distance between targeted x1 neuron and all putative "follower" x1 neurons (mean: 139 ± 35 µm). m. Scatter plot showing the relationship between distance and response in putative "follower" x1 neurons. Blue line shows the regression line. 11% of all assessed putative "follower" x1 neurons are within 50µm of the targeted x1 neurons. n. Average response from scatter plot in 'm'. Black line –mean over moving window of 15um. Shaded area – sem. o. Average response in non-targeted x1 neurons from photo-stimulation of single x1 neuron with (black trace) and without (green trace) exclusion of neurons within a 50µm

radius of the targeted neuron (pink shaded region in Ex. Data. Fig. 10l-n). Shaded area – sem. N = 8 mice.

p. Quantification of data from Ex. Data Fig.10o at various time periods after each photo-stimulation pulse. n.s: not significant, Kruskal-Wallis test with Dunn's correction for multiple comparisons, error bars - sem. N = 8 mice. q. x1 integration dimension activity with activation of one neuron (blue) versus five neurons (red). N = 8 mice. Shaded area – sem. r. Quantification of average z-scored activity of projected x1 dimension neurons with one neuron (blue) versus five neurons (red) during baseline or various inter stimulus intervals. N = 8 mice, *p = 0.0239, **p = 0.0063, **p = 0.0074, *p = 0.0341, Kruskal-Wallis test with Dunn's correction for multiple comparisons, error bars - sem.

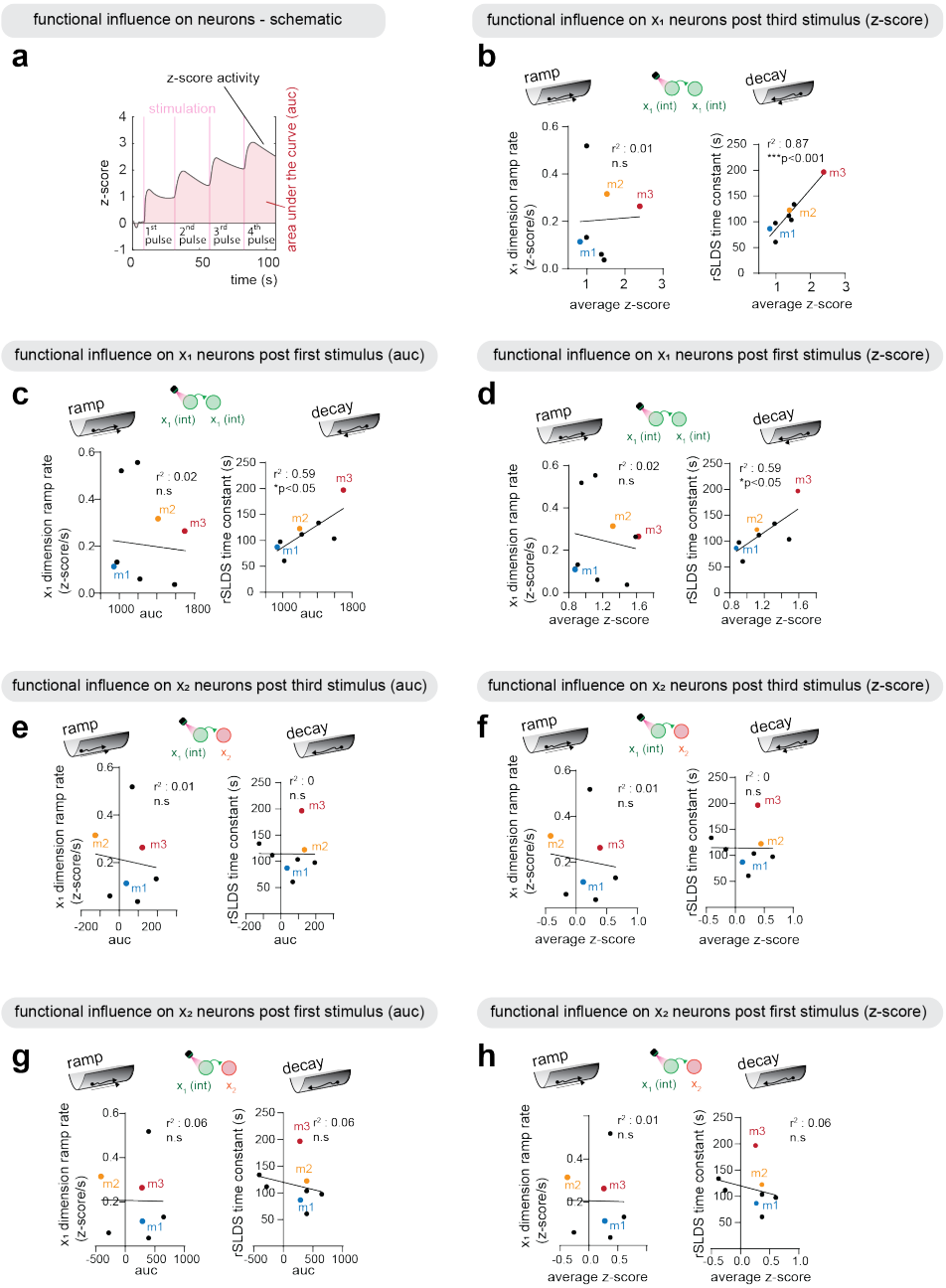
Extended Data Figure 11



Extended Data Fig. 11 | Deriving network time constant for model simulations

a. Analytical derivation of network time constant from connectivity matrix of purely excitatory recurrent neural network.

Extended Data Figure 12



Extended Data Fig. 12 | Additional quantifications of the correlation between functional connectivity and the stability of the decay and ramp.

a. Illustration of different quantification approaches to the change in activity of non-targeted x1 neurons from Main Figure 4b as either the average z-score activity following different stimulus pulses, or the area under the curve (auc). Red vertical lines, photostimulation pulses. b. Left: Correlation between the rate of ramping of the integration dimension obtained from observation of aggression and average z-score of non-targeted x1 neurons measured using the average z-score post third stimulus (r2: 0.01, n.s, n = 8 mice). Right: Correlation between rSLDS time constant obtained from observation of aggression and average z-score across non-targeted x2 neurons measured using the average z-score post third stimulus (r2: 0.87, ***$p<0.001$, n = 8 mice). c. Same as b) but calculated from non-targeted x1 neurons measuring the auc of activity post first stimulus. d. Same as c), calculated from non-targeted x1 neurons measuring the average z-score activity. e. Same as c) but calculated from non-targeted x2 neurons measuring the AUC of activity post third stimulus. f. Same as e) but calculated using the average z-score activity. g. Same as e) but calculated post first stimulus. h. Same as g) but calculated using the average z-score activity.

References

1 Vyas, S., Golub, M. D., Sussillo, D. & Shenoy, K. V. Computation through neural population dynamics. *Annu Rev Neurosci* **43**, 249-275 (2020). https://doi.org:10.1146/annurev-neuro-092619-094115

2 Khona, M. & Fiete, I. R. Attractor and integrator networks in the brain. *Nat Rev Neurosci* **23**, 744-766 (2022). https://doi.org:10.1038/s41583-022-00642-0

3 Langdon, C., Genkin, M. & Engel, T. A. A unifying perspective on neural manifolds and circuits for cognition. *Nat Rev Neurosci* **24**, 363-377 (2023). https://doi.org:10.1038/s41583-023-00693-x

4 Inagaki, H. K. *et al.* Neural algorithms and circuits for motor planning. *Annu Rev Neurosci* **45**, 249-271 (2022). https://doi.org:10.1146/annurev-neuro-092021-121730

5 Nair, A. *et al.* An approximate line attractor in the hypothalamus encodes an aggressive state. *Cell* **186**, 178-193 e115 (2023). https://doi.org:10.1016/j.cell.2022.11.027

6 Yang, T. *et al.* Hypothalamic neurons that mirror aggression. *Cell* **186**, 1195-1211 e1119 (2023). https://doi.org:10.1016/j.cell.2023.01.022

7 Kennedy, A. *et al.* Stimulus-specific hypothalamic encoding of a persistent defensive state. *Nature* **586**, 730-734 (2020). https://doi.org:10.1038/s41586-020-2728-4

8 Zeng, H. What is a cell type and how to define it? *Cell* **185**, 2739-2755 (2022). https://doi.org:https://doi.org/10.1016/j.cell.2022.06.031

9 Tye, K. M. & Uchida, N. Editorial overview: Neurobiology of behavior. *Current Opinion in Neurobiology* **49**, iv-ix (2018). https://doi.org:https://doi.org/10.1016/j.conb.2018.02.019

10 Luo, L. Architectures of neuronal circuits. *Science* **373**, eabg7285 https://doi.org:10.1126/science.abg7285

11 Barack, D. L. & Krakauer, J. W. Two views on the cognitive brain. *Nat Rev Neurosci* **22**, 359-371 (2021). https://doi.org:10.1038/s41583-021-00448-6

12 Hulse, B. K. & Jayaraman, V. Mechanisms underlying the neural computation of head direction. *Annu Rev Neurosci* **43**, 31-54 (2020). https://doi.org:10.1146/annurev-neuro-072116-031516

13 Durstewitz, D., Koppe, G. & Thurm, M. I. Reconstructing computational system dynamics from neural data with recurrent neural networks. *Nat Rev Neurosci* (2023). https://doi.org:10.1038/s41583-023-00740-7

14 Sylwestrak, E. L. *et al.* Cell-type-specific population dynamics of diverse reward computations. *Cell* **185**, 3568-3587 e3527 (2022). https://doi.org:10.1016/j.cell.2022.08.019

15 Rajan, K., Harvey, C. D. & Tank, D. W. Recurrent network models of sequence generation and memory. *Neuron* **90**, 128-142 (2016). https://doi.org:10.1016/j.neuron.2016.02.009

16 Liu, M., Nair, A., Linderman, S. W. & Anderson, D. J. Periodic hypothalamic attractor-like dynamics during the estrus cycle. *bioRxiv* (2023). https://doi.org:10.1101/2023.05.22.541741

17 Inagaki, H. K., Fontolan, L., Romani, S. & Svoboda, K. Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature* **566**, 212-217 (2019). https://doi.org:10.1038/s41586-019-0919-7

18 Daie, K., Svoboda, K. & Druckmann, S. Targeted photostimulation uncovers circuit motifs supporting short-term memory. *Nat Neurosci* **24**, 259-265 (2021). https://doi.org:10.1038/s41593-020-00776-3

19 Carrillo-Reid, L., Han, S., Yang, W., Akrouh, A. & Yuste, R. Controlling visually guided behavior by holographic recalling of cortical ensembles. *Cell* **178**, 447-457 e445 (2019). https://doi.org:10.1016/j.cell.2019.05.045

20 Carrillo-Reid, L., Yang, W., Bando, Y., Peterka, D. S. & Yuste, R. Imprinting and recalling cortical ensembles. *Science* **353**, 691-694 (2016). https://doi.org:10.1126/science.aaf7560

21 Kim, S. S., Rouault, H., Druckmann, S. & Jayaraman, V. Ring attractor dynamics in the Drosophila central brain. *Science* **356**, 849-853 (2017). https://doi.org:10.1126/science.aal4835

22 Green, J., Vijayan, V., Mussells Pires, P., Adachi, A. & Maimon, G. A neural heading estimate is compared with an internal goal to guide oriented navigation. *Nat Neurosci* **22**, 1460-1468 (2019). https://doi.org:10.1038/s41593-019-0444-x

23 Mei, L., Osakada, T. & Lin, D. Hypothalamic control of innate social behaviors. *Science* **382**, 399-404 (2023). https://doi.org:10.1126/science.adh8489

24 Lee, H. *et al.* Scalable control of mounting and attack by Esr1+ neurons in the ventromedial hypothalamus. *Nature* **509**, 627-632 (2014). https://doi.org:10.1038/nature13169

25 Karigo, T. *et al.* Distinct hypothalamic control of same- and opposite-sex mounting behaviour in mice. *Nature* **589**, 258-263 (2021). https://doi.org:10.1038/s41586-020-2995-0

26 Remedios, R. *et al.* Social behaviour shapes hypothalamic neural ensemble representations of conspecific sex. *Nature* **550**, 388-392 (2017). https://doi.org:10.1038/nature23885

27 Linderman, S. W. *et al.* Bayesian learning and inference in recurrent switching linear dynamical systems. *Pr Mach Learn Res* **54**, 914-922 (2017).

28 Carrillo-Reid, L. & Yuste, R. Playing the piano with the cortex: role of neuronal ensembles and pattern completion in perception and behavior. *Current Opinion in Neurobiology* **64**, 89-95 (2020). https://doi.org:https://doi.org/10.1016/j.conb.2020.03.014

29 Emiliani, V. *et al.* Optogenetics for light control of biological systems. *Nature Reviews Methods Primers* **2**, 55 (2022). https://doi.org:10.1038/s43586-022-00136-4

30 Russell, L. E. *et al.* All-optical interrogation of neural circuits in behaving mice. *Nature Protocols* **17**, 1579-1620 (2022). https://doi.org:10.1038/s41596-022-00691-w

31 Kim, D. W. *et al.* Multimodal analysis of cell types in a hypothalamic node controlling social behavior. *Cell* **179**, 713-728 e717 (2019). https://doi.org:10.1016/j.cell.2019.09.020

32 Knoedler, J. R. *et al.* A functional cellular framework for sex and estrous cycle-dependent gene expression and behavior. *Cell* **185**, 654-671 e622 (2022). https://doi.org:10.1016/j.cell.2021.12.031

33 Yang, C. F. *et al.* Sexually dimorphic neurons in the ventromedial hypothalamus govern mating in both sexes and aggression in males. *Cell* **153**, 896-909 (2013). https://doi.org:10.1016/j.cell.2013.04.017

34 Jazayeri, M. & Afraz, A. Navigating the neural space in search of the neural code. *Neuron* **93**, 1003-1014 (2017). https://doi.org:10.1016/j.neuron.2017.02.019

35 Ghosh, K. K. *et al.* Miniaturized integration of a fluorescence microscope. *Nat Methods* **8**, 871-878 (2011). https://doi.org:10.1038/nmeth.1694

36 Lo, L. *et al.* Connectional architecture of a mouse hypothalamic circuit node controlling social behavior. *Proc Natl Acad Sci U S A* **116**, 7503-7512 (2019). https://doi.org:10.1073/pnas.1817503116

37 Sadtler, P. T. *et al.* Neural constraints on learning. *Nature* **512**, 423-426 (2014). https://doi.org:10.1038/nature13665

38 Dana, H. *et al.* High-performance calcium sensors for imaging activity in neuronal populations and microcompartments. *Nature Methods* **16**, 649-657 (2019). https://doi.org:10.1038/s41592-019-0435-6

39 Marshel, J. H. *et al.* Cortical layer–specific critical dynamics triggering perception. *Science* **365**, eaaw5202 (2019). https://doi.org:10.1126/science.aaw5202

40 Goldman, M. S., Compte, A. & Wang, X. J. in *Encyclopedia of Neuroscience* (ed Larry R. Squire) 165-178 (Academic Press, 2009).

41 Brunel, N. Is cortical connectivity optimized for storing information? *Nat Neurosci* **19**, 749-755 (2016). https://doi.org:10.1038/nn.4286

42 Hashikawa, Y., Hashikawa, K., Falkner, A. L. & Lin, D. Ventromedial hypothalamus and the generation of aggression. *Front Syst Neurosci* **11**, 94 (2017). https://doi.org:10.3389/fnsys.2017.00094

43 Yamamoto, R., Ahmed, N., Ito, T., Gungor, N. Z. & Pare, D. Optogenetic Study of Anterior BNST and basomedial amygdala projections to the ventromedial hypothalamus. *eNeuro* **5** (2018). https://doi.org:10.1523/ENEURO.0204-18.2018

44 Minakuchi, T. *et al.* Independent inhibitory control mechanisms for aggressive motivation and action. *Nat Neurosci* (2024). https://doi.org:10.1038/s41593-023-01563-6

45 Franconville, R., Beron, C. & Jayaraman, V. Building a functional connectome of the Drosophila central complex. *Elife* **7** (2018). https://doi.org:10.7554/eLife.37017

46 Sebastian Seung, H. Continuous attractors and oculomotor control. *Neural Netw* **11**, 1253-1258 (1998). https://doi.org:10.1016/s0893-6080(98)00064-1

47 Seung, H. S. How the brain keeps the eyes still. *Proc Natl Acad Sci U S A* **93**, 13339-13344 (1996). https://doi.org:10.1073/pnas.93.23.13339

48 Robinson, D. A. Integrating with neurons. *Annu Rev Neurosci* **12**, 33-45 (1989). https://doi.org:10.1146/annurev.ne.12.030189.000341

49 Mountoufaris, G., Nair, A., Yang, B., Kim, D.-W. & Anderson, D. J. Neuropeptide signaling is required to implement a line attractor encoding a persistent internal behavioral state. *bioRxiv*, 2023.2011.2001.565073 (2023). https://doi.org:10.1101/2023.11.01.565073

50 Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78-84 (2013). https://doi.org:10.1038/nature12742

51 Yang, B., Karigo, T. & Anderson, D. J. Transformations of neural representations in a social behaviour network. *Nature* **608**, 741-749 (2022). https://doi.org:10.1038/s41586-022-05057-6

52 Paxinos, G. & Franklin, K. B. *Paxinos and Franklin's the mouse brain in stereotaxic coordinates*. (Academic Press, 2019).

53 Lin, D. *et al.* Functional identification of an aggression locus in the mouse hypothalamus. *Nature* **470**, 221-226 (2011). https://doi.org:10.1038/nature09736

54 Segalin, C. *et al.* The Mouse Action Recognition System (MARS) software pipeline for automated analysis of social behaviors in mice. *eLife* **10**, e63720 (2021). https://doi.org:10.7554/eLife.63720

55 Dubbs, A., Guevara, J. & Yuste, R. moco: Fast motion correction for calcium imaging. *Frontiers in Neuroinformatics* **10** (2016). https://doi.org:10.3389/fninf.2016.00006

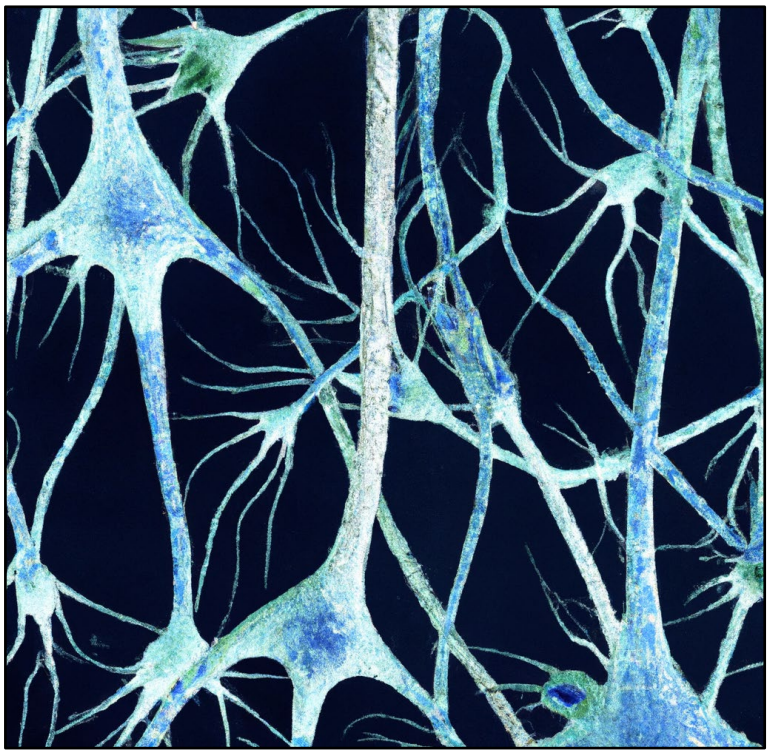
56 Zhou, P. *et al.* Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. *eLife* **7**, e28728 (2018). https://doi.org:10.7554/eLife.28728

57 Maheswaranathan, N., Williams, A., Golub, M., Ganguli, S. & Sussillo, D. Vol. 32 (eds H. Wallach *et al.*) (2019).

58 Syeda, A. *et al.* Facemap: a framework for modeling neural activity based on orofacial tracking. *Nat Neurosci* **27**, 187-195 (2024). https://doi.org:10.1038/s41593-023-01490-6

Chapter IV

IMPLEMENTATION



"പ്രകൃതി ഓരോ അറ്റോംയോടുംപോലും സ്നേഹിച്ചപ്പോൾ, അവയൊക്കെ ചേർന്ന് ഈ ഭ്രമമായ ജീവിതത്തിന്റെ ഒരു പാട്ടുപോലെയായി."

Sugathakumari, Ambalamani, 1994

Translation: "When nature loved even each atom, they all came together like a song in this illusory life."

Chapter IV

A line attractor encoding a persistent internal state requires neuropeptide signaling

This chapter details efforts to probe the mechanisms underlying line attractor dynamics and uncovers a non-canonical implementation of attractor dynamics based on neuropeptides

Published as and adapted from George Mountoufaris, **Aditya Nair,** Bin Yang, Dong-Wook Kim, Amit Vinograd, Sam Kim, Scott W. Linderman and David J. Anderson. A line attractor encoding a persistent internal state requires neuropeptide signaling. *Cell*. (2024)

Summary

Internal states drive survival behaviors, but their neural implementation is poorly understood. Recently we identified a line attractor in the ventromedial hypothalamus (VMH) that represents a state of aggressiveness. Line attractors can be implemented by recurrent connectivity or neuromodulatory signaling, but evidence for the latter is scant. Here we show that neuropeptidergic signaling is necessary for line attractor dynamics in this system, using a novel approach combining cell type-specific CRISPR/Cas9-based gene editing with single-cell calcium imaging. Co-disruption of receptors for oxytocin and vasopressin in adult VMH Esr1^{+} neurons that control aggression suppressed attack, reduced persistent neural activity and eliminated line attractor dynamics, while only slightly reducing overall neural activity and sex- or behavior-specific tuning. These data identify a requisite role for neuropeptidergic signaling in implementing a behaviorally relevant line attractor in mammals. Our approach should facilitate mechanistic studies in neuroscience that bridge different levels of biological function and abstraction.

Introduction

Innate survival behaviors such as aggression, mating, feeding and defense are driven by internal motivational or affective states[1-3], which are experienced in humans as subjective feelings[4,5]. How and where such internal states are encoded in the brain, and how they are causally related to overt behavior, is emerging as a major topic in circuit and systems neuroscience[6,7].

The study of internal states has been pursued via two approaches that have until recently remained relatively separate. One, a "bottom-up" approach, employs genetically or pharmacologically based manipulations of genes and neural circuits[6,8,9] aimed at providing causal explanations for behavioral and psychological internal states[10-12]. The other, a "top-down" approach, identifies internal states computationally in high-dimensional neural population activity[13,14]. The latter has revealed attractors as a mechanism for encoding low-dimensional variables underlying cognitive functions[15-19]. More recently, such models have been applied in behavioral neuroscience as well[20-22]. To test the causal role of such attractors it is important to understand their neural implementation at the level of cell types and genes. This in turn requires integration of these two approaches[23], which has been accomplished in very few systems[24,25].

Persistent neural activity (on a timescale of seconds to minutes) is a characteristic feature of neural integrators and attractor dynamics[18,26,27]. Two alternative (but not mutually exclusive) implementation mechanisms are typically invoked to explain such persistence: recurrent fast synaptic connectivity or slow neuromodulation[28]. While there is evidence of recurrent connectivity underlying a ring attractor that encodes head direction in *Drosophila*[24,25,29], to our knowledge there is no evidence of any neuromodulator that controls attractor dynamics in any system.

Neuropeptides comprise a class of evolutionarily conserved neuromodulators[30,31] that control behavior-specific internal motive states associated with mating[32,33], aggression[34-36], social

attachment[37], as well as other behaviors. Neuropeptides are well known to modulate synaptic strength and neural circuit properties such as patterns of oscillation[38-40], but their role in implementing neural integrator and attractor dynamics has not been extensively studied in vertebrates. Experiments in *C. elegans* have identified neuropeptides that control persistent states of locomotor activity[41,42], but whether they influence dynamical manifolds identified in that system[43] is not yet clear .

A powerful approach to this question is to combine cell type-specific genetic perturbations of neuromodulatory signaling with simultaneous large-scale recording of neural activity in the same brain region and genetically defined cell type. While these experimental modalities have been successfully integrated in *C. elegans*[42,44], *D. melanogaster*[45,46] and larval zebrafish[47], they have been difficult to combine in mammalian systems, for technical reasons (Supplementary Figure S1A).

Here we use a novel viral-based strategy that integrates cell type-specific CRISPR/Cas9-based multiplex gene editing[48,49] with single unit-resolution calcium imaging of neural dynamics in freely behaving adult animals[50], which we call "CRISPRoscopy". This method, when combined with dynamical systems modeling[51,52], allows investigation of the effects of local inactivation of different neuromodulatory systems on neural population coding, dynamics and behavior, in the same brain region and cell type during naturalistic behaviors.

As a proof-of-concept application of this approach, we have examined the role of oxytocin (OXT) and arginine vasopressin (AVP) signaling in a population of ventromedial hypothalamic (VMH) neurons that control aggression[53,54]. We chose these peptides for several reasons. First, they have been widely implicated in the control of social behaviors[37,55] (although the role of OXT in social behaviors, such as aggression, has been controversial[56-58]). Second, VMH neurons are known to express receptors for OXT and AVP[59,60] and infusion of the latter into VMH can enhance aggression in hamsters[61]. Third, aggression is an instinctive and phylogenetically widespread social behavior that expresses an internal

affective state[62,63]. Finally, dynamical systems modeling[51,52] of population activity from estrogen receptor-1 (Esr1)-expressing neurons in the ventrolateral subdivision of the VMH (VMHvl^{Esr1})[64] has revealed an approximate line attractor (or leaky integrator). This attractor is intrinsic to VMHvl[65] and represents a scalable, persistent aggressive internal state[21].

Here we show that genetic perturbation of OXT and AVP receptors in VMHvl^{Esr1} neurons disrupts aggressive behavior, persistent activity and line attractor dynamics, while only modestly affecting overall neuronal activity and population coding of behavior or intruder sex[66]. These data provide evidence of a requirement for neuropeptides in line attractor dynamics and strengthen the link between such dynamics and an internal affective state.

Results

Oxtr/Avpr1a-mediated neuropeptidergic signaling in VMH is required for male territorial aggression

While OXT, AVP and their receptors have been studied extensively in rodent aggression using pharmacologic reagents and zygotic gene knockouts[32,56,75-78], there are no reports of a specific requirement in offensive aggression for either *Oxtr* or *Avpr1a* in murine VMHvl. Using a newly developed, Cre-dependent gene editing strategy, we investigated the effect of *Oxtr/Avpr1a* co-editing on social behaviors in mice injected in VMH bilaterally with the OAR-gRNA LV (control) and a Cas9 AAV (experimental group), using a standard resident intruder (RI) assay. We used single-housed, sexually experienced wild-type C57BL/6N resident males pre-selected for adequate aggressiveness (Figure 2F, see Methods). Control animals were co-injected bilaterally with the scrambled gRNA (Scr gRNA) and Cas9 viruses (control group). Experimental mice displayed a notable reduction in aggression towards male intruders, as evidenced by significant decreases in the number and time-varying probability of attack bouts, the total time spent attacking and the average duration of each attack bout; in addition, the latency to the 1st attack bout was significantly increased (Figure 2Ai-ii and

Supplemental Figure S1A, Bi, ii). These behavioral effects were not due to defects in locomotor activity since average velocity during attack episodes was similar between experimental and control mice (Supplemental Figure S1Biii). In contrast, the time spent in close investigation (sniffing) of male intruders did not differ significantly between experimental and control male residents (Figure 2Aiii). Experimental males also did not differ significantly from controls in their sniffing or mounting behavior towards female intruders (Figure 2B and Supplemental Figure S1C).

These data suggest that *Oxtr* and/or *Avpr1a* expressed in VMHvl neurons play a requisite and selective role in aggressive behavior. Furthermore, they motivated us to analyze next how disrupting these receptors specifically in Esr1 neurons affects behavior, neural activity, population coding and network dynamics *in vivo*.

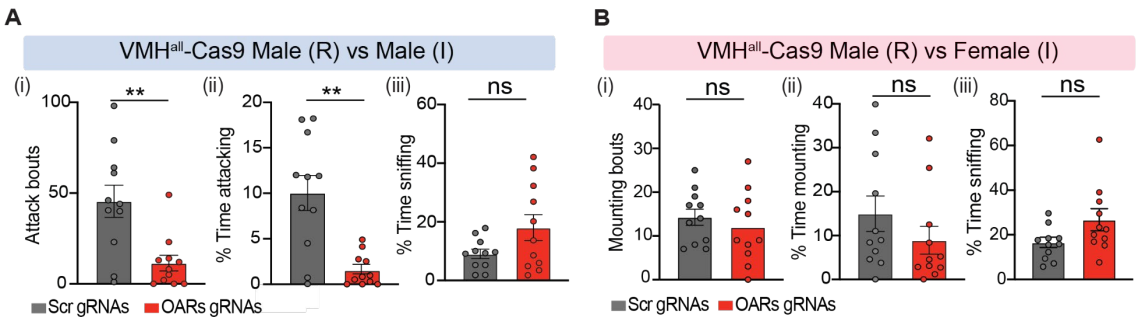


Figure 1 | CRISPR/Cas9 based co-perturbation of Oxtr and Avpr1a reduces territorial aggression in males. a) Quantification of the male directed behaviors in experimental and control males. b) Quantification of female directed behaviors in experimental and control males. n= 11 mice per group. Statistics: Mann-Whitney test was performed **p≤0.01.

Single cell "CRISPRoscopy" imaging of VMHvl^{Esr1} neurons with co-disruption of *Oxtr/Avpr1a*

To investigate how co-editing of *Oxtr/Avpr1a* affects activity in individual VMHvl^{Esr1} neurons, we imaged Ca^{2+} activity using a miniature head-mounted microscope[50] in Esr1-2A-CRE males co-injected with the experimental or control virus pairs. We call this approach "CRISPRoscopy." To avoid reducing aggressive behavior, we performed calcium imaging and *Oxtr/Avpr1a* co-disruption unilaterally (Figure 2A). Because there are virtually no commissural connections between VMHvl, unilateral loss-of-function manipulations are typically compensated by the unmanipulated side[82,84]. Indeed, unilateral injected experimental animals displayed no deficits in sniffing or aggression compared to controls (Supplementary Figure 2A). This design allowed us to determine the effects of *Oxtr/Avpr1a* co-disruption on neural activity and dynamics in behaviorally normal animals.

Effect of *Oxtr/Avpr1a* co-editing on VMHvl^{Esr1} activity and intruder sex representations

Our previous single unit calcium imaging studies have shown that socially experienced males contain distinct VMHvl^{Esr1} subpopulations that are activated by males vs. female intruders, respectively[66,82,84]. This tuning separation was also clear in raster plots of VMHvl^{Esr1} units imaged in control vs experimental (*Oxtr/Avpr1a* co-edited) males (Figure 2B). To quantify the proportion of intruder sex-tuned neurons, we measured unit activity during the first two minutes after the introduction of a male or female intruder in two ways: either by z-scoring (relative to the cell's mean fluorescence over the entire recording period), or by the change in fluorescence relative to the mean pre-intruder baseline[82,84] (in units of σ; see Methods). During interactions with a male intruder, the experimental cumulative distribution function (ECDF) and mean activity of all units (pooled from n=4 control and n=7 experimental animals) were slightly but significantly decreased in experimental mice (Figure 2Ei, Fi;

Supplementary Figure 2Di). However, the mean activity among cells considered as "active" (≥ 2σ above baseline[66]) did not differ between control and experimental mice (Supplemental Figure S2E). During interactions with females there was no significant difference in activity (measured in σ above baseline) between experimental and control animals (Figure 2Eii, Fii), although z-scored activity showed a slight but significant increase (Supplementary Figure 2Dii).

Next, we measured the percentage of male-selective (activity ≥ 2σ during male but not female interactions) and mixed selectivity (activity ≥ 2σ during both male and female interactions) neurons within the $Esr1^{+}$ population during male-male interactions. The percentage of all male-activated neurons (selective or mixed) was slightly smaller in experimental than control mice (34.6% vs. 38.3%, respectively; Figure 2G and Supplemental Figure S2F), while there was a ~56% reduction in the small fraction of male-selective neurons (6.3 ± 2% vs. 14.4 ± 4%, respectively; Figure 2G and Supplemental Figure S2F). Conversely, during female interactions the fraction of female-selective and mixed selectivity neurons was increased by ~4% and ~44%, respectively, in experimental mice (Figure 2G and Supplemental Figure S2F). Together these data are suggestive of a shift in sex-specific tuning from male-selective to mixed selectivity, a conclusion consistent with choice probability analysis (see below).

To determine whether this shift affected the ability to accurately decode intruder sex at the population level, we performed dimensionality reduction using partial least-squares (PLS) regression. This analysis revealed a clear separation of responses during encounters with males vs. females in both control and experimental animals (Figure 2C; Supplemental Figure S2B). In addition, linear SVM decoders trained on imaging data from either control or experimental mice correctly predicted intruder sex with virtually 100% accuracy (Figure 2D). Thus in socially experienced animals, co-targeting of *Oxtr*/*Avpr1a* in $VMHvl^{Esr1}$ cells does not disrupt the population coding of intruder sex[66], despite the altered sex-selectivity of some of these units.

