

Stability and Control: Signatures of the Conscious Brain

by

Adam J. Eisen

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of
DOCTORATE IN COMPUTATIONAL NEUROSCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2026

© 2026 Adam J. Eisen. All rights reserved.

The author hereby grants to MIT a nonexclusive, worldwide, irrevocable, royalty-free license to exercise any and all rights under copyright, including to reproduce, preserve, distribute and publicly display copies of the thesis, or release the thesis under an open-access license.

Authored by: Adam J. Eisen
Department of Brain and Cognitive Sciences
August 4, 2026

Certified by: Earl K. Miller
Picower Professor of Neuroscience, Thesis Supervisor

Certified by: Ila R. Fiete
Professor of Brain and Cognitive Sciences, Thesis Supervisor

Accepted by: Mark Harnett
Graduate Officer, Department of Brain and Cognitive Sciences

THESIS COMMITTEE

THESIS SUPERVISORS

Earl K. Miller

*Picower Professor of Neuroscience
Department of Brain and Cognitive Sciences, MIT*

Ila R. Fiete

*Professor of Brain and Cognitive Sciences
Department of Brain and Cognitive Sciences, MIT*

THESIS READERS

Robert Desimone

*Doris and Don Berkey Professor
Department of Brain and Cognitive Sciences, MIT*

Laura D. Lewis

*Athinoula A. Martinos Associate Professor
Department of Electrical Engineering and Computer Science, MIT*

Dani S. Bassett

*J. Peter Skirkanich Professor
Department of Bioengineering, University of Pennsylvania*

Stability and Control: Signatures of the Conscious Brain

by

Adam J. Eisen

Submitted to the Department of Brain and Cognitive Sciences
on August 4, 2026 in partial fulfillment of the requirements for the degree of

DOCTORATE IN COMPUTATIONAL NEUROSCIENCE

ABSTRACT

The brain is composed of interconnected cortical areas that dynamically coordinate to support conscious functioning. Consciousness is reliably disrupted by general anesthetics, providing a means through which the signatures of consciousness can be identified by contrasting across states. We leveraged dynamical systems theory to explore changes in neural dynamics between consciousness and anesthetic-induced unconsciousness. We focused on two fundamental dynamical systems properties: *dynamic stability*, the robustness of dynamics to perturbation, and *control*, how inputs can be coordinated with system dynamics to steer behavior. To evaluate these phenomena, we developed and applied two complementary methodologies. First, Delayed Linear Analysis for Stability Estimation (DeLASE) infers dynamic stability from delay-embedded neural recordings. We applied DeLASE to local field potentials (LFPs) recorded from non-human primates during propofol infusion and found that propofol anesthesia destabilized cortical dynamics relative to the awake state. This destabilization generalized to recordings under anesthetics with distinct molecular mechanisms (ketamine and dexmedetomidine), suggesting a shared dynamical signature of anesthetic-induced unconsciousness. Second, we constructed JacobianODE, a deep learning framework that estimates the time-varying Jacobian of nonlinear dynamics from data, and used it to characterize how subsystems within a complex network control one another. We applied this control-theoretic analysis to the same multi-area cortical recordings during propofol anesthesia. We found a broad disruption of intracortical control, both in terms of how easily brain areas could drive each other toward novel states, and how well they could control each other to produce reliable dynamics. Together, these contributions offer rigorous, data-driven tools for probing stability and control in neural dynamics. More broadly, they frame anesthetic-induced unconsciousness as a joint breakdown of the stability and directional control required for consciousness.

Thesis supervisor: Earl K. Miller
Title: Picower Professor of Neuroscience

Thesis supervisor: Ila R. Fiete
Title: Professor of Brain and Cognitive Sciences

Acknowledgments

This thesis would not have been possible without the support of many tremendous, wonderful people. I'm so grateful to have had all these people in my life, and would not be here without them. I will now do my best to acknowledge them commensurately.

To my advisors, Earl Miller and Ila Fiete. When I came to MIT five and a half years ago, I was just someone who loved math and was really excited about the intersection of math and neuroscience and consciousness. Despite my passions, I had very little idea of what I was doing or where to go with that. You welcomed me, allowed me the freedom to delve deeply into the questions I was interested in, helped guide me when I was getting lost, and along the way helped shape how I think, approach the world, and how I do science. I am incredibly grateful to have had you both as mentors. You've provided so much wisdom over the years that I will hear your voices in my head when I'm taking on any challenges both in science and in life.

To my committee members: Bob Desimone, Laura Lewis, and Dani Bassett. Thank you for your time, thoughtful advice, and guidance. You've seen several iterations of this work and have helped guide it towards where it is today.

To Leo Kozachkov: I couldn't have asked for a better person to help me navigate the start of my PhD. Beyond setting me on the path of dynamic stability and inducting me into the world of dynamical systems, you've been a wonderful mentor, collaborator, and friend, and I continue to learn so much from you.

To Mitchell Ostrow: thank you for helping to make the slog of doctoral work a lot more fun. I'm continuously inspired by your drive and passion, and I'm so glad that we were able to work together on so many projects over the years.

To Sarthak Chandra: I lost count of how many times I wandered into your office confused about some math problem and came away with a refreshed perspective. Thank you for your patience, guidance, and for helping me develop both of the methods presented in this work.

To Alex Bardon: thank you for being a calm and thoughtful presence amidst the chaos of large scale data analysis and journal reviews. It was a true pleasure to be able to work together.

To Scott Brincat and Jefferson Roy in the Miller Lab: thank you for being an ever-present

source of wisdom and perspective. We are all so lucky to have you in the lab!

To Emery Brown: thank you for your mentorship and for our many thoughtful and grounding conversations. I feel very fortunate to have had the opportunity to learn from you.

To all my other collaborators on these works: André Bastos, Jake Donoghue, Meredith Mahnke, John Tauber, Jesus Ballesteros, and Yumiko Ishizawa; thank you for helping make this science possible.

To my previous mentors and math professors: Mike Yan, Lisa Strug, Lewis Kay, Hue-sun Chan, Yi-Hsuan Lin, Abdol-Reza Mansouri, Serdar Yüksel, Andrew Lewis, Fady Alajaji, and Bahman Ghahesifard; thank you for helping develop my love for science, computing, and a passion for understanding the world with math.

To everyone from the Miller and Fiete labs, past and present: thank you for making the labs such a great place to do a PhD. I'm amazed by how many brilliant people are always around to learn from, and every single time I asked someone for help it was met with interest and thoughtful consideration. I'm thankful I got to work and learn alongside you all.

To the BCS community and my cohort: I have always thought that the fact that everyone here is interested in the mind makes for an amazing community. It's been a pleasure to learn and grow with you all, and a special thanks to the folks from my cohort. Sharing in this journey together has truly made it feel lighter.

To my friends from Boston and Cambridge: I feel so fortunate to have found such a beautiful community here and I'm so grateful for all the adventures and antics. To say I am sad to be leaving Boston is an understatement, but I know that I will now always feel like I have a home here with you all.

To my wonderful friends from Canada: thank you for supporting me unconditionally. It wasn't easy moving away from the community I grew up with, but the last few years have really emphasized how strong connections can continue to grow over distance.

To the Schachter family in Boston, Asher, Leora, Noah and Audrey: I'm very lucky to have had such a wonderful family here locally for support. It made a huge difference to know that help was nearby if I needed it, and that I could come by for shabbat dinner or passover or any other time and be welcomed with open arms.

To my partner, Agata Pietrzak: you've been an unwavering support over the past four and a half years. A few years ago as I sat frantically coding at my desk, Agata coined the term "crazy scientist mode", which I think is the state I've existed in for the last few months at least. But that never really fazed Agata, and whether in Boston where we met, or in Tampa, where we now spend a lot of time, she's never stopped believing in what I was capable of doing. Fortunately, we still broke out of crazy scientist mode often enough to have a ton of fun together these last few years too.

To my family, my parents Paul and Aviva Eisen, my sister Sarah Eisen and her partner Maurice Nayman: thank you for being a source of continuous support and encouragement. Even when I wanted to be a professional jazz musician, my parents unwaveringly supported me, and when I moved from that to engineering to applied math to neuroscience, they didn't miss a beat. They have truly enabled me to become the most authentic version of myself. Everything that I am today is because of their love and support.

To my grandma Judy Schachter and my nana Joyce Eisen, who attended my thesis defense remotely from Toronto, as well as my papa Abe Eisen and grandpa Harry Schachter, who are no longer with us: I'm continuously inspired by the amazing things you have both overcome and accomplished in your lives. For your entire lives you have embodied a passion for learning, knowledge, and the importance of using your aptitudes to help leave the world better than how you found it. I know I will always feel your presence and guidance in my life.

Finally, I want to dedicate this thesis to my grandfather, Dr. Harry Schachter, who was a professor of biochemistry and medical doctor at the University of Toronto. Every weekend, I would sleep over after Friday night shabbat dinner, and on Saturday morning my grandfather would teach me everything he knew about science, math, history, music, and anything else. He often used to say "the brain is the last frontier", and I think he may still be right about that.

Contents

<i>List of Figures</i>	16
<i>List of Tables</i>	19
Introduction	21
1 Propofol anesthesia destabilizes neural dynamics across cortex	24
1.1 Abstract	25
1.2 Introduction	25
1.3 Results	27
1.3.1 Dynamical systems approach: DeLASE method	27
1.3.2 DeLASE tracks changes in stability in simulated neural networks from partial observations	31
1.3.3 Delayed linear dynamical systems models capture neural dynamics . .	31
1.3.4 Propofol anesthesia destabilizes cortical neural dynamics	33
1.3.5 Destabilized dynamics explain sensory responses in anesthetized cortex	35
1.3.6 Sensory-evoked trajectories are consistent with destabilized dynamics under anesthesia	37
1.3.7 Increasing inhibition in random recurrent neural networks destabilizes them	39
1.4 Discussion	40

2	Similar destabilization of neural dynamics under different general anesthetics	43
2.1	Abstract	43
2.2	Introduction	44
2.3	Results	47
2.3.1	All three anesthetic agents destabilize neural dynamics	47
2.3.2	Slower recoveries to sensory perturbation matched decreases in dynamic stability	51
2.3.3	Dynamic stability is reflected across lower frequencies	53
2.4	Discussion	54
2.4.1	Possible mechanisms of destabilization	55
2.4.2	Slow timescales and low-frequency power during anesthesia	55
2.4.3	Connection to other signatures of anesthetic-induced unconsciousness	56
2.4.4	Unique dynamics under ketamine	57
2.4.5	Stability tracks fluctuating recovery under dexmedetomidine	57
2.4.6	Limitations of the study	58
2.4.7	Future directions	59
2.4.8	Conclusion	59
3	Characterizing control between interacting subsystems with deep Jacobian estimation	60
3.1	Abstract	60
3.2	Introduction	61
3.3	Related work	63
3.4	JacobianODE: learning Jacobians from data	64

3.4.1	Jacobian linearization	64
3.4.2	Parameterizing differential equations via the Jacobian	65
3.4.3	Loss functions	66
3.5	Jacobian estimation in dynamical systems	67
3.6	Control-theoretic analyses in a task-trained RNN	68
3.6.1	Characterizing control between subsystems with Jacobian linearization	68
3.6.2	Estimating the Jacobian of a task-trained RNN	70
3.6.3	Reachability in the task-trained RNN across contexts	71
3.6.4	Controlling the task-trained RNN	72
3.7	Discussion	73
4	Disruption of interareal control during propofol anesthesia	75
4.1	Abstract	75
4.2	Introduction	76
4.3	Jacobian-based directional control estimation	77
4.3.1	Jacobian-based directional control estimation in fully observed dynamical systems	79
4.3.2	Deep Jacobian estimation: learning Jacobians from data	80
4.3.3	A latent JacobianODE pipeline for partially observed dynamical systems	81
4.3.4	Conditional training	83
4.4	Results	84
4.4.1	Validation on synthetic systems with ground-truth dynamics	84
4.4.2	JacobianODE models capture neural dynamics and show more nonlinear geometry in awake state relative to anesthesia	86

4.4.3	Broad-spectrum destabilization during propofol anesthesia	86
4.4.4	Propofol anesthesia reduces directional reachability and controllability	87
4.4.5	Anesthesia reduces coupling strength and local coupling dimensionality	89
4.4.6	Changes in reachability and controllability are driven by multi-factorial anesthesia-induced effects	90
4.4.7	Anesthesia disrupts the hierarchy of cortical organization into senders and receivers	90
4.5	Discussion	91
4.5.1	Future directions	92
5	General discussion	93
5.1	Main contributions	93
5.2	Limitations and future directions	94
5.3	Closing remarks	95
A	Methods: Propofol destabilizes neural dynamics across cortex	97
A.1	Experimental Model Details	97
A.2	Method Details	98
A.2.1	LFP Preprocessing	98
A.2.2	Dynamics and stability estimation approach	99
A.2.3	Picking hyperparameters	102
A.2.4	Implementation of VAR	102
A.2.5	Estimating stability in simulated systems	104
A.2.6	Linear dynamics	104
A.2.7	Randomly connected RNN dynamics	105

A.2.8	Linear dynamics with inputs	106
A.2.9	Neural trajectories of stimulus responses	107
A.2.10	Increasing inhibition in RNNs	109
A.2.11	Example dynamical systems	110
A.2.12	Pharmacokinetics analysis	111
A.3	Quantification and Statistical Analysis	113
A.3.1	Prediction quality of dynamical models	113
A.3.2	Changes in stability in neural dynamics	114
A.3.3	Distances between stability distributions	115
A.3.4	Changes to characteristic root frequencies	116
A.3.5	Destabilization in simulated networks	116
B	Methods: Similar destabilization of neural dynamics	118
B.1	Experimental Model Details	118
B.1.1	Anesthetic protocols	118
B.2	Method Details	120
B.2.1	LFP recording and preprocessing	120
B.2.2	Dynamic stability analysis: DeLASE framework	121
B.2.3	Session epochs	129
B.2.4	VAR and baseline comparisons	129
B.2.5	Cross-window prediction quality analysis	131
B.2.6	Identifying unresponsive windows	132
B.2.7	Resting state stability analysis	132

B.2.8	Correlation with physiological markers	132
B.2.9	Surrogate data generation and analysis	134
B.2.10	Single-area stability and similarity	134
B.2.11	AIC-weighted stability estimates	135
B.2.12	Robustness across % of characteristic roots	136
B.2.13	Stability analysis with different windows	136
B.2.14	Responsiveness-tracking in dexmedetomidine	137
B.2.15	Sensory response and autocorrelation analysis	137
B.2.16	Power-spectral analyses	140
B.3	Quantification and Statistical Analysis	140
B.3.1	Testing session independence	140
B.3.2	Changes in stability in neural dynamics	141
B.3.3	Effect sizes for changes in stability	141
B.3.4	Estimating confidence intervals	142
B.3.5	Post-hoc statistical power analysis	142
B.3.6	Comparing autocorrelation values	143
B.3.7	Autocorrelation-instability correlations	143
B.3.8	Band power-instability correlations	144
B.3.9	Multiple comparison in time-resolved analyses	145
B.3.10	Comparing mean band power	145
C	Methods: Characterizing control between interacting subsystems	146
C.1	JacobianODE technical details	146

C.1.1	Parameterizing the time derivative via the Jacobian	146
C.1.2	Path integrating to generate predictions	148
C.1.3	Loop closure loss	149
C.2	Control-theoretic analysis details	149
C.2.1	Gramian computation	149
C.2.2	Extension to non-autonomous dynamics	151
C.2.3	Extension to more than two subsystems	151
C.3	Supplementary results	152
C.3.1	Control and communication capture different phenomena	152
C.3.2	Derivative estimation is not implied by function estimation	152
C.3.3	Full benchmark dynamical systems results	153
C.3.4	Ablation studies	153
C.3.5	NeuralODEs achieve improved performance at the cost of increased inference time	158
C.3.6	Comparison to echo state networks	160
C.3.7	Analyzing the dimensionality of Jacobian manifolds	161
C.4	Experimental details	161
C.4.1	Dynamical systems data	161
C.4.2	Task-trained RNN	163
C.4.3	NeuralODE details	163
C.4.4	Weighted linear Jacobian details	164
C.4.5	Model details	164
C.4.6	Lyapunov spectrum computation	165

C.4.7	Iterative Linear Quadratic Regulator (ILQR)	165
C.4.8	Training details	165
C.4.9	Information about computing resources and efficiency	169
C.4.10	Statistical details	169
C.4.11	Derivation of loop closure loss bound	170
D	Methods: Disruption of interareal control during propofol anesthesia	173
D.1	Supplementary results	173
D.1.1	Linearity of the learned dynamics and linear baseline analysis	173
D.1.2	Interareal control reorganization across integration horizons	177
D.2	Proofs	178
D.2.1	Jacobian-parameterized time derivative	178
D.2.2	Invariance of the pairwise reachability Gramian trace under isometric direct-sum diffeomorphism	180
D.2.3	Bounded error under generalized teacher forcing	181
D.3	Modeling details	183
D.3.1	Diffeomorphic encoder	183
D.3.2	JacobianODE-parameterized dynamics	187
D.3.3	Loop-closure loss	189
D.3.4	Gramian computation	190
D.3.5	Lyapunov exponent computation	193
D.3.6	Training details	193
D.3.7	Hyperparameter selection	194
D.4	Experimental details	197

D.4.1	Dynamical systems data	197
D.4.2	Non-human primate electrophysiology	199
D.4.3	Selection of resting-state windows	200
D.5	Statistical analyses	201
D.5.1	Sign-agreement gate	201
D.5.2	Pooled signed-rank test	201
D.5.3	Benjamini–Hochberg correction across the 4×4 grid	201
D.6	Per-pair statistical tables	202
D.7	Per-area Lyapunov: leading exponent and within-area spectra	205
D.8	Per-NHP session bookkeeping	206
D.9	Software and reproducibility	206
	<i>References</i>	207

List of Figures

1.1	Introduction: stability and instability, and hypothesis candidates	26
1.2	DeLASE: measuring stability through linear delayed dynamical models . . .	29
1.3	Validating DeLASE on simulated and empirical neural data	32
1.4	Propofol anesthesia destabilizes cortical neural dynamics	34
1.5	Destabilized dynamics explain sensory responses to tones and puffs during anesthetic unconsciousness	36
1.6	Low-dimensional sensory-evoked trajectories are consistent with destabilized dynamics under anesthesia	38
1.7	Increasing inhibition in random recurrent neural networks destabilizes them .	40
2.1	A dynamical systems approach to understanding neural dynamics during anesthesia	45
2.2	Propofol, ketamine, and dexmedetomidine destabilize neural dynamics	48
2.3	Destabilized dynamics explain sensory responses to tones during anesthetic- induced unconsciousness	52
2.4	Changes in dynamic stability during anesthetic unconsciousness are reflected across the low frequency spectrum	54
3.1	Schematic overview of control-theoretic framework applied to neural interactions	62
3.2	Analytical framework for pairwise interacting subsystem control	64
3.3	Jacobian estimation with JacobianODE models	65
3.4	JacobianODE surpasses benchmark models on chaotic dynamical systems . .	69

3.5	JacobianODE accurately infers trained RNN Jacobians	70
3.6	JacobianODE reveals differential interareal reachability	72
4.1	A control theoretic framing of interareal control.	78
4.2	Latent JacobianODE pipeline for partially observed multi-area neural recordings	80
4.3	Validation on synthetic systems with ground-truth dynamics	85
4.4	Trajectory prediction, encoder geometry, and Lyapunov spectra under propofol	88
4.5	Directional reorganization of interareal control under propofol	89
A.1	Core results hold across hyperparameter combinations	103
A.2	Dimensionality reduction across delay embedding hyperparameters	108
A.3	Extended results demonstrating propofol destabilizes neural dynamics	112
A.4	Neural dynamics prediction, separated by NHP	113
A.5	Visual depiction of section definitions used for statistics computation	115
A.6	p-values for statistical comparisons between the stability of RNNs with scaled inhibitory weights	117
B.1	DeLASE stability estimates are robust to hyperparameter choices	123
B.2	Model validation, behavioral confirmation, and supplementary stability analyses	125
B.3	Validation of DeLASE, robustness checks, and regional destabilization results	130
B.4	Dynamic instability tracks real-time responsiveness	133
B.5	Supplementary sensory and spectral analyses	138
C.1	Communication versus control in linear systems and the challenge of Jacobian estimation	153
C.2	Full dynamical systems prediction results	154

C.3	JacobianODEs achieve highly efficient Jacobian inference	159
D.1	Linearity diagnostics and linear delay-embedded baseline	174
D.2	Interareal control summaries across integration horizons	178

List of Tables

3.1	Mean Frobenius norm error on Jacobian estimation	68
B.1	DeLASE model hyperparameters selected for each recording session	127
B.2	Linear mixed-effects model results assessing instability changes and session independence.	141
B.3	Post-hoc statistical power analysis results	142
B.4	Statistical results for the correlation between dynamic instability and sensory response autocorrelation	143
B.5	Statistical results for the correlation between dynamic instability and spectral power	144
C.1	Trajectory prediction metrics (MSE and R^2) for each system and noise level	155
C.2	Jacobian estimation metrics (MSE and R^2) for each system and noise level .	156
C.3	Mean 2-norm error on Jacobian estimation	157
C.4	Mean Frobenius norm error for different model ablations on the Lorenz system	157
C.5	Mean Frobenius norm error for different model ablations on the task-trained RNN	158
C.6	Mean Frobenius norm error for different model types on the task-trained RNN data across observation noise levels	158
C.7	Hyperparameters used and model details for each system and noise level . .	169
C.8	Trajectory training time for each system and noise level	170

D.1	Per-session sweep grid	195
D.2	Fixed architecture and training settings	195
D.3	Per-session model-selection criteria	196
D.4	Sample-systems run identifiers	196
D.5	Per-session chosen hyperparameters	197
D.6	Per-pair reach $\log \text{tr } \mathbf{W}$	202
D.7	Per-pair reach $\log \text{tr } \mathbf{W}^{-1}$	203
D.8	Per-pair ctrl $\log \text{tr } \mathbf{W}$	204
D.9	Per-pair ctrl $\log \text{tr } \mathbf{W}^{-1}$	204
D.10	Per-pair Jacobian block Frobenius norm	205
D.11	Per-pair Jacobian block participation ratio	205
D.12	Per-area Lyapunov leading exponent	206
D.13	Per-animal session and window counts	206

Introduction

“Consciousness is everything you experience. It is the tune stuck in your head, the sweetness of chocolate mousse, the throbbing pain of a toothache, the fierce love for your child and the bitter knowledge that eventually all feelings will end.”

– Christof Koch¹

To understand consciousness is to understand the full range of humanity, all of the beauty, pain and mundanity that humans are capable of experiencing. However, this vast, all-encompassing nature of consciousness means that understanding its origins is an extraordinarily challenging problem.^{2–4} Accordingly, the pursuit to understand consciousness has vexed philosophers (and much more recently, psychologists and neuroscientists) since at least the time of ancient Greece and the late Vedic period.^{5–10} The problem can broadly be separated into understanding two kinds of consciousness: “access consciousness” (the availability of information for use), and “phenomenal consciousness” (experience itself),^{11,12} though even this distinction is not without controversy.¹³ Much deep thought has gone into creating theories of consciousness capable of capturing even a small part of its manifold complexities.^{14–24}

In this thesis, I do not present a theory of consciousness. Instead, I aim to rest on the shoulders of mathematical giants, and leverage the ensuing tools to characterize what changes in the brain between consciousness and unconsciousness. Thus the approach in this thesis might be best described as a data-driven approach to understanding consciousness.

The brain is a complex system that changes over time. As such, it is subject to the laws of dynamical systems theory, which captures and characterizes how systems change over time.²⁵ Having existed for centuries, dynamical systems theory has developed a rich and rigorous foundation in which to situate an understanding of the world. The theory has been applied across a huge range of systems, including a long history of use in neuroscience.^{26–36}

We will focus on two essential concepts in dynamical systems theory: dynamic stability (henceforth, stability) and control. Stability is the robustness of a system to perturbations. Control is the ability of one system to influence the behavior of another system. The two are naturally interrelated, and together constitute a useful toolset for characterizing neural dynamics.

Consciousness rests on the interplay of reliability and flexibility. We are extraordinarily consistent in our cognitive computations (e.g., we would never mistake the coffee mug on our

desk for anything else), and yet remarkably flexible (e.g., we can generate pieces of music that have never been heard before). Our brains are stable enough that we can control our attention, focus, respond, and act appropriately, and yet *unstable* enough that we can drive ourselves to think differently, to reappraise, to be creative, and to imagine. These dynamic attributes are exactly what is captured by the mathematical tools of stability and control.

Armed with this mathematical toolset, we approach an understanding of consciousness by performing a data-driven analysis of neural dynamics. We constructed rigorous approaches for characterizing dynamical features of neural data, we validated our approaches on simulated data, and finally we applied them to neural data and interpret the results.

Crucially, we investigated these dynamical systems concepts in the context of anesthetic-induced unconsciousness. Anesthesia enables the induction of a transition from consciousness to unconsciousness and its reversal, providing a rich setting in which to study changes in neural dynamics. We furthermore studied the universal impacts of three different unconscious-rendering anesthetics: propofol, ketamine, and dexmedetomidine. Our goal is to understand the dynamical mechanisms through which unconsciousness is induced by anesthesia. Through an understanding of how consciousness is disrupted, we seek to gain insight into the neural signatures of consciousness itself.

We found that dynamic stability decreased during propofol, ketamine and dexmedetomidine. We furthermore found that the capacity for interareal control decreased during propofol. This was true both when considering the ability of areas to control each other to produce reliable dynamics, and how well they could drive each other to reach novel states.

A complete and formal understanding of consciousness is currently out of reach. My hope is simply that by leveraging what little certainty can be gained through the elegance of mathematical formalism, consciousness, and by extension ourselves, may become infinitesimally less mysterious.

Organization of the thesis

This thesis is divided into four chapters. In Chapter 1, we introduce a rigorous approach, Delayed Linear Analysis for Stability Estimation (DeLASE), for measuring dynamic stability in partially observed high-dimensional nonlinear systems. We applied DeLASE to neural activity from two non-human primates before, during, and after induction into unconsciousness by propofol anesthesia. We found that neural dynamics were destabilized relative to the awake state, and that this destabilization was evident in neural responses to sensory stimuli. This chapter is derived from Eisen & Kozachkov [37], which was published in *Neuron* in 2024. In Chapter 2, we extend this analysis to include neural data from ketamine and dexmedetomidine in addition to propofol. We found that neural dynamics were destabilized relative to the awake state in all three anesthetics, pointing to a shared signature of unconsciousness. This

chapter is derived from Eisen et al. [38], which was published in *Cell Reports* in 2026. In Chapter 3, we develop a nonlinear control theoretic framework for pairwise characterization of control between interacting subsystems, and a deep learning framework for learning the Jacobian of dynamical systems from data. We validated it on high-dimensional chaotic systems, and then used it to evaluate interareal control in a task-trained recurrent neural network. This chapter is derived from Eisen et al. [39], which was published as a spotlight paper at *NeurIPS* in 2025. Finally, in Chapter 4, we apply the JacobianODE framework to evaluate interareal control in neural data from propofol anesthesia. We found that anesthesia decreased the ability of areas to control each other, both in terms of driving towards novel states and stabilizing along existing trajectories. This chapter is derived from Eisen et al. [40], which is available as a preprint on *bioRxiv* as of 2026.

Chapter 1

Propofol anesthesia destabilizes neural dynamics across cortex

Authors: Adam J. Eisen^{*}, Leo Kozachkov^{*}, André M. Bastos, Jacob A. Donoghue, Meredith K. Mahnke, Scott L. Brincat, Sarthak Chandra, John Tauber, Emery N. Brown, Ila R. Fiete^{**}, and Earl K. Miller^{**}.

Statement of contribution and publication

*The work presented in this chapter was published in **Neuron** in 2024.*

^{}A.J.E. and L.K. contributed equally to this work.*

*^{**}I.R.F. and E.K.M. contributed equally to this work.*

Conceptualization, A.J.E. and L.K.; methodology, A.J.E., L.K., S.C., J.T., and I.R.F.; simulations and data analysis, A.J.E.; macaque surgeries, recordings, and data preprocessing, A.M.B., J.A.D., M.K.M. and S.L.B.; writing – original draft, A.J.E., L.K., I.R.F. and E.K.M.; writing – review & editing, A.J.E., L.K., A.M.B., J.A.D., M.K.M., S.L.B., S.C., E.N.B., I.R.F., E.K.M.; funding acquisition, E.N.B., I.R.F. and E.K.M.; supervision, E.N.B., I.R.F. and E.K.M.