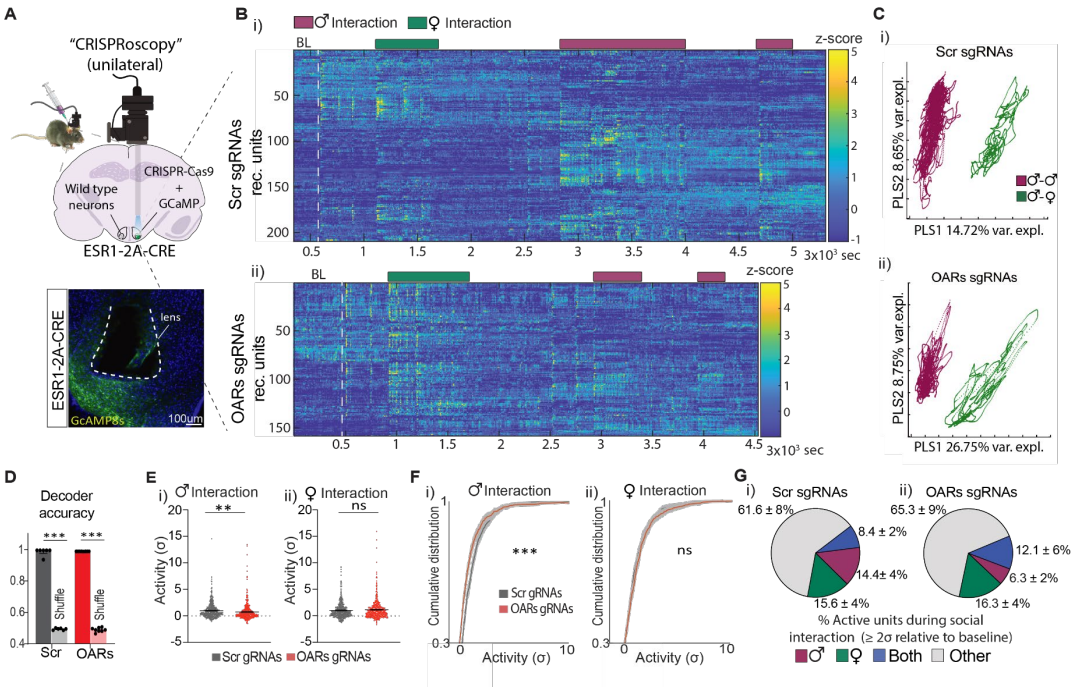


Figure 2 | Single cell "CRISPRscope" imaging of VMHvl^{Esr1} neurons with co-disruption of *Oxtr/Avpr1a.* A) A graphic illustration of "CRISPRoscopy." Male ESR1-2A-CRE residents unilaterally co-injected with a Cre-dependent Cas9 AAV and Cre-dependent OARs-GCaMP8s AAV or Cre-dependent Scr-GCaMP8s AAV in the VMHvl (top). Coronal histology section showing the expression of GCaMP8s (green) in VMHvl (bottom). Section counterstained with Dapi (blue). B) Example of micro-endoscope single unit z-scored responses towards female and male intruders from a control (i) and an experimental (ii) male resident during a recording session. C) Example of VMHvl^{Esr1} ensemble representations of intruder sex, for a control (i) and experimental (ii) male, projected onto the first two axes of a PLS regression against intruder sex. Traces are colored by intruder sex identity. The percentage of variance explained by the first two PLS components is noted for each mouse. D) Accuracy of frame-wise decoders predicting the sex of the intruder trained on VMHvl^{Esr1}neural activity in control and experimental animals. E) Average single VMHvl^{Esr1}unit (σ) responses and F) cumulative distribution of VMHvl^{Esr1} activity (σ) relative to pre-intruder baseline, towards male (i) or female (ii) intruders in control and experimental mice during 1 minute of interaction. G) Percentage of male- or female selective or co-active VMHvl^{Esr1} units (≥ 2σ above the pre-intruder baseline) in control and experimental mice. n=5 control, n=7 Oxtr/Avp1ra targeted animals. Nested Mann-Whitney test was performed except (E), where nested Kolmogorov–Smirnov test was used in (F). **p≤0.01 ***p≤0.001 ****p≤0.0001

Effect of *Oxtr/Avpr1a* co-editing on behavior representation in VMHvl during social encounters

Next, we examined the effect of co-disruption of *Oxtr/Avpr1a* on VMHvl^{Esr1} neuronal activity during the different behavioral phases of social interactions with males or females: appetitive (sniffing) or consummatory (attack or mounting, respectively). The ECDF and average single unit activity during male-directed sniffing or attack was slightly but significantly lower in experimental than in control mice (Figure 3Ai, ii, Ci, ii and Supplementary Figure S3Ai, Bi-ii). In contrast, during social interactions with females, activity during sniffing and mounting was similar (Figure 3Aiii, iv, Ciii, iv and Supplementary Figure S3Aii, Biii, iv). As an additional approach, we quantified the average activity in peri-event time histograms (PETHs) for each type of behavior (see Methods). The mean activity during male-directed sniffing or attack was slightly but significantly lower in experimental than in control mice (Supplemental Figure S3Ci, ii), while it was significantly higher during female-directed sniffing, but unchanged during mounting (Supplemental Figure S3Ciii, iv). In summary, the activity of VMHvl^{Esr1} units during different social behaviors was either unchanged or only modestly different between experimental and control animals, with statistically significant decreases or increases during male- vs. female-directed behaviors, respectively.

We next examined the proportion of behavior-selective active units (defined as units with activity $\geq 2\sigma$ above pre-intruder baseline during, e.g., sniff but not attack or vice-versa)[66,82,84]. As we previously showed, a relatively small fraction of VMHvl^{Esr1} neurons was selective for sniff or attack (~2.5-10%), with the majority showing mixed behavioral selectivity (Figure 3B)[21,66,82,84]. In experimental mice, during male interactions the fraction of sniff-selective units was reduced by ~40% relative to controls (3.9 ± 2% in OARs gRNAs mice vs. 8.9 ± 3% in Scr gRNAs mice) while the small proportion of attack-selective units was reduced by ~70% (1.8 ± 0.7% vs. 6 ± 0.9%; Figure 3B, Di). The fraction of neurons exhibiting mixed behavioral selectivity (i.e., active during both behaviors) was moderately reduced (~22%;

Figure 3Di). Overall, there was a ~38% reduction in the fraction of active units during all male-directed behaviors (from 36.2% in control to 22.4% in experimental mice). In contrast, the fraction of female-directed sniffing- or mount-selective units was similar between control and experimental males (Figure 3 Dii).

As an alternative metric of a cell's behavioral tuning, we calculated its choice probability (CP): a cell was considered "tuned" to one of two pairwise-compared behaviors if it exhibited a CP > 0.7 that was significantly different ($p < 0.5$) from shuffled data[66]. We observed an ~80% and ~70% reduction in the percentage of attack- and sniff-tuned units, respectively, in experimental vs. control mice during male-male interactions (Figure 3Ei), with a concomitant ~18% increase in the fraction of units exhibiting mixed selectivity (CP for sniff vs. attack ≤ 0.7; Figure 3Ei, gray bars and Supplementary Figure S3D). In contrast, the percentage of units tuned to sniffing females (vs. sniffing males) was increased in experimental mice (28.8 ± 3% vs. 41.1 ± 3%; Figure 3Eiii). We also observed a slight increase in the fraction of sniff female (vs. mount female) -tuned units and mount female (vs. attack male) -tuned units in experimental vs. control animals (Figure 3Eii). This analysis suggests that perturbation of normal OXTR/AVPR1a-mediated signaling decreases the relative number of units tuned to specific male-directed behaviors and increases the fraction of female behavior-tuned and mixed-selectivity cells.

Despite these shifts in behavior-selective tuning, there was no significant difference between control and experimental mice in the performance of linear decoders trained to distinguish attack from sniffing based on $VMHvl^{Esr1}$ activity (Supplemental Figure S3E). Thus, the population coding of social behavior by $VMHvl^{Esr1}$ neurons[66] is unaffected by the co-disruption of *Oxtr/Avpr1a.*

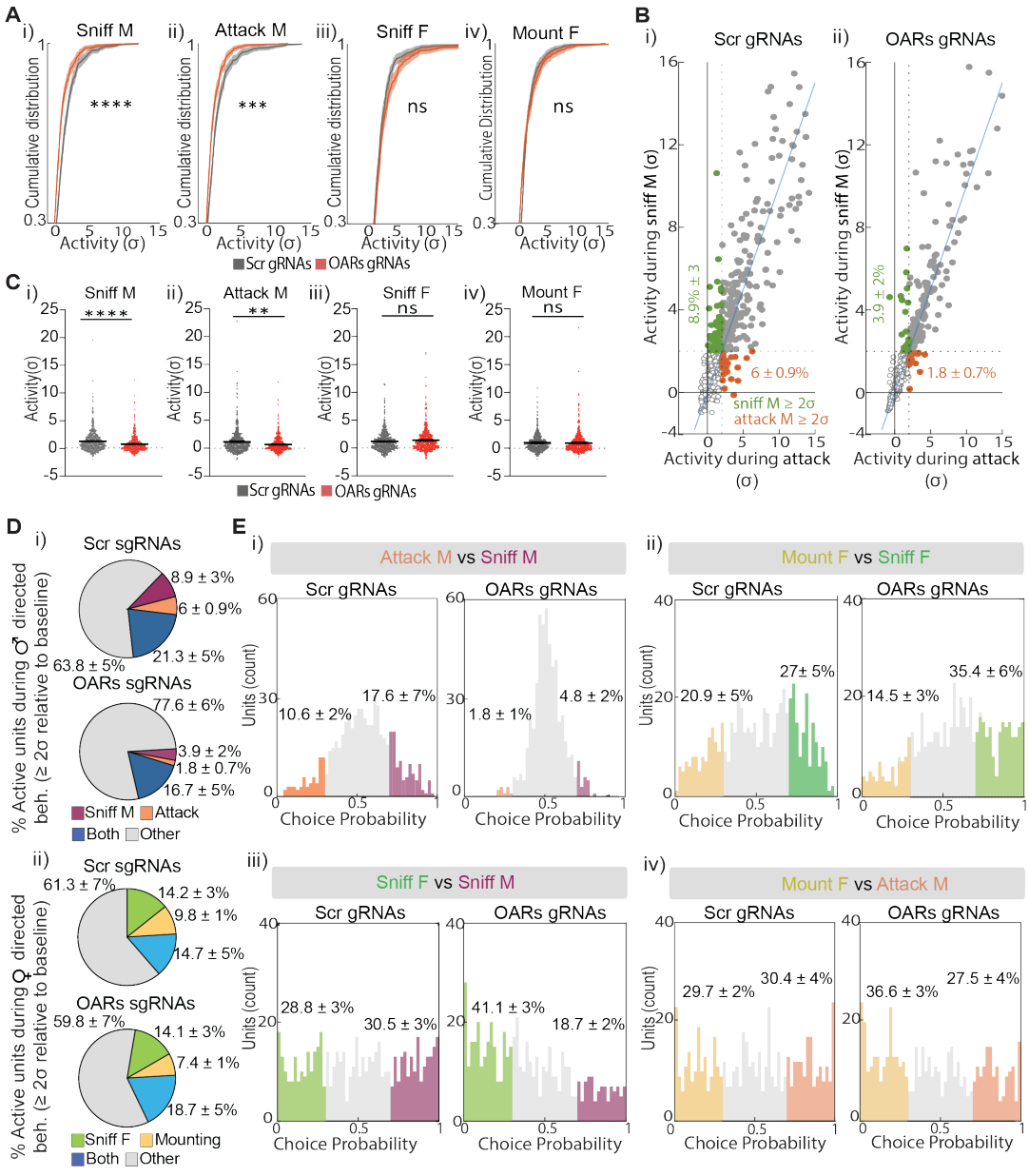


Figure 3 | Social behavior-selective activity and tuning of VMHvl^{Esr1} neurons with co-disruption of *Oxtr*/*Avpr1a*. A) Cumulative distribution of VMHvl^{Esr1}activity (σ) during male- (sniffing and attack) or female (sniffing and mounting) -directed behaviors in control and experimental mice relative to the pre- intruder baseline. B) Scatter plots of single VMHvl^{Esr1} unit activity (σ) during male-directed sniffing or attack in control and experimental mice. Green data points depict units with ≥ 2σ activity during male-directed sniffing and < 2σ activity during attack, relative to the pre intruder baseline

activity. Red data points depict units with ≥ 2 activity during attack and < 2σ activity male-directed sniffing, relative to the pre intruder baseline activity. C) Average activity (σ) of single VMHvl^{Esr1} units during male (sniffing and attack) or female (sniffing and mounting)-directed behaviors in control and experimental mice relative to the pre- intruder baseline. D) Percentage of VMHvl^{Esr1} units active (defined as ≥ 2σ relative to pre intruder baseline) during male- (i) or female-directed behaviors (ii) in control and experimental mice. E) Choice probability histograms of VMHvl^{Esr1} tuning during male- and female-directed behaviors in control and experimental mice. Statistics: Values plotted as means in (A) (± S.E.M), (D) and (E). Nested Kolmogorov–Smirnov test was used in (A), whereas the Nested Mann-Whitney test was performed in (C). **p≤0.01 ***p≤0.001 ****p≤0.0001

VMHvl^{Esr1} line attractor dynamics require Oxtr/Avpr1a-mediated signaling

In addition to the level of activity and degree of feature- (behavior or sex) specific tuning, neural dynamics can play an important role in the neural coding of cognitive function or internal state[18]. Using unsupervised linear dynamical systems modeling[51,52], we recently discovered an approximate line attractor in VMHvl neural state space that encodes a low-dimensional, scalable representation of aggressiveness[21]. This line attractor is implemented by a subset of male-tuned VMHvl^{Esr1} neurons (~20-25%) that are male-tuned, and whose collective activity ramps up as social interactions escalate to attack and thereafter decays with a long (~100s) time constant[21], reflecting persistent activity in this subset. Since neuromodulatory signaling has been implicated in some forms of persistent neural activity[28,85-87], we investigated whether line attractor dynamics during social behaviors are altered when OXTR/AVPR1a-mediated signaling is perturbed.

We fit recurrent switching linear dynamic system (rSLDS) models[52] individually to data from each experimental and control mouse, using data from animals with at least 31 imaged units. In both the control and experimental groups, the fit models reduced the dimensionality of the data to 5 latent factors and three states (S1-S3) capturing ~85% of the observed variance in neural activity. Model performance (cvR$^{2)}$ was similar between

control and experimental mice (Supplementary Figure S4D). During male-male interactions, attack behavior occurred during a single rSLDS state in both control and experimental males. In control mice, the time constant (τ) of the 1st rSLDS dimension (derived from the first eigenvalue of the fit dynamics matrix; see Methods) was significantly higher than that of the 2nd dimension (~100-120s vs. ~40s; Figure 4A and Supplementary Figure S4E). This yielded a line attractor score (calculated as the $\log_2$ of the ratio of the τ 's of the 1st and 2nd dimensions) of ~1.6 (Figure 4F, gray bar), similar to that observed in mice without CRISPR/Cas9 gene editing[21]. In contrast, in experimental mice the first two rSLDS dimensions had statistically indistinguishable τ values (<50s; Figure 4C) due to a reduced 1st dimension τ (Supplementary Figure S4E), and consequently a line attractor score close to zero (Figure 4F, red bar). However, during mounting behavior towards female intruders, co-perturbation of Oxtr/Avpr1a-mediated signaling did not significantly reduce the 1st dimension τ values or the line attractor score in experimental mice compared to controls (Supplementary Figure S4F), consistent with the fact that most mating attractor-weighted neurons are female-tuned[21].

To visualize neural dynamics in state space during male-male interactions, we generated 2D flow field graphs spanned by the first two PCs of the rSLDS models. The flow fields are comprised of arrows that indicate the rate and direction of change in neural population activity at different points in state space during social interactions (Supplementary Figure S4C). The 2D flow field of control mice (Scr gRNAs) revealed a roughly linear region of low vector flow constituting the line attractor, along which the neural population vector progressed during an inter-male interaction (Figure 4Gi and Supplemental Figure S4Ci, dashed black lines). In a 3D dynamic landscape, where the length of the flow-field vectors at each position in neural state space is converted into the height of the landscape (and represented as a heat scale), in control animals the population activity vector (PAV) progressed slowly along a trough-like structure (the line attractor) as aggression escalated (Figure 4Hi). In contrast, in experimental mice (OARs gRNAs) this line attractor was absent and was replaced by a point attractor (appearing as a circle in 2D and a cone in 3D), from which the population activity vector made transient excursions during bouts of

sniffing or attack (Figure 4Gii, Hii and Supplemental Figure S6Cii). This point attractor is called a "trivial" fixed point, corresponding to the resting state of a system in which population activity decays to a global minimum in the absence of inputs. This minimum (bottom of the cone in Figure 4Hii) represents the location of the PAV during baseline behavior in solitary mice. These data indicate that Oxtr/Avpr1a-mediated signaling is required for the emergence of line attractor dynamics in VMHvl during male-directed aggression but not during female-directed mounting[21].

To investigate in more detail how co-disruption of *Oxtr/Avpr1a* perturbs VMHvl^{Esr1} attractor dynamics, we projected the weighted average of neuronal activity in the 1st rSLDS dimension (which generates the line attractor) onto the time axis and overlayed behavior annotations. As reported previously[21], in control mice this activity ramped up during the progression from male-directed sniffing to attack, eventually reaching a plateau where it decayed slowly between attack bouts towards a single intruder and remained elevated between sequential trials with different intruders (Figure 4B, Ei and Supplementary Figure S4A, "post intruder"). In contrast, in experimental mice 1st dimension neural activity decayed rapidly between attack bouts, displaying a "sawtooth" profile, and was significantly lower during the post-intruder (i.e., inter-trial) interval (Figure 4D, Eii and Supplementary Figure S4B). The observation that 1st dimension activity in experimental mice is transiently elevated during aggressive episodes but does not remain stable across attack bouts and trials, indicates that these neurons are still activated during attack, but do not integrate recent activity in the same way as normal mice.

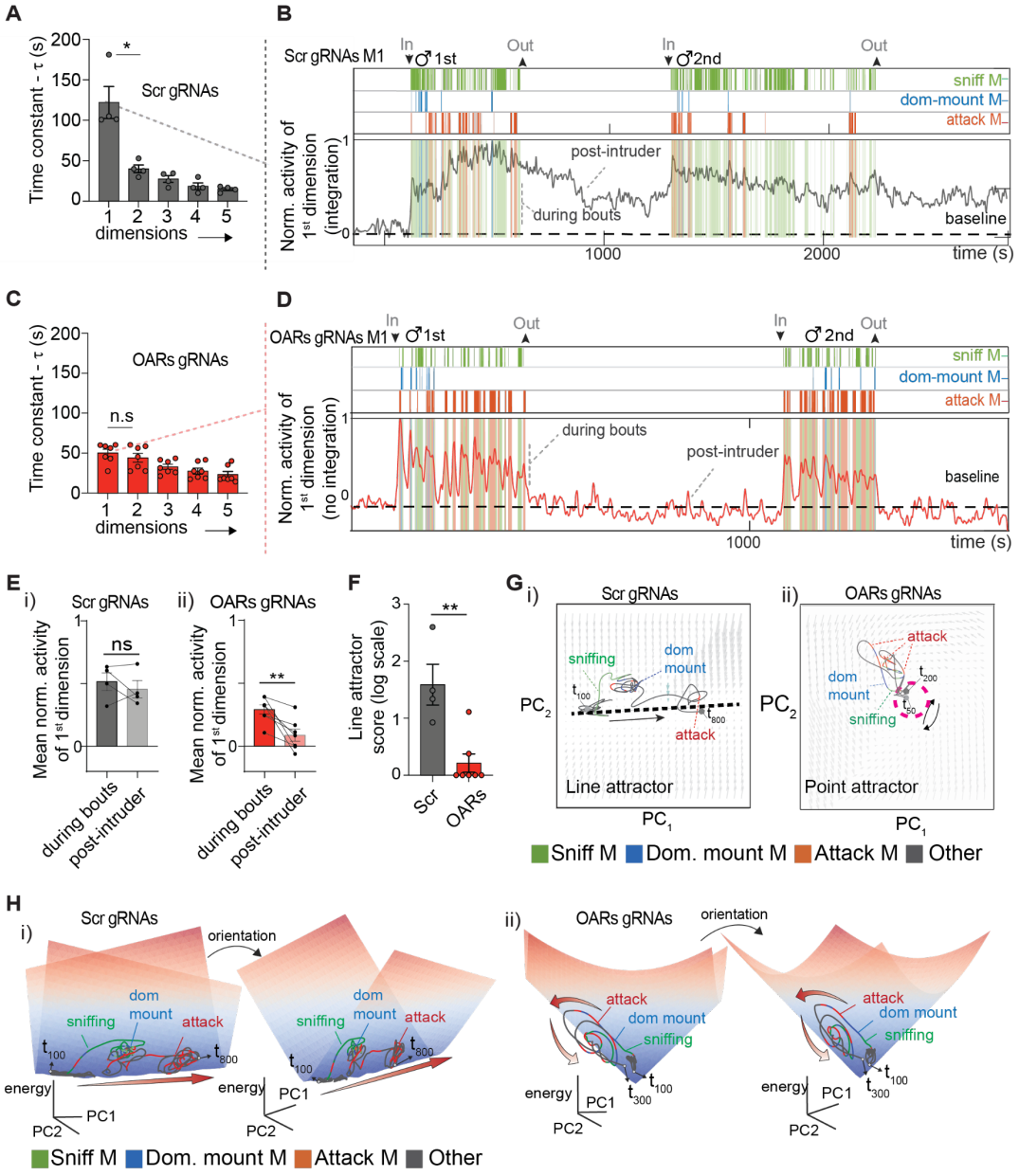


Figure 4 | VMHvlEsr1 line attractor dynamics require Oxtr/Avpr1a-mediated signaling

A) Average time constant (τ) of all dimensions, arranged in decreasing order in control mice. B) Normalized activity projection onto the time axis of the longest time constant (integration dimension) for VMHvlEsr1 units in control mouse M1. C) Average time constant of all dimensions, arranged in decreasing order for VMHvlEsr1 units in experimental mice. D) Normalized activity projection onto the time axis of the 1st dimension (no integration) for VMHvlEsr1 units in experimental mouse M1.

E) Mean normalized VMHvlEsr1 activity of the 1st dimension during all behavioral bouts and after removing the intruder male (post-intruder) in control (i) and experimental (ii) mice. F) Line attractor score for VMHvlEsr population activity in control and experimental mice. n = 4 control and n =7 in experimental group. G) Neural state space with the trajectories projected over time within the inferred flow field rSLDS states for VMHvlEsr1 in control (i) and experimental (ii) mouse. H) Inferred 3D dynamic landscape in VMHvl control mouse M1(i) and experimental mouse M1(ii). Different views of line (i) and point attractors (ii) are shown. Red arrows depict the neural trajectory associated with attack. All sniffing, dominant mounting and attack bouts following the introduction of an intruder male are depicted in the behavioral raster plots in (B) and (D). Statistics: Kruskal-Wallis test was performed in (A) and (C). Paired t-test was performed inin (E) and Mann-Whitney test in (F) *p≤0.05 **p≤0.01.

Oxtr/Avp1ra-mediating signaling controls VMHvl ^{Esr1} persistent neural activity

We next investigated the dynamics of individual VMHvl^{Esr1} neurons caused by the co-disruption of *Oxtr/Avpr1a*. We focused initially on cells that were strongly weighted by the 1st rSLDS dimension (stem plots in Figure 5Aii). In raster plots from control mice, these units exhibited activity that persisted across inter-attack bout intervals and decayed slowly after removing the intruder male, visible as a "smearing" of rasters over time (Figure 5Ai, Scr gRNAs). In contrast, analogous units from experimental mice exhibited activity time-locked to attack bouts, visible as a vertical stripe-like pattern (Figure 5Ai, OAR gRNAs). To quantify these dynamics, we computed the average autocorrelation half-width (ACHW)[66], an approximate measure of the decay constant[88,89], for each 1st dimension-weighted unit. The mean ACHW of these neurons across the entire social interaction was significantly shorter in experimental (9.86 ± 1 s) than in control mice (28.6 ± 1.23 s), by ~20 seconds (Figure 5Bi and Supplementary Figure S5A).

A reduction in the average ACHW was also observed among 2nd rSLDS dimension cells, as well as in the VMHvl^{Esr1} population as a whole (Figure 5Bii, C). However, the difference in mean ACHW between experimental and control animals was larger for 1st dimension-

weighted units (~66%) than for 2nd dimension and total Esr1+ neurons (~61% and 50.9%, respectively). In contrast, the average and cumulative distribution of ACHWs among all VMHvl^{Esr1} neurons were only slightly reduced (by ~22%) during male-female interactions (Supplementary Figure S5B). Thus, the decay time of individual VMHvl^{Esr1} units is faster, on average, in experimental than in control mice during male-male social interactions, especially among neurons that contribute to the 1st rSLDS dimension. This reduction in ACHW is consistent with the loss of line attractor dynamics caused by co-disruption of *Oxtr/Avpr1a* and may be a cause or a consequence of this loss.

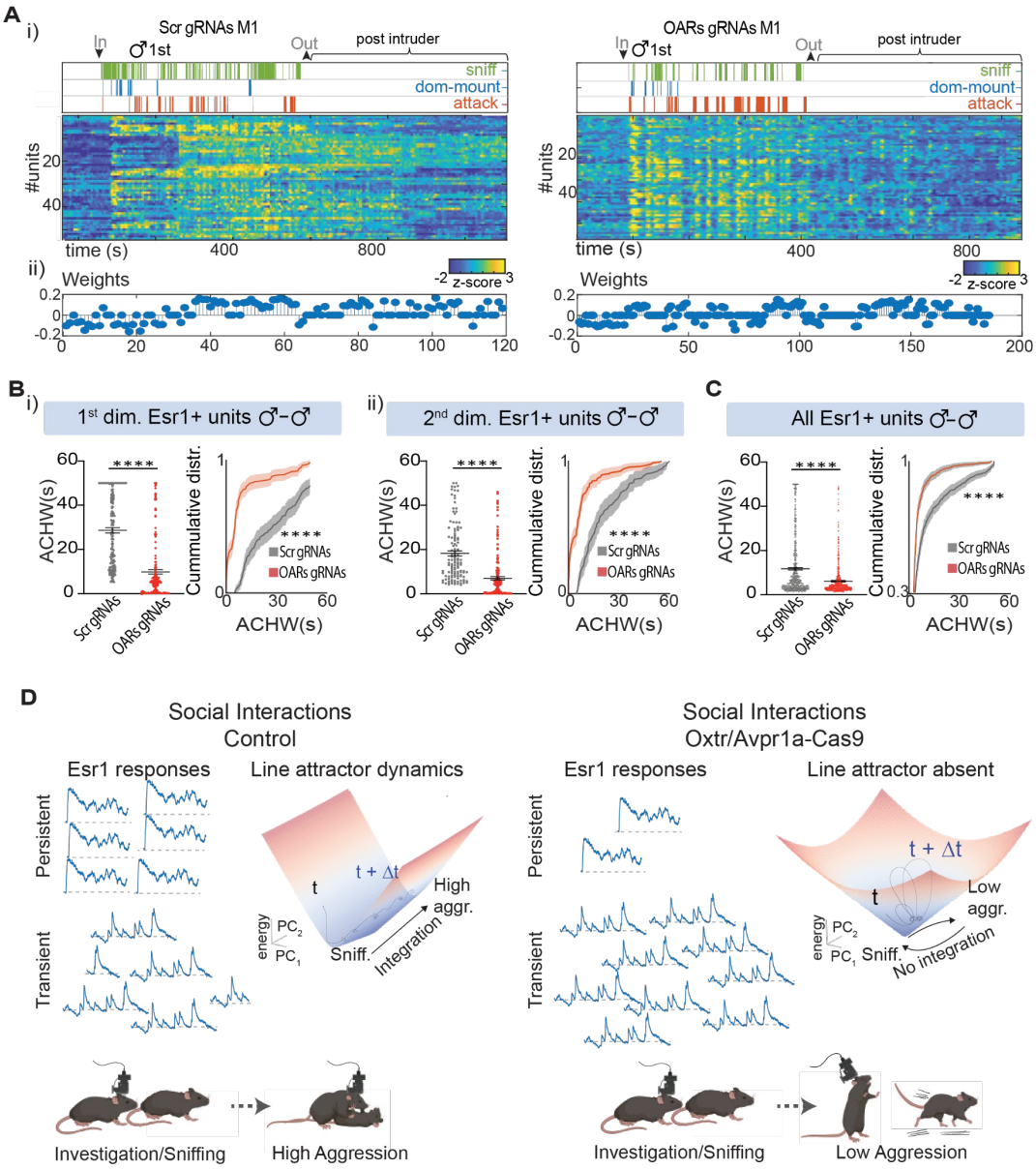


Figure 5 | Oxtr/Avp1ra-mediating signaling controls VMHvl ^{Esr1} persistent neural activity. A) Behavioral raster plot and the corresponding neural activity (z-score) of individual VMHvl^{Esr1} units that contribute to the 1st dimension in control (left) and experimental (right) mouse (i). The absolute rSLDS weight of neurons contributing to the 1st dimension is shown as a stem plot (ii). B) Average single unit and cumulative distribution of neuronal persistence measured by ACHW of individual VMHvl^{Esr1} units that contribute to the 1st dimension (i) and the 2st dimension (ii) during the entire duration of male-male interactions. n=4 control, n=7 experimental animals. C) Average single unit

and cumulative distribution of ACHW of all units in control and experimental mice during the first 1-2 minutes of male-male interactions. n=5 control, n=7 experimental animals. D) A graphic illustration depicting our working model. Oxtr and Avpr1a-mediated signaling control aggression escalation by regulating VMHvl^{Esr1} single cell dynamics (transient and persistent activity) and line attractor dynamics (left) during male-male interactions. However, co-perturbation of *Oxtr* and *Avpr1a* results in a strong reduction of the slow neural dynamics, an increase of transient responses of VMHvl^{Esr1} neurons, and the absence of a line attractor (replaced by a trivial point attractor). Under such conditions, male animals fail to display a high level of territorial aggression against male conspecifics (right). Statistics: Values plotted as mean ± S.E.M. Nested Mann-Whitney test was performed for (Bi, Ci, Aii inset). Kolmogorov test was performed (Aii, Bii, Cii) *p≤0.05 **p≤0.01 ****p≤0.0001

OXT and AVP evoke persistent responses in VMHvl^{Esr1} neurons *ex vivo*

The observation that co-disruption of *Oxtr/Avpr1a* shortened the average decay time of Esr1^{+} units raised the question of whether adding these peptides to VMHvl would, conversely lengthen this decay. Because it is not technically feasible to apply drugs or peptides directly at the site of miniscope imaging, we utilized the *ex vivo* VMH brain slice preparation (Figure 1F). We imaged calcium activity in slices perfused with a cocktail of OXT & AVP, and fit an rSLDS model to the data (Supplementary Figure S5C). This fit model captured 85% of the observed variance in neural activity. The time constant of dimension x1 (~90s) was ~8-9 -fold greater than that of x2 (~15s; Supplemental Figure S5D), similar to that observed *in vivo* (Figure 5A, C). Stem plots revealed that the neurons highly weighted by dimension x1 were distinct from those weighted by x2 (Supplementary Figure S5E). Plotting the time-varying weighted average activity of x1 and x2 neurons revealed that the former exhibited slowly decaying responses to OXT+AVP application, while the latter exhibited more transient responses (Supplementary Figure S5F-J). The two populations could also be identified independently of rSLDS analysis by quantifying the decay time of OXT+AVP mediated VMHvl^{Esr1} population responses (Supplementary Figure S7K). Interestingly, the inclusion of synaptic transmission blockers (20 μM CNQX and 10 μM MK-801) only

slightly increased the decay rate during later phases of the peptide response (Supplementary Figure S5L). Thus, persistent activity can be evoked by OXT+AVP in Esr1^{+} neurons within VMHvl slices *ex vivo*, suggesting that it does not require long-range interconnections with anatomically distant structures.

Discussion

Using a novel approach that integrates CRISPR/Cas9 gene editing[48,49] with miniscope imaging[50] and dynamical systems analysis[51,52], we show that OXT and/or AVP receptors are required for persistent neural activity, line attractor dynamics and aggression in VMHvl^{Esr1} neurons[64,84,90] (Figure 5D). These results in turn suggest that neuropeptides may control certain behaviors, at least in part, through an influence on population neural dynamics. Our approach should help unify molecular and circuit-level approaches with "manifold"-level approaches[13,14,23] to understanding the neural control of behavior, emotion and cognition.

The impact of Oxtr/Avpr1a co-disruption on VMHvl^{Esr1} activity, tuning and dynamics

In principle, the observed reduction in aggression could be a consequence of reduced activity of VMHvl^{Esr1} neurons. Indeed, fiber photometry and CRISPRoscopy revealed a statistically significant decrease in average activity during sniffing or attacking males. The percentage of attack- and sniff-tuned cells was also reduced. These relatively modest effects, however, seem unlikely to account for the strong reduction in aggressiveness caused by bilateral disruption of *Oxtr* and *Avpr1a*. Nevertheless, as aggression requires a high level of VMHvl^{Esr1} activity[64,82], we cannot exclude that these effects contribute to the behavioral phenotype. In contrast, co-disruption of *Oxtr/Avpr1a* caused a virtually complete elimination of the line attractor. Consistent with this finding, the average ACHW of VMHvl^{Esr1} neurons weighted on the integration (1st) dimension was strongly reduced. More specific

perturbations will be required to test decisively whether selective elimination of line attractor dynamics affects aggression.

How co-disruption of *Oxtr/Avpr1a* eliminates the line attractor is not clear. Our between-subject comparisons do not allow us to distinguish whether the reduced ACHW reflects a change in the dynamics of individual cells, or inactivation of a subset of cells with slow dynamics. However, persistent calcium responses evoked by bath application of OXT+AVP to VMHvl^{Esr1} neurons *ex vivo* were strongly reduced by co-disruption of *Oxtr/Avpr1a*. Persistent activity induced by OXT *ex vivo* has been demonstrated in hippocampal neurons [87]

Notably, only a slight reduction in persistence was observed for VMHvl^{Esr1} units that were active during male-female interactions. These female-tuned units are largely distinct from the units active during male-male interactions[66,82] . Consistent with this, a line attractor observed during male mounting of females[21] was not perturbed by co-editing of *Oxtr* and *Avpr1a* and there was no deficit in male-female mating behavior. Together, these data argue that the effect of the *Oxtr/Avpr1a* perturbation on neural dynamics is unlikely due to indiscriminate changes in VMH cytoarchitecture or function.

A line attractor dependent on neuropeptide signaling

Most theoretical studies of attractor dynamics to date have assumed that they reflect recurrent fast synaptic connectivity[17-19], as seen in the *Drosophila* ring attractor system[24,29]. Our finding that neuropeptide signaling is required for line attractor dynamics and persistent activity is consistent with recurrent neural network (RNN) modeling of persistent activity in VMH circuits, which indicated that only models incorporating both recurrence and slow neuromodulation were accounted for the experimental observervations[12,65]. OXT can cause increased excitability in anterior VMHvl neurons[91]; it could also strengthen recurrent connectivity within this nucleus[12,92] or between interconnected regions[93]. We note, however,

that neither OXT nor AVP are synthesized by VMHvl^{Esr1} neurons; therefore, their source(s) must be extrinsic to this nucleus[91].

A role for neuropeptides in implementing slow attractor dynamics that control innate behaviors is appealing for several reasons. First, it it may overcome the dependence of purely glutamatergic attractor networks on fine-tuned synaptic connectivity[17,18,27]. This requirement for fine-tuning makes attractors very "fragile", i.e., highly sensitive to experimental or physiological disruption (but see[94,95]). RNN modeling has indicated that VMHvl networks incorporating slow neuromodulatory transmission can reproduce observed network time-constants over a much wider range of synaptic connectivity densities than purely glutamatergic networks[65], suggesting that neuropeptidergic based line attractors may be less dependent on precise patterns of synaptic connectivity. Second, neuropeptidergic signaling can yield decay constants on the timescale of 100s of seconds. Although sufficiently fine-tuned glutamatergic attractor networks can in theory persist indefinitely[18], persistent activity in other experimentally described attractors has typically been observed for just a few seconds[19,96] , making it unclear if they can sustain activity on longer timescales. Finally, slow dynamics may be better suited to encode long-lasting and escalating affective states, such as aggressiveness, than the attractors invoked to compute functions like gaze stabilization[17], working memory[107] and head direction[24].

Neuropeptides may be also be advantageous for implementing line attractors or leaky integrators because their expression can be modulated by hormones[97] and neural activity[19]. For example, longitudinal single-cell calcium imaging studies from a female specific VMHvl subpopulation that controls sexual receptivity[20,98,99] has revealed an estrus cycle-dependent line attractor[20]. Finally, neuropeptide receptor expression is more restricted than that of receptors for biogenic amines[100]. This specificity could allow the regulation of different (and potentially competing) attractors within a local network[21] by distinct neuropeptides.

The results described demonstrate how the power of multiplex gene editing technology[48,49] can combined with single cell imaging to identify mechanisms that implement attractor

dynamics[18]. They also suggest that population neural dynamics may mediate the behavioral functions of some neuropeptides[38,40]. In this way, this approach may enable mechanistic explanations in neuroscience that unify different levels of abstraction and of biological organization[23].

Limitations of the study

A technical limitation of our approach is that two different viruses were used to deliver gRNAs-GCaMP and Cas9. Consequently, not all GCaMP$^+$ cells captured in our imaging analysis are necessarily Cas9$^+$ (and therefore mutant for *Oxtr/Avpr1*). Our data indicate co-infection rates of ~65-80% at the injection site depending on the viruses used. Another limitation is our reliance on between-subject comparisons of experimental vs. control mice. Our rSLDS analysis showed that there was no effect of *Oxtr/Avpr1a* co-editing on male mounting of females but do not exclude a possible effect on intromission. Finally, our results do not distinguish the individual roles of OXTR and AVPR1a, nor do they establish the cellular source and release dynamics of the endogenous peptide(s) that activates these receptors *in vivo* during aggression. Further studies will be required to elucidate these biologically important details.

Experimental model and subject details

All procedures were performed in accordance with NIH guidelines and approved by the Institutional Animal Care and Use Committee (IACUC) at the California Institute of Technology (Caltech). We used *Esr1*$^{Cre/+}$ [64] transgenic mice. Animals were housed and maintained on a reverse 12 h light-dark cycle with food and water ad libitum. We used wild-type (WT) C57BL/6N male mice (experimental), C57BL/6N female mice or BALB/c females (for sexual experience), and BALB/c male mice (intruders) were obtained from Charles River (Burlington, MA). Behavior was tested during the dark cycle.

Viruses

The following AAVs were used in this study, with injection titers as indicated. Viruses with a high original titer were diluted with clean PBS on the day of use. AAV1-Syn-Flex-GCaMP7f (2.1 x e^{13}) was purchased from Addgene. AAVDJ/8-EFS-NC-SpCas9-HA -NLS-Poly(A) (pBK694) (2.00x e^{13}) and AAV9-EFS-NC-SpCas9-HA -NLS-Poly(A) (2.17x e^{13}) purchased from Duke viral vector core. The AAV9-EFS-NC-DIO-SpCas9-myc-NLS-Poly(A) (2.31x+e^{13}), AAV9-hPGK-DIO-SpCas9-myc-NLS-Poly(A) (2.31x+e^{13}), AAV1-gRNAs$_{scramble}$-hSyn flex-GCaMP8s-wpre (2.75x e^{12}) and AAV1-gRNAs$_{Oxtr/1Avpr1a}$ -hSyn flex- GCaMP8s-wpre (2.44x e^{12}) were packaged at the HHMI Janelia Research Campus virus and the Duke Viral core facilities. Viruses with hPGK or an Efs promoter were used interchangeably to drive the expression of Cas9. In the illustration in Figure 2Aii and 3Aii only the Efs promoter is depicted for simplicity reasons. The RNAs$_{scramble}$-ubc DsRed (TU/ML 1.5X e^{8}) and RNAs$_{Oxtr/Avpr1a}$ ubc DsRed (OXTR/AVPR1a) lentiviruses (TU/ML 3.50 X e^{8}) were also packaged at the HHMI Janelia Research Campus virus facility.

Method details

Screening for aggressor male and resident intruder assay

All experimental male mice (‘‘residents’’) were individually housed for two weeks and received sexual experience (for at least one week). Previously it has been reported that ~20-25% of inbreeding C57BL/6N male animals fail to display territorial aggression against conspecific male intruders during the RI assay [93]. We pre-screened males for baseline aggression using resident-intruder testing sessions to identify and exclude no aggressors from our analysis. Animals that attacked two constitutively presented intruders were termed aggressors and added to the pool of animals for CRISRP/Cas9-based gene editing surgeries. On the experimental day, the preselected male residents were transported in their home cage to a novel behavioral testing room (under infrared light), where they acclimated for 5-10 min. An unfamiliar group housed BALB/c mouse (‘‘intruder’’) was then placed in the resident's home cage, and residents were allowed to interact with it for period of time.

Acute brain slices preparation

Briefly, male adult mice were anesthetized with isoflurane and transcardially perfused with cold NMDG-ACSF (adjusted to pH 7.3–7.4) containing CaCl2 (0.5 mM), glucose (25 mM), HCl (92 mM), HEPES (20 mM), KCl (2.5 mM), kynurenic acid (1 mM), MgSO4 (10 mM), NaHCO3 (30 mM), NaH2PO4 (1.2 mM), NMDG (92 mM), sodium L-ascorbate (5 mM), sodium pyruvate (3 mM), thiourea (2 mM), bubbled with carbogen gas (95% O_2 and 5% CO_2). The brain was sectioned at 250 μm using a vibratome (VT1000S, Leica Microsystems) on ice and was incubated in 34oc for 12 min, in NMDG – ACSF. Then transfer the sections to room temperature in aCSF/HEPES-GSH solution (adjusted to pH 7.3–7.4,) containing CaCl2(2 mM), glucose (25 mM), kCl (2.5 mM), HEPES (20 mM), NaCl (92 mM), MgSO4 (2 mM), NaHCO3 (30 mM), NaH2PO4 (1.2 mM), sodium L-ascorbate (5 mM), sodium pyruvate (3 mM), thiourea (2 mM), and Glutathione Monoethyl Ester (0.5-1mM)-before proceeding with Ca^{2+} imaging.

Peptide perfusion and two-photon calcium imaging experiments

Solutions of 400nM of Oxt and/or Avp, 20 μM CNQX and 10 peptides μM (Tocris) were prepared in aCSF and perfused with a rate of 1-2ml/min through a microfluidics chamber containing the brain slices. Calcium imaging was performed using a custom-modified Ultima two-photon laser scanning microscope (Bruker). The primary beam path was equipped with galvanometers driving a Chameleon Ultra II Ti:Sapphire laser (Coherent) and used for GCaMP imaging (920 nm). GCaMP emission was detected with photomultiplier-tube (Hamamatsu). Images were acquired with an Olympus20X XLUMPLFLN Objective, 1.00 NA, 2.0 mm WD. All image acquisition was performed using PrairieView Software (Version 5.3) with a framerate of ~1.2Hz.

Behavior recording

All behavioral experiments were performed in conventional mouse housing cages (home cage or new cage) under red lighting, using the previously described behavior recording setup [109]. The behavior video's top and front views were acquired at 30 Hz using the video recording software, StreamPix7 (Norpix).

Behavior annotations

Behavior videos were processed using an automated behavior classification system to generate frame-by-frame annotations of attack, mounting and sniffing behavior [110]. The output of the classifier and behavior videos were loaded into a MATLAB based MATLAB-based behavior annotation interface and then manually corrected by trained individuals to produce a final set of annotations[110]. A 'baseline' period of 5-minutes was recorded at the start of every recording session during which the animal was alone in its home cage. Six behaviors were annotated during the resident intruder assays: sniff (face, body, genital-directed sniffing), towards male or female intruders, and attack-, mount- directed behavior against male or female intruders. Post-surgery, male residents were exposed to a male intruder first, and subsequently to a female intruder. RI assay with a male or a female intruder

was performed on separate days, except during CRISPRoscopy. For quantifying the interval (s) between behavioral bouts in Figures 2, 3, S2 and S3, animals that show ≤1 mount or attack bout were excluded. 15 min for male-male interaction was scored in Figure 2 and S2. In Figure 3 and S3 ~15-20 min of RI was scored for male -male interaction. 10 min of male-female interaction was scored during the RI assays in Figure 2, 3, S2 and S3.

In addition to the classification of behaviors, automated pose estimation was performed on behavior videos to obtain key points of interacting mice[110]. The velocity of the resident mouse was calculated as the change in positions of centroids of the head and hips, computed across two consecutive frames as previously performed[84]. The distribution of this feature was computed for both experimental and control animals to obtain the data shown in Supplementary Figure S2C(iii).

Stereotaxic surgery

Surgeries were performed on socially and sexually experienced adult male *Esr1*$^{Cre/+}$ mice mice 8–12 weeks old. Virus injection and implantation were performed as described previously[66,84]. Briefly, animals were anesthetized with isoflurane (5% for induction and 1.5% for maintenance) and placed on a stereotaxic frame (David Kopf Instruments). The virus was injected into the target area using a pulled-glass capillary (World Precision Instruments) and a pressure injector (Micro4 controller, World Precision Instruments) at a 20 nl /min flow rate. The glass capillary was left in place for 5 -10 minutes following injection before withdrawal. The injection volumes were ~400-500nl for bilateral injection in mice used for behavioral analysis and CRISRPometry. For micro endoscope recordings, we performed unilateral ~200nl injections. The Stereotaxic injection coordinates were based on the Paxinos and Franklin atlas (posterior VMHvl, anterior–posterior: −4.68, medial–lateral: ±0.73, dorsal–ventral: −5.73). For single fiber optogenetic and fiber photometry experiments (optogenetics: diameter 200 μm, N.A., 0.22; fiber photometry: diameter 400 μm, N.A., 0.48; Doric lenses) were then placed above the virus injection sites (fiber photometry: 150 μm above) and fixed on the skull with dental cement (Metabond,

Parkell). For micro-endoscope experiments, virus injection and lens implantation were performed on the same day Lenses with a baseplate were slowly lowered into the brain and fixed to the skull with dental cement. Mice were habituated with weight-matched dummy micro-endoscopes (Inscopix) for at least one week before behavior testing. Mice were head-fixed on a running wheel 3-4 weeks after lens implantation, and a miniaturized micro-endoscope (nVista, Inscopix) was attached to the baseplate for imaging. Mice were singly housed after surgery and were allowed to recover for at least 4 weeks before behavioral testing.

Histology

Once the behavioral experiments were finished, virus expression and implant placement were histologically verified on all mice. Mice lacking correct virus expression or implant placement were excluded from the analysis. Mice were transcardially perfused with 1x PBS at room temperature, followed by 4% paraformaldehyde (PFA) (diluted from 16% EM grade PFA). Brains were extracted and post-fixed in 4% PFA 16-24h at 4°C, followed by 24 hours in 30% sucrose/PBS at 4 °C. Brains were embedded in OCT mounting medium, frozen on dry ice and stored at −80°C for subsequent sectioning. Brains were sectioned into 60 μm slices on a cryostat (Leica Biosystems). Sections were washed with 1× PBS and mounted on Superfrost slides, then incubated for 15 minutes at room temperature in DAPI/PBS (0.5 μg/ml) for counterstaining, rewashed and coverslipped. Sections were imaged with an epifluorescent microscope (Olympus VS120)

For some epitope staining 30 um sections were cut from either fresh-frozen tissue or post-fixed 2h 4%PFA on ice, immersed in 30% sucrose:1xPBS 4C 2h before embedding in OCT. For Cas9 immunostaining, a cocktail of antibodies against the Cas9-fused HA or Myc epitope and the Cas9 protein itself was used. Animals were stained after 9-12 weeks post-injection. Estimates of the co-indefectibility between the Cas9 and the gRNA expressing viruses were made at the center of the injection site

Microendoscope recordings

On the day of imaging, mice were habituated for at least 5-10 minutes after installing the micro endoscope in their home cage before the start of the behavior tests. Imaging data were acquired at 30 Hz with 2× spatial down sampling; light-emitting diode power (0.1–0.5) and gain (1–7×) were adjusted depending on the brightness of GCaMP expression as determined by the image histogram according to the user manual. A transistor–transistor logic (TTL) pulse from the Sync port of the data acquisition box (DAQ, Inscopix) was used for synchronous triggering of StreamPix7 (Norpix) for video recording. Imaging sessions typically lasted 1 h (20–25 min interactions per sex).

Micro-endoscopic data extraction

Preprocessing and Calcium data extraction was performed similarly to what has been previously described[82]. Briefly, data were 2x downsampled, motion corrected, and a spatial band-pass filter was applied to remove the out-of-focus background. Next, filtered imaging data were temporally downsampled to 10 Hz. Calcium traces were extracted and deconvolved using the CNMF-E[111] with the following parameters: patch_dims = [42, 42], gSig = 3, gSiz = 13, ring_radius = 19, min_corr ∼0.57-0.62, min_pnr = ∼5.5-6, deconvolution: foopsi with the ar1 model. Every extracted unit's spatial and temporal components were manually inspected (SNR, PNR, size, motion artifacts, decay kinetics, etc.). Traces of units were either z-scored or normalized in units of σ relative to the baseline fluorescence (during 7sec or more) of the neuron before the first trial of resident-intruder interactions, as previously described[82,84], Distinct hypothalamic control of same-and opposite-sex mounting behavior in mice. In Supplementary Figure 5 C, the z-scored value during a behavioral bout for each unit was normalized by subtracting the mean of a 2-3 sec baseline before the onset of the bout. The average normalized activity was quantified for a period of 15 sec. A total of 585 units (n=5 mice) from control and 546 units (n=7 mice) from experimental mice were recorded.

Quantification and statistical analysis

Miniscope neural data analysis

Choice probability

Choice probability (CP) analysis was used as before[84] to measure a cell's tuning, defined here as how well two conditions could be predictively discriminated from a single cell's activity[112]. The CP of a given cell for a pair of behavioral conditions was computed by constructing a histogram of that cell's $\Delta F(t)/F_0$ values under each of the two conditions. These two histograms were plotted against each other to generate a ROC (receiver-operating characteristic) curve. The integral of the area under this ROC curve generated the CP value for each cell with respect to each of the two behavioral conditions. This CP value is bounded from 0 to 1, where a CP of 0.5 indicates that the neuron's activity cannot distinguish between the two conditions. As in previous studies, the statistical significance of choice probabilities was determined relative to chance. We shuffled behavioral bout timings for each of the two compared conditions and computed the choice probability for this shuffled data. Shuffling was repeated 100 times for each of the two behaviors, from which we calculated the mean and s.d. (σ) of the 'shuffled' choice probabilities.

As significant, we considered any observed choice probabilities $>2\sigma$ above the shuffled mean and imposed an additional choice probability threshold > 0.7 as previously described[84]. The colored bars indicate the neurons that show a strong and statistically significant choice probability, and grey bars indicate cells for which the choice probability was either activated < 2 σ above (not responsive) the shuffled mean or was considered which choice probability not significantly higher than chance or choice probability ≤ 0.7 for that neuron.

Dimensionality reduction for visualizing intruder sex

Low-dimensional representations for visualizing changing ensemble dynamics over time were constructed using partial least squares (PLS) regression (MATLAB). For PLS, all

traces were concatenated and regressed against a $1 \times T$ vector with entries valued at −1 (if a male intruder was present), 1 (if there was a female intruder), or 0 (otherwise).

Decoding intruder sex from neural data

We constructed a frame-wise linear SVM decoders (as described previously[66,84]) to distinguish intruder sex. Training data was constructed from the set of $N \times 1$ (N = neurons) population activity vectors from all frames occurring during social interaction in each mouse. Equal numbers of frames of male and female interaction were used during decoder training to ensure chance decoder performance of 50%. Shuffled decoder data were generated by the training the decoder on the same neural data but with behavior labels randomly assigned to each behavior bout (n=5 control and n=7 *expreimental mice*). This training data, along with intruder sex labels, was then used to train a linear SVM decoder. Accuracy was evaluated using a stratified fivefold cross-validator. Decoding was repeated 100 times, with decoder performance reported as the mean accuracy per imaged animal. For significance testing, the mean accuracy of the decoder trained on shuffled data (repeated 500 times per imaged animal) was computed to compare against the decoder accuracy trained on actual data.

Decoding behavior from neural activity

We constructed frame-wise linear SVM decoders (as described previously[66,84]) to discriminate male directed sniffing and attack from imaged control and experimental VMHvl^{Esr1} units. Briefly manual annotations of sniffing behavior and attack behavior for each intruder male mouse were used to provide training labels of behavior type in control and experimental mice. Bar graphs of decoder accuracy (Figure S5E) were generated to discriminate sniffing and attack from imaged activity on individual frames of a behavior (sampled at 15 Hz). Equal numbers of sniff and attack frames (frame-wise decoder) were used during decoder training, to ensure chance decoder performance of 50%. 'Shuffled' decoder data were generated by training the decoder on the same neural data, but with sniff and attack behavior annotations randomly assigned to each behavior bout.

Decoding was repeated 20 times for each intruder and each imaged mouse, and decoder performance was reported as the average accuracy across imaged mice for control and experimental mice. For significance testing, the mean accuracy of the decoder trained on shuffled data was computed across mice, in each condition, and shuffling was repeated 1,000 times. Significance was determined across imaged mice using the Mann–Whitney U test between the mean accuracy of the decoders trained on real versus shuffled data.

Statistical analysis

Data were processed and analyzed using Python, MATLAB and GraphPad (GraphPad PRISM v.9). Data were analyzed using two-tailed, nested non-parametric tests. Wilcoxon signed-rank test (paired, non-parametric Mann–Whitney *U*-test) was used for binary paired samples. Kolmogorov–Smirnov test was used for non-paired samples plotted as ECDF graphs. N.s. $P > 0.05$, $*P < 0.05$, $**P < 0.01$, $***P < 0.001$, $****P < 0.0001$.

Dynamical system models of neural data

As previously described in published work[21], we modeled neural activity using recurrent switching linear dynamical systems (rSLDS). Briefly, rSLDS is a generative model that breaks down non-linear time series data into sequences of linear dynamical modes. The model relates three sets of variables: a set of discrete states (z), a set of continuous latent factors (x) that captures the low-dimensional nature of neural activity, and the activity of recorded neurons (y) during male directed behavior (sniffing, dominant mounting and attack). For how the model is formulated, see Nair, et al., 2023[21]. Model accuracy is evaluated using a forward simulation metric as described in Nair et al., 2023[21]. Briefly: given the observed neural activity at time t, we predict the trajectory of the population activity vector over an ensuing short time interval Δt using the model, then compute the mean squared error (MSE) between that trajectory and the observed data at time $t + \Delta t$. This MSE is calculated across all dimensions of the latent space and repeated for all times t. This error metric is normalized to a 0-1 range in each animal across the whole recording and is computed across cross-validation folds to obtain a bounded measure of model performance.

Code used to fit rSLDS on neural data is available in the SSM package: (https://github.com/lindermanlab/ssm)

Code to generate flow fields and energy landscapes from fit dynamical systems is available at (https://github.com/DJALab/VMHvl_MPOA_dynamics)

Visualization of attractor dynamics as 3D landscape

Conversion of the flow-fields obtained from rSLDS into a 3D landscape for visualization by calculating the dynamic velocity at each point in neural state space and using it as the height of a 3D landscape. Dynamic velocity was calculated as previously reported in Nair et al., 2023[21].

Estimation of time constants & calculation of line attractor score.

We estimated the time constant of each mode of linear dynamical systems using eigenvalues λ_a of the dynamics matrix of that system as: $\tau_a = \left|\frac{1}{\log(|\lambda_a|)}\right|$ as derived by Maheswaranathan et al., 2019 [113]. We used a line attractor score computed as $\log_2 \frac{t_n}{t_{n-1}}$ where t_n is the largest time constant of the dynamics matrix of a dynamical system andt_{n-1} is the second largest time constant. In the case of point attractors, the line attractor score is zero due to the similar magnitudes of the first two largest time constants, and it is greater than one for systems that possess a line attractor.

Supplemental Figure 1

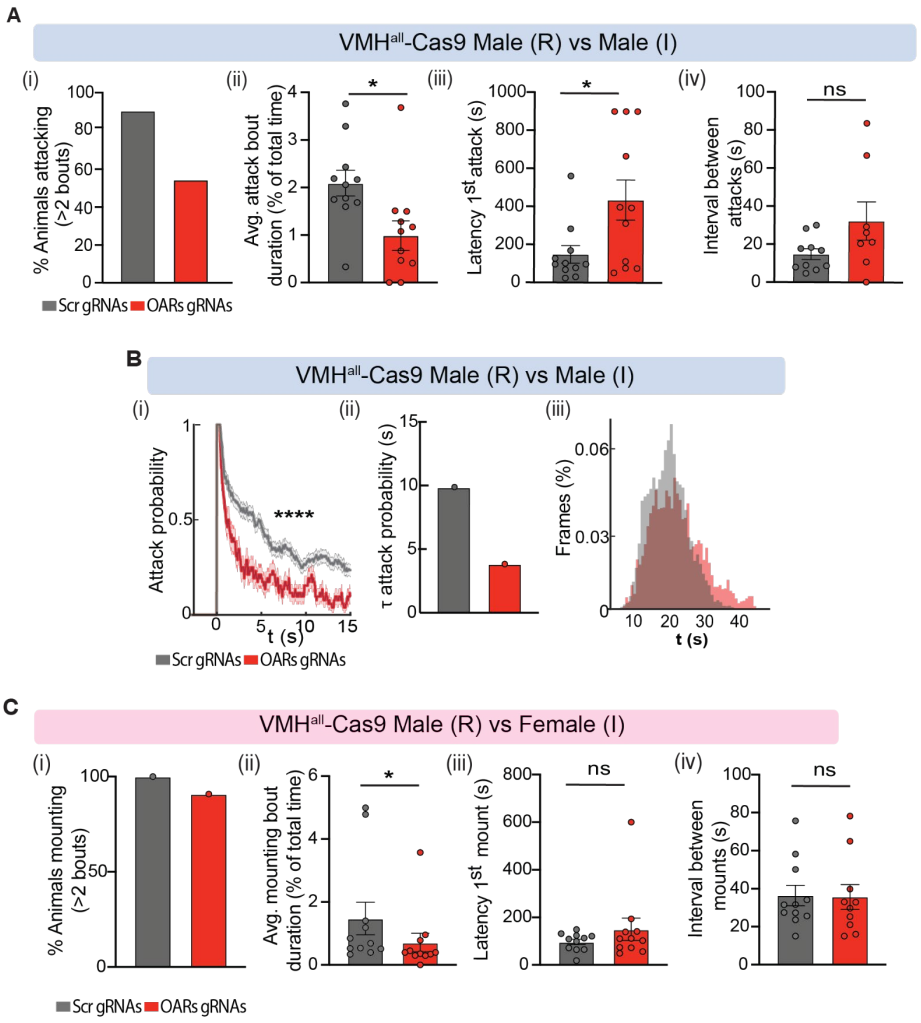


Figure S1. Experimental validation and behavioral characterization of CRISPR/Cas9 based Oxtr and Avpr1a co-perturbations, related to Figure 1.

A) Quantification of the percentage of animals attacking (i), the average duration of each attack bout (ii), the latency of the 1st attack bout (iii), and the interval between attack bouts (iv) against a male intruder. n=11 mice per group. B) Quantification of the number and time-varying probability of attack bouts against a male intruder (i-ii). n=11 mice per group. Quantification of the average velocity during attack in control (n=9 mice) and experimental mice (n=8 mice) (iii). C) Quantification of the percentage of animals mounting (i), the average duration of each mounting bout (ii), the latency of the 1st mounting bout (iii), and the interval between mounting bouts (iv) in control and experimental mice during male-female interactions. n=11 mice per group. Statistics: nested Mann-Whitney test was performed, and values were plotted as mean ± S.E.M. *$p \leq 0.05$. ****$p \leq 0.0001$.

Supplemental Figure 2

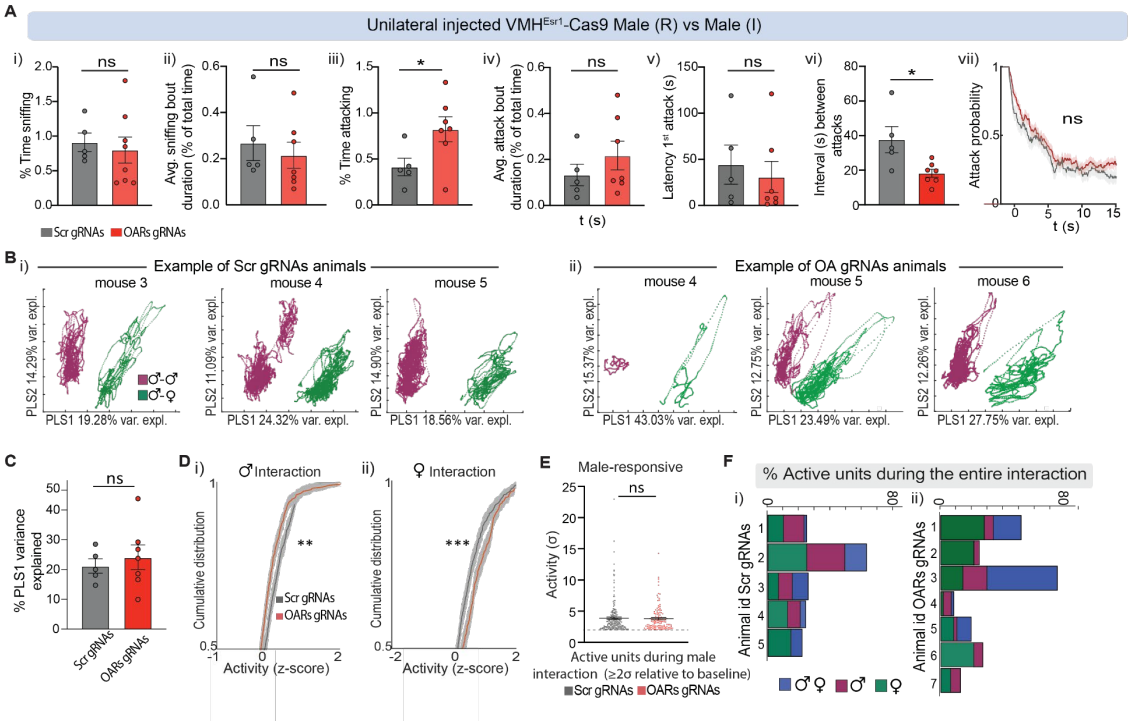


Figure S2. Effect of Oxtr/Avpr1a co-editing on intruder sex-specific representations and tuning. A) Quantification of the total time spent sniffing (i), the average duration of each sniffing bout (ii), the total time spent attacking (iii), the average duration of each attack bout (iv), the latency of the first attack (v), interval (s) between attack bouts (vi) and the attack probability (vii). n=5 control and n=7 experimental unilaterally injected mice. B) VMHvl^{Esr1} ensemble representations of intruder sex, for control (i) and experimental (ii) mice, projected onto the first two axes of a PLS regression against intruder sex. Traces are colored by intruder sex identity. The percentage of variance explained by the first two PLS components is noted for each male resident. C) Quantification of the PLS1 variance explained (which accounts for intruder sex) in control and experimental mice. D) Cumulative distribution of z-scored activity of all VMHvl^{Esr1} units during 1 min interaction with male (i) or female (ii) intruders in control and experimental mice. E) Average single unit activity (σ) of male responses (≥ 2σ relative to the pre-intruder baseline) between control and experimental mice. F) Percentage of male- or female selective or co-active units (≥ 2σ relative to the pre-intruder baseline), per imaged control (i) or experimental (ii) mouse. n=5 control, n=7 experimental animals.

Supplemental Figure 3

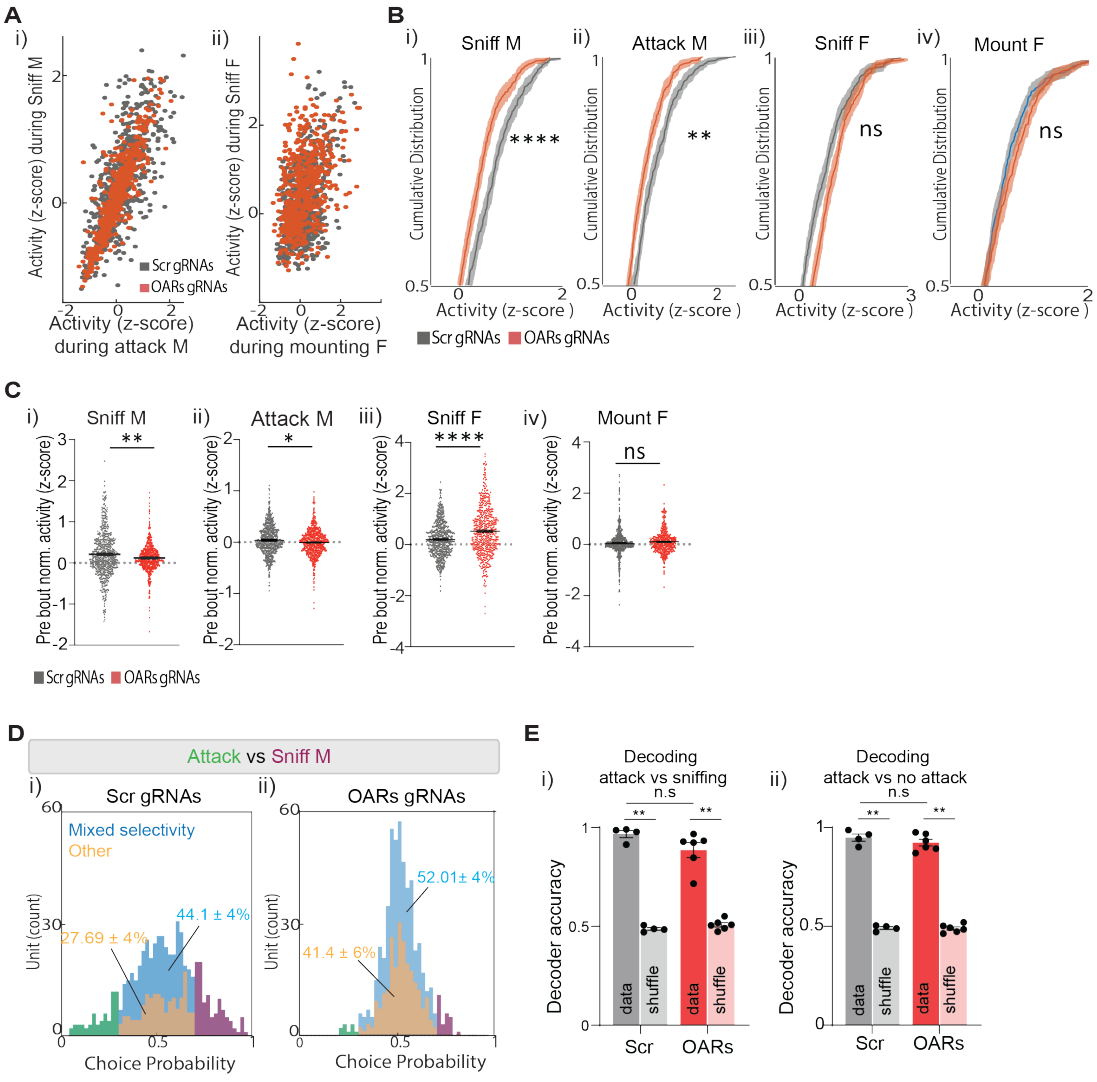


Figure S3. Effect of Oxtr/Avpr1a co-editing on behavior specific single unit activity and tuning, related to Figure 3.

A) Scatter plot of the average single VMHvl^{Esr1} unit activity (z-score) during male (i) and female (ii) directed behaviors in control and experimental mice. B) Cumulative distribution of VMHvl^{Esr1} activity (z-score) during male-directed (sniffing and attack; i & ii) and female-directed behaviors (sniffing and mounting; iii-iv) in control and experimental mice. C) Average z-scored VMHvl^{Esr1} activity normalized to the pre-behavior bout period during male-directed sniffing (i), attack (ii), female-directed sniffing (iii) and mounting (iv). D) Choice probabilities histograms of male-directed behaviors (attack vs. sniffing) in control and experimental mice. Mixed-tuned units (blue) and units tuned for other behaviors or not active (yellow) are highlighted. E) Accuracy of frame-wise decoders predicting attack vs sniffing (i) and attack vs no attack (ii) trained on VMHvl^{Esr1}neural activity in control or experimental animals. Decoders were trained and tested on held out data from each group separately. Statistics: Values plotted as mean ± S.E.M. in (B). Nested Kolmogorov–Smirnov test was used in (B), whereas nested Mann-Whitney test was performed in (C) and (E). *$p \leq 0.05$ **$p \leq 0.01$ ****$p \leq 0.0001$

Supplemental Figure 4

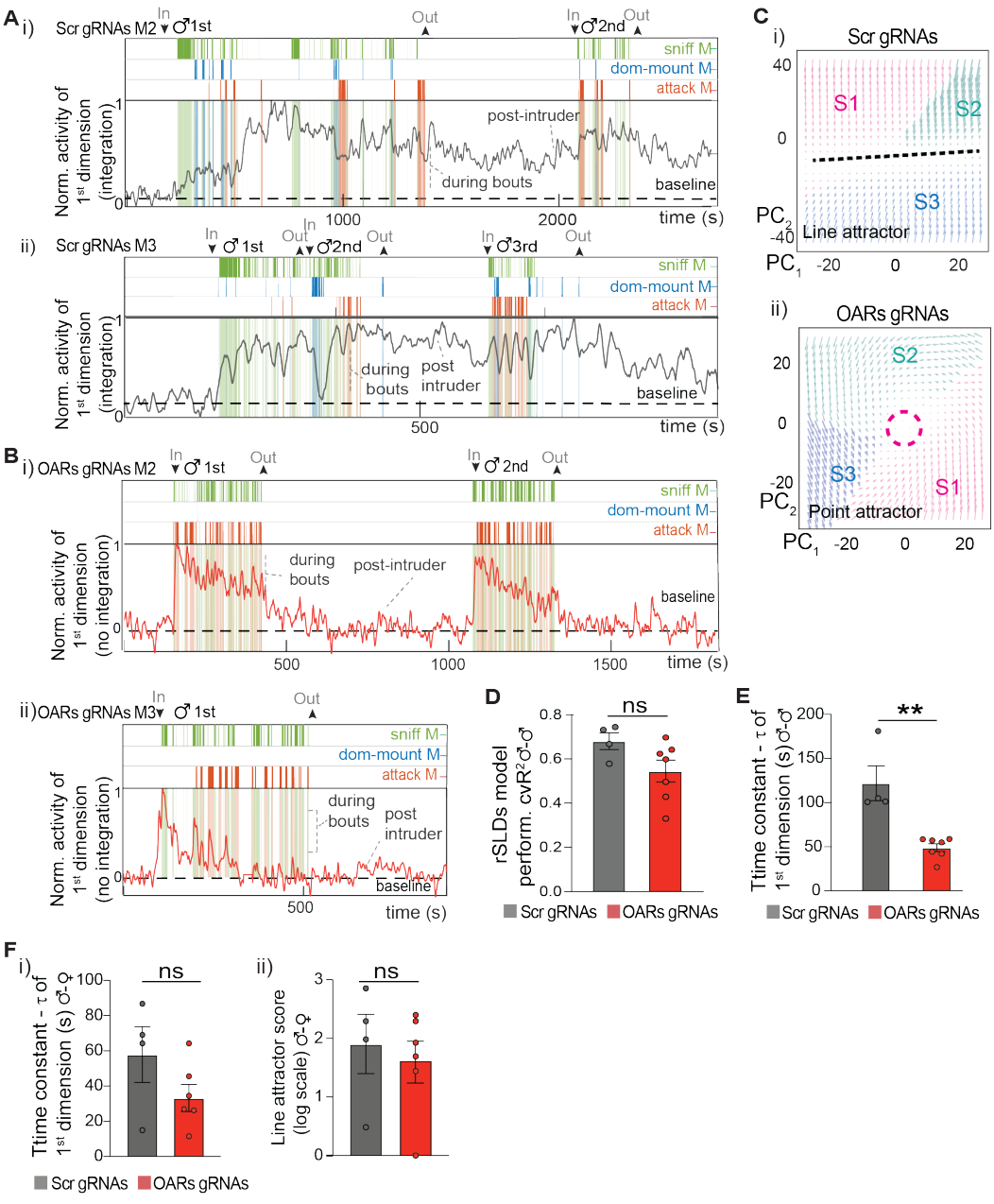


Figure S4. VMHvl^{Esr1} line attractor dynamics require Oxtr/Avpr1a-mediated signaling, related to Figure 4.