1.1 Abstract

Every day, hundreds of thousands of people undergo general anesthesia. One hypothesis is that anesthesia disrupts dynamic stability, the ability of the brain to balance excitability with the need to be stable and controllable. To test this hypothesis, we developed a method for quantifying changes in population-level dynamic stability in complex systems, **Delayed Linear Analysis for Stability Estimation (DeLASE)**. Propofol was used to transition animals between the awake state and anesthetized unconsciousness. DeLASE was applied to macaque cortex local field potentials (LFPs). We found that neural dynamics were more unstable in unconsciousness compared to the awake state. Cortical trajectories mirrored predictions from destabilized linear systems. We mimicked the effect of propofol in simulated neural networks by increasing inhibitory tone. This in turn destabilized the networks, as observed in the neural data. Our results suggest that anesthesia disrupts dynamical stability that is required for consciousness.

1.2 Introduction

The pharmacological action and neurophysiological response of propofol is well understood, but how it renders unconsciousness is not. Propofol boosts inhibition through GABA_A receptors and significantly alters cortical dynamics.^{41–46} This could disrupt the cortical communication on which consciousness depends,^{41,42,47} but the exact link to theories of consciousness is not clear. Many theories of consciousness have focused on the representation and network structure involved in integrating information or linking together cortical representations.^{14,15,17–20,48} For example, one prominent theory of consciousness posits that awareness follows from an “ignition” that produces widespread cortical spiking much like a few claps can lead to a whole audience applauding.^{16,48,49} However, overly excitable and unstable states are uncontrollable, indicative of pathological conditions.^{50,51} Thus, we hypothesize that a key factor in consciousness is *dynamic stability*. Brain states should be sufficiently excitable for generation of widespread activity and information integration. But they also need to be controllable and stable, reliably producing the same computations.^{36,52–55}

Stability has long been known to be critical for brain function but early computational work investigated it in the context of convergence to a single state, involving a kind of “freezing” of neural activity.^{56–58} However, normal neural activity is rarely so stationary; rather it constantly evolves through dynamic trajectories.^{34,59} Thus, stability, and hence consciousness, needs to be understood in terms of a dynamic brain.^{36,54} Here, we approach the analysis of anesthetic unconsciousness through the lens of *dynamic stability* (henceforth stability) (Figure 1.1A), a fundamental concept in dynamical systems theory and control. Essentially, dynamic stability is a measure of the robustness of a dynamical system. The system needs to be able to recover from disturbances (e.g., distractions, random fluctuations in activity) to its normal state.

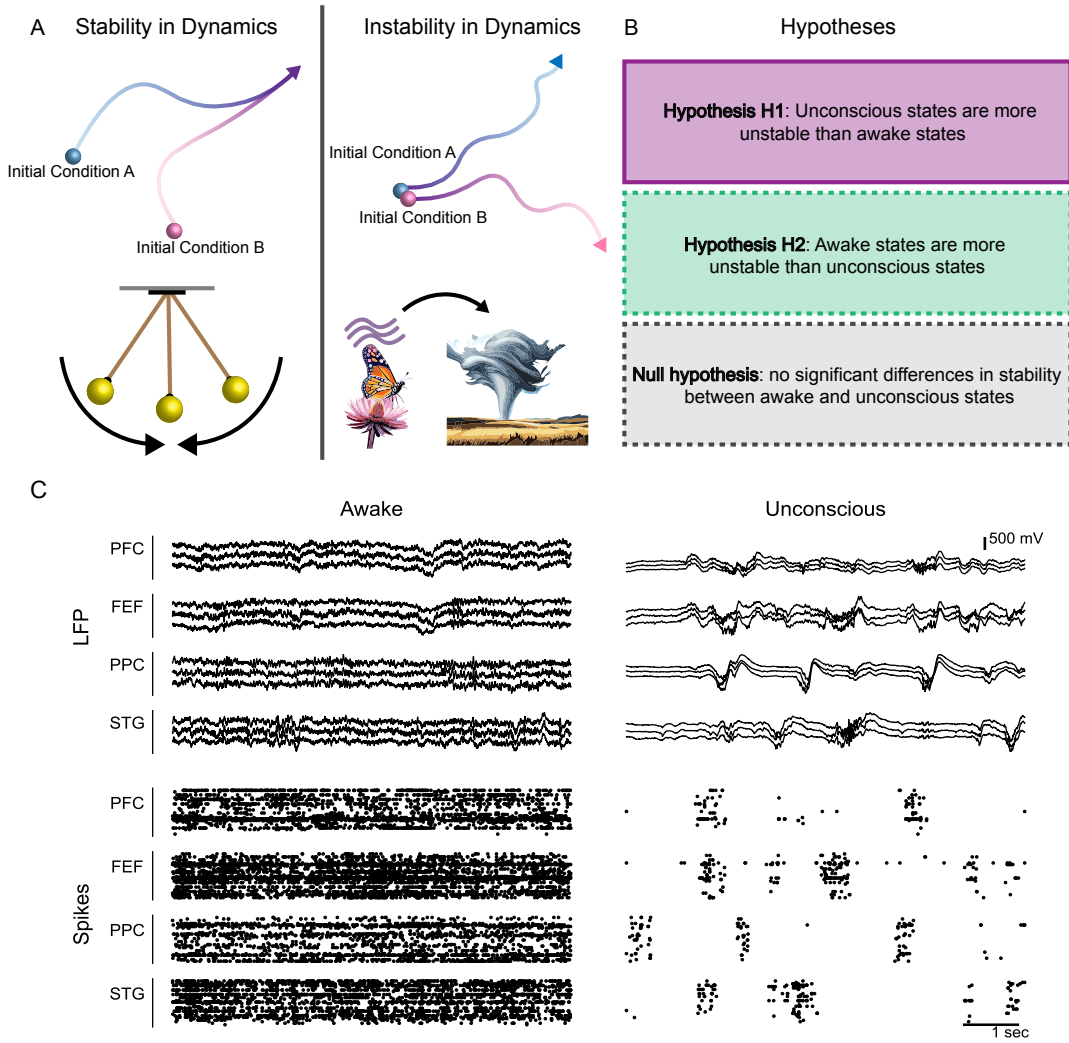


Figure 1.1: Introduction: stability and instability, and hypothesis candidates. (A) (left, top) Depiction of stability in dynamics: starting from two distinct initial conditions, system trajectories converge. (left, bottom) Diagram of a stable system—a pendulum with friction—that will converge to the bottom position regardless of the starting point. (right, top) Depiction of instability in dynamics: starting from two similar initial conditions, system trajectories diverge to distinct paths. (right, bottom) Cartoon of an unstable system—the weather—where a small perturbation, like a butterfly’s wings flapping, may cause a large-scale change such as a tornado.⁶⁰ (B) Three hypotheses regarding the impact of propofol anesthesia on neural dynamics: dynamics can be more unstable, more stable, or show no significant change compared to awake dynamics. (C) Sample neural data from the propofol dataset: (top row) LFPs during the awake (left) and unconscious (right) states. (bottom row) Spike rasters during awake (left) and unconscious (right) states. In the awake state, LFP signals are lower amplitude with higher frequency activity, and spiking is consistent without coordinated bursting. In the unconscious state, LFP signals display low-frequency, hypothesized to underlie loss of consciousness.^{41,42} Spiking during the unconscious state shows up-state/down-state bursting patterns.

Previous work on cortical stability during anesthesia has produced contradictory results, suggesting that anesthesia either destabilizes^{61,62} or excessively stabilizes^{63–65} neural dynamics (Figure 1.1B). This could be due to a paucity of studies using high-density intracortical electrophysiology and the inability to therefore apply sufficiently rich dynamical tools to assess stability. Thus, we used a dataset of local field potential (LFP) recordings with hundreds of electrodes from multiple brain regions in two non-human primates (NHPs, specifically adult rhesus macaque monkeys) as they lost and regained consciousness due to propofol anesthesia (Figure 1.1C).

We introduce a new approach - Delayed Linear Analysis for Stability Estimation (DeLASE). DeLASE directly quantifies changes in stability in neural data. We show that this method produces high-quality models of nonlinear circuit dynamics while maintaining the simplicity and tractability of a linear dynamical system. We validated the model’s estimates of changes in dynamic stability in systems for which the ground truth stability is known. We found that propofol-induced unconsciousness is associated with destabilized neural dynamics.

1.3 Results

1.3.1 Dynamical systems approach: DeLASE method

There are three major challenges to a dynamical systems analysis of neural data.

- a) **Partial observation:** Neural dynamics are high-dimensional. Any one sample of neural data can only hope to capture a few of the dimensions of the overall system. This means we are restricted to partial observation.
- b) **High-dimensionality:** These partial observations are nonetheless high-dimensional. Electrophysiology can employ hundreds of electrodes and a high LFP sampling rate. This generates thousands of neural data points every second. Therefore, analysis of neural dynamics requires a computationally efficient model-fitting procedure to handle vast amounts of data.
- c) **Nonlinearity:** Neural dynamics are highly nonlinear. Nonlinear models are computationally intensive. Any model must be optimized for computational tractability while still capturing nonlinear dynamics.

In this section, we propose a new method that leverages advances in data-driven dynamical systems analysis with strong theoretical guarantees to solve each of these challenges.

Partial observations

Takens’ Delay Embedding Theorem is a powerful result from dynamical systems theory which shows how the full attractor of a dynamical system can be reconstructed from partial observation of data by appending lagged copies of the subset of observed variables onto the existing measurements.⁶⁶ In other words, remarkably, it is possible to trade time (multiple time observations of a subset of variables) for space (the full attractor of the system) to characterize a nonlinear dynamical system. As an example of this delay embedding principle,^{67–78} the famous “butterfly attractor”⁶⁰ of the three-dimensional Lorenz system can be reconstructed from observing a single dimension over the full time period (Figure 1.2A). The key takeaway is that information is gained through considering the trajectory history of a system, in conjunction with its current state. As we hope to convey in this paper, this has tremendous implications for neural data, which is likely capturing a very small fraction of the information contained in the full neural system. Delay embeddings have been widely harnessed as a tool for elucidating dynamical insights from partially-observed data,^{79–86} including examples in the context of neural data.^{73,87–92}

The trajectory history also enables reconstruction of the time-resolved dynamics from partial observations.^{88,93–98} For instance, if we observe a snapshot of a pendulum’s position but not its velocity (a partial observation), we cannot predict its next state (Figure 1.2B). By including the position history, the upcoming dynamics of the pendulum become clear. We harness the information provided by the trajectory history of neural data in a similar way to predict future states.

High-dimensionality and nonlinearity

There are many efficient tools for accurate prediction of linear systems. However, neural and neural circuit dynamics are highly nonlinear, and nonlinear prediction is often both challenging and computationally expensive. Here we leverage a second deep insight from dynamical systems, the Koopman operator theory, which shows that a *nonlinear* dynamical system can be represented without approximation as an *infinite dimensional linear* system^{99–101} (Figure 1.2C). The major challenge in practically exploiting this theoretical insight is to find large but *finite*-dimensional representations that allow the Koopman operator theory to approximately hold.

Given that incorporating trajectory history through delay embedding reconstructs the underlying attractor from partial observation, we might surmise that the dimension-expansion from delay embeddings constitutes a reasonable finite-dimensional representation in which the Koopman theory approximately holds.^{67,68,72,73,102–104} An approach that exploits exactly this insight is the Hankel Alternative View of Koopman (HAVOK),⁶⁷ which uses a decorrelated and low-rank representation of the delay embedding matrix (known as eigen-time-delay coordinates) as an embedding space in which to estimate the Koopman operator. The

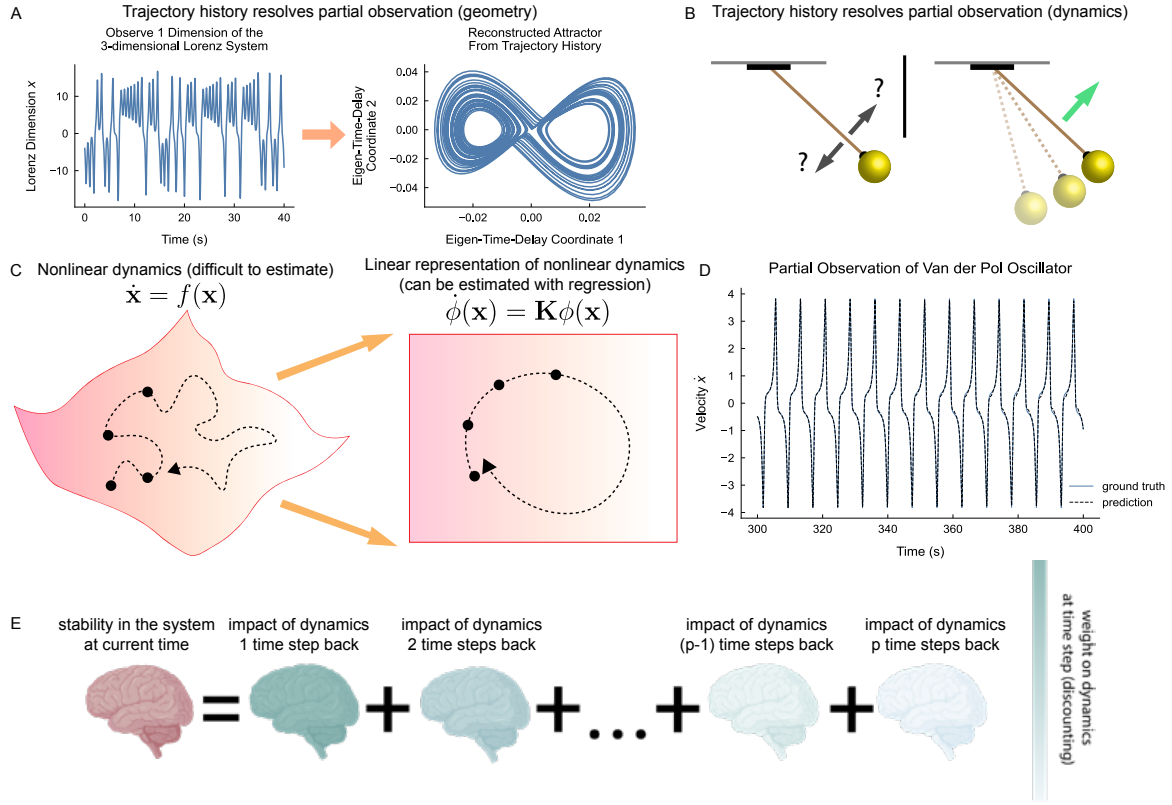


Figure 1.2: DeLASE: measuring stability through linear delayed dynamical models. (A) Trajectory history enables attractor reconstruction from partial observation. (left) With only a single observation of the three-dimensional Lorenz system, it is not clear what the attractor geometry is. (right) However, by performing a delay embedding, the famous butterfly attractor geometry can be reconstructed. We plot the first two eigen-time-delay coordinates of the data, equivalent to performing principal component analysis (PCA) whitening on the delay embedding matrix. (B) Trajectory history enables prediction in partially observed systems. (left) With only the position of a pendulum, no prediction of future states can be made. (right) The trajectory history illuminates the impact of the unobserved variable (the pendulum velocity) on the observed variable (the position). (C) A cartoon depiction of the Koopman operator. A flow on a nonlinear manifold can be expanded (in this case through time-delay embedding) into a high-dimensional embedding in which there exists a linear representation for the dynamics. (D) Dynamical modeling of the partially-observed second dimension of the nonlinear Van der Pol oscillator. The HAVOK model was able to achieve a fully autonomous linear representation of the nonlinear oscillator. (E) A cartoon of the approach for assessing stability from delay differential equations is illustrated. The impacts of modeled dynamics are weighted based on how far back they are from the current state.

resulting decorrelated representation used to estimate the dynamics is relevant for neural data, which can often be highly correlated. Under certain conditions, the HAVOK approach has been shown to find an optimal finite-dimensional space for representing the Koopman operator.⁶⁸ HAVOK is a variant of Dynamic Mode Decomposition (DMD),^{105–107} an approach to estimating the Koopman operator which has been explored in many varieties,^{102,105,108–117} including applications to neuroscience.^{73,115,117–121} To demonstrate the predictive power of HAVOK models, we take a partial observation of the Van der Pol Oscillator (a two-dimensional

nonlinear system).¹²² HAVOK models are able to autonomously reproduce this nonlinear time series with purely linear dynamics (Figure 1.2D). By “autonomously reproduce”, we mean that future predictions are generated by previous ones.