A) Normalized activity projection onto the time axis of the longest time constant (integration dimension) for VMHvl^{Esr1} units in control and B) in experimental mice. C) Schematic illustrating inferred dynamics shown as flow fields, with the line attractor illustrated by a black dashed line in control (i) and the point attractor with a purple dashed circle in experimental mouse (ii). The different rSLDS states (S1-S3) are depicted with different colors. D) Quantification of the performance score for the rSLDS model for each mouse in the control and experimental group during male-male interactions. E) Quantification of the average time constant (τ) of the 1st dimension in control and experimental mice during male-male interactions. n=4 control and n=7 mutant. F) Quantification of the average time constant (τ) of the 1st dimension (i) and the line attractor score for VMHvl^{Esr} population activity in control and experimental mice during male-female interactions. n=4 control and n=6 mutant. All sniffing, dominant mounting and attack bouts following the introduction of an intruder male are depicted in the behavioral raster plots in (A) and (B). Statistics: Mann-Whitney test was performed in (D, E, F). **$p \leq 0.01$

Supplemental Figure 5

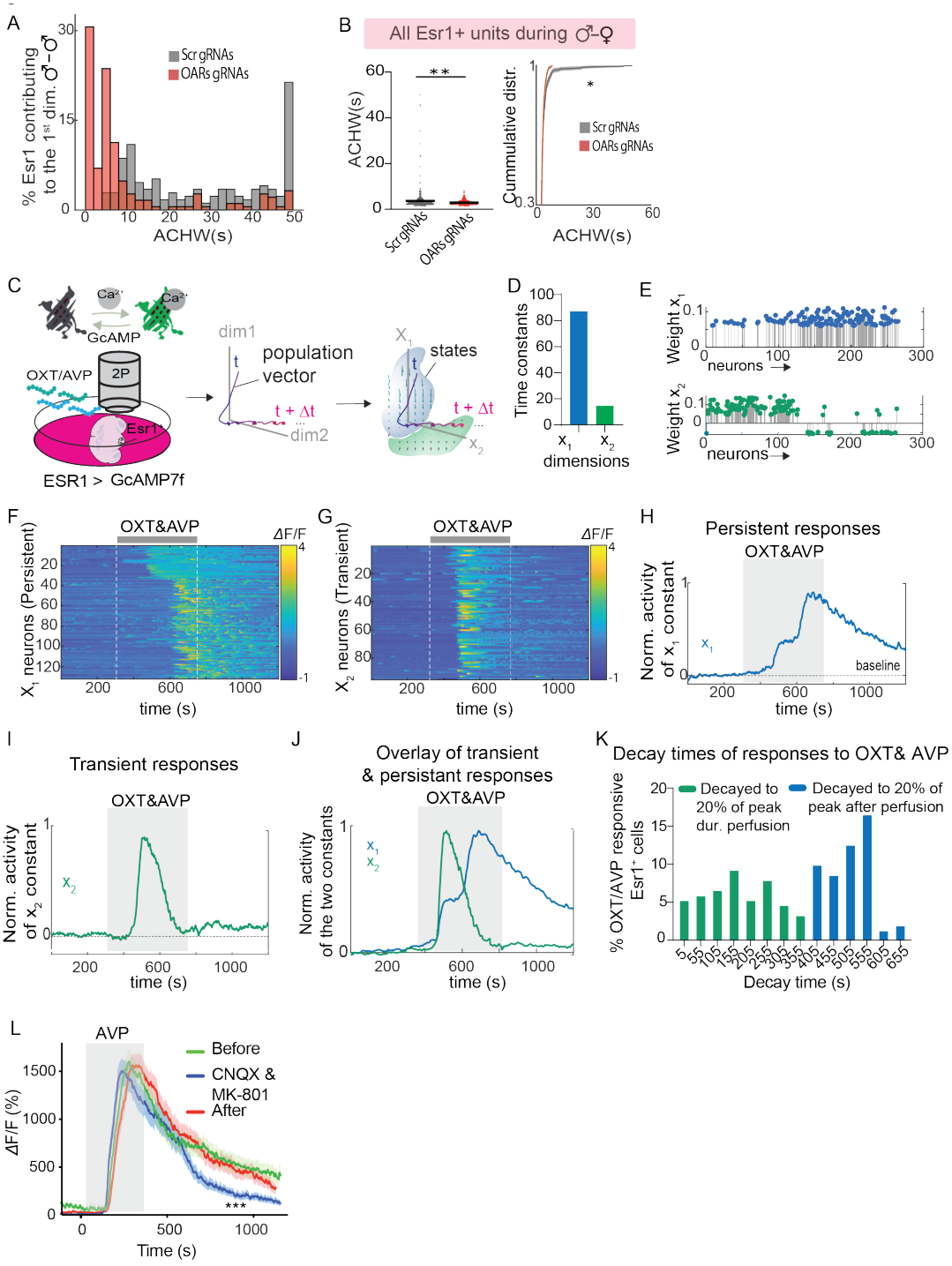


Figure S5 | OXT and AVP evoke persistent responses in VMHvl^{Esr1} neurons *ex vivo,* related to Figure 5.

A) Distribution plot illustrating the ACHW of VMHvl^{Esr1} units contributing to the 1st dimension during the entire duration of male-male interactions. n=4 control, n=7 experimental animals. B) Average single unit and cumulative distribution of ACHW from all VMHvl^{Esr1} recorded units in control and experimental mice during the first 1-2 minutes of male-female interactions. n=5 control, n=7 experimental animals. C) rSLDS modeling of Esr1^{+} cells calcium responses to 400nM OXT & AVP. Acute brain slices from male ESR1-2A-CRE animals that express Cre-dependent GCaMP7f were used. D) Average time constant of the identified two dimensions (X_1, X_2), arranged in decreasing order for acute brain slices. E) Absolute rSLDS weight of neurons contributing to X_1 (top, blue) and X_2 (bottom, green) dimensions sorted by choice probability values for activity during the OXT and AVP perfusion window. F) Heat maps of OXT and AVP induced slow persistent X_1 (left) and G) transient X_2 responses (right) VMHvl^{Esr1} neurons. H) Projection of population activity onto the time axis of the slow persistent X_1 and I) the transient X_2 dimensions. J) Overlay projection of population activity onto the time axis of the slow persistent X_1 and the transient X_2 dimensions. K) Distribution plot illustrating the decay times of Esr1 calcium responses to a 400nM OXT and AVP cocktail. Cells whose responses decayed to 20% of the peak within the peptide perfusion window were categorized as "transient" (green), while those responses that decayed to 20% of the peak after the perfusion window were labeled as "persistent" (blue). L) AVP (100nM)-mediated calcium average responses of VMH slices in the presence of synaptic transmission blockers (20uM CNQX and 10uM MK-801; n = 45 cells). Statistics: Kruskal-Wallis test was performed, corrected with Dunn's multiple comparison during time points 0s-1000s. *p≤0.05 **p≤0.01 ***p≤0.001

References

1. Sternson, S.M. (2013). Hypothalamic survival circuits: blueprints for purposive behaviors. *Neuron 77*, 810-824. 10.1016/j.neuron.2013.02.018.

2. Anderson, D.J., and Adolphs, R. (2014). A framework for studying emotions across species. *Cell 157*, 187-200. 10.1016/j.cell.2014.03.003.

3. Zych, A.D., and Gogolla, N. (2021). Expressions of emotions across species. Curr Opin Neurobiol *68*, 57-66. 10.1016/j.conb.2021.01.003.

4. Damasio, A., and Carvalho, G.B. (2013). The nature of feelings: evolutionary and neurobiological origins. *Nature Reviews Neuroscience 14*, 143-152. 10.1038/nrn3403.

5. Dolan, R.J. (2002). Emotion, cognition, and behavior. *Science 298*, 1191-1194.

6. Flavell, S.W., Gogolla, N., Lovett-Barron, M., and Zelikowsky, M. (2022). The emergence and influence of internal states. *Neuron 110*, 2545-2570. 10.1016/j.neuron.2022.04.030.

7. Malezieux, M., Klein, A.S., and Gogolla, N. (2023). Neural circuits for emotion. *Annu Rev Neurosci 46*, 211-231. 10.1146/annurev-neuro-111020-103314.

8. Luo, L., Callaway, E.M., and Svoboda, K. (2018). Genetic Dissection of Neural Circuits: A Decade of Progress. *Neuron 98*, 256-281. 10.1016/j.neuron.2018.03.040.

9. Deisseroth, K. (2014). Circuit dynamics of adaptive and maladaptive behaviour. *Nature 505*, 309-317. 10.1038/nature12982.

10. Xu, S., Yang, H., Menon, V., Lemire, A.L., Wang, L., Henry, F.E., Turaga, S.C., and Sternson, S.M. (2020). Behavioral state coding by molecularly defined paraventricular hypothalamic cell type ensembles. *Science 370*. 10.1126/science.abb2494.

11. Augustine, V., Lee, S., and Oka, Y. (2020). Neural Control and Modulation of Thirst, Sodium Appetite, and Hunger. *Cell 180*, 25-32. 10.1016/j.cell.2019.11.040.

12. Kennedy, A., Kunwar, P.S., Li, L.Y., Stagkourakis, S., Wagenaar, D.A., and Anderson, D.J. (2020). Stimulus-specific hypothalamic encoding of a persistent defensive state. *Nature 586*, 730-734. 10.1038/s41586-020-2728-4.

13. Jazayeri, M., and Afraz, A. (2017). Navigating the Neural Space in Search of the Neural Code. *Neuron 93*, 1003-1014. 10.1016/j.neuron.2017.02.019.

14. Ebitz, R.B., and Hayden, B.Y. (2021). The population doctrine in cognitive neuroscience. *Neuron 109*, 3055-3068. 10.1016/j.neuron.2021.07.011.

15. Shenoy, K.V., Sahani, M., and Churchland, M.M. (2013). Cortical control of arm movements: a dynamical systems perspective. *Annu Rev Neurosci 36*, 337-359.

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

17. Seung, H.S. (1996). How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences 93*, 13339-13344.

18. Khona, M., and Fiete, I.R. (2022). Attractor and integrator networks in the brain. *Nat Rev Neurosci*. 10.1038/s41583-022-00642-0.

19. Inagaki, H.K., Chen, S., Daie, K., Finkelstein, A., Fontolan, L., Romani, S., and Svoboda, K. (2022). Neural algorithms and circuits for motor planning. *Annu Rev Neurosci 45*, 249-271. 10.1146/annurev-neuro-092021-121730.

20. Liu, M., Nair, A., Linderman, S.W., and Anderson, D.J. (2023). Periodic hypothalamic attractor-like dynamics during the estrus cycle. *bioRxiv*. 10.1101/2023.05.22.541741.

21. Nair, A., Karigo, T., Yang, B., Ganguli, S., Schnitzer, M.J., Linderman, S.W., Anderson, D.J., and Kennedy, A. (2023). An approximate line attractor in the hypothalamus encodes an aggressive state. *Cell 186*, 178-193 e115. 10.1016/j.cell.2022.11.027.

22. Sylwestrak, E.L., Jo, Y., Vesuna, S., Wang, X., Holcomb, B., Tien, R.H., Kim, D.K., Fenno, L., Ramakrishnan, C., Allen, W.E., et al. (2022). Cell-type-specific population dynamics of diverse reward computations. *Cell 185*, 3568-3587 e3527. 10.1016/j.cell.2022.08.019.

23. Langdon, C., Genkin, M., and Engel, T.A. (2023). A unifying perspective on neural manifolds and circuits for cognition. *Nat Rev Neurosci 24*, 363-377. 10.1038/s41583-023-00693-x.

24. Hulse, B.K., and Jayaraman, V. (2020). Mechanisms underlying the neural computation of head direction. *Annu Rev Neurosci 43*, 31-54.

25. Kim, S.S., Rouault, H., Druckmann, S., and Jayaraman, V. (2017). Ring attractor dynamics in the Drosophila central brain. *Science 356*, 849-853. 10.1126/science.aal4835.

26. Inagaki, H.K., Chen, S., Ridder, M.C., Sah, P., Li, N., Yang, Z., Hasanbegovic, H., Gao, Z., Gerfen, C.R., and Svoboda, K. (2022). A midbrain-thalamus-cortex circuit reorganizes cortical dynamics to initiate movement. *Cell 185*, 1065-1081 e1023. 10.1016/j.cell.2022.02.006.

27. Goldman, M., Compte, A., and Wang, X.-J. (2007). Neural integrators: recurrent mechanisms and models. *New Encyclopedia of Neuroscience*, 1-26.

28. Major, G., and Tank, D. (2004). Persistent neural activity: prevalence and mechanisms. *Curr Opin Neurobiol 14*, 675-684. 10.1016/j.conb.2004.10.017.

29. Turner-Evans, D.B., Jensen, K.T., Ali, S., Paterson, T., Sheridan, A., Ray, R.P., Wolff, T., Lauritzen, J.S., Rubin, G.M., Bock, D.D., and Jayaraman, V. (2020). The Neuroanatomical ultrastructure and function of a biological ring attractor. *Neuron 108*, 145-163 e110. 10.1016/j.neuron.2020.08.006.

30. Hokfelt, T., Broberger, C., Xu, Z.Q., Sergeyev, V., Ubink, R., and Diez, M. (2000). Neuropeptides--an overview. *Neuropharmacology 39*, 1337-1356. 10.1016/s0028-3908(00)00010-1.

31. Russo, A.F. (2017). Overview of Neuropeptides: Awakening the Senses? *Headache 57 Suppl 2*, 37-46. 10.1111/head.13084.

32. Wang, P., Wang, S.C., Liu, X., Jia, S., Wang, X., Li, T., Yu, J., Parpura, V., and Wang, Y.F. (2022). Neural functions of hypothalamic oxytocin and its regulation. ASN Neuro *14*, 17590914221100706. 10.1177/17590914221100706.

33. Koppan, M., Nagy, Z., Bosnyak, I., and Reglodi, D. (2022). Female reproductive functions of the neuropeptide PACAP. *Front Endocrinol* (Lausanne) *13*, 982551. 10.3389/fendo.2022.982551.

34. Asahina, K., Watanabe, K., Duistermars, B.J., Hoopfer, E., González, C.R., Eyjólfsdóttir, E.A., Perona, P., and Anderson, D.J. (2014). Tachykinin-expressing neurons control male-specific aggressive arousal in drosophila. *Cell 156*, 221-235. 10.1016/j.cell.2013.11.045.

35. Siegel, A., Roeling, T.A., Gregg, T.R., and Kruk, M.R. (1999). Neuropharmacology of brain-stimulation-evoked aggression. *Neuroscience & Biobehavioral Reviews 23*, 359-389.

36. Zelikowsky, M., Hui, M., Karigo, T., Choe, A., Yang, B., Blanco, M.R., Beadle, K., Gradinaru, V., Deverman, B.E., and Anderson, D.J. (2018). The neuropeptide Tac2 controls a distributed brain state induced by chronic social isolation stress. *Cell 173*, 1265-1279 e1219. 10.1016/j.cell.2018.03.037.

37. Lim, M.M., and Young, L.J. (2006). Neuropeptidergic regulation of affiliative behavior and social bonding in animals. *Horm Behav 50*, 506-517. 10.1016/j.yhbeh.2006.06.028.

38. Marder, E. (2012). Neuromodulation of neuronal circuits: Back to the future. *Neuron 76*, 1-11. 10.1016/j.neuron.2012.09.010.

39. Nusbaum, M.P., Blitz, D.M., and Marder, E. (2017). Functional consequences of neuropeptide and small-molecule co-transmission. *Nat Rev Neurosci 18*, 389-403. 10.1038/nrn.2017.56.

40. Bargmann, C.I., and Marder, E. (2013). From the connectome to brain function. *Nature Methods 10*, 483-490. 10.1038/nmeth.2451.

41. Flavell, S.W., Pokala, N., Macosko, E.Z., Albrecht, D.R., Larsch, J., and Bargmann, C.I. (2013). Serotonin and the neuropeptide PDF initiate and extend opposing behavioral states in C. elegans. *Cell 154*, 1023-1035. 10.1016/j.cell.2013.08.001.

42. Ji, N., Madan, G.K., Fabre, G.I., Dayan, A., Baker, C.M., Kramer, T.S., Nwabudike, I., and Flavell, S.W. (2021). A neural circuit for flexible control of persistent behavioral states. *Elife 10*. 10.7554/eLife.62889.

43. Kato, S., Kaplan, H.S., Schrödel, T., Skora, S., Lindsay, T.H., Yemini, E., Lockery, S., and Zimmer, M. (2015). Global Brain Dynamics Embed the Motor Command Sequence of Caenorhabditis elegans. *Cell*, 656-669. 10.1016/j.cell.2015.09.034.

44. Dag, U., Nwabudike, I., Kang, D., Gomes, M.A., Kim, J., Atanas, A.A., Bueno, E., Estrem, C., Pugliese, S., Wang, Z., et al. (2023). Dissecting the functional organization of the C. elegans serotonergic system at whole-brain scale. *Cell*. 10.1016/j.cell.2023.04.023.

45. Root, C.M., Ko, K.I., Jafari, A., and Wang, J.W. (2011). Presynaptic facilitation by neuropeptide signaling mediates odor-driven food search. *Cell 145*, 133-144. 10.1016/j.cell.2011.02.008.

46. Watanabe, K., Chiu, H., Pfeiffer, B.D., Wong, A.M., Hoopfer, E.D., Rubin, G.M., and Anderson, D.J. (2017). A circuit node that integrates convergent input from neuromodulatory and social behavior-promoting neurons to control aggression in Drosophila. *Neuron 95*, 1112-1128 e1117. 10.1016/j.neuron.2017.08.017.

47. Andalman, A.S., Burns, V.M., Lovett-Barron, M., Broxton, M., Poole, B., Yang, S.J., Grosenick, L., Lerner, T.N., Chen, R., Benster, T., et al. (2019). Neuronal dynamics regulating brain and behavioral state transitions. *Cell 177*, 970-985 e920. 10.1016/j.cell.2019.02.037.

48. Heidenreich, M., and Zhang, F. (2016). Applications of CRISPR-Cas systems in neuroscience. *Nat Rev Neurosci 17*, 36-44. 10.1038/nrn.2015.2.

49. Doudna, J.A., and Charpentier, E. (2014). Genome editing. The new frontier of genome engineering with CRISPR-Cas9. *Science 346*, 1258096. 10.1126/science.1258096.

50. Ziv, Y., Burns, L.D., Cocker, E.D., Hamel, E.O., Ghosh, K.K., Kitch, L.J., El Gamal, A., and Schnitzer, M.J. (2013). Long-term dynamics of CA1 hippocampal place codes. *Nat Neurosci 16*, 264-266. 10.1038/nn.3329.

51. Glaser, J., Whiteway, M., Cunningham, J.P., Paninski, L., and Linderman, S. (2020). Recurrent switching dynamical systems models for multiple interacting neural populations. *Advances in neural information processing systems 33*, 14867-14878.

52. Linderman, S., Johnson, M., Miller, A., Adams, R., Blei, D., and Paninski, L. (2017). Bayesian learning and inference in recurrent switching linear dynamical systems. (*PMLR*), pp. 914-922.

53. Zha, X., and Xu, X.H. (2021). Neural circuit mechanisms that govern inter-male attack in mice. *Cell Mol Life Sci 78*, 7289-7307. 10.1007/s00018-021-03956-x.

54. Hashikawa, Y., Hashikawa, K., Falkner, A.L., and Lin, D. (2017). Ventromedial Hypothalamus and the Generation of Aggression. *Frontiers in systems neuroscience 11*, 94. 10.3389/fnsys.2017.00094.

55. Froemke, R.C., and Young, L.J. (2021). Oxytocin, Neural Plasticity, and Social Behavior. *Annu Rev Neurosci 44*, 359-381. 10.1146/annurev-neuro-102320-102847.

56. de Jong, T.R., and Neumann, I.D. (2018). Oxytocin and Aggression. *Curr Top Behav Neurosci 35*, 175-192. 10.1007/7854_2017_13.

57. Berendzen, K.M., Sharma, R., Mandujano, M.A., Wei, Y., Rogers, F.D., Simmons, T.C., Seelke, A.M.H., Bond, J.M., Larios, R., Goodwin, N.L., et al. (2022). Oxytocin receptor is not required for social attachment in prairie voles. *Neuron*. 10.1016/j.neuron.2022.12.011.

58. Anpilov, S., Shemesh, Y., Eren, N., Harony-Nicolas, H., Benjamin, A., Dine, J., Oliveira, V.E.M., Forkosh, O., Karamihalev, S., Hüttl, R.-E., et al. (2020). Wireless Optogenetic Stimulation of Oxytocin Neurons in a Semi-natural Setup Dynamically Elevates Both Pro-social and Agonistic Behaviors. *Neuron 107*, 644-655.e647. 10.1016/j.neuron.2020.05.028.

59. Bale, T.L., Dorsa, D.M., and Johnston, C.A. (1995). Oxytocin receptor mRNA expression in the ventromedial hypothalamus during the estrous cycle. *J Neurosci 15*, 5058-5064. 10.1523/JNEUROSCI.15-07-05058.1995.

60. Dumais, K.M., and Veenema, A.H. (2016). Vasopressin and oxytocin receptor systems in the brain: Sex differences and sex-specific regulation of social behavior. *Front Neuroendocrinol 40*, 1-23. 10.1016/j.yfrne.2015.04.003.

61. Delville, Y., Mansour, K.M., and Ferris, C.F. (1996). Testosterone facilitates aggression by modulating vasopressin receptors in the hypothalamus. *Physiol Behav 60*, 25-29. 10.1016/0031-9384(95)02246-5.

62. Lischinsky, J.E., and Lin, D. (2020). Neural mechanisms of aggression across species. *Nat Neurosci 23*, 1317-1328. 10.1038/s41593-020-00715-2.

63. Anderson, D.J. (2016). Circuit modules linking internal states and social behaviour in flies and mice. *Nat Rev Neurosci 17*, 692-704. 10.1038/nrn.2016.125.

64. Lee, H., Kim, D.W., Remedios, R., Anthony, T.E., Chang, A., Madisen, L., Zeng, H., and Anderson, D.J. (2014). Scalable control of mounting and attack by Esr1+ neurons in the ventromedial hypothalamus. *Nature 509*, 627-632. 10.1038/nature13169.

65. *Vinograd, A., *Nair, A., Linderman, S.W., and Anderson, D.J. (2024). Intrinsic dynamics and neural implementation of a hypothalamic line attractor encoding an internal behavioral state. *bioRxiv*. 10.1101/2024.05.21.595051.

66. Remedios, R., Kennedy, A., Zelikowsky, M., Grewe, B.F., Schnitzer, M.J., and Anderson, D.J. (2017). Social behaviour shapes hypothalamic neural ensemble representations of conspecific sex. *Nature 550*, 388-392. 10.1038/nature23885.

67. Kim, D.-W., Yao, Z., Graybuck, L.T., Kim, T.K., Nguyen, T.N., Smith, K.A., Fong, O., Yi, L., Koulena, N., Pierson, N., et al. (2019). Multimodal analysis of cell types in a hypothalamic node controlling social behavior. *Cell in press*.

68. Ragnauth, A.K., Goodwillie, A., Brewer, C., Muglia, L.J., Pfaff, D.W., and Kow, L.M. (2004). Vasopressin stimulates ventromedial hypothalamic neurons via oxytocin receptors in oxytocin gene knockout male and female mice. *Neuroendocrinology 80*, 92-99. 10.1159/000081844.

69. Inenaga, K., Karman, H., Yamashita, H., Tribollet, E., Raggenbass, M., and Dreifuss, J.J. (1991). Oxytocin excites neurons located in the ventromedial nucleus of the Guinea-pig hypothalamus. *J Neuroendocrinol 3*, 569-573. 10.1111/j.1365-2826.1991.tb00318.x.

70. Song, Z., and Albers, H.E. (2018). Cross-talk among oxytocin and arginine-vasopressin receptors: Relevance for basic and clinical studies of the brain and periphery. *Front Neuroendocrinol 51*, 14-24. 10.1016/j.yfrne.2017.10.004.

71. Kabadi, A.M., Ousterout, D.G., Hilton, I.B., and Gersbach, C.A. (2014). Multiplex CRISPR/Cas9-based genome engineering from a single lentiviral vector. *Nucleic Acids Res 42*, e147. 10.1093/nar/gku749.

72. Sarin, S., Zuniga-Sanchez, E., Kurmangaliyev, Y.Z., Cousins, H., Patel, M., Hernandez, J., Zhang, K.X., Samuel, M.A., Morey, M., Sanes, J.R., and Zipursky, S.L. (2018). Role for Wnt Signaling in Retinal Neuropil Development: Analysis via RNA-Seq and In Vivo Somatic CRISPR Mutagenesis. *Neuron 98*, 109-126 e108. 10.1016/j.neuron.2018.03.004.

73. Sentmanat, M.F., Peters, S.T., Florian, C.P., Connelly, J.P., and Pruett-Miller, S.M. (2018). A Survey of Validation Strategies for CRISPR-Cas9 Editing. *Sci Rep 8*, 888. 10.1038/s41598-018-19441-8.

74. Brinkman, E.K., Chen, T., Amendola, M., and van Steensel, B. (2014). Easy quantitative assessment of genome editing by sequence trace decomposition. *Nucleic Acids Res 42*, e168. 10.1093/nar/gku936.

75. Ferris, C.F. (2005). Vasopressin/oxytocin and aggression. Molecular mechanisms influencing aggressive behaviours, 190-200.

76. DeVries, A.C., Young, W.S., 3rd, and Nelson, R.J. (1997). Reduced aggressive behaviour in mice with targeted disruption of the oxytocin gene. *J Neuroendocrinol 9*, 363-368. 10.1046/j.1365-2826.1997.t01-1-00589.x.

77. Wersinger, S.R., Caldwell, H.K., Martinez, L., Gold, P., Hu, S.B., and Young, W.S., 3rd (2007). Vasopressin 1a receptor knockout mice have a subtle olfactory deficit but normal aggression. *Genes Brain Behav 6*, 540-551. 10.1111/j.1601-183X.2006.00281.x.

78. Egashira, N., Tanoue, A., Matsuda, T., Koushi, E., Harada, S., Takano, Y., Tsujimoto, G., Mishima, K., Iwasaki, K., and Fujiwara, M. (2007). Impaired social interaction and reduced anxiety-related behavior in vasopressin V1a receptor knockout mice. *Behav Brain Res 178*, 123-127. 10.1016/j.bbr.2006.12.009.

79. Ghosh, K.K., Burns, L.D., Cocker, E.D., Nimmerjahn, A., Ziv, Y., Gamal, A.E., and Schnitzer, M.J. (2011). Miniaturized integration of a fluorescence microscope. *Nat Methods 8*, 871-878. 10.1038/nmeth.1694.

80. Tuladhar, R., Yeu, Y., Tyler Piazza, J., Tan, Z., Rene Clemenceau, J., Wu, X., Barrett, Q., Herbert, J., Mathews, D.H., Kim, J., et al. (2019). CRISPR-Cas9-based mutagenesis frequently provokes on-target mRNA misregulation. *Nat Commun 10*, 4056. 10.1038/s41467-019-12028-5.

81. Hashikawa, K., Hashikawa, Y., Tremblay, R., Zhang, J., Feng, J.E., Sabol, A., Piper, W.T., Lee, H., Rudy, B., and Lin, D. (2017). Esr1(+) cells in the ventromedial hypothalamus control female aggression. *Nat Neurosci 20*, 1580-1590. 10.1038/nn.4644.

82. Yang, B., Karigo, T., and Anderson, D.J. (2022). Transformations of neural representations in a social behaviour network. *Nature 608*, 741-749. 10.1038/s41586-022-05057-6.

83. Lerner, T.N., Shilyansky, C., Davidson, T.J., Evans, K.E., Beier, K.T., Zalocusky, K.A., Crow, A.K., Malenka, R.C., Luo, L., Tomer, R., and Deisseroth, K. (2015). Intact-Brain Analyses Reveal Distinct Information Carried by SNc Dopamine Subcircuits. *Cell 162*, 635-647. 10.1016/j.cell.2015.07.014.

84. Karigo, T., Kennedy, A., Yang, B., Liu, M., Tai, D., Wahle, I.A., and Anderson, D.J. (2021). Distinct hypothalamic control of same- and opposite-sex mounting behaviour in mice. *Nature 589*, 258-263. 10.1038/s41586-020-2995-0.

85. Egorov, A.V., Hamam, B.N., Fransen, E., Hasselmo, M.E., and Alonso, A.A. (2002). Graded persistent activity in entorhinal cortex neurons. *Nature 420*, 173-178. 10.1038/nature01171.

86. Heys, J.G., Schultheiss, N.W., Shay, C.F., Tsuno, Y., and Hasselmo, M.E. (2012). Effects of acetylcholine on neuronal properties in entorhinal cortex. *Front Behav Neurosci 6*, 32. 10.3389/fnbeh.2012.00032.

87. Liu, J.J., Eyring, K.W., Konig, G.M., Kostenis, E., and Tsien, R.W. (2022). Oxytocin-modulated ion channel ensemble controls depolarization, integration and burst firing in CA2 pyramidal neurons. *J Neurosci 42*, 7707-7720. 10.1523/JNEUROSCI.0921-22.2022.

88. Cavanagh, S.E., Towers, J.P., Wallis, J.D., Hunt, L.T., and Kennerley, S.W. (2018). Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. *Nat Commun 9*, 3498. 10.1038/s41467-018-05873-3.

89. Murray, J.D., Bernacchia, A., Freedman, D.J., Romo, R., Wallis, J.D., Cai, X., Padoa-Schioppa, C., Pasternak, T., Seo, H., Lee, D., and Wang, X.J. (2014). A hierarchy of intrinsic timescales across primate cortex. *Nat Neurosci 17*, 1661-1663. 10.1038/nn.3862.

90. Yang, C.F., Chiang, M.C., Gray, D.C., Prabhakaran, M., Alvarado, M., Juntti, S.A., Unger, E.K., Wells, J.A., and Shah, N.M. (2013). Sexually dimorphic neurons in the ventromedial hypothalamus govern mating in both sexes and aggression in males. *Cell 153*, 896-909. 10.1016/j.cell.2013.04.017.

91. Osakada, T., Yan, R., Jiang, Y., Wei, D., Tabuchi, R., Dai, B., Wang, X., Zhao, G., Wang, C.X., Liu, J.J., et al. (2024). A dedicated hypothalamic oxytocin circuit controls aversive social learning. *Nature 626*, 347-356. 10.1038/s41586-023-06958-w.

92. Shao, Y.Q., Fan, L., Wu, W.Y., Zhu, Y.J., and Xu, H.T. (2022). A developmental switch between electrical and neuropeptide communication in the ventromedial hypothalamus. *Curr Biol 32*, 3137-3145 e3133. 10.1016/j.cub.2022.05.029.

93. Stagkourakis, S., Spigolon, G., Liu, G., and Anderson, D.J. (2020). Experience-dependent plasticity in an innate social behavior is mediated by hypothalamic LTP. *Proc Natl Acad Sci U S A 117*, 25789-25799. 10.1073/pnas.2011782117.

94. Koulakov, A.A., Raghavachari, S., Kepecs, A., and Lisman, J.E. (2002). Model for a robust neural integrator. *Nat Neurosci 5*, 775-782. 10.1038/nn893.

95. Goldman, M.S., Levine, J.H., Major, G., Tank, D.W., and Seung, H.S. (2003). Robust persistent neural activity in a model integrator with multiple hysteretic dendrites per neuron. *Cereb Cortex 13*, 1185-1195. 10.1093/cercor/bhg095.

96. Inagaki, H.K., Fontolan, L., Romani, S., and Svoboda, K. (2019). Discrete attractor dynamics underlies persistent activity in the frontal cortex. *Nature 566*, 212-217. 10.1038/s41586-019-0919-7.

97. Simerly, R.B. (1990). Hormonal control of neuropeptide gene expression in sexually dimorphic olfactory pathways. *Trends Neurosci 13*, 104-110. 10.1016/0166-2236(90)90186-e.

98. Inoue, S., Yang, R., Tantry, A., Davis, C.H., Yang, T., Knoedler, J.R., Wei, Y., Adams, E.L., Thombare, S., Golf, S.R., et al. (2019). Periodic Remodeling in a Neural Circuit Governs Timing of Female Sexual Behavior. *Cell 179*, 1393-1408 e1316. 10.1016/j.cell.2019.10.025.

99. Yin, L., Hashikawa, K., Hashikawa, Y., Osakada, T., Lischinsky, J.E., Diaz, V., and Lin, D. (2022). VMHvll(Cckar) cells dynamically control female sexual behaviors over the reproductive cycle. *Neuron 110*, 3000-3017 e3008. 10.1016/j.neuron.2022.06.026.

100. Teitler, M., and Herrick-Davis, K. (1994). Multiple serotonin receptor subtypes: molecular cloning and functional expression. *Crit Rev Neurobiol 8*, 175-188.

101. Li, Y., Mathis, A., Grewe, B.F., Osterhout, J.A., Ahanonu, B., Schnitzer, M.J., Murthy, V.N., and Dulac, C. (2017). Neuronal Representation of Social Information in the Medial Amygdala of Awake Behaving Mice. *Cell 171*, 1176-1190.e1117. 10.1016/j.cell.2017.10.015.

102. Kim, I.H., Kim, N., Kim, S., Toda, K., Catavero, C.M., Courtland, J.L., Yin, H.H., and Soderling, S.H. (2020). Dysregulation of the Synaptic Cytoskeleton in the PFC Drives Neural Circuit Pathology, Leading to Social Dysfunction. *Cell Rep 32*, 107965. 10.1016/j.celrep.2020.107965.

103. Hunker, A.C., Soden, M.E., Krayushkina, D., Heymann, G., Awatramani, R., and Zweifel, L.S. (2020). Conditional Single Vector CRISPR/SaCas9 Viruses for Efficient Mutagenesis in the Adult Mouse Nervous System. *Cell Rep 30*, 4303-4316 e4306. 10.1016/j.celrep.2020.02.092.

104. Melzer, S., Newmark, E.R., Mizuno, G.O., Hyun, M., Philson, A.C., Quiroli, E., Righetti, B., Gregory, M.R., Huang, K.W., Levasseur, J., et al. (2021). Bombesin-like peptide recruits disinhibitory cortical circuits and enhances fear memories. *Cell 184*, 5622-5634 e5625. 10.1016/j.cell.2021.09.013.

105. Castro, D.C., Oswell, C.S., Zhang, E.T., Pedersen, C.E., Piantadosi, S.C., Rossi, M.A., Hunker, A.C., Guglin, A., Moron, J.A., Zweifel, L.S., et al. (2021). An endogenous opioid circuit determines state-dependent reward consumption. *Nature 598*, 646-651. 10.1038/s41586-021-04013-0.

106. Soden, M.E., Yee, J.X., and Zweifel, L.S. (2023). Circuit coordination of opposing neuropeptide and neurotransmitter signals. *Nature 619*, 332-337. 10.1038/s41586-023-06246-7.

107. Swiech, L., Heidenreich, M., Banerjee, A., Habib, N., Li, Y., Trombetta, J., Sur, M., and Zhang, F. (2015). In vivo interrogation of gene function in the mammalian brain using CRISPR-Cas9. *Nat Biotechnol 33*, 102-106. 10.1038/nbt.3055.

108. Zhang, Y., and Looger, L.L. (2023). Fast and sensitive GCaMP calcium indicators for neuronal imaging. *J Physiol.* 10.1113/JP283832.

109. Hong, W., Kennedy, A., Burgos-Artizzu, X.P., Zelikowsky, M., Navonne, S.G., Perona, P., and Anderson, D.J. (2015). Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning. *Proc Natl Acad Sci U S A 112*, E5351-5360. 10.1073/pnas.1515982112.

110. Segalin, C., Williams, J., Karigo, T., Hui, M., Zelikowsky, M., Sun, J.J., Perona, P., Anderson, D.J., and Kennedy, A. (2021). The Mouse Action Recognition System (MARS) software pipeline for automated analysis of social behaviors in mice. *Elife 10*. 10.7554/eLife.63720.

111. Zhou, P., Resendez, S.L., Rodriguez-Romaguera, J., Jimenez, J.C., Neufeld, S.Q., Giovannucci, A., Friedrich, J., Pnevmatikakis, E.A., Stuber, G.D., Hen, R., et al.

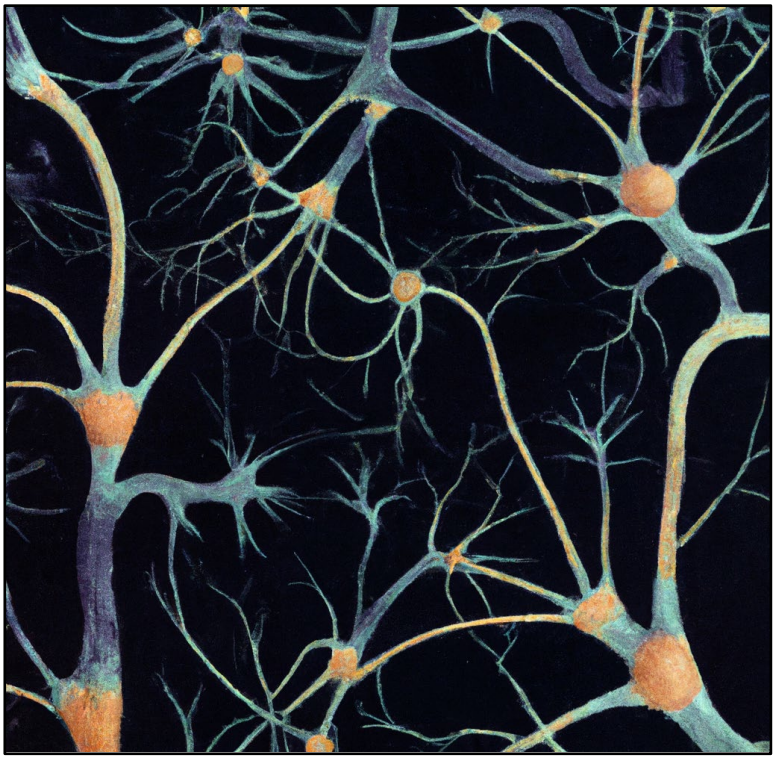
(2018). Efficient and accurate extraction of in vivo calcium signals from microendoscopic video data. *Elife 7*. 10.7554/eLife.28728.

112. Shadlen, M.N., Britten, K.H., Newsome, W.T., and Movshon, J.A. (1996). A computational analysis of the relationship between neuronal and behavioral responses to visual motion. *J Neurosci 16*, 1486-1510. 10.1523/JNEUROSCI.16-04-01486.1996.

113. Maheswaranathan, N., Williams, A.H., Golub, M.D., Ganguli, S., and Sussillo, D. (2019). Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics. *Adv Neural Inf Process Syst 32*, 15696-15705.

Chapter V

GENERALIZATION



ഒരേ വാക്കു, പല മനസ്സുകളിൽ അനവധി ഭാവങ്ങൾ സൃഷ്ടിക്കുന്നു.”

Vallathol Narayana Menon, Sahityamanjari, 1925

Translation: “The same word creates countless emotions in many minds.”

Chapter V

Encoding of female mating dynamics by a hypothalamic line attractor

This chapter shows the line attractor dynamics also generalize to other internal states, specifically the state of sexual arousal in female mice

Published as Mengyu Liu* **Aditya Nair***, Nestor Coria, Scott W. Linderman, David J. Anderson. Encoding of female mating dynamics by a hypothalamic line attractor. *Nature* (2024)

Summary

Females exhibit complex, dynamic behaviors during mating with variable sexual receptivity depending on hormonal status[1–4]. However, how their brains encode the dynamics of mating and receptivity remains largely unknown. The ventromedial hypothalamus, ventro-lateral subdivision contains estrogen receptor type 1-positive neurons that control mating receptivity in female mice[5,6]. Unsupervised dynamical systems analysis of calcium imaging data from these neurons during mating uncovered a dimension with slow ramping activity, generating a line attractor in neural state space. Neural perturbations in behaving females demonstrated relaxation of population activity back into the attractor. During mating population activity integrated male cues to ramp up along this attractor, peaking just before ejaculation. Activity in the attractor dimension was positively correlated with the degree of receptivity. Longitudinal imaging revealed that attractor dynamics appear and disappear across the estrus cycle and are hormone-dependent. These observations suggest that a hypothalamic line attractor encodes a persistent, escalating state of female sexual arousal or drive during mating. They also demonstrate that attractors can be reversibly modulated by hormonal status, on a timescale of days.

*co-first author.

Introduction

Mating is a complex social interaction whose success is essential to species' survival. In rodents, female mating receptivity has been considered as a binary behavior defined by lordosis[7–10], a reflexive acceptance posture. In fact, however, female receptivity is highly dynamic, exhibiting variability both within a mating interaction and across different physiological states[11]. Nevertheless, the female's important contribution to the dynamics of successful mating has been under-appreciated and under-studied, relative to the male's.

Recent progress has identified circuits that control female receptivity[1,3,4,12]. The ventrolateral subdivision of the ventromedial hypothalamic nucleus (VMHvl) contains a subset of Esr1^{+} neurons that controls mating behaviors in female mice[5,6,13–16]. Recent findings have revealed hormone-dependent changes in the anatomy and physiology of these neurons. The axonal arborizations of VMHvl progesterone receptor (PR)-expressing neurons in AVPV increase in receptive females, in an estrogen-dependent manner[17]. In addition, a small subset of Esr1^{+} neurons defined by expression of the cholecystokinin A receptor (*Cckar*)[6,18] has been shown to be necessary and sufficient for female receptivity and to exhibit estrus cycle-dependent changes in excitability *ex vivo*, and in response dynamics during the investigation phase of mating interactions *in vivo*[6]. While these studies have identified important circuit-level changes associated with the state of receptivity, how the dynamics of female behavior during mating are encoded in the brain is largely unknown.

To address this issue, we have characterized neural population representations in female VMHvl during interactions with males across the estrus cycle, using longitudinal miniscope imaging of calcium activity[19]. We imaged a subpopulation of Esr1^{+} neurons that are Npy2r^{-} that we call "α cells," which causally control sexual receptivity[5]; these cells overlap with the aforementioned Cckar^{+} cells[6,7]. Unsupervised modeling of VMHvl α cell activity using a dynamical systems approach[20] revealed an approximate line attractor in neural state space, which disappeared during non-receptive phases of the estrus cycle and

was hormone-dependent. Analysis of female mating behavior and line attractor dynamics suggest that the attractor integrates male contact cues and may represent a persistent, escalating internal state of female sexual arousal or receptivity during mating.

Results

Dynamics of female behaviors in mating

Female mating behavior has been studied more extensively in rats than in mice[4,21]. To detail murine female mating behavior under our standard conditions, we manually annotated video recordings of sexually receptive females interacting with a male (Figure 1a). We identified 10 female motor behaviors and classified them as appetitive (approaching and sniffing the male), accepting (lordosis and wiggling), or resisting (darting, top up, kicking and turning), based on the behavior's apparent intent[6]. The behaviors were dynamic, with the probability of accepting behaviors gradually increase, while resistance behaviors initially increased and then slowly decreased (Figure 1b and Extended Data Fig. 1a). Thus, receptivity is not binary but graded and dynamic.

We categorized mating behaviors as "self-initiated" (appetitive and check genital area) or "male-responsive" (accepting, resistance and staying in response to make attempts). Females spent six times longer displaying male-responsive (83.9%) vs. self-initiated behaviors (16.1%) (Figure 1c). Due to this asymmetry, we also quantified male-initiated behaviors (sniffing, mounting and intromission). Male spent 11.3 times more time displaying self-initiated mating behaviors than females (Figure 1d), indicating males largely drive mating interaction. The number and duration of male copulation bouts and inter-bouts intervals varied across interactions (Figure 1e and Extended Data Fig. 1b).

Next, we compared female behaviors during male copulation bouts vs. inter-bout intervals (Figure 1f). Behaviors were classified as "social" (accepting, resisting, and appetitive), or

"disengaged" (rearing, digging, and chewing). During copulation bouts, females primarily exhibited social behaviors (62% of each bout's duration) and rarely disengaged behaviors (Figure 1g, left). During inter-bout intervals (IBIs), when females were separated from the males, they continued social behaviors initiated during the preceding bout (Figure 1g, right, 23% of IBI duration). A behavior probability plot aligned to copulation offset showed females continued accepting or resisting behaviors and performed appetitive approaches and sniffing during IBIs (Figure 1h and Extended Data Fig. 1c).

These results demonstrate that female behaviors during mating are dynamic and primarily driven by male-initiated behaviors. The persistence of female social behaviors observed during pauses in copulation (Figure 1h) suggests a corresponding persistent internal state of mating receptivity or engagement. The ramping dynamics of "accepting" behaviors (Fig. 1b) further suggests that escalation may be a property of this internal state. Persistence and escalation (or "scalability") are features of internal states underlying other dynamic social behaviors, like male aggression[22]. We next investigate how these state properties are instantiated by neural activity and dynamics.

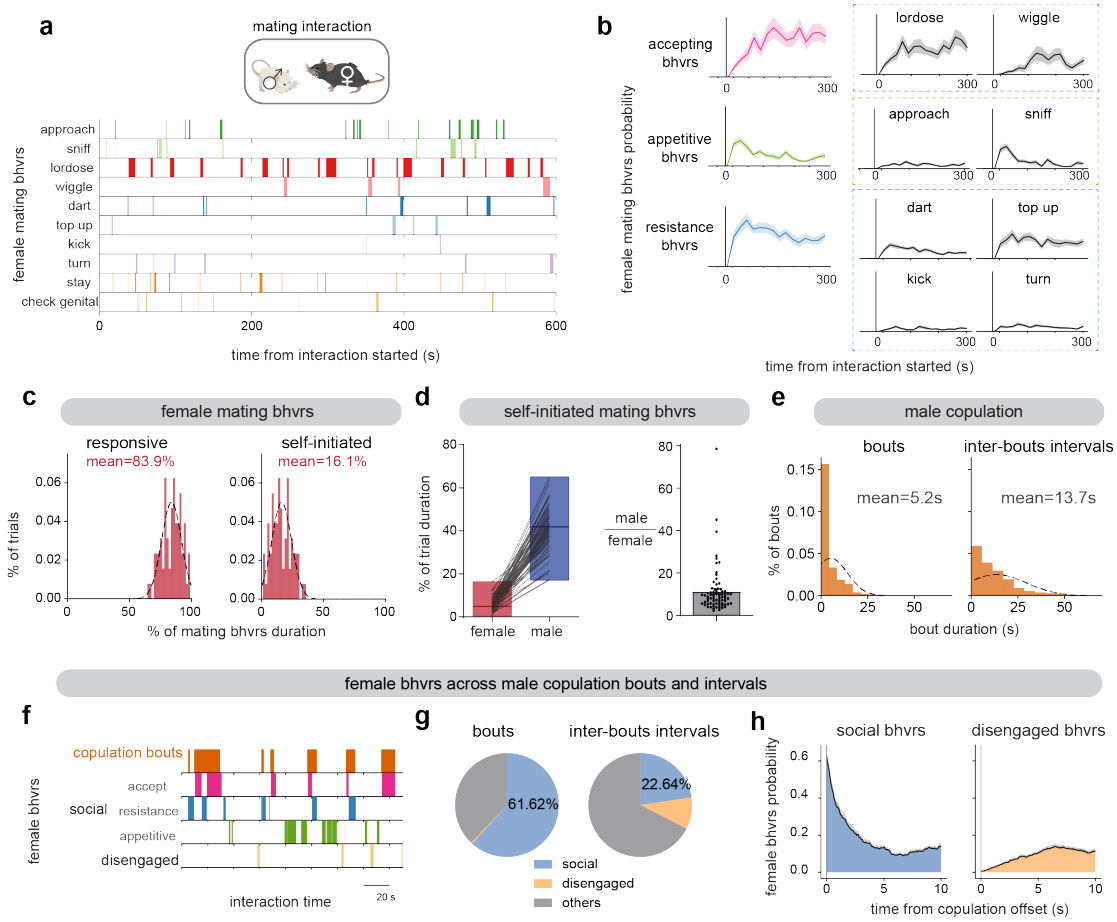


Figure 1. | Dynamics of female behaviors during mating interaction.

a, Raster plot of 10 female mating behaviors during one interaction with a male. b, The probability of mating behaviors during every 20 seconds (n=74 trials, N=28 mice). Behaviors were grouped as accept (comprising lordose, wiggle), appetitive (comprising approach, sniff), and resistance (comprising dart, top up, kick and turn), data presented as mean ± SEM. c, Distribution of the percentage of time females displayed responsive vs. self-initiated mating behaviors over the total mating behavior time in each trial (n=74 trials). Female self-initiated mating behaviors comprised appetitive behaviors and check genital. Female responsive mating behaviors comprised accepting, resistance behaviors and staying. d, Left, percentage of time female or male displayed self-initiated mating behaviors in each trial. Box boundaries range from min to max, with a line at the median. Right, male self-initiated mating time over female in each trial (n=74 trials). Data presented as mean ± SEM. Male self-initiated mating behaviors included male sniffing, mounting and

intromission. e, Distribution of the durations of male copulation bouts (left, n=1685) and inter-bout intervals (right) (n=1611). Male copulation included mounting and intromission. f, Raster plot of female behaviors during copulation bouts and inter-bout intervals. Social behaviors comprised accepting, resistance, and appetitive behaviors; non-social disengaged behaviors comprised rearing, digging and chewing. g, Percentage of time female displaying social behaviors in each male copulation bout or inter-bout interval; "others" indicates all behaviors other than the defined social behaviors or non-social disengaged behaviors during interaction. h, Female behavior probability aligned to male copulation offsets.

Tuning of female VMHvl neurons in mating

To uncover how the dynamics of female mating behavior are encoded in neural activity, we imaged VMHvl$^{Esr1+,Npy2r-}$ ("α") cells[5] in freely moving sexually receptive females interacting with males, using a miniature head-mounted microscope[19] (Figure 2a). Initially, we analyzed responses observed in the first minute of exposure to either a male or female conspecific and observed distinct subsets tuned to intruder sex, with male-preferring cells ~4 times more abundant, contributing to clear separation of sex in principle component space (Extended Data Fig. 1e,f,h). Correspondingly, the averaged population response to a male was ~4 times higher than that to a female (Extended Data Fig. 1g), consistent with prior bulk calcium imaging studies[5].

We next analyzed imaging data from females acquired during free mating interactions with a male, over 5-10 min observation periods (Fig. 2b; 16-207 units/mouse, mean 89±12= units/mouse; N = 15 mice). Choice probability[23] indicated that a relatively small percentage of VMHvl α cells (~2-13%) were "tuned" to specific mating behaviors (Figure 2c and Extended Data Fig. 2a, b), while the majority exhibited "mixed selectivity" (Figure 2c and Extended Data Fig. 2a, b, gray bars), indicating relatively weak behavior-specific tuning at the single neuron level.

To further investigate the relationship between mating behaviors and neural activity, we fit Generalized Linear Models (GLMs) to each neuron using female mating behaviors (Figure 2d), male mating behaviors (Extended Data Fig. 2c), or both female and male behaviors (Extended Data Fig. 2d). Across all animals, only ~8% of the variance in neural activity was explained by female mating behaviors (Figure 2e, mean cvR^2=0.08, N=15 mice); ~14% by male mating behaviors (Extended Data Fig. 2c, mean cvR^2=0.14) and ~15% of the variance was explained by combined female and male behaviors (Extended Data Fig. 2d, mean cvR^2=0.15).

Taken together, these single-cell analyses indicated that a large fraction of the variance in VMHvl α-cell neural activity during female mating could not be explained by behavior using a GLM. Nevertheless, a trained SVM linear decoder could distinguish mating behaviors with an accuracy higher than chance (Extended Data Fig. 2f-i), suggesting some relationship between neural activity or dynamics and behavior. To examine whether local interactions between neurons could improve the fit of our GLMs, we included coupling filters[24,25] in addition to the behavioral variables (Figure 2f). The introduction of neuronal coupling dramatically increased variance explained by the GLM, suggesting that local circuit interactions contribute more than behavior to neuronal variance in VMHvl α cell activity (Figure 2f, g, mean cvR^2=0.46, N = 15 mice). Because GLMs fit using low-dimensional coupling matrices (as obtained here) can reflect slowly evolving neural dynamics, we were motivated to analyze next the dynamics of VMHvl α cell activity.

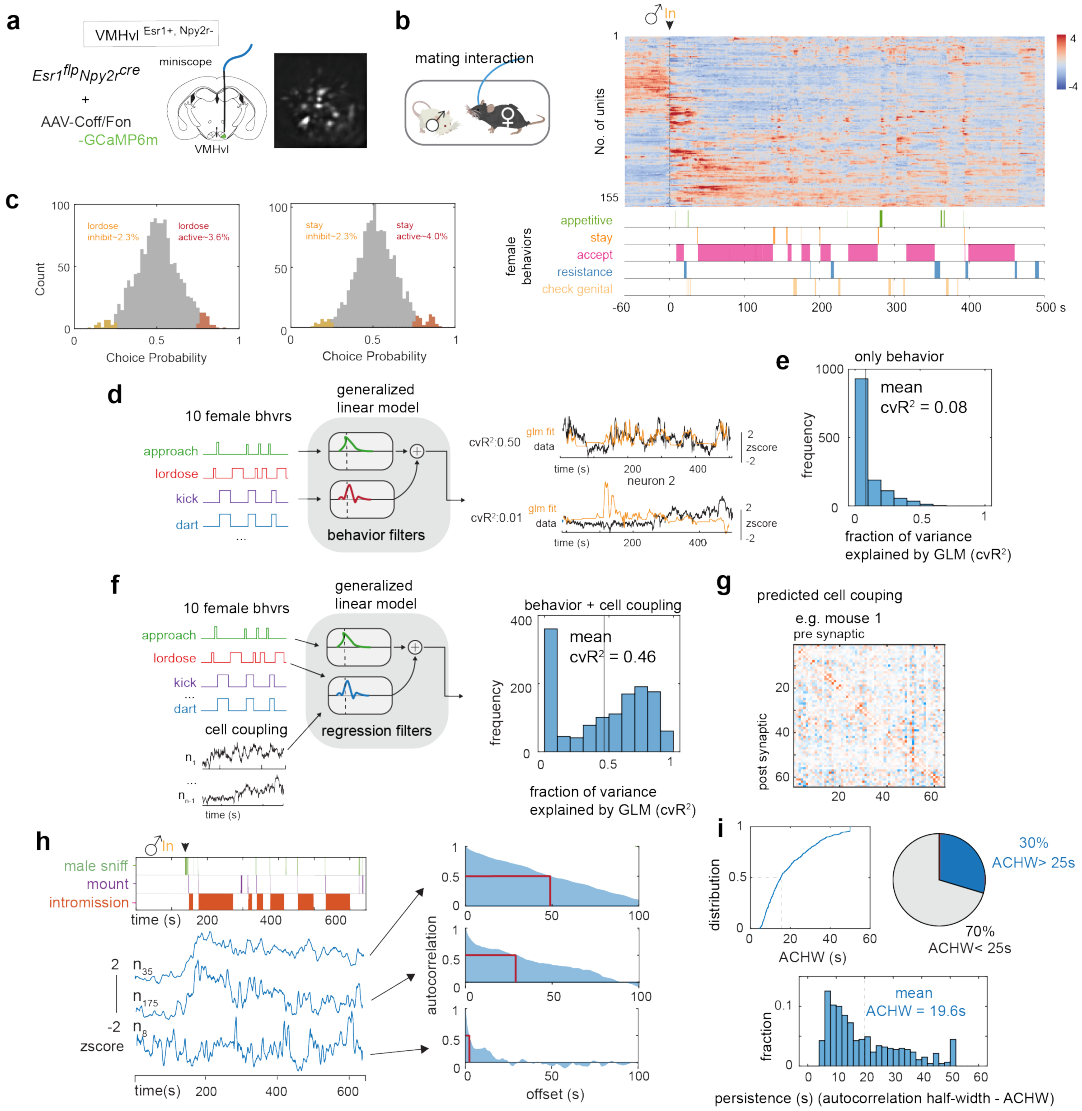


Figure 2 | Tuning properties of female VMHvl neurons during mating.

a, Left, schematic illustrating miniscope imaging of female VMHvlEsr1+,Npy2r- (α) cells; Right, an example imaging plane. b, Diagram illustrating mating interaction test. Single cell responses during mating interaction (top) and their corresponding behaviors (bottom), from one example female. Units were sorted by temporal correlation. Color scale indicates z-scored activity. c, Choice Probability (CP) histograms and percentages of tuned cells. cutoff: CP>0.7 or <0.3 and >2σ. (N = 15 mice). d, Left, schematic showing the generalized linear model (GLM) used to predict neural activity from behavior; Right, example fit of selected neurons with cvR2 (0.50 and 0.01). e,

Distribution of cvR2 across all mice for GLMs trained using only behavior (left, N = 15 mice, mean: 0.08) and f, using behavior with cell coupling (right, N = 15 mice, mean: 0.46). g, Predicted cell coupling (relative strength of connectivity) between neurons in one example mouse. See also Extended Data Fig. 3. h, Left, example VMHvl neurons in female showing a range of persistent activity (z-scored ΔF/F); Right, auto-correlation half-width (ACHW) as a measure of persistent activity, for example units shown. i, Left, cumulative distribution of ACHW for all units; Right, distribution of the number of neurons with ACHW > 25s (N = 4 mice).

Neural dynamics in female VMHvl

We first examined the dynamics of single neuron activity by measuring the half-width of each neuron's autocorrelation function[23] (ACHW), an approximation of the neuron's time constant[26,27] (Figure 2h). This analysis identified individual cells that exhibited apparent persistent activity across the mating interaction. Thirty percent of cells displayed ACHWs longer than 25 seconds (mean ACHW: 20s, Figure 2i), a duration longer than the mean copulation IBI (13.7s, Figure 1e, *right*). Notably, the ACHW of the same female cells was significantly lower when the male was confined in a perforated enclosure (pencil cup), than during free mating interaction (mean ACHW for pencil cup: 14.3 ± 0.42s, mean ACHW for free interaction: 19.64 ± 0.58s, Extended Data Fig. 3a-c), suggesting that the latter cannot be fully explained by persistent male odors.

Given that the single cell analysis revealed evidence of persistent neural activity, we considered whether a systems-level approach could capture low-dimensional features of population neural dynamics. To this end we fit an unsupervised dynamical systems model (recurrent Switching Linear Dynamical Systems, rSLDS[20,22]) directly to neural data during individual trials (Figure 3a, d, e; N=15 mice, mean variance explained (calculated as cvR^2 between observed and predicted neural trajectories: 64.08 ± 6.8%).

Applying rSLDS analysis to VMHvl α-cell activity revealed an "integration dimension" n state S2 with a relatively long time constant, in comparison to the other dimensions

(110.7±13.6s; Figure 3b, red dot, 3c, N = 15 mice). Examining the $\log_2$ ratio of the two longest time constants to calculate a so-called "line attractor score"[22] revealed that the fit dynamical system contains a line attractor (see Figure 3e), which is aligned in neural state-space to the integration dimension. The integration dimension could also be uncovered using supervised targeted dimensionality reduction (Extended Data Fig. 3d, e), confirming that slow integration dynamics is indeed a property of a subset of VMHvl neurons and not dependent on the method used.

Projecting neural activity into the integration dimension revealed ramping activity that began to increase at the onset of sniffing, mounting, or intromission (depending on the trial), and which continued to increase across multiple mating bouts and IBIs (Figure 3d and Extended Data Fig. 3f-q). The continuous "ramping up" in the integration dimension observed over a long-time scale during mating is unexpected. It contrasts with studies of bulk calcium activity in female VMHvl^{Cckar} neurons, which showed a unidirectional decrease from the start of mounting until ejaculation[6].

To quantitatively reveal the variable integrated by the integration dimension, we modeled the dimension as a neural integrator using a single-state linear dynamical system, allowing the model to flexibly use male behaviors (male-sniff, mount and intromission) to move activity along the integrator (Extended Data Fig. 3r). We find that the model fits with high accuracy (cvR^2: 0.88 ± 0.02) and possess a large intrinsic time constant (>200s), revealing that it does indeed function as an integrator (Extended Data Fig. 3s). To dissect how the different inputs contribute to the model, we obtained the transformed input from the model (Extended Data Fig. 3s) and found that it peaks following every male contact (Extended Data Fig. 3t, bottom). Thus, male-contact driven input, in combination with sustained input during male-engagement behaviors such as intromission, is used to integrate activity over time (Extended Data Fig. 3t, bottom, 3u).

Analysis of the traces of individual neurons contributing to the integration dimension indicated that some single cells exhibited ramping-like activity (Extended Data Fig. 4a,b, 56% of neurons r^2 >0.5), but that different cells peaked at different times during the mating interaction (Extended Data Fig. 4c, d, orange arrowheads). This suggests that the robust ramping activity seen in the integration dimension (Extended Data Fig. 4c, upper) is a property of the population and not solely a collective property of all single neurons.

To visualize the flow field of the rSLDS-fit dynamical system we projected it into a 2-dimensional state space using PCA (Figure 3e). This projection revealed a stable region (white area) comprising a linear array of "slow points" that approximated a line attractor, which is primarily contributed by the integration dimension of the model (Figure 3d-g). Mapping annotated behaviors onto the neural trajectory in this state space indicated that the population vector entered the line attractor following initial close contact with the male and progressed along it during successive male intromission bouts (Figure 3f, g). This progression reflects the ramping seen in the integration dimension discovered by rSLDS (Figure 3d and Extended Data Fig. 3f-q). The pattern of fixed points discovered by rSLDS could also be uncovered by independently fitting recurrent neural networks (RNNs) to neural data using FORCE[28,29], revealing that the putative line attractor is a feature of neural data and not an artifact of the rSLDS method (Extended Data Fig. 5a-b). Notably, in some animals the neural vector exhibited brief, loop-like excursions orthogonal to the attractor dimension during IBIs (Figure 3g; Extended Data Fig. 3i, m, q), suggesting "attractiveness" of the observed fixed points against either natural perturbations orthogonal to the attractor, or noise.

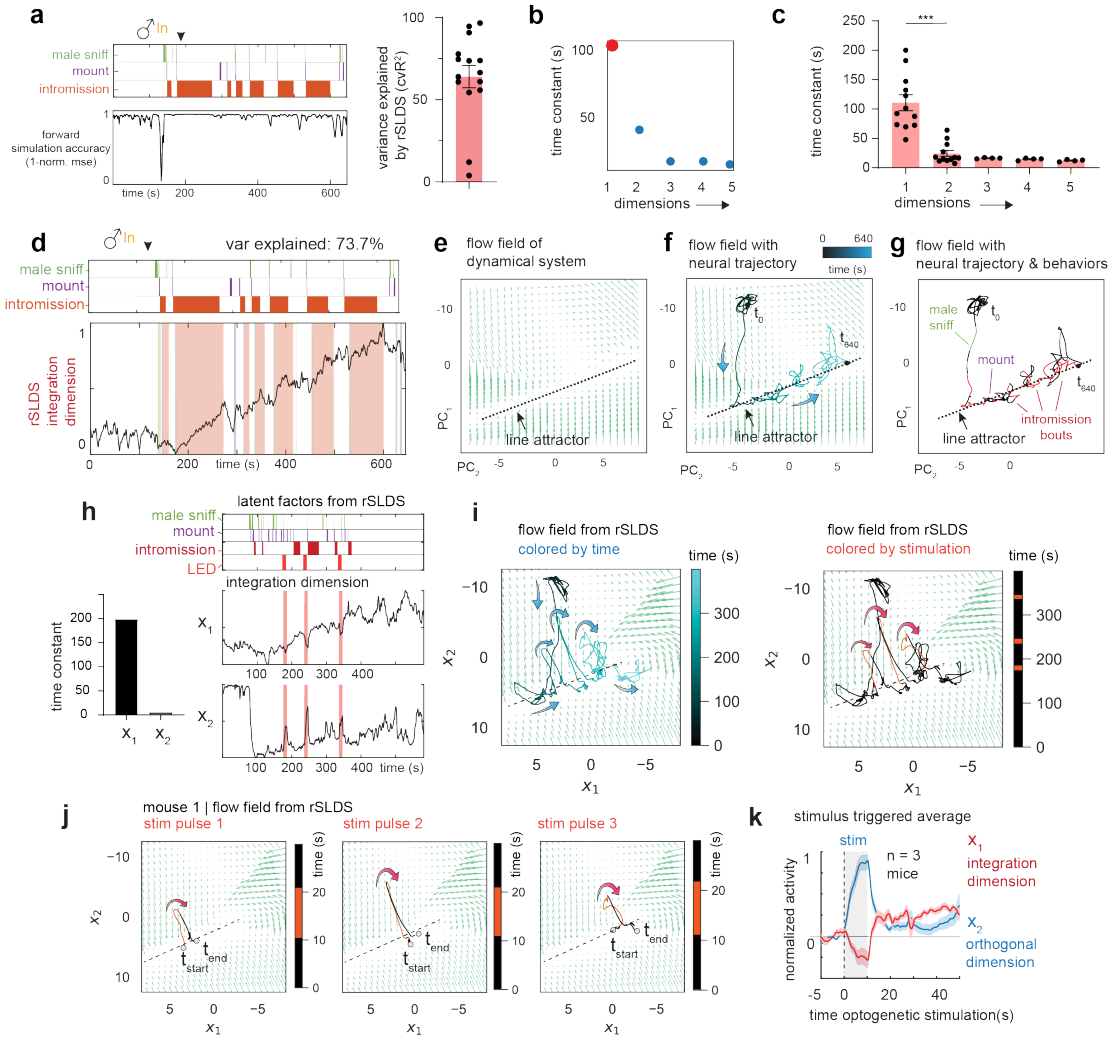


Figure 3 | An approximate line attractor in female VMHvl during mating.

a, Left, rSLDS model performance measured by forward simulation accuracy (calculated as [1-normalized mean squared error, mse])21 in an example mouse. Right, variance explained by rSLDS model fit without an input term (see Methods) for all mice (N = 15 mice, mean = 64.08%). The variance explained by the two outliers can be increased by incorporating an input term. b, Time constants reveal a single dimension with a large time constant. c, Distribution of time constants across animals fit by the rSLDS. Time constants sorted by magnitude in each animal (****p<0.0001, N = 15 mice, mean time constant of dim1: 110.7 ± 13.6, dim2: 24.5 ± 5.1, p value = 6.5e-05). d, Dynamics of the integration dimension reveals a ramping dimension, aligned to male mating behaviors in an example trial. e, Flow field of VMHvl α dynamical system colored by

rSLDS states. f, Flow field of VMHvl α dynamical system showing neural trajectories in state space. g, Neural state space of VMHvl α dynamical system highlighting regions where fixed points are present (dash line). h: Left: Time constants of latent factors from rSLDS model. Right: Projection of rSLDS latent factor activity from rSLDS model trained on neural data from unperturbed periods (i.e., excluding LED stimulation and 20s post-stimulation interval). i: Left: Flow field and neural trajectories from rSLDS model colored by time. Right: Neural trajectory colored by stimulation periods. j: Flow field and neural trajectories for each of the three stimulation periods for mouse 1. Note that trajectories are pushed away from the attractor during stimulation and then return to line attractor following stimulation offset, as predicted by the flow field. This approach also tests the validity of the extrapolated regions of the flow field uncovered by rSLDS. k: Stimulus triggered average of response in integration dimension (x1) and orthogonal dimension (x2) upon optogenetic stimulation.