Therefore, we use HAVOK to construct efficient and accurate dynamical models of partially observed neural circuits, resolving the challenges of high-dimensionality and nonlinearity.

Estimating stability from delayed dynamical systems models

We use the accurate dynamical models constructed with HAVOK to estimate the stability of the observed dynamics. We rearrange the dynamics equation, shifting from a Koopman representation in eigen-time-delay coordinates to a delay differential equation in the original neural space. Delay differential equations make explicit the dependence of the future states of the systems on past states. The stability of a system described by a delay differential equation is determined by the roots of its corresponding characteristic equation.¹²³

These roots, known as characteristic roots, are complex-valued numbers. The real part corresponds to the (inverse) characteristic timescale at which the system will respond to a perturbation along a particular direction. The sign in front of the timescale determines whether the system will grow (a positive sign) or decay (a negative sign) in response to a perturbation. For instance, a negative root of -100 s^{-1} will decay quickly in response to perturbation (in this case, with a characteristic timescale of 0.01 s). A negative root of -10 s^{-1} is more unstable than a negative root of -100 s^{-1} , because the response will decay more slowly—namely, with a timescale of 0.1 s. The imaginary part of the characteristic root is the frequency of the perturbation response.

To numerically approximate a finite portion of the (infinitely-many) roots of a given delay differential equation, we harness the TRACE-DDE algorithm.¹²⁴ This algorithm broadly estimates stability by discounting the impact of the previous time steps based on how far back they are from the current time (Figure 1.2E).

DeLASE, our approach to directly estimating changes in stability in neural data, consists of four primary steps:

- 1) Performing a grid search across key HAVOK hyperparameters (the size of the delay embedding, the rank of the eigen-time-delay coordinates) to identify the optimal parameters for all the dynamic states being compared.
- 2) Fitting HAVOK dynamical systems models to the data.
- 3) Using the TRACE-DDE algorithm to extract characteristic roots from the delay differential equations representation of the models as an estimate of stability.
- 4) Compare the estimated stability from each of the dynamic states.

1.3.2 DeLASE tracks changes in stability in simulated neural networks from partial observations

To validate the DeLASE procedure, we attempted to accurately predict changes in stability from partial observations of systems for which the ground truth stability is known.

We began by considering noise-driven linear dynamical systems of varying stability. We picked 10 different degrees of stability of such systems, all of which were close to the transition to unstable dynamics (as awake and unconscious neural states are hypothesized to be).^{52,61,63,125} We altered the degree of stability by varying the maximum eigenvalue of the linear system matrix. For the stability analysis, we randomly chose 10% of the dimensions of the system for partial observation and fit HAVOK models. The correlation of the mean instability estimated from DeLASE with the ground truth stability is 1 (Figure 1.3A).

To further validate the DeLASE procedure on more brain-like systems, we considered numerous randomly connected recurrent neural networks (RNNs) with a gain parameter g . The gain parameter scales the synaptic weights and induces a transition to chaos in these networks.²⁷ We ran DeLASE on a randomly chosen partial observation of approximately 1% of the dimensions from each network. DeLASE predicted relative changes in the stability of RNNs with a correlation of 0.993 (Figure 1.3B). For reference, we compare DeLASE to vector autoregression (VAR), a simple linear dynamical systems model, which has been used previously to estimate stability in neural data.^{63,64,126} Both methods perform somewhat comparably on stability estimation where ground-truth is known (i.e., on simulated data). However, on actual neural data, DeLASE consistently outperforms VAR on next time-step prediction. This is explained in the following section.

1.3.3 Delayed linear dynamical systems models capture neural dynamics

Good estimates of stability depend on good models of dynamics. In other words, if we are to trust a statistical method for estimating dynamical variables from data via model fitting, then that model should do a good job of predicting the future data. We chose HAVOK models because they use a history of neural states to construct simpler linear dynamics from a nonlinear system (like the brain). Our first step towards estimating changes in stability in real neural data, therefore, was to confirm that HAVOK was a good model for capturing neural dynamics.

We compared HAVOK to three other models. (1) A persistence baseline model. This model predicts that the neural state at each time step will be identical to the previous time step. For instance, given a time series with values: [1, 4, 5, 7], when predicting the second time-point the persistence baseline is 1 and when predicting the third one, the persistence baseline is 4. Better models of neural dynamics should thus be able to outperform the persistence baseline

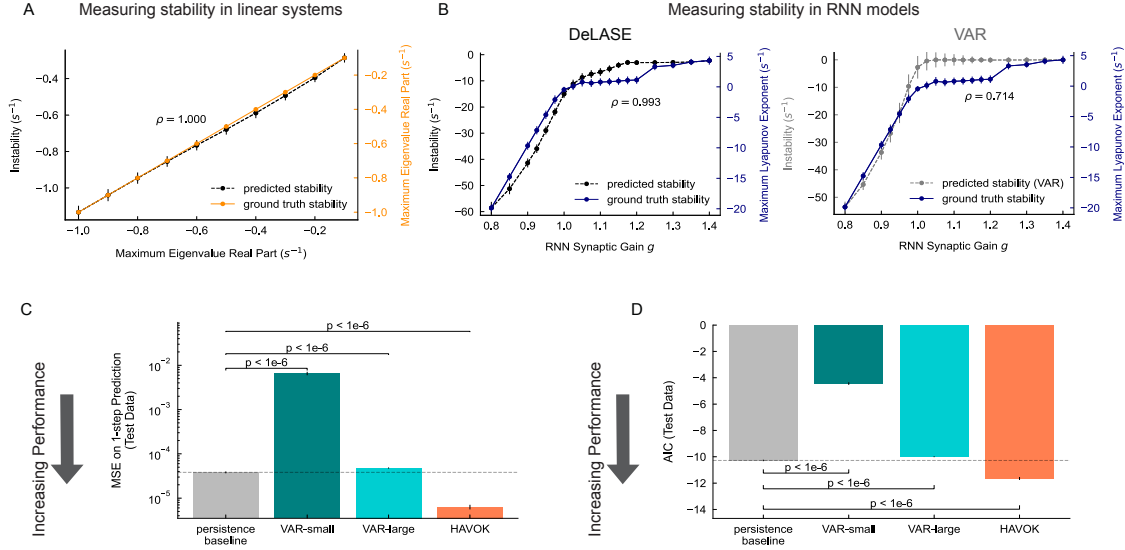


Figure 1.3: Validating DeLASE on simulated and empirical neural data. Data are represented as mean $\pm$ SEM. (A) Linear systems. The x-axis and right y-axis show the maximal real part of the dynamics matrix eigenvalues (orange line). The left y-axis displays instability estimated from DeLASE (black dotted line), using the maximal 10% of the real parts of characteristic roots from delay differential equations analysis (averaged over 10 simulations). DeLASE was fit to 10 out of 100 system dimensions and achieved a Pearson correlation of 1. (B) RNN Models. (left) The x-axis is the gain parameter, which scales synaptic weights and increases network instability. The right y-axis shows the maximum Lyapunov exponent (blue curve), averaged over 10 simulations, determining stability. The full system state and dynamics are needed to compute this in the standard approach we used.¹²⁷ The left y-axis displays instability estimated from DeLASE (black dotted line), with the same approach as (A). DeLASE was fit 10 out of 1024 system dimensions and achieved a Pearson correlation of 0.993. (right) The VAR stability estimation approach is shown, fit to a large temporal window and using the maximal 10% of real parts of eigenvalues. VAR was fit to 10 out of 1024 dimensions and achieved a Pearson correlation of 0.714. Both methods struggle between synaptic gains of 1 and 1.2, where networks transition from quasiperiodic to chaotic dynamics, which are difficult to distinguish.¹²⁸ (C) Comparison of different models' one-step predictions on neural data (LFPs from four recorded areas, sampled at 1 kHz). The persistence baseline, predicting the next state as identical to the previous one, is plotted as the gray bar and dotted line. HAVOK models were the only ones to surpass this baseline in prediction tasks. (D) Similar to (C), but using AIC, showing that HAVOK's superior performance was not due to having more parameters than other models. See also Figure A.4.

model. (2) + (3) Two forms of vector autoregression (VAR) models, which were previously used to study stability in propofol-induced unconsciousness.^{63,64} Here we use 1st-order VAR (VAR(1)) models, which generate predictions for the next state using only the most recent state - but also take into account the dynamics across the training set. One VAR model (VAR-small) used 500 milliseconds of neural data for training, as in previous work.⁶³ The other VAR model (VAR-large) used a 15 second window. VAR-large was included because we used 15 second windows for HAVOK. All models were tested on a window of data of equal size to the training window, and temporally immediately following the training window. HAVOK, in contrast to VAR, predicts future states by taking into account the history of multiple preceding timesteps.

We computed the mean-squared error (MSE) of one-step model predictions averaged over all sessions (Figure 1.3C). The VAR-small models were significantly outperformed by the persistence baseline models ($p < 1e-6$, one-sided Wilcoxon signed rank-test). The VAR-large models achieved an MSE much closer to the persistence baseline but were still outperformed ($p < 1e-6$, one-sided Wilcoxon signed rank-test). Only the HAVOK models were capable of beating the persistence baseline ($p < 1e-6$, one-sided Wilcoxon signed rank-test). Akaike Information Criterion (AIC) was also used to assess the models' prediction quality relative to the number of parameters. AIC penalizes models for complexity and thus guards against overfitting data. HAVOK models have more parameters. Again, only the HAVOK models outperformed the persistence baseline models on one-step prediction (Figure 1.3D, $p < 1e-6$, one-sided Wilcoxon signed rank-test). All results held when considering the prediction quality for each NHP separately (Figure A.4).

Thus, through the inclusion of the history of neural states, HAVOK models form accurate dynamical models of the multi-electrode activity while preserving the tractability and interpretability of linear methods.

1.3.4 Propofol anesthesia destabilizes cortical neural dynamics

To determine the impact of propofol anesthesia on the stability of neural dynamics, we analyzed multi-electrode activity recorded from two non-human primates (NHPs).⁴² Electrodes were placed in four areas: ventrolateral prefrontal cortex, frontal eye fields, posterior parietal cortex and auditory cortex. We found that propofol destabilized neural activity.

We first characterized the stability of every brain region separately using DeLASE (Figure 1.4A). Characteristic roots were used as a quantification of instability. One component is the timescale at which the system will respond to a perturbation along a particular direction (see "Estimating stability from delayed dynamical systems models"). The faster the response, the more stable the system. While we later analyze neural responses to experimental perturbations in the form of stimuli, these perturbations are hypothetical and used to describe the general system response. We analyzed the instability values from the upper 10% of the distribution of characteristic roots (Figure 1.4A). We used the upper 10% because they have the most impact on dynamic stability. For both animals and across all four cortical regions, propofol anesthesia reliably destabilized neural dynamics (blue and gray curves, $p < 0.001$ for all combinations of NHP/area, one-sided Wilcoxon signed-rank test). The same was true for all areas considered together as a single system (purple curves, $p < 0.001$ for both NHPs, one-sided Wilcoxon signed-rank test). Note that for each curve, the instability values increased after propofol induction. This means the system was slower to respond (i.e., less stable).