Perturbations of the female line attractor

Definite evidence for the attractive nature of the fixed points discovered by rSLDS requires performing neural perturbations orthogonal to the line attractor. Such perturbations for line attractors have yet to be performed for freely behaving animals[30]. To achieve this, we performed optogenetic inhibition of the VMHvl network while concurrently imaging $VMHvl^{Esr1}$ α neurons (Extended Data Fig. 5c-d). We found that transient optogenetic inhibition creates consistent transient off-manifold responses in neural state space during the period of photostimulation, with the neural trajectory returning to the nearest fixed point along the line attractor post inhibition (Figure 3h-k, Extended Data Fig. 5e-j). Using forward simulations of the model fit to data from the unperturbed period (excluding data during and 20s after stimulation), we find that the dynamical system model is able to predict neural trajectories in the held-out post-stimulus period, revealing the predictive nature of the flow field (Extended Data Fig. 5f-h). Moreover, by providing this inhibition at different positions along the line attractor, we show that different fixed points along the line attractor revealed by rSLDS are indeed attractive (Figure 3i-j, Extended Data Fig. 5i-j).

The presence of a line attractor suggested a mechanism to stably maintain population activity in a particular state during interruptions or pauses in male mating behavior. To test this hypothesis, we first examined activity during copulation inter-bout intervals, when the male is physically separated from the female. Notably, we found that the average value of the integration dimension during copulation IBIs was relatively high, similar to and statistically indistinguishable from that measured during the preceding copulation bout (Figure 4a). Accordingly, it was not possible, using activity in this dimension, to train a decoder to distinguish videoframes containing copulation bouts vs. IBIs with accuracy greater than chance (Figure 4b).

To further probe the stability of the identified line attractor in female VMHvl, we next carried out behavioral perturbation experiments to non-invasively and transiently interrupt male mating (Figure 4c). After several successful intromission bouts had been performed, we remotely abrogated copulation by optogenetic activation of VMHdm^{Sf1+} cells in males, which promoted an abrupt defensive state[31,32]. During the laser-on period, males stopped all mating behaviors, including singing, and displayed no active approach to the female. The induced mating pauses lasted for several minutes (1-5 mins), which were much longer than the natural mating pauses (Figure 1f; average IBI = 13.7s). Nevertheless, activity in the integration dimension in the female brain remained elevated for minutes while the male was prevented from resuming mating (Figure 4d), consistent with the persistent activity we observed during the natural male copulation pauses (Figure 4a).

Together, these data indicated that VMHvl α cell activity displays line attractor dynamics during mating, and that while male contact is integrated along the line attractor (Extended Data Fig. 3r-u), the stability of the system in the integration dimension does not require continuous male contact-dependent sensory input. In further support of this conclusion, in a cohort of naturally cycling females exhibiting variable receptivity (see below), we obtained some trials with high male intromission rates but low female receptivity behavior (Figure 4e, colored dots, Videos). Notably, analysis of those trials revealed relatively little if any ramping in the integration dimension (Figure 4f). These data suggest that ramping

does not simply reflect accumulated mechanosensory inputs derived from male intromission.

Lastly, we sought to identify a correlate of the ramping activity revealed by line attractor dynamics in VMHvl (Figure 4j). Males display sequential mating behaviors with escalating intensity from sniffing to mounting, intromission and finally ejaculation, reflecting an escalating internal state of sexual arousal. We examined activity in the integration dimension during ejaculation and observed that it peaked just prior to ejaculation, and immediately dropped thereafter (Figure 4j, h), although this drop is characteristic of bulk calcium activity as well[6].

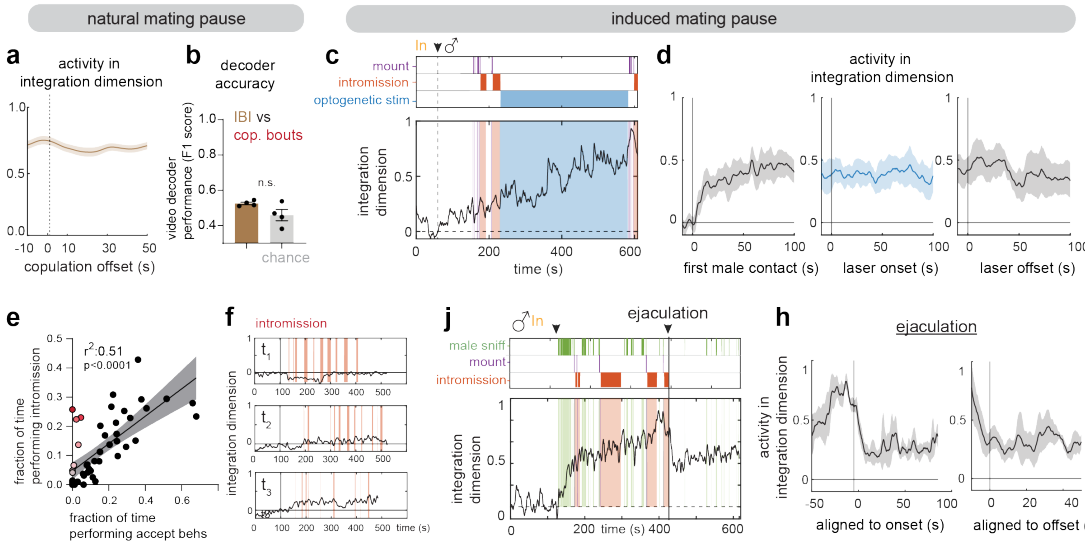


Figure 4 | Female line attractor encoded a persistent and ramping state during mating.
a, Behavior triggered average of the normalized activity of the integration dimension aligned to the offset of male copulation. Data presented as mean ± SEM. b, Videoframe behavioral decoder performance trained on neural data from copulation bouts vs. inter-bout intervals (N=4 mice, p = 0.2, Mann-Whitney U test, mean value of data:0.52 ± 0.007, shuffle: 0.49 ± 0.03). c, Dynamics of the integration dimension in an example female combined with optogenetic inhibition of mating behaviors in the interacting male. d, Behavior triggered average of the normalized activity of the

integration dimension aligned to first male contact (left), optogenetic mating inhibition onset (middle) and inhibition offset (right) (N = 4 mice). Data presented as mean ± SEM. e, Scatter plots of time spent performing intromission and time spent performing accept behaviors to identify trials with high intromission and low receptivity (colored dots). Data presented as mean ± SEM. ****p<0.0001, Linear regression: R2=0.51, p value = 7.09e-12. f, Example traces of the integration dimension for trials with intromission but low receptivity (identified from e). g, Dynamics of the integration dimension, aligned to male mating behaviors in an example trial with ejaculation. h, Behavior triggered average of the normalized activity of the integration dimension aligned to the ejaculation onset and offset (N=4 mice), data presented as mean ± SEM.

Attractor encodes sexual receptivity

We considered whether the line attractor observed during mating reflects or encodes the level of female receptivity by altering receptivity in two different paradigms. First, we performed longitudinal imaging in multiple females (N = 4 mice) across their 4-5 day estrus cycle, during which receptivity changes (Extended Data Fig. 6a). In each animal, we were able to obtain data from 1 sexually receptive day and 2 unreceptive days and to align neurons from those recordings across days (Figure 5a). Consistent with previous studies[5,6,17], no change in average VMHvl α cell population activity (triggered on male mounting attempt) was apparent on receptive vs. unreceptive days (Extended Data Fig. 6b). However, raster plots revealed obvious differences in the pattern of single-unit activity on receptive vs. unreceptive days (Figure 4a, *right*).

To determine whether there were also differences in VMHvl α cell dynamics across the estrus cycle, we fit rSLDS models to data obtained on both receptive and unreceptive days, for each individual. Models fit to data from unreceptive days failed to identify a single dimension with a very long time constant, indicating the absence of a line attractor (Figure 5b, c). Accordingly, the first 2 PCs of rSLDS state-space did not exhibit integration-like activity, but rather relatively fast dynamics time-locked to male sniffing and mounting (PC2 in Figure 5d). In 2D flow-field projections, neural state space contained a single point

attractor, reflecting stable baseline activity prior to interaction with the male, from which the population vector made rapid excursions during sniffing and mounting, returning to the same point attractor after interaction (Figure 5e, f).

To compare neural dynamics on non-receptive vs. receptive days more directly, we projected neural activity from unreceptive days into the rSLDS model fit to data from the receptive day using the same neurons aligned across days (Figure 5g). The projected neural data failed to show ramping behavior in the 1st rSLDS dimension (Figure 5h, Extended Data 6c). Accordingly, in 2D projections of the flow-field the neural population activity vector remained at one end of the line attractor (Extended Data Fig. 6d). Importantly, although male mounting occurred on unreceptive days (Figure 5h, purple rasters), activity in the 1st rSLDS dimension was low during this behavior (Extended Data Fig. 6e), indicating that it is not sufficient to explain the ramping observed on receptive days.

These results suggested that a change in neural dynamics occurred between receptive and unreceptive days. This inference was supported by the lower ACHW of cells weighted on the 1st rSLDS dimension on unreceptive vs. receptive days (Extended Data Fig. 7a, b; distribution mean for ACHW on unreceptive days 16.1 ± 0.8s; on receptive days 25.2 ± 1.5s, $p<0.001$). This difference in mean ACHW was observed regardless of the order in which receptive and unreceptive days occurred in different mice (Extended Data Fig. 7c, f). Neurons that did not contribute to the 1st rSLDS dimension did not exhibit a change in ACHW (Extended Data Fig. 7d, g). Finally, we compared the ACHWs of each individual unit on receptive vs. non-receptive days. A scatterplot of these data revealed a subpopulation (39 ± 5%) of line attractor-weighted neurons whose ACHW was higher on receptive than on unreceptive days (Extended Data Fig. 7e, h, red datapoints). Indeed, incorporating these differences in ACHW into a mechanistic spiking network model can recapitulate our empirical results, exhibiting a loss of line attractor dynamics during unreceptive states (Extended Data Fig. 8d).

As a second independent test of the hypothesis that the line attractor encodes receptivity, we subjected a cohort of females to ovariectomy (OVX) to render them unreceptive, and performed longitudinal imaging in the OVX animals before vs. after hormone priming to restore receptivity (daily injection of estrogen (E) + progesterone (P) in oil; OVX/EP; controls injected with oil only). The results indicated that attractor dynamics disappeared following OVX and were reinstated following hormone-priming (Extended Data Fig. 9j). The females used in this cohort had also been imaged during their natural cycle and fit with rSLDS models. In some individuals, the model possessed a poor fit on receptive days (Extended Data Fig. 9k, "forward simulation accuracy"). Strikingly, in one such animal the fit of the rSLDS model was strikingly improved following OVX and hormone priming, compared to the fit obtained on her naturally receptive day (Extended Data Fig. 9k vs. l, m-n). Taken together, these data confirm a strong prediction of the hypothesis that the line attractor observed during mating encodes some aspect of mating receptivity.

The foregoing data left open the important question of whether the continuous low-dimensional variable instantiated by the line attractor reflects or encodes continuous variation in the degree of female receptivity. In males, differences in the time constant of the integration dimension are strongly correlated with differences in aggressiveness, across individual animals[19]. We therefore sought to examine line attractor parameters within a cohort of females exhibiting individual differences in receptivity across trials and days. To generate this cohort, we injected naturally cycling females with estrogen (but not progesterone) daily beginning 2 days before imaging and continued the injections during 3-7 days of repeated imaging of the same animals during daily mating tests (N = 6 mice). These injections increased the number of days on which females exhibited receptivity, while still allowing variation in the level of receptivity (as measured by the amount of accepting behaviors displayed in a given trial) in response to changing levels of endogenous sex hormones across the estrus cycle (Figure 5i, j and Extended Data Fig. 10a). This design afforded the opportunity to correlate quantitative variation in receptivity with variation in line attractor parameters.

We fit rSLDS models to imaging data from each animal and mating trial and plotted the average activity of the integration dimension over time (cf. Figure 5h, receptive day 3). Strikingly, the area under this curve (auc) was strongly and positively correlated with the percentage of time that females performed acceptance behaviors during each mating interaction (Figure 5k, l; r^2=0.62, p<0.0001, n = 50 trials). In contrast, other behaviors such as resistance and appetitive behaviors were not correlated with this measure (Extended Data Fig. 10b). Lastly, the population mean of neural activity did not show any correlation with the percentage of time that females performed acceptance behaviors, highlighting the value of rSLDS to identify physiologically distinct subpopulations whose activity is quantitatively correlated with receptivity during male mating (Extended Data Fig. 10c-d). Thus, these data indicate that variation in movement along the line attractor reflects variation in levels of sexual receptivity, across individuals and trials.

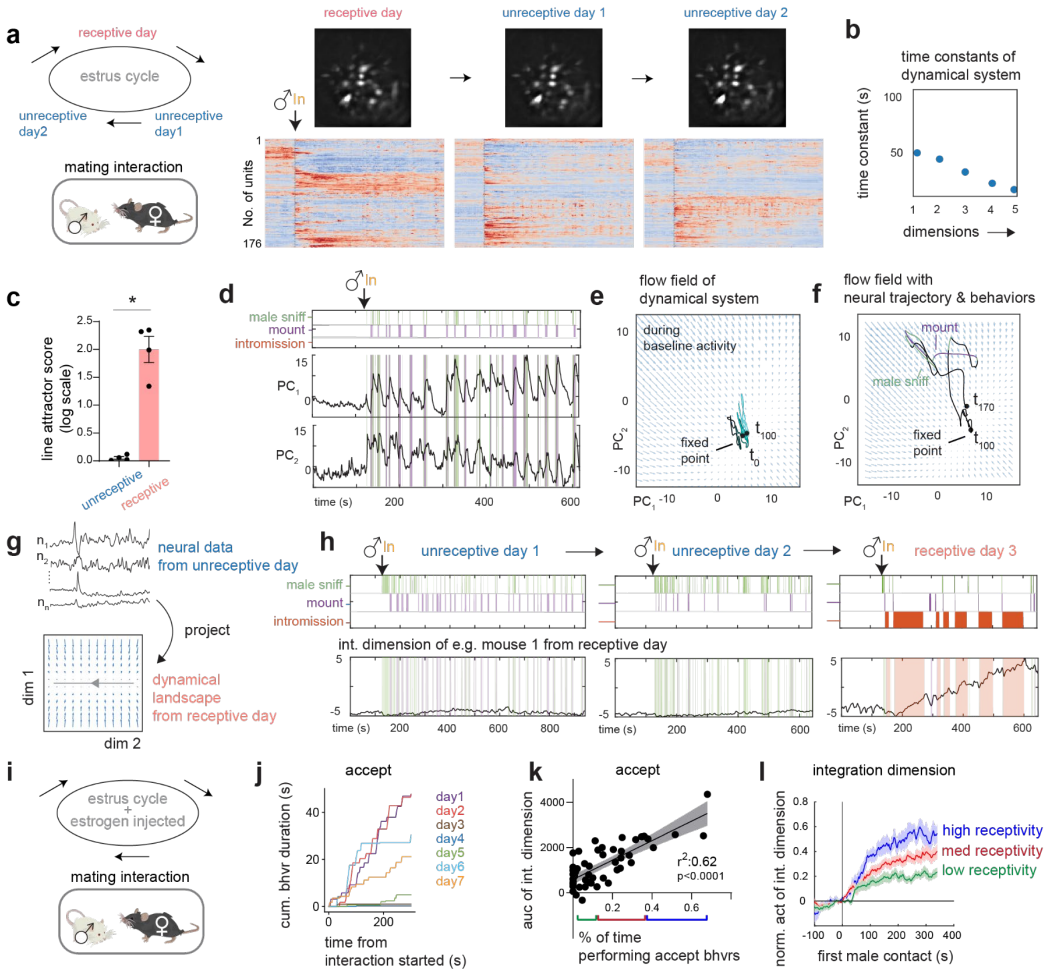


Figure 5 | Female line attractor dynamics encoded sexual receptivity across days.

a, Left, illustration for longitudinal imaging strategy across estrus states of naturally cycling females; Right, example longitudinal imaging planes and traces from one female. Units were sorted by temporal correlation. Color scale indicates z-scored activity. b, Time constants of VMHvl α dynamical system on one unreceptive day. c, Line attractor scores for dynamical systems fit during receptive and unreceptive days (N = 4 mice, mean ± sem - unreceptive: 0.05 ± 0.02, receptive: 1.9 ± 0.2, *p<0.05, Mann Whitney U test, p value: 0.02). d, Low dimensional principal components of VMHvl α rSLDS fit model on unreceptive day. Principal components show fast time locked dynamics and lack ramping and persistence. e, Flow field of VMHvl α dynamical system on unreceptive day. f, Same as e, showing neural trajectories in state space colored by time and behaviors. g, Schematic illustrating the projection of neural activity from an unreceptive day into fit dynamical system from a receptive day. h, Dynamics of integration dimension in VMHvl

discovered during a receptive day (same example trial as shown in Fig. 3d) compared to activity of the same dimension on unreceptive days. I, Illustration for longitudinal imaging strategy across estrus states of naturally cycling females with estrogen injection. j, Accepting behaviors displayed in mating interactions across days from one example female. k, The scatter plots of the integration dimension values and the amount of female accepting behaviors (****$p<0.0001$, linear regression, $R2=0.62$: , $n = 50$ trials) in each trials. Data presented as mean ± SEM. l, The integration dimension activity aligned to the first male contact, in high, medium or low receptivity trials, defined in k. ****$p<0.0001$, Wilcoxon matched-pairs signed rank test. Data presented as mean ± SEM.

Discussion

Using unsupervised analysis of neural data, we have discovered an approximate line attractor in a genetically defined subset of VMHvl^{Esr1} neurons that causally controls female mating receptivity. Activity in the attractor scales with individual differences in receptivity and is estrus cycle-dependent. To our knowledge, there is no prior example of a line attractor that appears and disappears with periodic changes in behavioral/hormonal state on a time scale of days.

Line attractor dynamics can afford internal states two important features: stability (persistence) and ramping (scalability). The stability of the line attractor across intromission bouts may function to maintain the female in a persistently aroused state during intermittent male copulatory behavior, facilitating its re-initiation following pauses. This interpretation is supported by our observation that female social behavior persists following natural interruptions in copulation (Figure 1f-h). The ramping activity seen during copulation may represent a continuous, scalable variable in the female brain. A reasonable interpretation is that this variable encodes the level of escalating female sexual arousal. We emphasize that this ramp-up was not visible in the bulk α cell calcium signal, but only in the integration dimension. This may explain why it was not reported in a study of mating-related VMHvl^{Cckar} neuronal activity using fiber photometry[6].

The idea that the line attractor encodes mating receptivity is supported by its presence or absence during receptive vs. unreceptive estrus cycle days or in ovariectomized females with vs. without hormone priming, respectively (Figure 5c). Importantly, however, it is not just a binary correlate of receptivity: the degree of movement along the attractor was highly correlated (r^2=0.62) with the level of receptivity as measured by the frequency of accepting behaviors (Figure 5k and Extended Data Fig. 10b). In contrast, the integration dimension was not well-correlated with other female behaviors.

Our previous work has shown that transcriptionally distinct subsets of VMHvl^{Esr1} neurons, called α and β cells, control female sexual receptivity and maternal aggression, respectively[5]. Here we show that the α cell population exhibits further heterogeneity at the physiological level, including subsets that contribute to the line attractor or to orthogonal dimensions. Whether these subpopulations are transcriptomically distinct is not yet clear[6,18]. Yin et al., reported that VMHvl^{Cckar} neurons (a subset of α cells) displayed receptivity-associated changes in their spontaneous activity, and responsivity to males during investigation[6]. These cells may contribute to the integration dimension neurons identified here.

Together, our data suggest that neural population dynamics represent the dynamics of female mating receptivity and can be reversibly sculpted by physiological state. They also generalize the concept that line attractors with slow dynamics encode internal states underlying innate social behaviors[21]. Because the molecular, cellular and connectional features of VMHvl are well-described[5,15,16,35–37], this system may be advantageous for understanding how hormones, genes, cell types and local circuitry contribute to emergent neural population dynamics.

Methods

Mice

All experimental procedures involving the use of live mice or their tissues were carried out in accordance with NIH guidelines and approved by the Institute Animal Care and Use Committee (IACUC) and the Institute Biosafety Committee (IBC) at the California Institute of Technology (Caltech). All mice in this study, including wild-type and transgenic mice, were bred at Caltech or purchased from Charles River Laboratory. Group housed C57BL/6N female or singly housed male mice (2-5 months) were used as experimental mice. *Npy2r*^{cre} mice (Jackson Laboratory stock no. 029285) (=N1), *Esr1*^{cre} mice, *Esr1*^{flp} mice (>N10), *Sf1*^{cre} mice (Jackson Laboratory stock no. 012462) were backcrossed into the C57BL/6N background and bred at Caltech. Heterozygous *Npy2r*^{cre}, *Esr1*^{cre} or double heterozygote *Esr1*$^{flp/+}$*Npy2r*$^{cre/+}$ mice were used for cell-specific targeting experiments and were genotyped by PCR analysis using genomic DNA from tail tissue. All mice were housed in ventilated micro-isolator cages in a temperature-controlled environment (median temperature 23°C, humidity 60%), under a reversed 11-h dark–13-h light cycle, with ad libitum access to food and water. Mouse cages were changed weekly.

Surgeries

Surgeries were performed on female Esr1$^{Flp/+}$Npy2r $^{Cre/+}$ females aged 2 months. Virus injection and implantation were performed as described previously[34,38]. Briefly, animals were anaesthetized with isoflurane (5% for induction and 1.5% for maintenance) and placed on a stereotaxic frame (David Kopf Instruments). Virus was injected into the target area using a pulled-glass capillary (World Precision Instruments) and a pressure injector (Micro4 controller, World Precision Instruments), at a flow rate of 20 nl min^{-1}. The glass capillary was left in place for 10 min following injection before withdrawal. Lenses were slowly lowered into the brain and fixed to the skull with dental cement (Metabond, Parkell).

Females were co-housed with a vasectomized male mouse after virus injection and lens implantation. Four weeks after lens implantation, mice were head-fixed on a running wheel and a miniaturized micro-endoscope (nVista, nVoke, Inscopix) was lowered over the implanted lens until GcaMP-expressing fluorescent neurons were in focus. Once GcaMP-expressing neurons were detected, a permanent baseplate was attached to the skull with dental cement. The co-housed vasectomized males were removed.

Virus injection and GRIN lens implantation

The following AAVs were used in this study, with injection titers as indicated. Viruses with a high original titer were diluted with clean PBS on the day of use. AAV-DJ-EF1a-Coff/Fon-GcaMP6m ($4.5e^{12}$, Addgene plasmid) was packaged at the HHMI Janelia Research Campus virus facility. “Coff/Fon” indicates Cre-OFF/FLP-ON virus. Stereotaxic injection coordinates were based on the Paxinos and Franklin atlas. Virus injection: VMHvl, AP: −1.6, ML: ±0.78, DV: −5.7; GRIN lens implantation: VMHvl: AP: −1.6, ML: −0.76, DV: −5.55 (⌀0.6 × 7.3 mm GRIN lens).

Vaginal cytology

To determine the estrus phases of tested females, vaginal smear cytology was applied on the same day as the behavior test. A vaginal smear was collected immediately after the behavioral test and stained with 0.1% crystal violet solution for 1 minute. Cell types in the stained vaginal smear were checked microscopically. In this study, the proestrus phase was characterized by many nucleated epithelial, some cornified epithelial and no leukocytes.

Hormone priming

Female mice were ovariectomized and estrus was induced by hormone priming. Estradiol benzoate (E2) and progesterone (PG) powder was dissolved in sesame oil. For primed females, 50ul 200ug/ml E2 was delivered subcutaneously on day-2 and day-1 at 3pm.

10mg/ml PG was delivered subcutaneously on the day of test at 10am. Behavior test was performed 4-6 hours after PG injection. For unprimed female, sesame oil was injected at the same time points as hormone injections. Vaginal smear cytology was applied on the same day as the behavior test to make sure that the females were completely primed or unprimed. For estrogen-injected females used in Fig. 3 and Fig. 4, 50ul 200ug/ml E2 was delivered subcutaneously everyday at 10am. Behavior tests were conducted after the first two days of injection.

Sex representation assay

All behavior tests were performed under red light. Group housed C57BL/6N male and female mice (2-4 months) were used for the test. Tested female was acclimated in her home cage under the recording setup[39] for 10 minutes. A toy, a female or a male was introduced to the tested female with a 90-second interval. Each interaction lasted for 1 minute before transitioning into the consummatory phase. The sequential representations were repeated for 3 times.

Mating assay

Singly housed sexually experienced C57BL/6N male were used for mating assay. Male mice used for test were initially co-housed with a female mouse for at least 1 week and singly housed at least 1 week before test. On the day of test. Male mouse was acclimated in his home cage under the recording setup. A random female mouse was placed into male cage until three male mounting bouts were observed. The tested female mice were acclimated in a new cage for 10 minutes before being introduced into the male cage. The male contact mating interaction lasted for 5-15 mins. At the end of the free interaction, a pencil cup was introduced to restrain the male. Then the imaging and behavior recording during the non-contact period continued for 3-5 mins.

Wireless optogenetic male mating inhibition assay

Singly housed sexually experienced Sf1-Cre$^{+/-}$ males were used in this test. All hardware and wireless devices for optogenetic stimulation were sourced from NeuroLux (Urbana, IL). Specifically, AAV2-EF1a-DIO-hChR2(H134R)-EYFP-WPRE-pA (4.2e^{12}, UNC vector core) was unilaterally injected into ventromedial hypothalamus (VMHdm) of the male mice at coordinates: AP: −1.5, ML: +0.4, DV: −5.6. Simultaneously, wireless optogenetic devices were implanted. A recovery period of three weeks followed the surgical procedures to allow for optimal viral vector expression and ensure the animals' well-being. Subsequently, a mating assay was performed, and when multiple successful copulations were observed, male mice were exposed to a wirelessly powered blue light photostimulation (473 nm, 1-5 minutes, 20Hz, 10W). During the stimulation, male mice promptly discontinued all mating-related behaviors, including vocalization, sniffing, mounting or intromission, instead exhibiting exploratory behaviors within the homecage and distancing themselves from the female mouse. Following the cessation of photostimulation, male mice typically resumed mating-related behaviors, either immediately or with a delay.

Behavior annotations

Behavior videos were manually annotated using a custom MATLAB-based behavior annotation interface. A 'baseline' period of 2 min when the animal was alone in its cage was recorded at the start of every recording session.

During female-male interaction, we manually annotated the following male mating behaviors: male sniff, mount, intromission and ejaculation.

For the same video, we annotated the following female mating behaviors: approach, sniff, lordose, wiggle, stay, dart, top up, kick, turn, check genital. 'Approach': Female faced male and walked to it without pausing. 'Sniff': Female actively sniffed male. 'Lordose': Female abdomen was on the ground and motionless or showing an arched back posture responding

to male mounting or intromission. ‘Wiggle’: Female continuously moving her head or body responding to male mounting or intromission. ‘Stay’: Female quietly stayed in place, but abdomen was not clearly on the ground, responding to male mounting or intromission. ‘Dart’: Female quickly ran away from male, responding to male mating behaviors. ‘Top up’: Female stood up to conceal the anogenital area, responding to male mating behaviors. ‘Kick’: Female kicked the male, responding to male mating behaviors. ‘Turn’: Female turned away from male, responding to mating mounting or intromission. ‘Check genital’: Female examined her genital area, usually after male mounting or intromission.

‘Lordose’, ‘wiggle’, ‘stay’, dart’, ‘top up’, ‘kick’ and ‘turn’ were grouped as responsive mating behaviors. ‘Approach’, ‘sniff’ and ‘check genital’ were grouped as female self-initiated mating behaviors.

‘Approach’ and ‘sniff’ were grouped as appetitive mating behaviors. ‘Lordose’ and ‘wiggle’ were grouped as accepting mating behaviors. ‘Dart’, ‘top up’, ‘kick’ and ‘turn’ were grouped as resistance mating behaviors.

All appetitive, accepting and resistance behaviors were grouped as social behaviors.

For the same video, we also annotated the following female non-social disengaged behaviors: rear, dig, chew. ‘Rear’: Female extended her body upright and attempted to explore outside the testing chamber. ‘Dig’: Female dig beddings. ‘Chew’: Female stood up and chewed with her mouth.

Fiber photometry

The fiber photometry setup was as previously described in earlier research with minor modifications. We used 470 nm LEDs (M470F3, Thorlabs, filtered with 470-10 nm bandpass filters FB470-10, Thorlabs) for fluorophore excitation, and 405 nm LEDs for isosbestic excitation (M405FP1, Thorlabs, filtered with 410–10 nm bandpass filters FB410-10, Thorlabs). LEDs were modulated at 208 Hz (470 nm) and 333 Hz (405 nm) and controlled by a real-time processor (RZ5P, Tucker David Technologies) via an LED driver

(DC4104, Thorlabs). The emission signal from the 470 nm excitation was normalized to the emission signal from the isosbestic excitation (405 nm), to control for motion artefacts, photobleaching and levels of GcaMP6m expression. LEDs were coupled to a 425 nm longpass dichroic mirror (Thorlabs, DMLP425R) via fiber optic patch cables (diameter 400 mm, N.A., 0.48; Doric lenses). Emitted light was collected via the patch cable, coupled to a 490 nm longpass dichroic mirror (DMLP490R, Thorlabs), filtered (FF01-542/27-25, Sem- rock), collimated through a focusing lens (F671SMA-405, Thorlabs) and detected by the photodetectors (Model 2151, Newport). Recordings were acquired using Synapse software (Tucker Davis Technologies). On the test day, after at least 5 minutes of acclimation under the recording setup, the female was first recorded for 5 minutes to establish a baseline. Then behavior assays were proceeded and fluorescence were recorded for the indicated period of time, as described in the text. All data analyses were performed in Python. Behavioral video files and fiber photometry data were time-locked. Fn was calculated using normalized (405 nm) fluorescence signals from 470 nm excitation. Fn(t) = 100 x [F470(t) – F405fit(t)] / F405fit(t). For the peri-event time histogram (PETH), the baseline value F_0 and standard deviation SD_0. was calculated using a -5 to -3 second window. Overlapping behavioral bouts within this time window were excluded from the analysis. Then PETH was calculated by [(Fn(t) – F_0]/ SD_0.

Micro-endoscopic imaging data Acquisition

Imaging data were acquired by nVista 3.0 (Inscopix) at 30 Hz with 2× spatial downsampling; light-emitting diode power (0.2–0.5) and gain (6–8×) were adjusted depending on the brightness of GcaMP expression as determined by the image histogram according to the user manual. A transistor–transistor logic (TTL) pulse from the Sync port of the data acquisition box (DAQ, Inscopix) was used for synchronous triggering of StreamPix7 (Norpix) for video recording.

For perturbation-imaging experiments, AAV5-hSyn-eNpHR3-mCherry (Addgene) was expressed together with GcaMP in VMHvl. Imaging data were acquired by nVoke 2.0

(Inscopix). One to three bouts of inhibitory LED stimulation (5 mW, continuous, 10 s) were performed during receptive mating trials.

Micro-endoscopic data extraction and preprocessing.

Miniscope data were acquired at 30 Hz using the Inscopix Data Acquisition Software as 2× down sampled .isxd files. Preprocessing and motion correction were performed using Inscopix Data Processing Software. Briefly, raw imaging data from three recording dates were concatenated. 2× spatially down sampled, motion corrected and temporally down sampled to 10 Hz. Further and exported as a .tiff image stack. A spatial band-pass filter was then applied to remove out-of-focus background. After preprocessing, calcium traces were extracted and deconvolved using the CNMF-E large data pipeline with the following parameters: patch_dims = [32, 32], gSig = 3, gSiz = 13, ring_radius = 19, min_corr = 0.75, min_pnr = 8. The spatial and temporal components of every extracted unit were carefully inspected manually (SNR, PNR, size, motion artefacts, decay kinetics and so on) and outliers (obvious deviations from the normal distribution) were discarded. The extracted traces were then z-scored before analysis.

Longitudinal imaging data extraction and preprocessing.

The females performed mating assay and were imaged for consecutive 3-7 days. ‘Receptive day’ was defined as female displayed accepting behaviors on the testing day, while ‘unreceptive day’ was defined as female did not display accepting behaviors on the testing day. Miniscope data from one receptive day and two unreceptive days were selected and concatenated to one .isxd file. Data was preprocessed and the traces were extracted as described in the last section. The three-day concatenated traces were z-scored, and then split to multiple traces for individual days.

Choice probability

Choice probability is a metric that estimates how well either of two different behaviors can be predicted/distinguished, based on the activity of any given neuron during these two behaviors. CP of single neurons was computed using previously described methods[38]. To compute the CP of a single neuron for any behavior pair, 1 s binned neuronal responses occurring during each of the two behaviors were used to generate a receiver operating characteristic curve. CP is defined as the area under the curve bounded between 0 and 1. A CP of 0.5 indicates that the activity of the neuron cannot distinguish between the two alternative behaviors. We defined a neuron as being capable of distinguishing between two behaviors if the CP of that neuron was >0.7 or <0.3 and was >2 σ or <−2 σ of the CP computed using shuffled data (repeated 1000 times).

Generalized linear model

To predict neural activity from behavior, we trained generalized linear models to predict the activity of each neurons k, as a weighted linear combination of 3 male behaviors: male sniffing, mounting and intromission as follows:

$$y_k(t) = \vec{x}(t)\vec{\beta} + \varphi$$

Here, $y_k(t)$ is the calcium activity of neuron k at time t, $\vec{x}(t)$ is a feature vector of 3 binary male behaviors at time lags ranging from t-D to t where D = 10s. $\vec{\beta}$ is a behavior-filter which described how a neuron integrates stimulus over a 10s period (example filters are shown in Extended Data Fig.2d-e). φ is an error term. The model was fit using 10-fold cross validation with ridge regularization and model performance is reported as cross-validated R^2 (cvR^2). To account for cell-cell interactions within the network, we also used the activity of simultaneously imaged neurons as regressors in addition to behavior as previously performed[24,25].

Dynamical system modelling

Recurrent-switching linear dynamical system (rSLDS) models[20,40] are fit to neural data as previously described[22]. Briefly, rSLDS is a generative state-space model that decomposes non-linear time series data into a set of linear dynamical systems, also called 'states'. The model describes three sets of variables: a set of discrete states (z), a set of latent factors (x) that captures the low-dimensional nature of neural activity, and the activity of recorded neurons (y). While the model can also allow for the incorporation of external inputs based on behavior features, such external inputs were not included in our first analysis.