Note that the absolute values of instability were variable across areas and NHPs. Because of the nature of the instability values, their absolute value will capture details of the differences in signals due to factors such as differences in their intrinsic dynamics, exact location of the

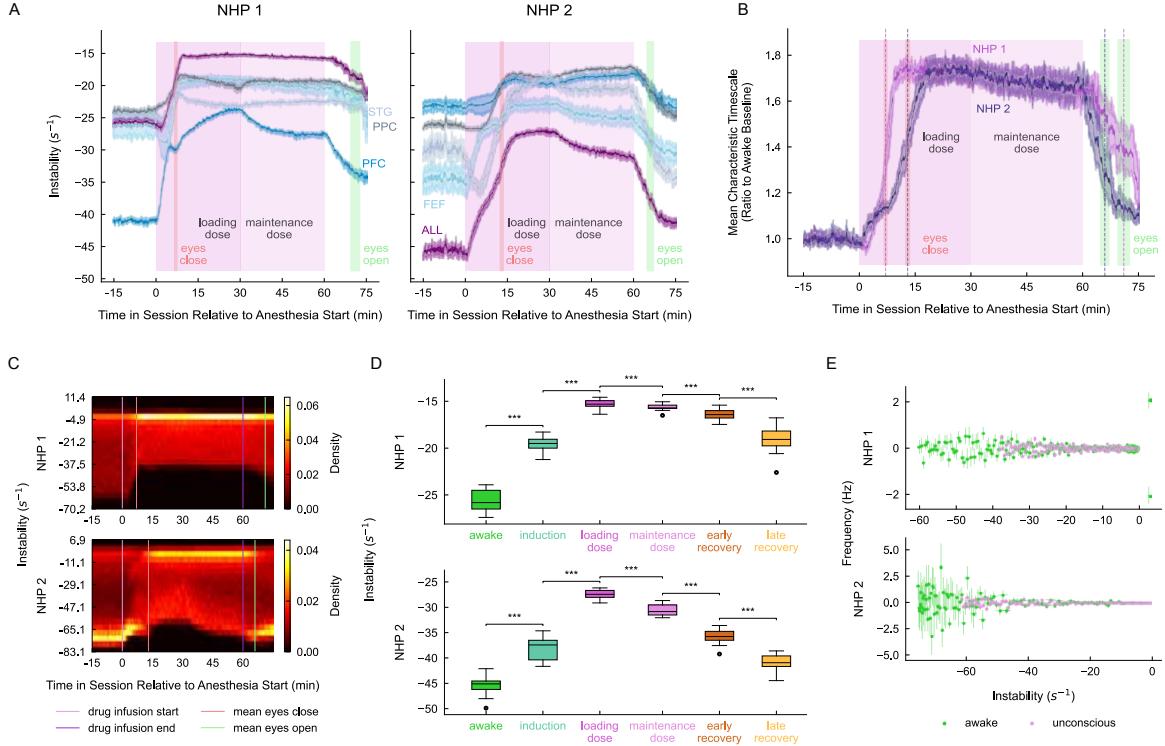


Figure 1.4: Propofol anesthesia destabilizes cortical neural dynamics. All characteristic root analyses refer to the maximal 10% of the distribution of characteristic roots extracted from delay differential equations analysis. All figures (except C) reported as mean $\pm$ SEM. (A) Instability across the session in both NHPs. DeLASE was fit to contiguous windows of 15 seconds across the session for each area individually. The x-axis is time relative to the start of anesthesia. The y-axis is the mean real part of characteristic roots (the inverse timescale of response) - a measure of instability. Instability increased during anesthesia for all areas in both NHPs. The red line indicates the mean moment of eyes closing (used as a proxy for loss of consciousness) and the green line indicates the mean moment of eyes opening (used as a proxy for return of consciousness). Areas are ventrolateral prefrontal cortex (PFC), frontal eye fields (FEF), posterior parietal cortex (PPC), and auditory cortex (STG). "ALL" refers to all areas considered together as a single system. (B) The mean timescales of response normalized to the awake baseline for "ALL" areas considered together. The awake baseline was computed for each session by taking the geometric mean of the timescales associated with the characteristic roots across all windows in the awake section of the session. Then, for each window, the geometric mean of the ratio of the timescales to the awake baseline was computed. After accounting for the baseline awake instability, the degree of destabilization in the timescales in each NHP was very similar. Vertical dotted lines are eyes close/eyes open, with color used to indicate the mean for each NHP. (C) The distribution of instability values across sessions in each NHP. The x-axis is time relative to anesthesia start and the y-axis is the real part of the characteristic roots. Color represents the density of characteristic roots in a particular bin. In both NHPs, the distribution shifted upwards to be more unstable during the unconscious state. (D) The mean instability grouped by section of the session, including at the higher loading dose (0.58 mg/kg/min for NHP 1 and 0.285 mg/kg/min for NHP 2) and the lower maintenance dose (0.32 mg/kg/min for NHP 1 and approx. 0.075 mg/kg/min for NHP 2). For full section definitions, see Appendix A.3.2 and Figure A.5. (E) Frequency information from computed roots. The x-axis is the real part of the characteristic root and the y-axis is the imaginary part of the characteristic root (converted to a frequency in Hz by dividing by 2π). See also Figures A.3, A.1 and A.5.

electrodes relative to the neurons, electrode impedance, etc. The critical measure, therefore, is the change in values over time. When normalized to baseline values, the change in instability across areas differed in magnitude but followed similar time-courses (Figure A.3A). When we considered all recorded areas as a single system, the ratio of change in instability values were remarkably similar across the two NHPs (Figure 1.4B). For the remainder of this paper we focus on models constructed from all areas as a single system (Figure 1.4B). Instability values during unconsciousness were nearly twice that of awake states (pre-propofol), indicating that the system’s responses to perturbation were much slower than the response during the awake state (Figure 1.4B).

To better visualize the change in instability values, we plot the distribution of the top 10% of values at each time point across all areas (Figure 1.4C). The entire distribution of instability values shifted upward during propofol infusion ($p < 0.01$ for both NHPs, Wilcoxon signed-rank test on the Cramér-von Mises criterion of awake-unconscious distributions for each session). This was also true when considering the instability value distribution below the top 10% (Figure A.3B). During recovery, the distribution of values shifted down to be more stable ($p < 0.001$, one-sided Wilcoxon signed-rank test), approaching that seen pre-propofol (Figure A.3C, $p < 0.001$ for both NHPs, one-sided Wilcoxon signed-rank test on the Cramér-von Mises criterion of awake-recovery distributions compared to unconscious-recovery for each session).

Instability changed throughout the session. It rose to a maximum during the large loading dose, then gradually declined through the maintenance dose and finally recovery phases (Figure 1.4D). The estimated integrated dosage of propofol was predictive of instability in the system (Figure A.3D, R^2 of 0.873 for NHP 1 and 0.853 for NHP 2 using a linear-log model).

Next, we considered another measure of system response to perturbation, the frequency of the system response (Figure 1.4E). During unconsciousness, a larger portion of the frequency components fell in the lower frequency bands than during the awake state (Figure A.3E, $p < 0.01$ for both NHPs, Wilcoxon signed-rank test on the proportion of frequencies in the delta band). Conversely, during the awake state, more fell into the highest populated frequency bands (Figure A.3E, $p < 0.01$ for both NHPs, Wilcoxon signed-rank test on the proportion of frequencies in the beta band for NHP 1 and gamma band and NHP 2). This is consistent with observations that propofol causes a large increase in lower frequency power and decrease in higher frequency power in cortex.⁴²

1.3.5 Destabilized dynamics explain sensory responses in anesthetized cortex

In the previous section, we established that propofol measurably destabilizes cortical neural dynamics. We now investigate how this destabilization changes sensory responses. We found that sensory responses under anesthesia are consistent with destabilized linear filters.

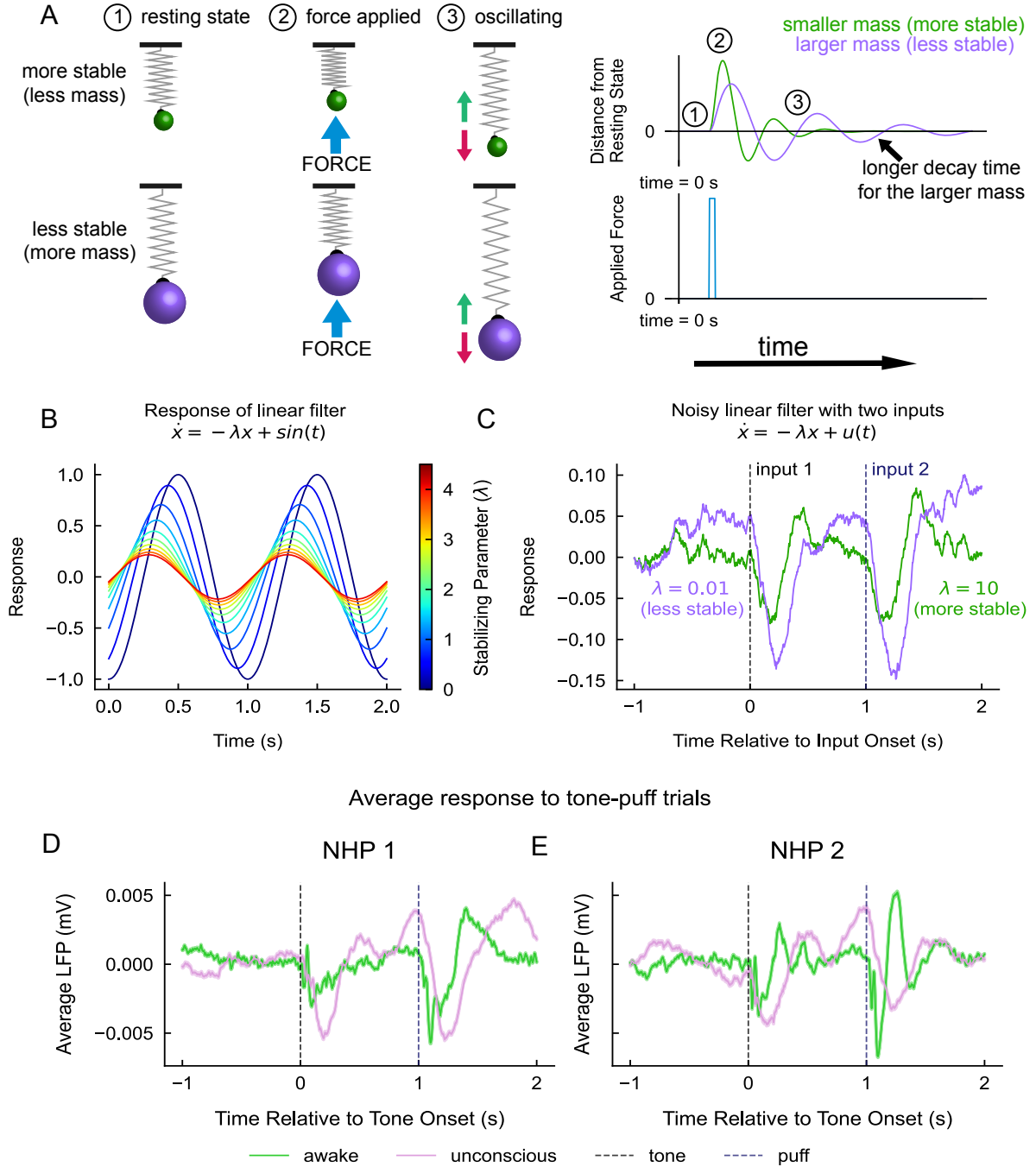


Figure 1.5: Destabilized dynamics explain sensory responses to tones and puffs during anesthetic unconsciousness. Data in (D) and (E) are represented as mean $\pm$ SEM. (A) Mass-spring oscillators with spheres of different masses hanging from identical springs. (1) Spheres perturbed upwards by equal force (2) oscillate back to the resting state (3). Smaller mass (green) decays faster than larger mass (purple). (B) Simulations show that increasing filter decay rate - thus stabilizing the filter - generates a phase shift and amplitude decrease for oscillatory input. (C) Response to two oscillatory inputs and noise for two decay parameters. Less stable (purple) yielded amplitude increase and phase shift relative to more stable (green). (D, E) Mean tone-evoked LFP response for each NHP ((D) NHP 1, (E) NHP 2). Anesthetic unconscious state shows a phase shift relative to awake state, matching destabilized linear filters.

When more stable systems are perturbed, they recover faster than less stable systems. An example of this is two systems in which spheres of different mass are hanging at rest from identical springs (Figure 1.5A). A smaller mass causes the system to be more stable. Thus, when perturbed with identical force, the spring with the smaller mass decays back to its resting state more quickly.

We can describe the pattern of response to perturbation with a simple linear filter of the form $\dot{x} = -\lambda x + u(t)$, where $u(t)$ is a perturbation. In this filter, the parameter λ controls the “stability” of the filter. For larger values of λ , the filter is more stable in that the filter will take less time to recover in response to perturbations. We considered the response of such a filter to a pure sinusoidal input ($u(t) = \sin(t)$) across a wide range of λ values (Figure 1.5B). Smaller values of λ corresponded to larger amplitude, phase shifted responses. To illustrate such responses in a more neural-like setting, we simulated two different linear filters, driven by a small amount of noise (Figure 1.5C). At times 0 and 1, sinusoidal inputs were provided to the linear filter. The purple curve - having a less stable parameter λ - shows a response with a phase shift relative to the green curve, which has a more stable parameter.

This simple model perfectly captured the qualitative change in neural responses to stimuli from the awake state to the unconscious state. Throughout the entire session, tone and puff stimuli were delivered to the NHPs. In our analyses, we focused on “tone-puff” trials consisting of a 500 ms tone at time 0, followed by a 500 ms delay, and then a 10 ms airpuff at time 1. We computed the mean cortical LFP response to the tones and puffs from all sessions (Figure 1.5D,E). There was a phase shift in the sensory responses under anesthesia and the responses are slower than during the awake state. This exactly matched the intuition from the simulated systems: less stable systems decay back to rest more slowly after being perturbed.

1.3.6 Sensory-evoked trajectories are consistent with destabilized dynamics under anesthesia

We already demonstrated that sensory responses to perturbations are slower in unconsciousness compared to the awake state. We now consider how this destabilization manifests in the neural state-space. To gain some intuition for what we might expect, we consider a conceptual diagram of what happens when stable systems with different timescales of response are perturbed (Figure 1.6A). In a system with more stability, a perturbation may cause a brief divergence from a baseline region of state-space (left). In a system with less stability, however, the perturbation induces a prolonged rotational divergence from the baseline region, depicted with slowly decaying oscillations (right).

We simulated two two-dimensional linear systems with different levels of stability (Figure 1.6B). At times 0 and 1, the system was perturbed with constant inputs, matching the structure of the tone and puff stimuli presented to the NHPs. As suggested in the conceptual diagram, the

more stable system (shorter governing timescale) was perturbed and then quickly returned back to the baseline. In contrast, the less stable system (longer governing timescale) responded to perturbations with slowly decaying oscillations.