The model Is formulated as follows: At each timepoint, there Is a discrete state $z_t \in \{1, \dots, K\}$ that depends recurrently on the continuous latent factors (x) as follows:

$$p(z_{t+1} \mid z_t = k, x_t) = \text{softmax}\{R_k x_t + r_k\} \quad (1)$$

where $R_k \in \mathbb{R}^{K \times K}$ and $r_k \in \mathbb{R}^K$ parameterizes a map from the previous discrete state and continuous state to a distribution over the next discrete states using a softmax link function. The discrete state z_t determines the linear dynamical system used to generate the latent factors at any time t:

$$x_t = A_{z_t} x_{t-1} + b_{z_t} + \epsilon_t \quad (2)$$

where $A_k \in \mathbb{R}^{d \times d}$ is a dynamics matrix and $b_k \in \mathbb{R}^D$ is a bias vector, where D is the dimensionality of the latent space and $\epsilon_t \sim N(0, Q_{z_t})$ is a Gaussian-distributed noise (aka innovation) term.

Lastly, we can recover the activity of recorded neurons by modelling activity as a linear noisy Gaussian observation $y_t \in \mathbb{R}^N$ where *N* is the number of recorded neurons:

$$y_t = Cx_t + d + \delta_t \quad (3)$$

For $C \in \mathbb{R}^{N \times D}$ and $\delta_t \sim N(0, S)$, a Gaussian noise term. Overall, the system parameters that rSLDS needs to learn consists of the state transition dynamics, library of linear dynamical system matrices and neuron-specific emission parameters, which we write as:

$$\theta = \{\{A_k b_k, Q_k, R_k, r_k\}_{k=1}^{K}, C, d, S\} \quad (4)$$

We evaluate model performance using both the evidence lower bound (ELBO) and the forward simulation accuracy (FSA) (Fig. 3a) described in Nair et al., 2023[22] as well as by calculating the variance explained by the model on data. Briefly, given observed neural activity in the reduced neural state space at time t, we predict the trajectory of population activity over an ensuing small time interval Δt using the fit rSLDS model, then compute the mean squared error (MSE) between that trajectory and the observed data at time t+ Δt. This MSE is computed across all dimensions of the reduced latent space and repeated for all times t across cross validation folds. This error metric is normalized to a 0-1 range in each animal across the whole recording to obtain a bounded measure of model performance. The FSA can intuition of where model performance drops during the recording. In addition to MSE, we also calculate the Pearson's correlation coefficient (R^2) between the predicted and observed data for each dimension following the forward simulation. By taking the average correlation coefficient across dimensions, we can obtain a quantitative estimate of variance explained by rSLDS on observed data.

The number of states and dimensions used for the model are determined using 5-fold cross validation. Visualization of the dynamical system using principal components analysis is performed as described previously [22].

For neural perturbation experiments, the rSLDS model was trained on data from unstimulated periods of time (i.e., excluding data during and 20s immediately after stimulation) and then tested on data from stimulated periods along with a 20s post-stimulus period (Extended Data Fig. 5e,f).

Code used to fit rSLDS on neural data is available in the SSM package: (https://github.com/lindermanlab/ssm)

Estimation of time constants

We estimated the time constant of each dimension of linear dynamical systems using eigenvalues λ_a of the dynamics matrix of that system, derived previously[41] as:

$$\tau_a = \left| \frac{1}{\log(|\lambda_a|)} \right|$$

Decoding of behavior using support vector machines

We trained frame-wise decoders to discriminate various pairs of behaviors as shown in Extended Data Fig. 4, from the population activity of all neurons during a mating interaction. We first created 'trials' from bouts of each behavior by merging all bouts that were separated by less than five seconds and balanced data to ensure chance performance of the model to be 50%. We then trained a linear SVM to identify a decoding threshold that maximally separates the two behaviors and tested the accuracy of the trained decoder on held-out frames. 'Shuffled' decoder data was generated by setting the decoding threshold on the same "trial", but with the behavior annotations randomly assigned to each behavior bout. We repeated shuffling 20 times. We report performances of actual and shuffled decoders as the average F1 score of the fit decoder, on data from all other "trials" for each

mouse. For significance testing, the mean accuracy of the decoder trained on shuffled data was computed across mice, with shuffling repeated 1000 times for each mouse.

Dynamical system modelling using FORCE learning.

We trained a recurrent neural network (RNN) to reproduce activity of individual neurons during mating interactions using FORCE as previously described[29,42]. The dynamics of each unit in the RNN is governed by the following equation:

$$\tau \frac{dx_i}{dt} = -x_i(t) + g(\sum_{j=1}^{N} J_{ij} r(x_j(t)) + H(t)$$

Here, τ is the time constant of the system (0.5s as used previously[42]), H is the total external input to neurons (consisting of a weighted combination of male-sniff, mounting and intromission), J is a heterogeneous matrix of recurrent connections whose strength is determined by the parameter g. For chaotic networks, we use g = 1.5 as used previously[29,42]. The elements of the matrix J are modified through recursive least squares as described in[29,43], by reducing an error term $e_i(t) = z_i(t) - f_i(t)$. Here $f_i(t)$ is the experimental calcium trace and $z_i(t) = \sum_j J_{ij} r_j(t)$. The network contains the same number of units as in experimental data (between 100-200 neurons) and dynamics were solved using Euler's method (dt = 0.05s).

To estimate the fixed points of the RNN, we reverse engineered the trained RNN with fixed point analysis[44] using gradient-based optimization. The estimated slow points of the dynamical system were then projected into the same neural state space as determined by rSLDS to determine the similarity in attractor landscapes discovered by the two methods (Extended Data Fig. 5a,b).

Modeling of integration dynamics to reveals inputs to line attractor

To reveal the input received by the integration dimension (Extended Data Fig. 3r-u), we modelled activity of this dimension using a single-state linear dynamical system model as:

$$x_t = A_z x_{t-1} + b_{z_t} + \epsilon_t + W u_t$$

here *x* refers to weighted activity of the integration dimension and *W* is a matrix to used linearly combine behavioral inputs (male-sniff, mounting and intromission) to the integration dimension. The model was fit as described above for rSLDS and model performance was evaluated using variance explained with cross-validation.

Mechanistic modelling of spiking neural networks

We constructed a model population of N = 1,000 standard current-based leaky integrate-and-fire neurons as previously performed[45]. We modelled an excitatory spiking network with feedback inhibition designed to account for finite size effects and runaway excitation. In this network, each neuron has membrane potential x_i characterized by dynamics:

$$\tau_m \frac{dx_i}{dt} = -x_i(t) + g\left(\sum_{j=1}^{N} W p_i(t) - g_{inh} I_{inh}(t)\right) + w_i s(t)$$

where $\tau_m = 20ms$ is the membrane time constant, *W* is the synaptic weight matrix, s is an input term representing external inputs and p represents recurrent inputs. To model spiking, we set a threshold (θ = 0.1), such that when the membrane potential $x_i(t) > \theta$, $x_i(t)$ is set to zero and the instantaneous spiking rate $r_i(t)$ is set to 1.

Inhibition was modelled as recurrent inhibition from a single graded input I_{inh} representing an inhibitory population that receives equal input from and provides equal input to, all excitatory units. The dynamics of I_{inh} evolves as:

$$\tau_I \frac{dI_{inh}}{dt} = -I_{inh}(t) + \frac{1}{N}\sum_{n=1}^{N} r_N(t),$$

where $\tau_I = 50ms$ is the decay time constant for inhibitory currents.

We designed the synaptic connectivity matrix to include a subnetwork of 200 neurons (20% of the network), designated the integration subnetwork as suggested by empirical measurements, with a connectivity density of 12% as opposed to 1% in the remaining network. Weights of the overall network were sampled from a uniform distribution:$W_{ij} \sim U(0,1/\sqrt{N})$, while weights of the subnetwork were sampled as $W_{ij} \sim U(0,1/\sqrt{N_p})$, where N_p = 200.

External input was provided to the network as a step function consisting of twenty pulses at 10 ISI. This stimulus drove a random 25% of neurons in each subnetwork.

Spiking-evoked input was modelled as a synaptic current with dynamics:

$$\tau_s \frac{dp_i}{dt} = -p_i(t) + r_i(t),$$

where τ_s is the synaptic conductance time constant, set to 20s for neurons in the integration subnetwork and 100ms for remaining neurons in the network.

Model dynamics were simulated in discrete time using Euler's method with a timestep of 1ms and a small gaussian noise term $\eta_i \sim N(0,1)/5$ was added at each time step. We used g = 2.5 and varied $g_{inh} = 4.25$ as suggested by measurements of inhibitory input to VMHvl[46] and used previously[45]. To simulate hypothesis 1 in Extended Data Fig. 12, we set the synaptic time constant for integration neurons to 100 ms. To simulate hypothesis 2, we changed the gain associated with input to each subnetwork, decreasing this quantity for

the integration subnetwork by 50% and increasing the same for the remaining neurons by 50%.

Calculation of auto-correlation half-width

We computed autocorrelation halfwidths by calculating the autocorrelation function for each neuron timeseries data (y_t) for a set of lags as described previously[23]. Briefly, for a time series (y_t), the autocorrelation for lag k is:

$$r_k = \frac{c_k}{c_0}$$

where c_k is defined as:

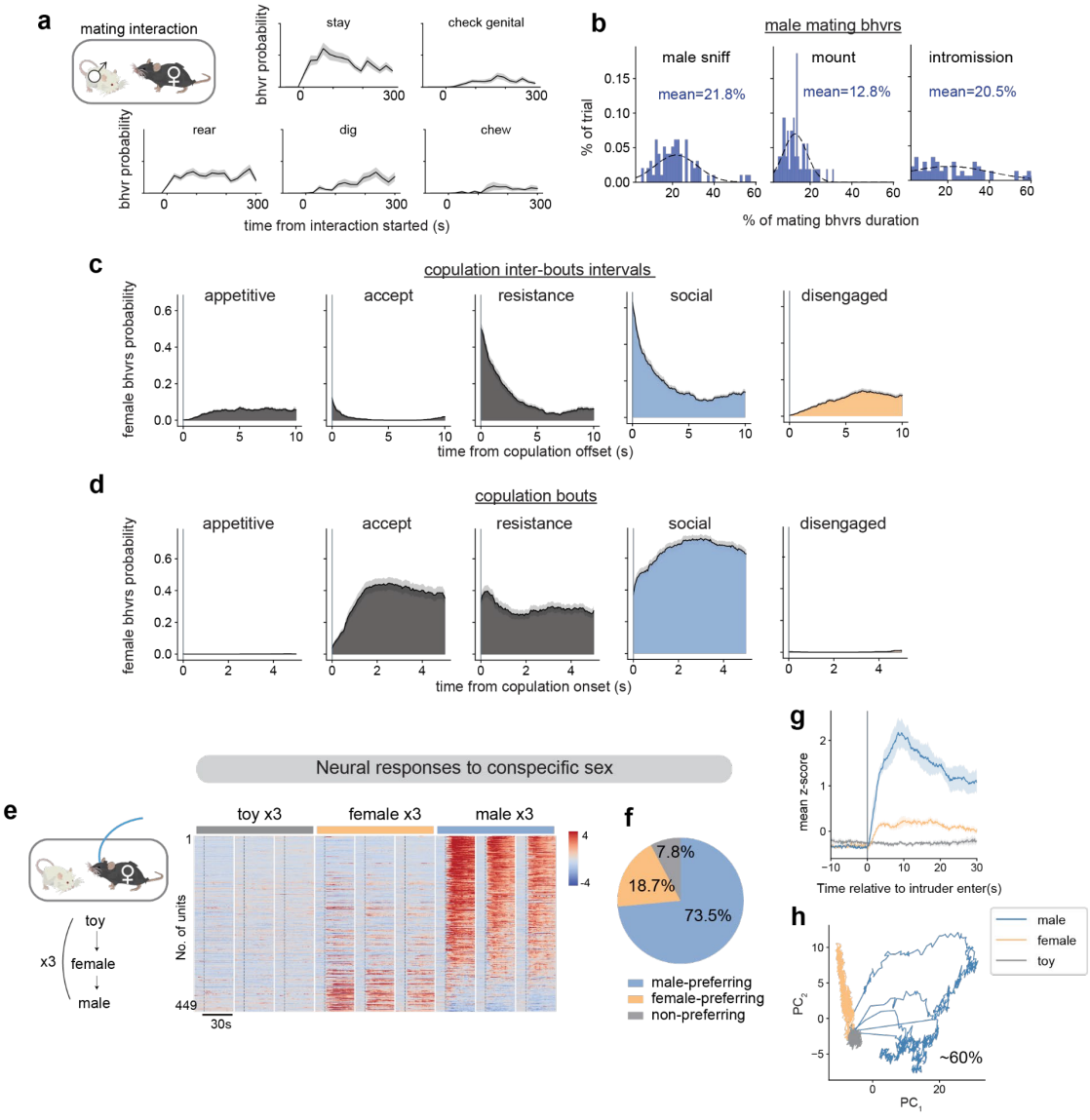
$$c_k = \frac{1}{T}\sum_{t=1}^{T-k}(y_t - \bar{y})(y_{t+k} - \bar{y})$$

and c_0 is the sample variance of the data. The half-width is found for each neuron as the point where the autocorrelation function reaches a value of 0.5.

Partial least squares regression to identify integration dynamics

To identify the integration dimension using an independent method, we also used partial least squares regression. Towards this, all traces were concatenated and regressed against a 1 x T vector designed such that the vector shows ramping activity upon entry of the male intruder (see Extended Data Figure 3d-e).

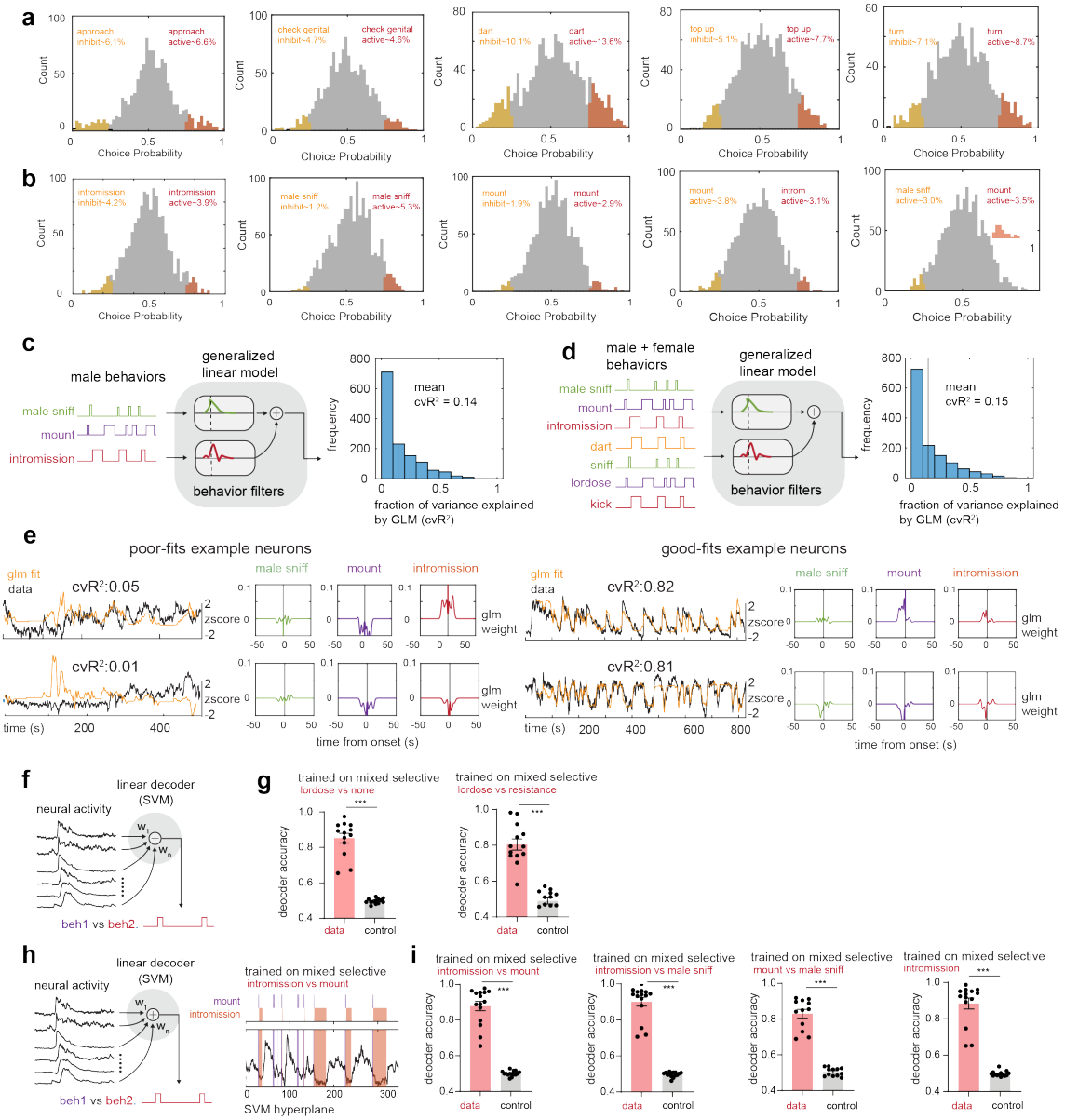
Extended Data (Supplemental) Figure 1



Extended Data Figure 1. Behavior dynamics and neural responses to conspecific sex.

a, The probability of female behaviors every 20s (n=74 trials, N=28 mice). b, Distribution of the percentage of time males displayed mating behaviors in each trial (n=74 trials). c, The probability of female behaviors aligned to male copulation offsets and d, copulation onsets. (e-h), Neural responses to conspecific sex. e, Left, diagram of sex representation test. Each intruder was presented for 1 min. Right, concatenated average responses to toy, female, or male (N = 8 mice). Color scale indicates z-scored activity. Units were sorted by temporal correlation. f, Percentages of male- or female-preferring cells (calculated by Choice Probability). g, Mean responses of female VMHvlEsr1 α cells to male, female and toy (N=8 mice). Data presented as mean ± SEM. h, PCA of neuronal responses to male, female and toy from one example female.

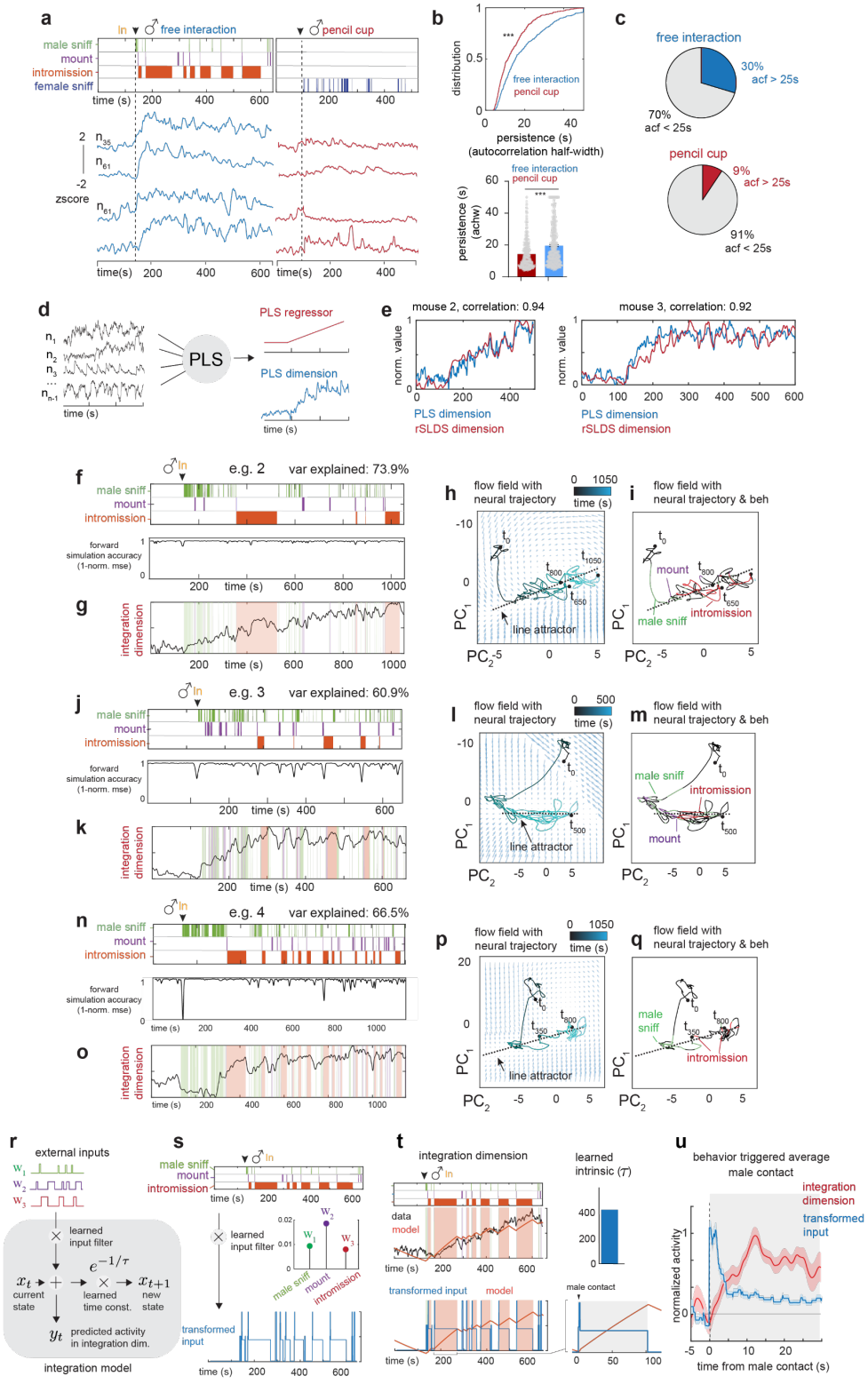
Extended Data Figure 2



Extended Data Figure 2. Neural tuning to conspecific sex and behavior.

a, Choice Probability (CP) histograms and percentages of tuned cells for female behaviors. cutoff: CP>0.7 or <0.3 and >2σ. N = 15 mice. b, Same as a, but for male behavior. c, Schematic showing generalized linear model (GLM) used to predict neural activity from male behaviors and distribution of cvR2 across all mice, or d, both male and female behaviors and distribution of cvR2 across all mice (N = 15 mice). e, Example GLM fits and behavior filters for poorly and well fit neurons. (f-h), Decoder analysis. f, schematic showing linear support vector machine (SVM) decoder trained on frames of male mating behaviors. g, performance of SVM trained to separate female behavior. Left, performance of SVM trained to separate frames of lordosis versus all remaining frames (***p<0.001, N = 15 mice, data: 0.85 ± 0.03, shuffle: 0.49 ± 0.003). Right, performance of SVM trained to separate frames of lordosis versus resistance behaviors (***p<0.001, N = 15 mice, data: 0.80 ± 0.03, shuffle: 0.48 ± 0.01). h, Same as a, but showing the SVM hyperplane for separating male behaviors (mount versus intromission) on right. (***p<0.001). (N = 15 mice). i, performance of SVM trained to separate intromission versus mount (data: 0.89 ± 0.02, shuffle: 0.49 ± 0.003), intromission versus male sniffing (data: 0.90 ± 0.02, shuffle: 0.49 ± 0.003), mount versus male sniffing (data: 0.83 ± 0.02, shuffle: 0.50 ± 0.006) and intromission versus remaining frames male sniffing (***p<0.001, N = 15 mice, mean data: 0.88 ± 0.03, shuffle: 0.48 ± 0.003).

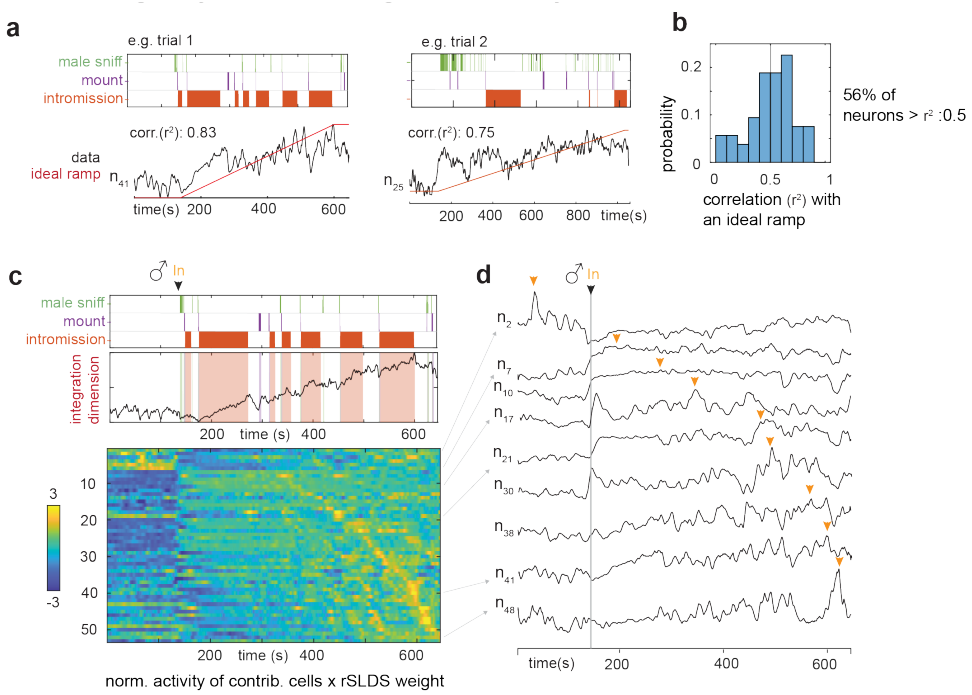
Extended Data Figure 3



Extended Data Figure 3. | Additional example trials with rSLDS model fit, additional information for Fig. 3.

a, Dynamics of persistently active neurons identified during receptive interaction with pencil-cup assay. b, Cumulative distribution & bar plot of ACHW for same neurons during free interaction vs pencil cup assay ****p<0.0001, Mann-Whitney U test, p value: 1.25e-11, N = 470 neurons from 5 mice. mean ACHW during pencil cup: 14.3 ± 0.42, free interaction: 19.6 ± 0.58. c, Pie chart indicating fraction of neurons with ACHW > 25s in free interaction and in pencil cup assay. d, Schematic illustrating partial least squares regression to extract integration dynamics in VMHvl. e, Comparison of rSLDS integration dimension and PLS dimension for two example mice showing a high correlation. (f-q), Additional example trials with rSLDS model fit. f, Recurrent switching linear dynamical systems (rSLDS) model fit forward simulation accuracy aligned to male behaviors in example trial 2. g, Dynamics of the integration dimension in trial 2. h, Flow field of VMHvl α dynamical system showing neural trajectories in state space, annotated by time from male encounter (t0) for trial 2. i, Neural state space of VMHvl α dynamical system highlighting behaviors and the region containing the line attractor for trial 2. j-m, the same as g-i for example trial 3. n-q, the same as j-i for example trial 4. r, Integration model used to dissect the contribution of intrinsic decay and external inputs (male behaviors; male-sniff, mount, and intromission). A single state LDS model is used to fit external inputs to predict activity in the integration dimension. s, Top: External inputs to integration model, middle: learned input filter showing weights that are multiplied with the external inputs. Bottom: transformed input obtained by multiplying external inputs with input filter. t, Top: Data and model prediction from LDS to predict activity in the integration dimension. The learned model has a large intrinsic time constant (right). Bottom: Transformed input (weighted input from three male behaviors) and model prediction overlayed with behaviors. u, Behavior triggered average of transformed input and integration dimension aligned to male contact. Male contact is present for the duration of the shaded region. Data presented as mean ± SEM.

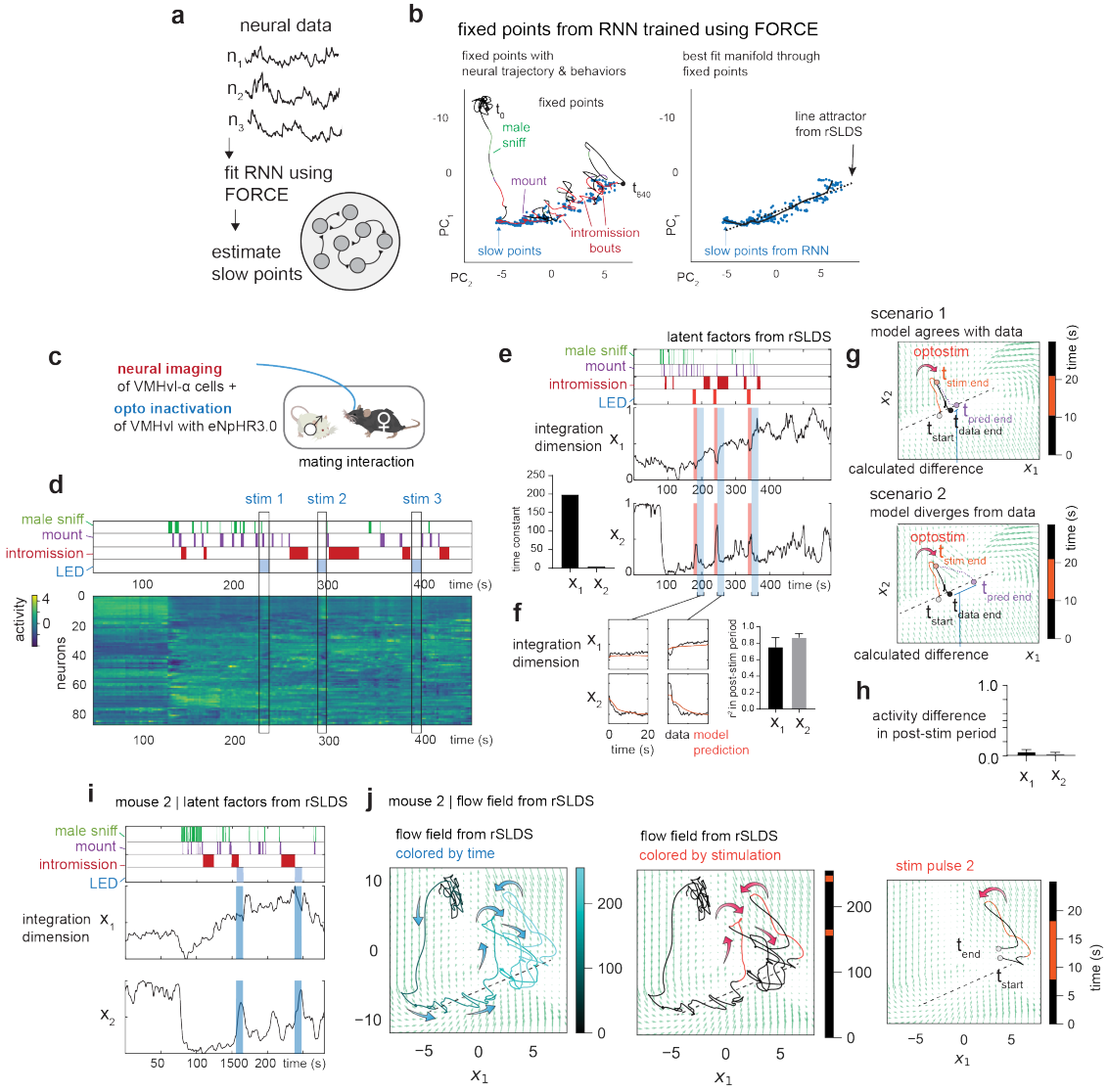
Extended Data Figure 4



Extended Data Figure 4. | Dynamics of single cell activity.

a, Correlation of example unit activity with an ideal ramp. b, Distribution of correlation of individual neuron activity with ideal ramp. c, *Upper*, relationship of male behavior to weighted average of all units contributing to integration dimension as a function of time. Data from the same example trial as shown in Fig. 3f. *Lower*, normalized activity (z-score) of individual units times rSLDS weight for each unit exhibiting a significant weight in the integration dimension, sorted by time to peak. d, Traces of example units from f, *lower*. Yellow arrow indicates peak of activity for each unit.

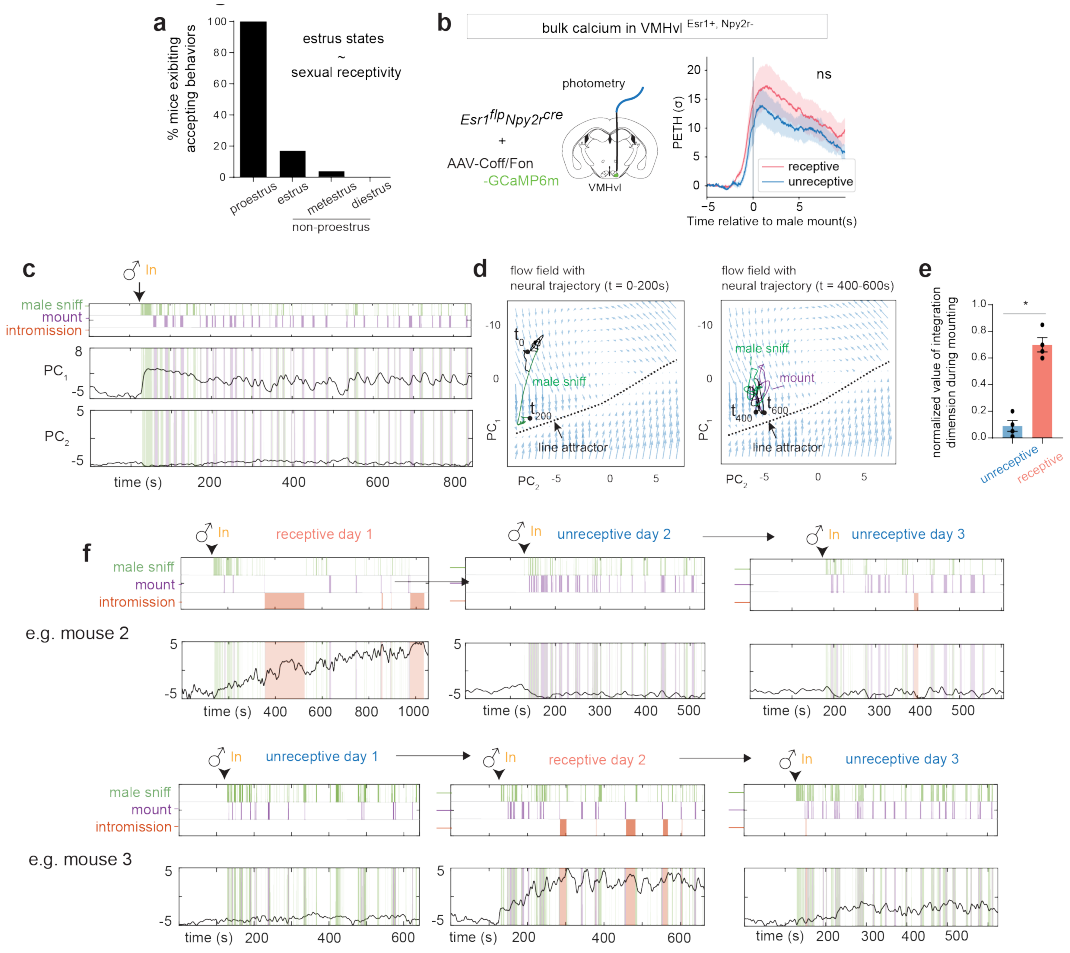
Extended Data Figure 5



Extended Data Figure 5. | Independent verification and neural perturbations of line attractor dynamics.

a, Cartoon illustrating approach of fitting RNNs to neural data using FORCE. b, Slow points and attractor manifold uncovered by FORCE, overlaid with line attractor uncovered by rSLDS. c, Paradigm for simultaneous neural perturbation & imaging during a mating interaction in females. GcaMP was expressed in VMHvl-α cells while halorhodopsin (eNpHR3.0) was expressed in all VMHvl neuron using a pan-neuronal driver. d, Neural data obtained from a female showing annotated male behaviors and optogenetic inhibition (LED). e, Left: Latent factors from two-dimensional rSLDS model fit to neural data. Reproduced for explanatory purposes from Figure 3h. Right: Time constants of the two longest-lived dimensions from rSLDS model fit to data from unperturbed periods (excluding stimulation period plus a 20s post-stimulus period). f, Left: Performance of model on held out data from 20s immediate post-stimulus period (taken from highlighted blue portions of graphs in e). g, Cartoon depicting quantification of flow field prediction following optogenetic perturbation. The flow field fit from unperturbed periods of time is used to predict the neural trajectory following perturbation (t-pred end, purple line). This trajectory is then compared to data (t-data end, black line). Scenario 1 illustrates when the model agrees with data, resulting in a low difference in activity along the line attractor (top). Scenario 2 illustrates when the model diverges from data resulting in a large deviation in final position along the line attractor (bottom). h, Quantification of flow field prediction following perturbation as the difference in activity level at the end of the 20s post-stimulus period between data and model in both x1 and x2 dimensions across mice (activity difference for x1: 0.05 ± 0.03, for x2: 0.03 ± 0.01, $n = 3$ mice). i, Latent factors from rSLDS of mouse 2 during neural perturbation. j, Flow field and neural trajectories for mouse 2. Note that trajectories are pushed away from the attractor during stimulation and then return to line attractor following stimulation offset, as predicted by the flow field.