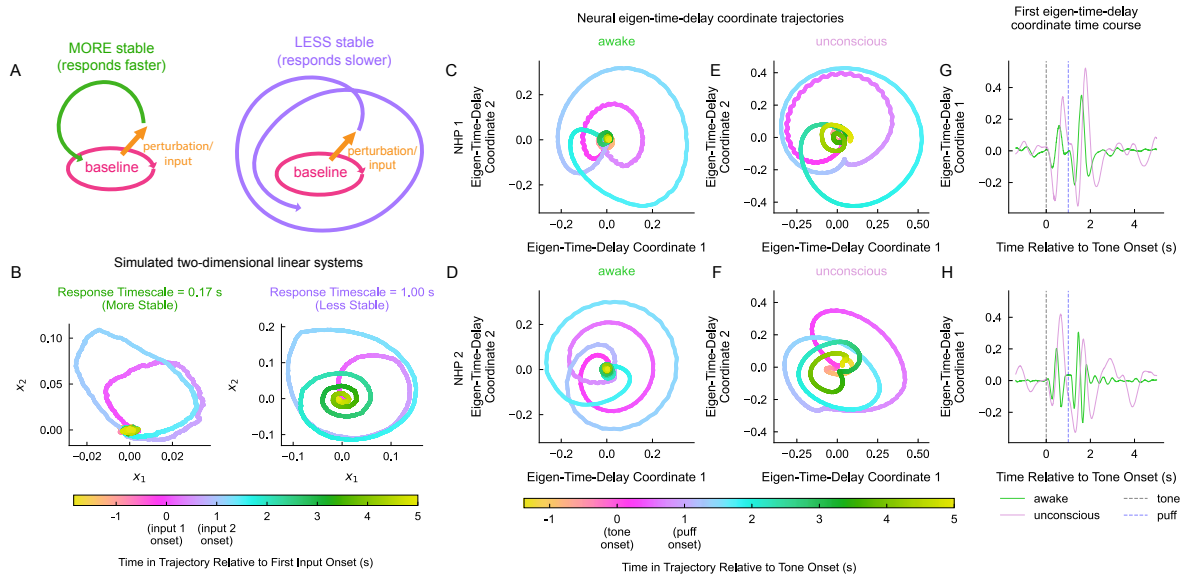


Figure 1.6: Low-dimensional sensory-evoked trajectories are consistent with destabilized dynamics under anesthesia. (A) Cartoon depiction of dynamic responses to perturbation in stable systems. Without perturbation (orange arrow), the system remains in a small state-space region (pink oval). In more stable systems, there is a quick recovery to baseline after perturbation (left, green arrow); less stable systems show slower recovery, potentially with decaying oscillatory dynamics (right, purple arrow). (B) Two-dimensional noise-driven linear systems simulated with two different intrinsic timescales. Systems were perturbed with 500 ms input at time 0 and 150 ms input at time 1, approximately matching the stimuli provided to the NHPs. The systems exhibited different responses based on stability (left: more stable, right: less stable). (C, D) State-space embeddings of the mean responses to tone-puff trials from NHP 1 (C) and NHP 2 (D) during the awake state. Neural responses from tone-puff trials were each delay embedded to illuminate the attractor structure and then averaged. PCA was performed on the mean delay embedded trajectories for visualization (yielding scaled eigen-time-delay coordinates - see Appendix A.2.9). Awake trajectories rapidly decayed in response to stimuli. (E, F) Same as (C, D), during unconscious state. Trajectories decayed slower with oscillatory structure. (G, H) First coordinate from awake and unconscious state tone-puff responses for NHPs. The unconscious response was slower. See also Figure A.2.

We performed dimensionality reductions for subspace coding on the LFPs from all brain areas collected during both the awake and unconscious states. Our dimensionality reduction approach involved first performing a time-delay embedding on neural data from tone-puff trials from all sessions (we use a delay interval of 20 milliseconds and 32 delays). As discussed, delay embedding can help with attractor reconstruction when the available data constitutes a partial observation from a higher-dimensional system. We then averaged each delay embedding coordinate across trials before performing principal component analysis (PCA) to obtain a visualization of neural trajectories in two dimensions. This is equivalent to computing (scaled, i.e. non-whitened) eigen-time-delay coordinates.

We visualized two-dimensional state-space trajectories of neural responses to tone-puff trials in eigen-time-delay coordinates (Figure 1.6C-F). During the awake state, the state-space trajectories display two clear responses to the two stimuli: the response to the tone, at time 0, followed by the response to the airpuff at time 1 (Figure 1.6C,D). These perturbations caused a deviation from the baseline region into other regions of state-space. The system quickly recovered from the perturbation, returning to the baseline state. By contrast, in unconsciousness, the perturbations due to tones and puffs caused prolonged deviations into state-space (Figure 1.6E,F). These state-space trajectories are characterized by a slow oscillatory decay back to baseline, as in the simulation. Furthermore, in the awake state, after being first perturbed by the tone, the system was able to decay back to the baseline region before the puff. However, in unconsciousness, the system did not recover from the tone before the puff was delivered. To highlight the different timescales between states, we also visualized the time course of the first eigen-time-delay coordinate (Figure 1.6G,H). The unconscious curve depicts a slower neural response to stimulus perturbation, as suggested by the results from the previous section. The results of this section were consistent across variations in the delay embedding parameters (Figure A.2).

1.3.7 Increasing inhibition in random recurrent neural networks destabilizes them

We now propose a simple mechanism through which propofol can induce destabilization in neural circuits: increased inhibition.

Propofol is known to act as an agonist at GABA_A inhibitory receptors, thus increasing inhibitory tone. To model the effects of propofol on neural circuits, we alter the connectivity of the randomly connected RNNs. These RNNs include a gain parameter g that scales the synaptic weights and induces a transition to chaos in these networks.²⁷ While it is therefore known that increasing the magnitude of *all* the network weights leads to a destabilization in these networks, the effects of changing *only* the inhibitory weights are unknown (Figure 1.7A,B).

We tested the impact of scaling the magnitude of the inhibitory weights in networks of varying baseline gain (thus impacting the baseline stability of the networks). We simulated the RNN dynamics with a small amount of noise. We then measured the network instability through the maximum Lyapunov exponent, computed using standard methods.¹²⁷ We found that across all baseline connectivity gains, increasing the inhibitory tone in the networks destabilized them (Figure 1.7C,D). Furthermore, decreasing the inhibitory tone in the networks stabilized them ($p < 0.05$ for all comparisons with one-sided Wilcoxon signed-rank test, see Figure A.6).

We emphasize that this is a somewhat surprising result, given that inhibitory connectivity is often thought of as suppressing activity. In fact, it has been suggested that the increased inhibitory tone during propofol anesthesia might shut off regions of the brain during uncon-

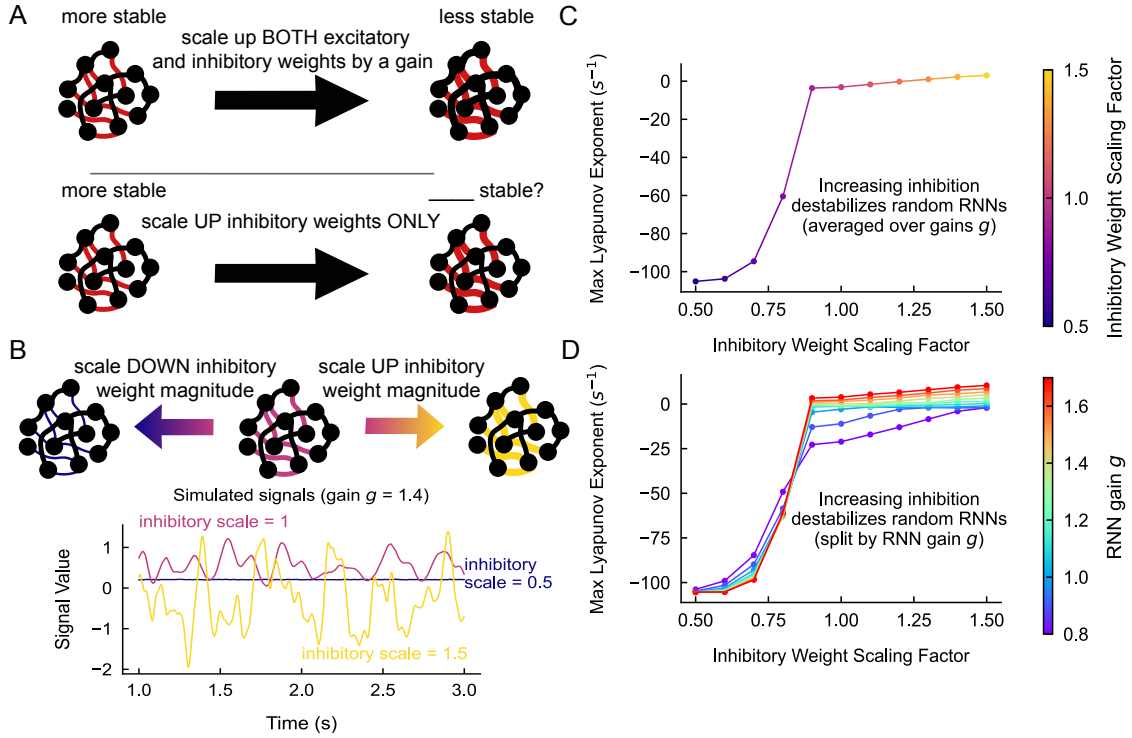


Figure 1.7: Increasing inhibition in random recurrent neural networks destabilizes them. Data in (C) and (D) are represented as mean $\pm$ SEM. (A) The effect of scaling only inhibitory weights in random RNNs is unknown. (B) In an RNN driven by noise, scaling up inhibitory weights led to high amplitude dynamics. Scaling down these weights suppressed amplitudes. (C) Mean instability (maximum Lyapunov exponent) of RNNs for different scaling of inhibitory weights, averaged over (10) simulations and different baseline gains. Increasing inhibitory weight magnitude destabilized networks, decreasing the inhibitory weight magnitudes stabilized them. (D) Same as (C), but split by baseline gain. Increasing inhibitory weights destabilized, while decreasing inhibitory weights stabilized networks. See also Figure A.6.

sciousness.^{46,129} Nevertheless, the demonstrated destabilization of RNNs through increased inhibitory connectivity lends support to propofol's action at GABA_A receptors underlying the observed destabilization and loss of consciousness during propofol anesthesia.

1.4 Discussion

Our results show that propofol anesthesia destabilizes cortical neural dynamics. We found that the stability of brain dynamics was an excellent marker for anesthetic depth, as it smoothly and monotonically varied with the depth of anesthetic state. We then examined neural responses to sensory inputs. We found a longer timescale for recovery from perturbation in unconsciousness, like that seen in destabilized linear systems. We also found that increasing inhibition (as propofol does) in artificial recurrent neural network models produced destabilization.

Propofol disrupts the balance between cortical excitation and inhibition. This balance is known to be critical for maintaining the stability of cortical dynamics.¹³⁰ Combined with our findings, this paints a picture in which propofol tampers with this balance, causing widespread cortical instabilities, and thereby disrupting the brain’s capacity for information processing. Overall, our analysis suggests a mechanism for anesthesia that involves destabilizing brain activity to the point where the brain loses the ability to maintain conscious awareness.

We have also demonstrated the efficacy of a novel approach, DeLASE, designed to directly estimate changes in neural stability from data. The approach brings together multiple aspects of dynamical systems theory, including delay embeddings, Koopman operators, and delay differential equations theory. The linearity of DeLASE enables it to be tractably deployed on ultra-high-volume and high-dimensional neural data. Its theoretically rigorous grounding enables it to provide robust stability estimates despite challenging features of the data such as nonlinearity and partial observation.

The extended response times to stimuli may contribute to the loss of conscious awareness during anesthesia. We discovered that during anesthetic unconsciousness, the response timescales were nearly twice as slow as during the awake state. Therefore, when faced with an input, such as a sensory one, the neural dynamics may not be capable of the synchronization dynamics across areas required to produce conscious awareness. This is consistent with the observation that sensory cortex responses to sensory stimuli are less affected by anesthesia than those in the higher cortex.¹³¹ Though the signal may be present in sensory areas, the full brain - including the higher-level areas thought to be necessary for conscious perception - do not converge to a combined stimulus-guided trajectory fast enough.

We wish to emphasize the use of delay embeddings as part of our methodology for analyzing the nature of neural dynamics across states. Delay embeddings are a widely-used tool known to improve the quality of dynamics and attractor reconstruction when observations are of a much smaller dimension than the dimensionality of the true system,^{66,67,73,84,85,87–89,92–94,102} as is always the case in neuroscience. Delay embedding the neural data was, in our case, not only beneficial for the estimation of dynamics, but also for the visualization of neural trajectories. Patterns in neural response trajectories emerged most clearly when the trajectory history was used to reconstruct the attractor (Figure 1.6, Figure A.2).

Our findings support the hypothesis that neural computations are instantiated by reliable dynamics in neural state-space. Given that computations such as working memory, motor control, and decision making are instantiated by dynamical features such as fixed points and attractors, destabilizing neural dynamics would undermine the ability of neural circuits to perform these computations.^{28,31,34–36,56,59,132–135}

While we focused on the comparison between conscious and unconscious states in this paper, we emphasize that our approach to stability estimation was completely agnostic to the nature of the mental state that generated the neural data. It could thus be applied to a wide variety of data from a range of states. In particular, a compelling potential application is to data from psychiatric and mood disorders. Conditions such as depression, anxiety, substance use

disorder, and schizophrenia can all be characterized as having distorted thinking patterns relative to neurotypical states; distortions which have been hypothesized to arise from changes to the stability landscape.^{52,136–141} Tracking changes in stability in neural dynamics over time for individuals with these conditions could help shape the course of treatment. It could also shed light on the mechanisms of interventions like psychedelics and meditation, which are thought to disrupt overly stable dynamics.^{142–144} Many clinical science studies have tracked changes in physiological measures across time - for example, over the course of treatment,^{145,146} such as neurofeedback.¹⁴⁷ When considered over the course of treatment, changes in these physiological measures were shown to correspond to decreases in symptom severity. The hypothesized deep connection between neural stability and psychiatric disorders suggests that measuring changes in stability could be an excellent approach for monitoring treatment efficacy.

Acknowledgments

This study was supported by The Office of Naval Research N00014-23-1-2768, The National Institute of Mental Health 1R01MH131715-01, The Simons Center for the Social Brain, The JPB Foundation, and The Picower Institute for Learning and Memory. We thank Chandrika Prakash Vyasarayani and Antonio Carlos Costa for many helpful discussions and thoughts.

Chapter 2

Similar destabilization of neural dynamics under different general anesthetics

Authors: Adam J. Eisen, Alexandra G. Bardon, Jesus J. Ballesteros, André M. Bastos, Jacob A. Donoghue, Meredith K. Mahnke, Scott L. Brincat, Jefferson E. Roy, Yumiko Ishizawa, Emery N. Brown, Ila R. Fiete*, and Earl K. Miller*.

Statement of contribution and publication

*The work presented in this chapter was published in **Cell Reports** in 2026.³⁸*

**I.R.F. and E.K.M. contributed equally to this work.*

Conceptualization, A.J.E., I.R.F., and E.K.M.; methodology, A.J.E. and I.R.F.; software, A.J.E.; formal analysis, A.J.E. and A.G.B.; investigation, J.J.B., A.M.B., J.A.D., M.K.M., S.L.B., and J.E.R.; data curation, S.L.B.; writing – original draft, A.J.E., I.R.F., and E.K.M.; writing – review & editing, A.J.E., A.G.B., S.L.B., I.R.F., and E.K.M.; visualization, A.J.E.; supervision, Y.I., E.N.B., I.R.F., and E.K.M.; funding acquisition, Y.I., E.N.B., I.R.F., and E.K.M.

2.1 Abstract

Different classes of anesthetics induce unconsciousness despite acting through distinct molecular mechanisms, raising the possibility that their convergent effects arise at the level of the dynamics of neural population activity. To explore this, we analyze intracortical electrophysiological recordings during infusions of propofol, ketamine, and dexmedetomidine, applying a rigorous method to estimate the stability of population dynamics during anesthesia. We find

that all three anesthetics, despite their molecular differences, similarly affect cortical states by reducing dynamical stability. The estimated destabilization is corroborated by the slower recovery from sensory perturbations and longer stimulus-induced autocorrelation times observed during the anesthetic infusions. Destabilization is most prevalent in the low-frequency band of the population dynamics, linking it to the well-known increase in low-frequency power during anesthesia. Together, these findings suggest that cortical destabilization may be a shared neural correlate of anesthetic-induced unconsciousness and provide a plausible link to the low-frequency oscillations observed during anesthesia.

2.2 Introduction

A remarkable property of anesthetic-induced unconsciousness is that it is achieved by different anesthetics through distinct molecular mechanisms. Anesthesia thus presents an exciting lens through which to explore consciousness,^{148–153} as seemingly disparate low-level mechanisms generate related effects at the level of arousal and awareness.¹⁵⁴

Propofol agonizes GABA_A (γ -aminobutyric acid type A) receptors, changing the balance between excitation and inhibition.^{41,42,45,46,155,156} Ketamine is an N-methyl-D-aspartate (NMDA) antagonist, which acts across the brain to significantly alter cortical dynamics.^{150,157–160} Dexmedetomidine, in contrast, is a highly selective agonist of α_2 -adrenergic receptors (primarily in locus coeruleus). It is thought to alter arousal through a pathway involved in sleep, in which agonizing α_2 -adrenergic receptors results in reduced firing of noradrenergic neurons and subsequently reduced inhibitory output from locus coeruleus.^{161–164} Despite the variety of molecular actions among these three anesthetic agents, at sufficient doses all result in loss of consciousness, and all three evoke stereotyped changes in the frequency spectrum of brain activity, most notably boosting low frequencies.^{42,46,154,163,165–173} Anesthetic-induced unconsciousness has also been widely linked to increased similarity between direct anatomical connectivity and functional connectivity, as well as with the collapse of the hierarchical structure of and variation in the functional organization of neural activity.^{55,174–178} It is unclear how these dissimilar molecular mechanisms can all lead to similar changes in both arousal and dynamical features of neural activity.

To bridge this gap, we take another dynamical systems perspective towards understanding anesthetic-induced unconsciousness, focusing our analysis on *dynamic stability*. Dynamic stability (henceforth stability) is a fundamental concept in dynamical systems theory and control.^{25,179} It is a measure of the robustness of a dynamical system, characterizing the rate at which the system recovers from disturbances (e.g., inputs, distractions, random fluctuations in activity) and returns to its baseline state (Figure 2.1A). Dynamic stability is thought to be closely linked to conscious state,^{37,49,52–54,61–65,180–190} and thus could provide a link between the oscillatory changes and unconsciousness induced by the anesthetic agents.

This notion of *dynamic* stability is distinct from what we might call *pattern* stability. The

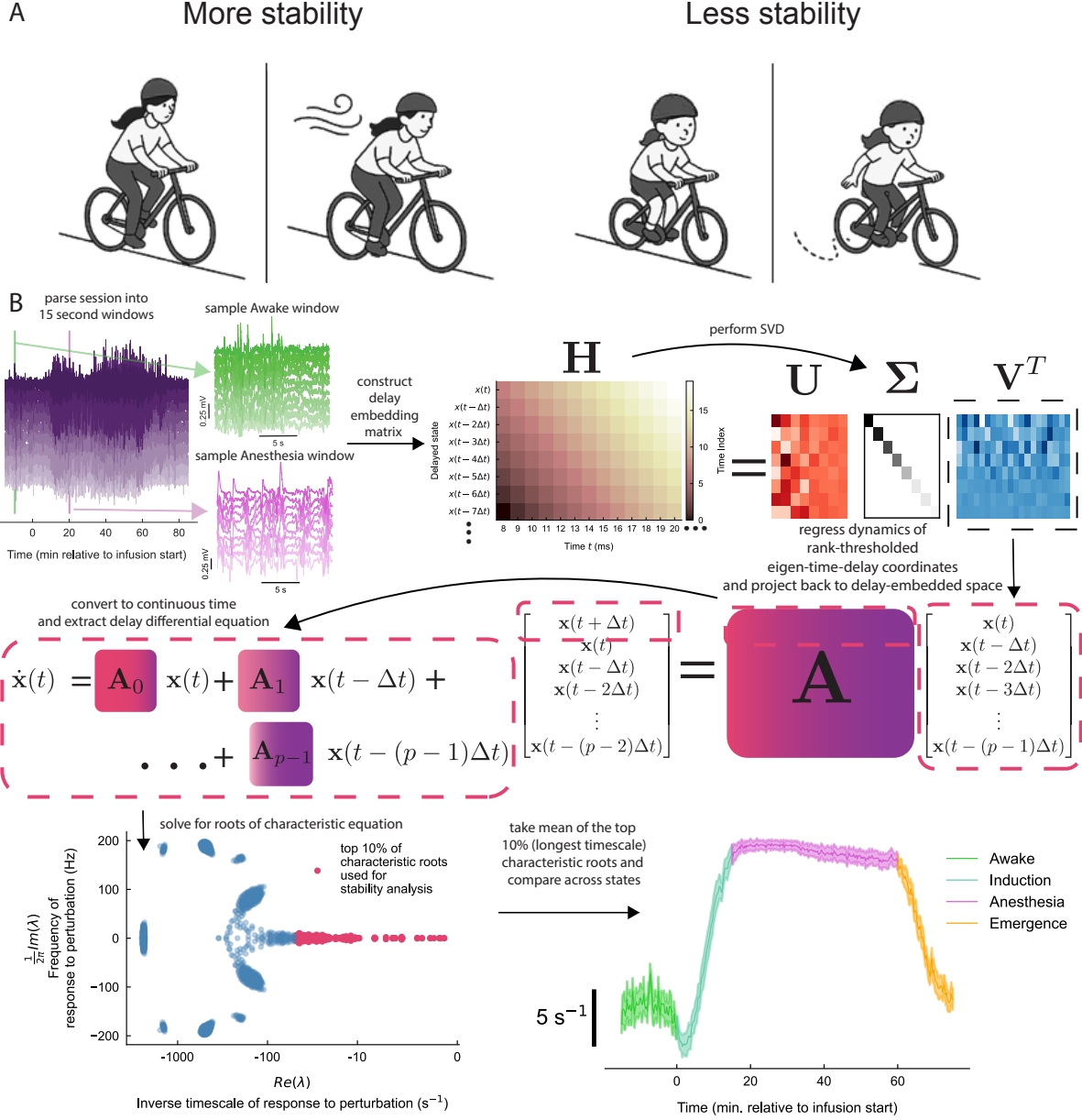


Figure 2.1: A dynamical systems approach to understanding neural dynamics during anesthesia. (A) Conceptual illustration of dynamic stability versus pattern stability. An expert cyclist (left) and a novice cyclist (right) both successfully ride along a straight line. When hit by a perturbation (a gust of wind), the expert (left) quickly corrects course, demonstrating high dynamic stability. The novice (right), however, wobbles and recovers slowly, indicating low dynamic stability. This analogy distinguishes the stable pattern of oscillations observed in the anesthetized brain from its low dynamic stability, which is characterized by a slow recovery from disturbances. (B) Schematic of the DeLASE (Delayed Linear Analysis for Stability Estimation) pipeline applied to electrophysiology data. (Top, left) Continuous LFP data from a session is parsed into 15-second windows. Sample LFP traces are shown for two windows, one from the Awake epoch and one from the Anesthesia epoch. For each window, a time-delay embedding (or Hankel) matrix, $\mathbf{H}$, is constructed from the multi-channel LFP data. For the purposes of demonstration, the color corresponds to the time index of the associated measurement. (Top, right) Singular value decomposition (SVD) is performed on $\mathbf{H}$. This results in the principal components ($\mathbf{U}$), the singular value matrix ($\mathbf{\Sigma}$), and the projections ($\mathbf{V}^T$) onto the principal components, also called the eigen-time-delay coordinates. These data are for illustrative purposes. The color represents the magnitude and sign of the values within the matrices $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}^T$. (Middle, right) A linear dynamical model is fit to the eigen-time-delay coordinates to approximate the system's nonlinear...

Figure 2.1: (Continued from previous page) dynamics. To represent the dynamics in the delay-embedded space, this matrix is projected back up using the principal components and singular values, yielding the dynamics matrix $\mathbf{A}$. (Middle, left) This discrete-time model $\mathbf{A}$ is converted into a continuous-time linear delay differential equation (DDE). (Bottom, left) The stability of this DDE is determined by solving for the roots λ of its characteristic equation. The real part of each root, $Re(\lambda)$ corresponds to an inverse timescale (s^{-1}) of the system’s response to perturbation. (Bottom, right) The mean of the top 10% of roots (those with the largest real parts, indicating the slowest recovery timescales) is computed for each window. This yields a time-resolved measure of dynamic instability that is tracked across the session’s different epochs (awake, induction, anesthesia, and emergence).

latter refers to consistency over time. To illustrate, consider two cyclists, one an expert, one a novice, both successfully riding along a straight painted line (Figure 2.1). Both are exhibiting the same stable pattern of forward motion (namely, moving forward in a straight line). When hit by a gust of wind (a disturbance), the expert quickly corrects course, showing high dynamic stability. The novice, however, wobbles and swerves widely before returning to the line, thus showing less dynamic stability. Similarly, while the anesthetized brain exhibits a stable *pattern* of low-frequency oscillations, this work investigates its *dynamic* stability—how quickly the cortex recovers from disturbances.

We applied a recent rigorous dynamical stability estimation method (delayed linear analysis for stability estimation, or DeLASE), previously applied to examine neural population dynamics during propofol anesthesia, to quantify the *dynamical stability* of cortical activity over time across anesthetics (Figure 2.1B).³⁷ DeLASE consists of two steps: building a good predictive model of the observed data from some nonlinear dynamical system, and then using the model to characterize the stability of the system. The model-building step leverages two discoveries: First, that the state and dynamics of a nonlinear dynamical system can be recovered from partial observations by considering a signal’s history simultaneously at multiple delay lags.^{66,69,70,74–79,81–83,86,87,89–91,93,94,191–195} For instance, if we observe a snapshot of a pendulum’s position but not its velocity (a partial observation), we cannot predict its next state. By including the position history, the upcoming dynamics of the pendulum become clear. This delayed approach to representing the system is also called time-delay coordinate embedding. Second, that it is possible to build a predictive model of a nonlinear dynamical system using linear regression in combination with the time-delay coordinates,^{67,68,73,102,104} as shown by Koopman operator theory.^{71,99–101,105–108,110–114} This is crucial given the brain’s nonlinear dynamics.^{35,134,196–199} DeLASE uses decorrelated, low-rank representations of a signal’s history to estimate the nonlinear dynamics of the data (Figure 2.1B, top).⁶⁷ To obtain stability estimates from the dynamical model, DeLASE first builds a delay differential equation (DDE) representation of the dynamics (Figure 2.1B, middle). It then uses an algorithm that modulates the weight of earlier time delay lags to estimate the solutions of the DDE’s characteristic equation (Figure 2.1B, bottom).^{123,124} Each solution corresponds to a characteristic timescale, which quantifies how slowly the system will respond to a perturbation. This timescale serves as a direct measure of dynamic stability. Longer timescales indicate a slower recovery and therefore less stability.

Through this approach, we seek global signatures of loss of consciousness that arise in the

brain independent of local molecular actions. By characterizing activity changes that are signatures of the high-level transition from consciousness to its absence, we can shed light on the necessary mechanisms of consciousness that are disrupted by anesthetic agents. It was demonstrated previously through the application of DeLASE that propofol destabilizes cortical neural dynamics.³⁷

Using intracortical electrophysiology in non-human primates (NHPs), we found that propofol, ketamine, and dexmedetomidine all induce a similar destabilization of cortical neural dynamics. This destabilization was evident in how the brain responds to sensory perturbations during the unconscious state and was reflected across the low-frequency spectrum. We suggest that the stereotypical increase in low-frequency power during anesthetic-induced unconsciousness reflects the brain’s slower recovery times in the more unstable state.

To validate our findings, we provide extensive supplementary analyses. Figure B.2 details model validation and comparison, behavioral analyses, and extended stability analyses. Figure B.3 provides comparison to alternate stability estimation methods, surrogate analyses, and stability analyses within individual areas. Figure B.1 assesses robustness across hyperparameters. Figure B.4 investigates the correspondence between dynamic stability and real-time responsiveness. Figure B.5 focuses on spectral analyses. Finally, Tables S1-S5 provide details on statistical analyses and hyperparameters.

2.3 Results

2.3.1 All three anesthetic agents destabilize neural dynamics

The LFPs were recorded from four NHPs before, during and after infusion of each anesthetic in different sessions, with two NHPs administered each anesthetic. For the propofol sessions, electrodes were placed in the ventrolateral prefrontal cortex, frontal eye fields, posterior parietal cortex, and auditory cortex. For the ketamine and dexmedetomidine sessions, electrodes were placed bilaterally in the ventrolateral and dorsolateral prefrontal cortex. We used all valid electrodes for our analysis.

To quantify the stability of cortical dynamics, we applied DeLASE to the recorded local field potentials (LFPs). This method quantifies stability through the system’s characteristic timescale. The Mean Characteristic Timescale (y-axis) is a measure of dynamic stability, representing the average time it takes for neural dynamics to return to baseline after a transient perturbation. In Figure 2.2A, we plot this timescale normalized to the awake baseline. A value greater than 1 indicates the system is less stable and recovers more slowly than in the awake state. A ratio of 1.5, for example, means the system takes 50% longer to recover compared to the awake state, indicating greater instability.