Extended Data Figure 6

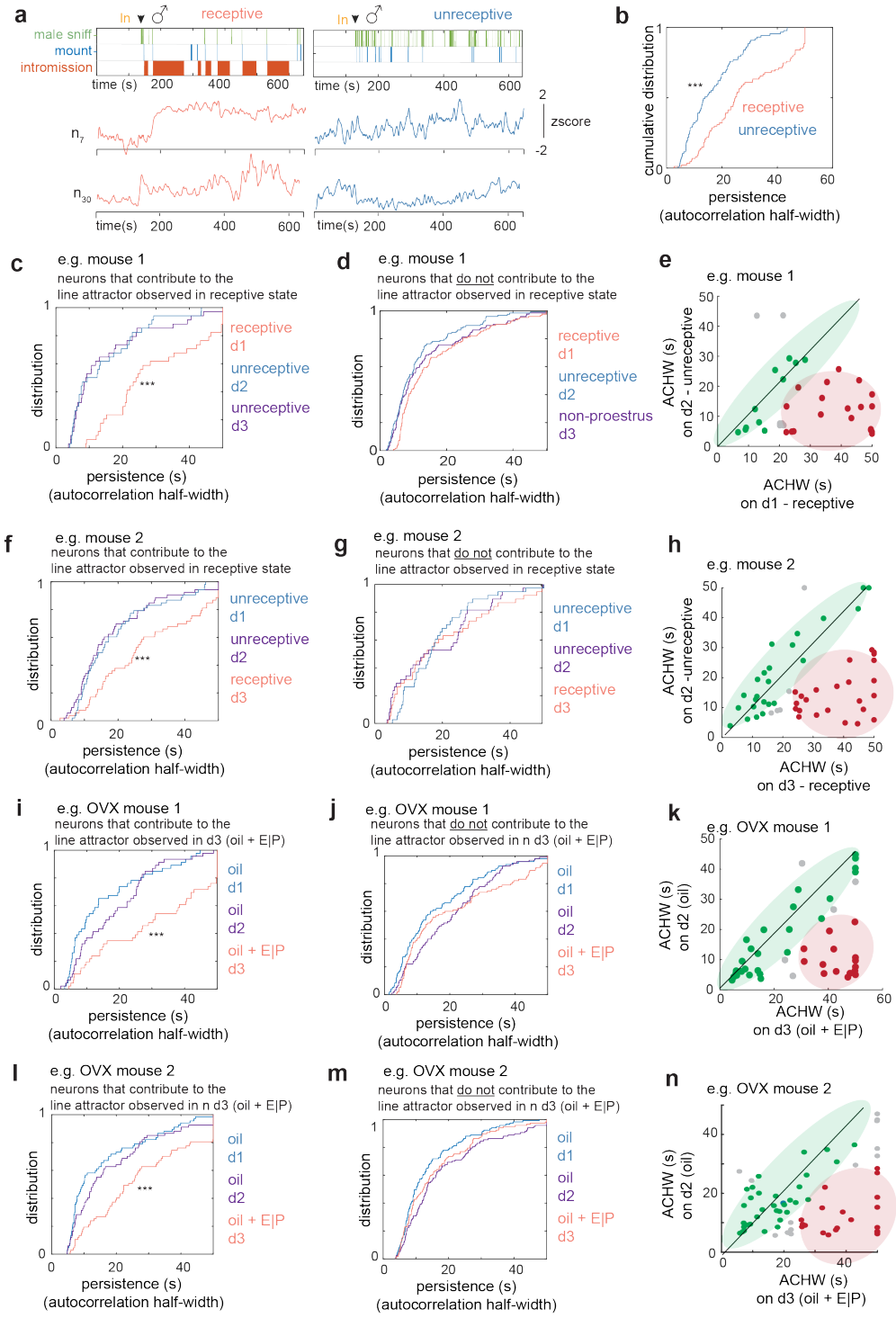


Extended Data Figure 6. Line attractor dynamics across the estrus cycle.

a, Correlation between female estrus states and the presence of sexual receptivity, measured by whether female displayed accepting behaviors during interaction with male. b, Photometry recording in female VMHvl α cells during receptive and unreceptive mating interactions. Data presented as mean ± SEM. c, Low dimensional PCs of VMHvl α dynamical system in receptive day with neural data projected from unreceptive day. d, Flow field of VMHvl α dynamical system in receptive day with neural trajectories projected from unreceptive for t = 0 to t = 200s (left) and t = 200s to t = 400s (right). e, Quantification of normalized value of integration dimension during male-mounting in unreceptive and receptive days (*p<0.05, N = 4 mice, mean value during unreceptive day: 0.09±0.04, receptive day: 0.69±0.05. Mann-Whitney U test, p value: 0.02). f,

Dynamics of integration dimension in two more example mice discovered during receptive day compared to activity of the same dimension on unreceptive days.

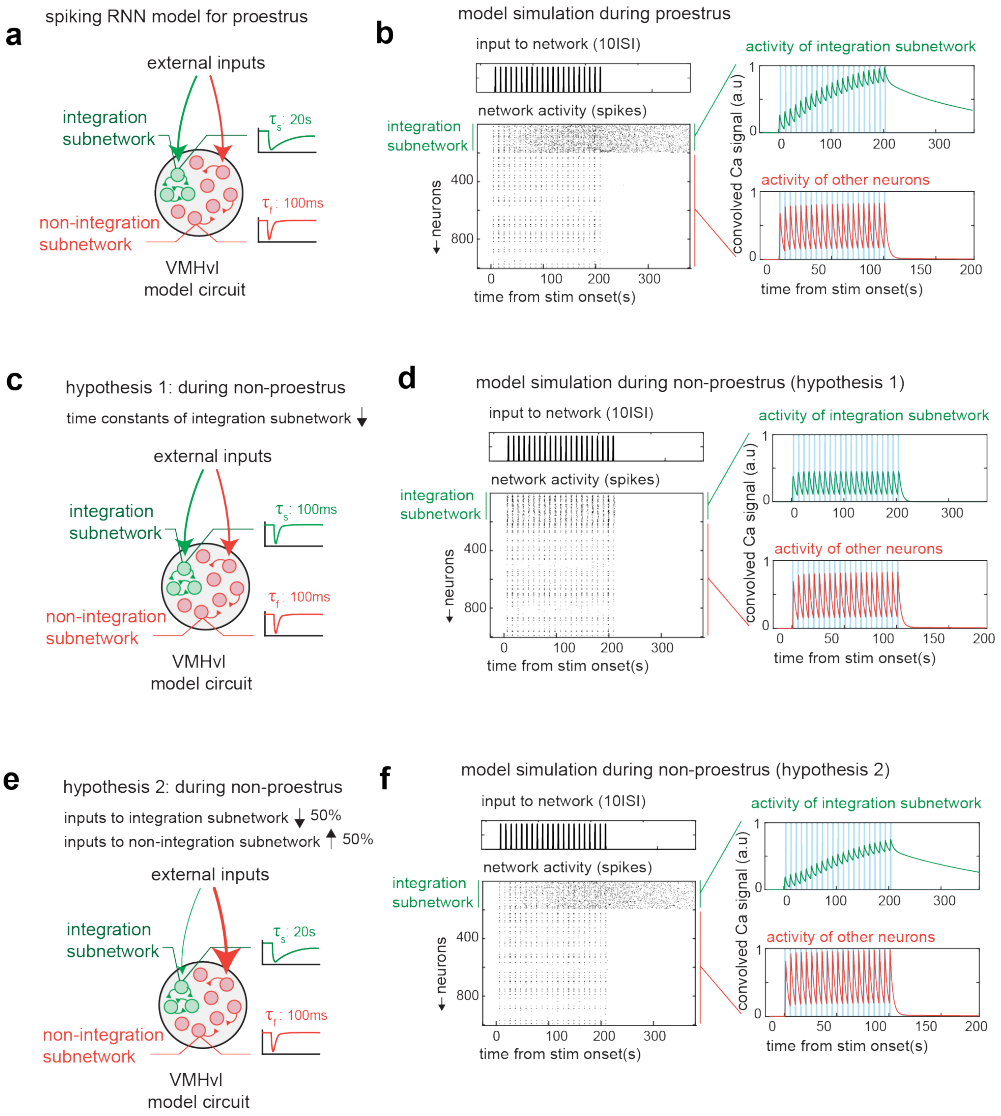
Extended Data Figure 7



Extended Data Figure 7. | Single cell persistence at receptive and unreceptive days.

a, Example units active during both receptive (red traces, left) and unreceptive (blue traces, right), showing persistence on receptive day and fast dynamics on the unreceptive days. b, Comparison of cumulative distribution of ACHWs to that of same neurons on unreceptive days. Data from example mouse 1. ***p<0.001, KS-test. c, Cumulative distribution of ACHWs for units with significant weights on integration dimension across receptive and unreceptive day, ***p<0.001, KS-test. Data from example mouse 1. d, Cumulative distribution of ACHWs for example mouse 1, for units that do not contribute to the integration dimension on the receptive day, compared on receptive vs unreceptive days. e, Scatter plot of ACHWs for units with significant weights on integration dimension for receptive day vs unreceptive day. Data from example mouse 1. (f-h) Same as c-e for example mouse 2. I, Cumulative distribution of ACHWs for units with significant weights on integration dimension across hormone primed (day 3) and non-primed days (days 2, 1). ***p<0.001, KS-test. Data from example OVX mouse 1. ***p<0.001, KS-test. j, Cumulative distribution of AHWs for example OVX mouse 1, for units that do not contribute to the integration dimension across hormone primed (day 3) and non-primed days (days 2, 1). k, Scatter plot of ACHWs for units with significant weights on integration dimension for hormone-primed day vs non-primed day. Data from example OVX mouse 1 (l-n) Same as i-j. for example OVX mouse 2.

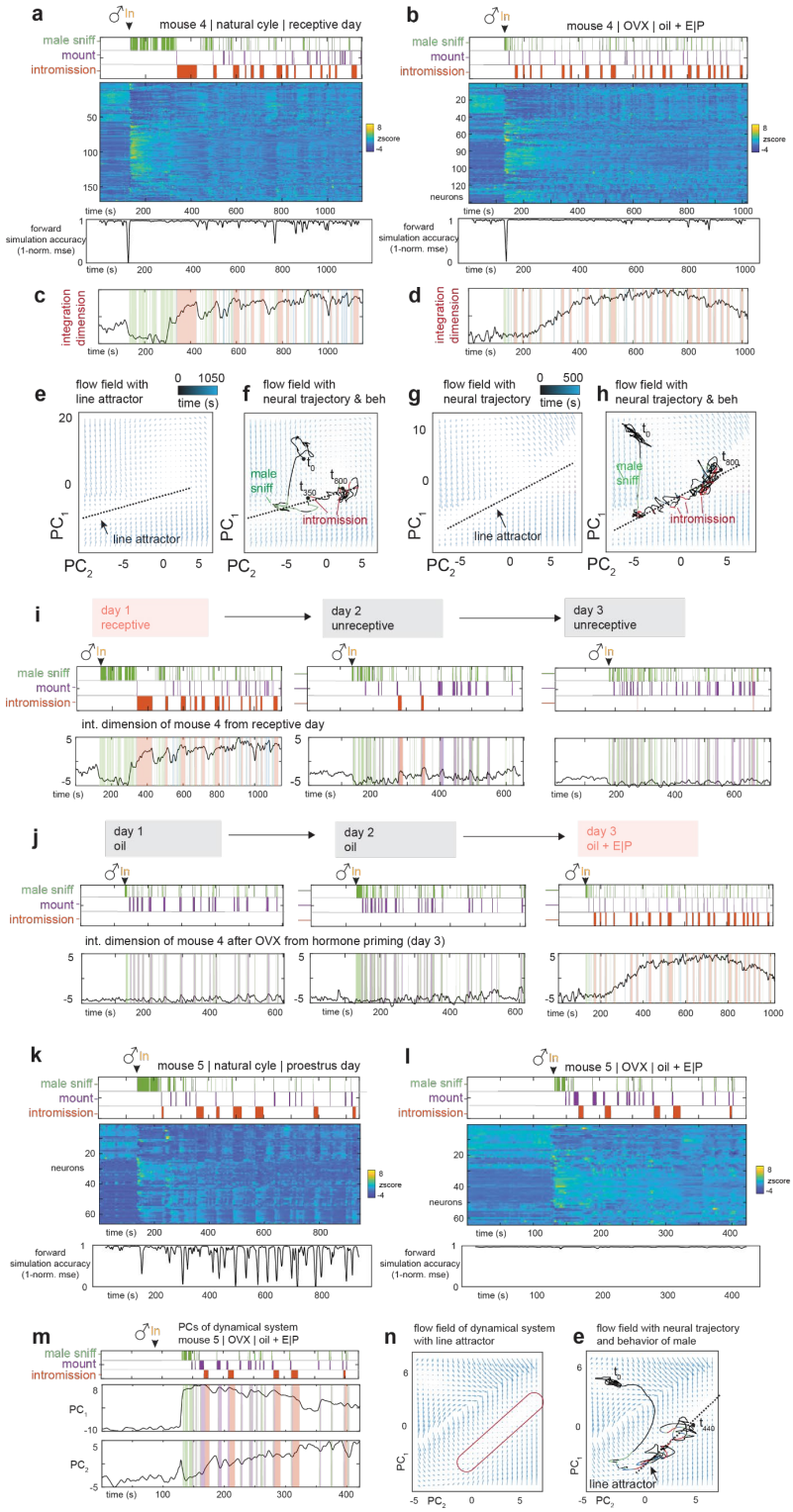
Extended Data Figure 8



Extended Data Figure 8. | Mechanistic model for loss of line attractor dynamics in unreceptive states.

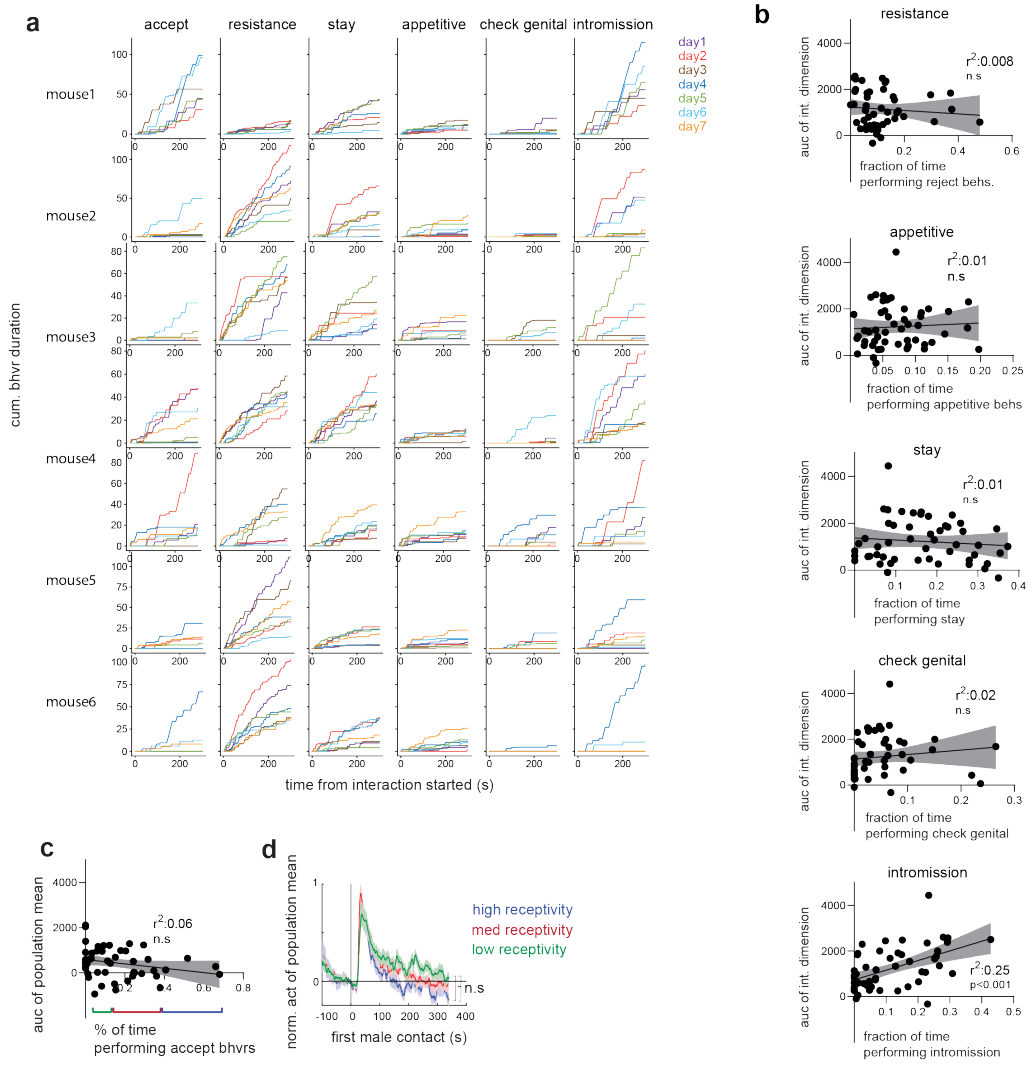
a, Schematic illustrating the construction of a spiking recurrent neural network (RNN) with a line attractor. The line attractor is created by allowing a subset of neurons to possess a larger intrinsic time constant (20s vs 100ms), and by denser connectivity within the subnetwork (12% versus 1% in remaining network). b, Model simulation during the proestrus phase with pulse like input delivered at 10s ISI. Right, activity of integration subnetwork (green) and other neurons (red). c, Schematic for hypothesis 1: we hypothesize that during non-proestrus, there is a reduction in the intrinsic constant of the integration subnetwork (from 20s to 100ms). d, Same as b but for hypothesis 1 during non-proestrus. e, Schematic for hypothesis 2: we test whether changes in the firing rate of different neuronal subsets can lead to the loss of attractor dynamics. To investigate this, we provide the integration subnetwork with 50% reduced input strength, while increasing the same for the remaining neurons. f, Same as b but for hypothesis 2 during non-proestrus.

Extended Data Figure 9



Extended Data Figure 9. | Population dynamics before and after OVX in the same female. (a, b,) Neural raster and behaviors and rSLDS model performance (measured as forward simulation error, see Methods) for one example mouse in receptive day of natural estrus cycle a, and same mouse on hormone primed day after OVX (day3, oil + E|P) b. (c, d,) Integration dimension identified by rSLDS on natural cycle receptive day c, and during hormone primed day after OVX d. (e, f,) Flow field e, and neural trajectories of dynamical system f, with line attractor highlighted of model fit during the receptive state of the estrus cycle. (g, h,) Same as e, f, for model fit during hormone primed day after OVX. I, Dynamics of integration dimension discovered during natural cycle receptive day compared to activity of the same dimension on unreceptive days. j, Dynamics of integration dimension in the same mouse discovered during hormone primed day (day 3) compared to the activity of the same dimension during non-primed days. (k, l,) Neural raster and behaviors and rSLDS model performance for mouse in proestrus day of natural estrus cycle k, and same mouse on hormone primed day after OVX (day3, oil + E|P) l. m, Principal components of mouse dynamic system fit during hormone primed day. (n, o,) Flow field n, and neural trajectories of dynamical system. o, with line attractor highlighted of model fit during the hormone primed day after OVX in mouse.

Extended Data Figure 10



Extended Data Figure 10. Longitudinal mating assay and correlation with attractor dynamics.

a, Behaviors displayed in mating interactions across days from all the recorded females. b, The scatter plots of the integration dimension values and the amount of female resistance behaviors (linear regression, R^2= 0.008), appetitive behaviors (R^2= 0.01), staying (R^2= 0.01), checking genital (R^2= 0.02) and male intromission (R^2= 0.25). Data presented as mean ± SEM. c, correlation of area under the curve (auc) of the population mean of all neurons with the percentage of time spent performing accepting behaviors. Data presented as mean ± SEM. d: activity of population mean from trials with varying degrees of receptivity defined in a), Mann-Whitney U test.

References:

1. Jennings, K. J. & de Lecea, L. Neural and hormonal control of sexual behavior. *Endocrinol. (United States)* **161**, 1–13 (2020).
2. Gutierrez-Castellanos, N., Husain, B. F. A., Dias, I. C. & Lima, S. Q. Neural and behavioral plasticity across the female reproductive cycle. *Trends Endocrinol Metab.* 1–17 (2022). doi:10.1016/j.tem.2022.09.001
3. Yin, L. & Lin, D. Neural control of female sexual behaviors. *Horm Behav* **151**, 105339 (2023).
4. Lenschow, C. & Lima, S. Q. In the mood for sex: neural circuits for reproduction. *Curr Opin Neurobiol* **60**, 155–168 (2020).
5. Liu, M., Kim, D.-W., Zeng, H. & Anderson, D. J. Make war not love: The neural substrate underlying a state-dependent switch in female social behavior. *Neuron* **110**, 841-856.e6 (2022).
6. Yin, L. *et al.* VMHvllCckar cells dynamically control female sexual behaviors over the reproductive cycle. *Neuron* **110**, 3000-3017.e8 (2022).
7. Pfaff, D.W., Diakow, C., Zigmond, R.E. and Kow, L. M. Neural and hormonal determinants of female mating behavior in rats. *Neurosci.* **3**, 621–646 (1974).
8. Pfaff, D. W., Gagnidze, K. & Hunter, R. G. Molecular endocrinology of female reproductive behavior. *Mol Cell Endocrinol.* **467**, 14–20 (2018).
9. Rodriguez-Sierra, J. F., Crowley, W. R. & Komisaruk, B. R. Vaginal stimulation in rats induces prolonged lordosis responsiveness and sexual receptivity. *J Comp Physiol Psychol.* **89**, 79–85 (1975).

10. Kelli L. Boyd, Atis Muehlenbachs, Mara H. Rendi, Rochelle L. Garcia, K. N. G.-C. *Female Reproductive System*. (Academic Press, 2018). doi:https://doi.org/10.1016/B978-0-12-802900-8.00017-8

11. Gutierrez-castellanos, N., Husain, B. F. A., Dias, I. C. & Lima, S. Q. Endocrinology & Metabolism Neural and behavioral plasticity across the female reproductive cycle. *Trends Endocrinol Metab.* 1–17 (2022). doi:10.1016/j.tem.2022.09.001

12. Micevych, P. E. & Meisel, R. L. Integrating neural circuits controlling female sexual behavior. *Front Syst Neurosci.* **11**, 1–12 (2017).

13. Pfaff, D. W. & Sakuma, Y. Deficit in the lordosis reflex of female rats caused by lesions in the ventromedial nucleus of the hypothalamus. *J Physiol.* **288**, 203–210 (1979).

14. Pfaff, D. W. & Sakuma, Y. Facilitation of the lordosis reflex of female rats from the ventromedial nucleus of the hypothalamus. *J Physiol.* **288**, 189–202 (1979).

15. Yang, C. F. *et al.* Sexually dimorphic neurons in the ventromedial hypothalamus govern mating in both sexes and aggression in males. *Cell* **153**, 896–909 (2013).

16. Hashikawa, K. *et al.* Esr1+ cells in the ventromedial hypothalamus control female aggression. *Nat Neurosci.* **20**, 1580–1590 (2017).

17. Inoue, S. *et al.* Periodic remodeling in a neural circuit governs timing of female sexual behavior. *Cell* **179**, 1–16 (2019).

18. Knoedler, J. R. *et al.* A functional cellular framework for sex and estrous cycle-dependent gene expression and behavior. *Cell* **185**, 654-671.e22 (2022).

19. Ziv, Y. *et al.* Long-term dynamics of CA1 hippocampal place codes. *Nat. Neurosci.*

16, 264–266 (2013).

20. Linderman, S. W. *et al.* Bayesian learning and inference in recurrent switching linear dynamical systems. *Proc 20th Int Conf Artif Intell Stat. AISTATS 2017* **54**, (2017).
21. Beach, F. A. Sexual attractivity, proceptivity, and receptivity in female mammals. *Horm Behav.* **7**, 105–138 (1976).
22. Nair, A. *et al.* An approximate line attractor in the hypothalamus that encodes an aggressive internal state. *Cell* **186**, 178–193 (2022).
23. Remedios, R. *et al.* Social behaviour shapes hypothalamic neural ensemble representations of conspecific sex. *Nature* **550**, 388–392 (2017).
24. Pillow, J. W. *et al.* Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* **454**, 995–999 (2008).
25. Weissbourd, B. *et al.* A genetically tractable jellyfish model for systems and evolutionary neuroscience. *Cell* **184**, 5854-5868.e20 (2021).
26. Cavanagh, S. E., Towers, J. P., Wallis, J. D., Hunt, L. T. & Kennerley, S. W. Reconciling persistent and dynamic hypotheses of working memory coding in prefrontal cortex. *Nat Commun.* **9**, 1–16 (2018).
27. Murray, J. D. *et al.* A hierarchy of intrinsic timescales across primate cortex. *Nat Neurosci* **17**, 1661–1663 (2014).
28. Mante, V., Sussillo, D., Shenoy, K. V. & Newsome, W. T. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).
29. Rajan, K., Harvey, C. D. D. & Tank, D. W. W. Recurrent network models of sequence generation and memory. *Neuron* **90**, 128–142 (2016).

30. Khona, M. & Fiete, I. R. Attractor and integrator networks in the brain. *Nat Rev Neurosci.* **23**, 744–766 (2022).

31. Kunwar, P. S. *et al.* Ventromedial hypothalamic neurons control a defensive emotion state. *Elife* **2015**, (2015).

32. Wang, L., Chen, I. Z. & Lin, D. Collateral pathways from the ventromedial hypothalamus mediate defensive behaviors. *Neuron* **85**, 1344–1358 (2015).

33. Manoli, D. S., Fan, P., Fraser, E. J. & Shah, N. M. Neural control of sexually dimorphic behaviors. *Curr Opin Neurobiol.* **23**, 330–338 (2013).

34. Karigo, T. *et al.* Distinct hypothalamic control of same- and opposite-sex mounting behaviour in mice. *Nature* **589**, 258–263 (2021).

35. Kim, D.-W. *et al.* Multimodal analysis of cell types in a hypothalamic node controlling social behavior. *Cell* **179**, 713-728.e17 (2019).

36. Lo, L. *et al.* Connectional architecture of a mouse hypothalamic circuit node controlling social behavior. *Proc Natl Acad Sci.* **116**, 7503–7512 (2019).

37. Knoedler, J. R. & Shah, N. M. Molecular mechanisms underlying sexual differentiation of the nervous system. *Curr Opin Neurobiol.* **53**, 192–197 (2018).

38. Yang, B., Karigo, T. & Anderson, D. J. Transformations of neural representations in a social behaviour network. *Nature* **608**, 741–749 (2022).

39. Hong, W. *et al.* Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning. *Proc Natl Acad Sci U. S. A.* **112**, E5351–E5360 (2015).

40. Linderman, S. W. *et al.* Recurrent switching linear dynamical systems. *arXiv Prepr.* (2016).

41. Maheswaranathan, N., Williams, A., Golub, M., Ganguli, S. & Sussillo, D. Reverse engineering recurrent networks for sentiment classification reveals line attractor dynamics. *Adv Neural Inf Process Syst.* (2019).

42. Hadjiabadi, D. *et al.* Maximally selective single-cell target for circuit control in epilepsy models. *Neuron* **109**, 2556-2572.e6 (2021).

43. Sussillo, D. & Abbott, L. F. Generating coherent patterns of activity from chaotic neural networks. *Neuron* **63**, 544–557 (2009).

44. Sussillo, D. & Barak, O. Opening the black box: low-dimensional dynamics in high-dimensional recurrent neural networks. *Neural Comput.* **25**, 626–649 (2013).

45. Kennedy, A. *et al.* Stimulus-specific hypothalamic encoding of a persistent defensive state. *Nature* **586**, 730–734 (2020).

46. Yamamoto, R., Ahmed, N., Ito, T., Gungor, N. Z. & Pare, D. Optogenetic Study of anterior BNST and basomedial amygdala projections to the ventromedial hypothalamus. *eNeuro* **5**, (2018).

EPILOGUE

A future for dynamical systems in encoding affective states

This thesis presents a new paradigm in understanding the computations that shape our emotional states. It reveals how emergent continuous attractor dynamics in the hypothalamus—born from a complex interplay of connectivity, neuromodulation, and cell-intrinsic properties—act as a canonical motif that encodes the persistence and scalability of diverse affective states. The insights gained in this thesis about the attractors' implementation also deepen our understanding of the computation and algorithmic processes shaping innate affective states, resonating across all three levels of understanding proposed by David Marr.

There is still much to uncover about this neural dynamics-based paradigm, particularly regarding the behavioral relevance of these signals. Are the discovered representations causal in enabling our affective states, or do they merely reflect feedback during those states? These questions also extend to the integration process along the line attractor. What is the modality (e.g., mechanosensory, olfactory) of input being integrated by the attractor, and which brain regions are relaying these inputs to the hypothalamus? A comprehensive understanding of this process is essential to fully grasp the computation of affective states.

While we have identified the importance of neuropeptides in creating these emergent dynamics, the precise mechanistic details of this process remain unknown. The line attractor may reflect local connectivity within the hypothalamus, facilitated by the recurrent release of neuropeptides, or it might depend on macro-level interactions between multiple brain regions. Given the extensive interconnectivity of the hypothalamus[1,2], macro-level interactions are likely to play a key role in creating and regulating the emergent dynamics we have discovered.

Perhaps the most exciting potential of these ideas lies in their ability to reveal aspects of neural dynamics that are impaired in mood and neuropsychiatric disorders[3-5]. The shape of the manifold underlying hypothalamic dynamics may be malleable[6], and specific mutations affecting neuromodulation or connectivity could alter the stability of the attractor. Such changes might manifest behaviorally as persistent alterations in social behavior, such as

ongoing fear or anxiety in models of post-traumatic stress disorder or depression[7]. Designing behavioral or neural interventions that can restore the attractor landscape could represent a promising avenue for therapeutic intervention.

There are still other important factors that will determine the success of the ideas presented in this thesis. The discovery of these computations related to internal states relies on new machine learning methods that extract these properties from neural data in an unsupervised manner. Making these methods accessible to the broader scientific community, beyond just computational neuroscience, will be crucial for this research to reach its full potential. Ongoing efforts in the lab to create intuitive and accessible software platforms show promise in meeting this need.

In closing, it is important to note that the studies presented in this thesis are deeply rooted in close collaborations with experimentalists, where theoretical and machine learning-driven insights have guided critical experiments, and those experiments, in turn, have fueled further modeling and theory. This synergistic cycle has been crucial to advancing our understanding of dynamical systems in affective states. I am optimistic that the lessons learned, and the collaborative framework established through these efforts will continue to unite diverse fields of neuroscience, paving the way for more paradigm-shifting discoveries in the future.

References

1 Lo, L. et al. Connectional architecture of a mouse hypothalamic circuit node controlling social behavior. *Proc Natl Acad Sci U S A* 116, 7503-7512 (2019).

2 Saper, C. B. & Lowell, B. B. The hypothalamus. *Curr Biol* 24, R1111-1116 (2014). https://doi.org:10.1016/j.cub.2014.10.023

3 LeDuke, D. O., Borio, M., Miranda, R. & Tye, K. M. Anxiety and depression: A top-down, bottom-up model of circuit function. *Ann N Y Acad Sci* 1525, 70-87 (2023). https://doi.org:10.1111/nyas.14997

4 Rolls, E. T. A non-reward attractor theory of depression. *Neurosci Biobehav Rev* 68, 47-58 (2016). https://doi.org:10.1016/j.neubiorev.2016.05.007

5 Rolls, E. T., Loh, M. & Deco, G. An attractor hypothesis of obsessive-compulsive disorder. *Eur J Neurosci* 28, 782-793 (2008). https://doi.org:10.1111/j.1460-9568.2008.06379.x

6 Nair, A. et al. An approximate line attractor in the hypothalamus encodes an aggressive state. *Cell* 186, 178-193 e115 (2023). https://doi.org:10.1016/j.cell.2022.11.027

7 Czeh, B., Fuchs, E., Wiborg, O. & Simon, M. Animal models of major depression and their clinical implications. *Prog Neuropsychopharmacol Biol Psychiatry* 64, 293-310 (2016). https://doi.org:10.1016/j.pnpbp.2015.04.004

INDEX