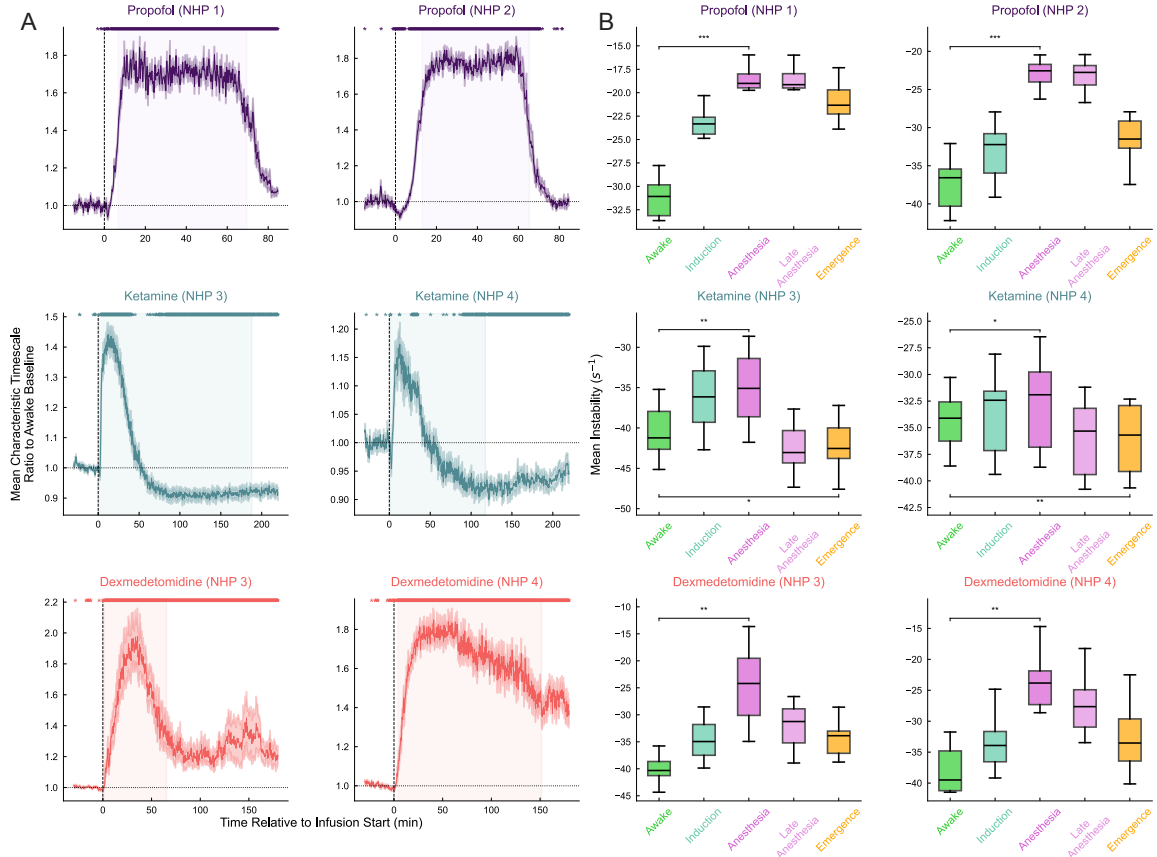


Figure 2.2: Propofol, ketamine, and dexmedetomidine destabilize neural dynamics. (A) The (geometric) mean timescales of response to perturbation normalized to the awake baseline for propofol (top), ketamine (middle) and dexmedetomidine (bottom). The left column corresponds to one NHP involved in the data collection and the right column corresponds to the other. Horizontal dotted line corresponds to a ratio of 1, i.e., when the mean timescale is equivalent to the awake baseline. The vertical dotted line corresponds to the start of the anesthetic infusion. Shaded regions indicate the period between the mean estimated time of loss of consciousness and the mean estimated time of the (first) return of consciousness. In each case, neural activity was robustly destabilized. * indicates $p < 0.05$. $n = 10$ for propofol NHP 1, $n = 11$ for propofol NHP 2, and $n = 8$ for each NHP in ketamine and dexmedetomidine, using two-sided Wilcoxon signed-rank test with Benjamini-Hochberg procedure. (B) The mean instability grouped by epoch of the session. The awake epoch was the part of the session before the infusion began. The induction epoch was defined as 15 min following the start of infusion for propofol, and 10 min following the start of infusion for ketamine and dexmedetomidine. The anesthesia epoch was defined as the period from 15 min post-infusion to 45 min post-infusion for propofol, and from 10 min post-infusion to 45 min post-infusion for ketamine and dexmedetomidine. The late anesthesia epoch was from 30 min following the infusion until the end of infusion (60 min following infusion start) for propofol, and from 45 min after infusion start until the final auditory oddball task began for ketamine and dexmedetomidine. Emergence was defined as 15 min following the end of infusion for propofol and as the final auditory oddball task for ketamine and dexmedetomidine. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. $n = 10$ for propofol NHP 1, $n = 11$ for propofol NHP 2, and $n = 8$ for each NHP in ketamine and dexmedetomidine, using one-sided Wilcoxon signed-rank test. Data are represented as mean $\pm$ SEM, with results averaged over sessions for each NHP. All characteristic root analyses refer to the maximal 10% of the distribution of characteristic roots extracted from the delay differential equations analysis. See also Figures B.2-B.4 and Tables B.1-B.3.

DeLASE employs the Hankel alternative view of Koopman (HAVOK) approach to model the dynamics of the neural data.⁶⁷ Models were fit to 15-second windows tiled across each session. HAVOK models achieved lower prediction error than other approaches that have been used to model electrophysiology data (Figure B.2A). They also achieved lower Akaike information criterion (AIC, lower is better), which penalizes models for complexity and thus prevents overfitting. The HAVOK dynamics models did not generalize well to other windows, particularly between epochs (e.g., Awake and Anesthesia). This illustrates the need for a tiled modeling approach, and points to a fundamental change in neural dynamics that occurs between states (Figure B.2B).

Mean Characteristic Timescales normalized to the awake baseline state (Figure 2.2A) as well as raw instability values (negative-inverse characteristic timescales) (Figure 2.2B) show a clear destabilization, beginning shortly after infusion of each anesthetic drug. The propofol sessions, which involved a continuous infusion, showed a sustained destabilization over the entire 60-minute infusion. The Anesthesia epoch was destabilized compared to the Awake epoch ($p < 0.001$, median difference Anesthesia - Awake Instability in s^{-1} : 12.970, 95% CI: [11.743, 13.649] for NHP 1, $p < 0.001$, median difference Anesthesia - Awake Instability in s^{-1} : 14.285, 95% CI: [13.557, 16.512] for NHP 2, one-sided Wilcoxon signed-rank test). All subjects were non-responsive during the Anesthesia epoch, though responsivity varied during the Late Anesthesia epoch in ketamine and dexmedetomidine session (Figure B.2C). This is because propofol was administered as a continuous infusion, while ketamine and dexmedetomidine were administered as boluses. For both ketamine and dexmedetomidine, the Anesthesia epoch was similarly destabilized compared to the Awake epoch ($p < 0.01$, median difference Anesthesia - Awake Instability in s^{-1} : 5.975, 95% CI: [3.365, 6.664] for NHP 3 with ketamine, $p < 0.05$, median difference Anesthesia - Awake Instability in s^{-1} : 2.201, 95% CI: [-0.622, 2.968] for NHP 4 with ketamine, $p < 0.01$, median difference Anesthesia - Awake Instability in s^{-1} : 16.420, 95% CI: [6.571, 22.117] for NHP 3 with dexmedetomidine, $p < 0.01$, median difference Anesthesia - Awake Instability in s^{-1} : 14.491, 95% CI: [12.851, 16.317] for NHP 4 with dexmedetomidine, one-sided Wilcoxon signed-rank test). Ketamine was unique in that it appeared to produce greater stability later in the session after initially destabilizing cortex ($p < 0.05$, median difference Emergence - Awake Instability in s^{-1} : -2.095, 95% CI: [-3.114, -0.844] for NHP 3, $p < 0.01$, median difference Emergence - Awake Instability in s^{-1} : -1.582, 95% CI: [-2.634, -0.354] for NHP 4, one-sided Wilcoxon signed-rank test). The destabilizations were also evident when considering how the distribution of instability values shifted during anesthetic infusion (Figure B.2D), and when restricting our analysis only to windows in which there were no stimuli (i.e., resting-state data, 3 second windows used due to data constraints, Figure B.2E). The ketamine and dexmedetomidine sessions involved an initial bolus infusion. They showed a rapid initial destabilization followed by a gradual decline (Figure 2.2A). This time course mirrored the typical metabolism of an anesthetic agent administered as a bolus. Likewise, the propofol sessions were a continuous infusion, and the destabilization time course reflected that. Accordingly, there was a high correlation between instability and physiological markers of anesthetic depth (e.g., eye movement and pupil diameter, Figure S1F). We note that all statistical tests treated separate sessions as independent samples, which was validated using a linear mixed-effects model (Table S1). See Supplementary Materials for the chosen hyperparameters for all primary models (Table S2).

A post-hoc statistical power analysis indicated that the study was largely well-powered to determine its primary finding of destabilization in all agents (Table S3).

To validate our findings, we performed several additional analyses (Figure B.3). First, we confirmed that our central result is not method dependent. A stability analysis using Vector Autoregressive (VAR) models, fit to both 500 millisecond (as in previous work^{63,64,190}) and 15 second windows, qualitatively reproduced the same significant destabilization during the anesthesia epoch when considering the top 10% of eigenvalues (Figure B.3A,B). We next validated the DeLASE method itself by using Iterative Amplitude Adjusted Fourier Transform (IAAFT) surrogates. This method preserves the spectral power of each channel but randomizes the phase. Using an Awake window and an Anesthesia window, we constructed an IAAFT surrogate for each electrode channel. Despite the identical power spectra, the surrogates had significantly different stability estimates (Figure B.3C). This analysis confirmed that our stability metric captures high-dimensional, cross-channel dynamics and is not trivially driven by the well-known changes in spectral power. Furthermore, to address potential experimental confounds, a similarity analysis of the time-series of mean instability measurements demonstrated that the observed dynamic profiles are robustly determined by the specific anesthetic agent administered, rather than being an artifact of the individual animal or the cortical areas recorded (Figure B.3D). This similarity was particularly notable given the variation in subjects and electrode coverage in the analyzed data. Finally, we confirmed that the destabilization effect was global and consistent, as the increase in characteristic timescale was observed across all individual brain regions recorded for all three anesthetics (Figure B.3E).

To ensure our findings were not dependent on specific parameter choices, we performed an extensive robustness analysis. First, we confirmed our findings hold across the entire hyperparameter space of model rank and time delays. An AIC-weighted model average showed a robust destabilization during anesthesia (Figure B.1A). Next, we verified that our choice of using the top 10% of characteristic roots was robust. Our results remained significant across a wide range of chosen thresholds between 1% and 30%. This was true for both our chosen best-fit models (Figure B.1B) and the AIC-weighted model average (Figure B.1C). Finally, we reproduced our entire analysis using both shorter (5-second) and longer (30-second) time windows and found that the central result of significant destabilization during anesthesia was qualitatively and statistically upheld (Figure B.1D,E). Together, these analyses confirm that our central findings are robust and not an artifact of a specific set of parameters.

Finally, we found that dynamic instability not only tracks the transition into unconsciousness but also covaries with real-time fluctuations in responsiveness. This was evident in the dexmedetomidine sessions, where animals fluctuated between states as they were coming out of anesthesia (Figure B.4A,B). When aggregated across all sessions, this effect was highly significant, with unresponsive windows showing substantially greater instability than responsive windows for both NHPs (Figure B.4C,D). To ensure this finding was not an artifact of our behavioral definition, we performed a robustness check. This analysis confirmed that the significant difference in stability between responsive and unresponsive states was independent of the specific behavioral accuracy threshold used to delineate responsive states (Figure B.4E).

2.3.2 Slower recoveries to sensory perturbation matched decreases in dynamic stability

In the previous section, we established that all three anesthetic agents destabilized (and in ketamine, later stabilized) cortical neural dynamics. We now investigate how this destabilization changed recovery from perturbation using auditory sensory stimulation.

More stable systems recover faster from perturbations and disturbances than less stable systems. An example of this is the expert and novice cyclist discussed in the introduction (Figure 2.1A). When perturbed with a gust of identical force, the expert recovers more quickly than the novice. Thus, the perturbations linger for longer in less stable systems.

We analyzed the evoked responses to sensory perturbation during Awake, Anesthesia, and Emergence epochs. In propofol sessions, NHPs were periodically exposed to auditory tones (500 milliseconds) throughout the session. In ketamine and dexmedetomidine sessions, NHPs underwent an auditory oddball task (five 50 millisecond tones spaced by 150 milliseconds). We averaged evoked responses across trials and electrodes for each session to obtain a time-series representation of the recovery from sensory perturbation. We then averaged across sessions to obtain a single time series.

The recoveries from sensory perturbation matched predictions from destabilized systems. For all the anesthetic agents, the sensory responses showed slower recovery from perturbation during the Anesthesia epoch compared to Awake (Figure 2.3A, Figure B.5A,B). During the Emergence epoch, there was a difference between anesthetic agents. For both propofol and dexmedetomidine, perturbation recovery remained slower compared to the Awake epoch, but faster than the Anesthesia epoch (Figure 2.3A, top & bottom). By contrast, for ketamine, after a period of slower recovery from perturbation, there was a period of faster recovery (i.e., increased stability) relative to Awake (Figure 2.3A, middle).

These results were quantified using an autocorrelation function (ACF) of response to the tones, computed over 1 second of data following the tone onset, and averaged across trials and electrodes within each session. For all three agents, the Anesthesia epoch was characterized by longer autocorrelation values (Figure 2.3B, significance computed using the two-sided Wilcoxon signed-rank test). During emergence from propofol and dexmedetomidine, the autocorrelation values dropped relative to Anesthesia but were still higher compared to Awake epochs. By contrast, during emergence from ketamine, autocorrelation values were smaller than those in the Awake epoch.

To provide a direct, quantitative link between our model-based stability estimates and the model-independent sensory response properties, we performed a correlation analysis. We correlated the time-resolved instability values from DeLASE with metrics derived from the ACF of the evoked LFP responses. These metrics involved integrating (i.e., computing the area under) the ACF over different time periods. The integrated autocorrelation function of the system's fluctuations is directly related to the diffusion constant, which characterizes the

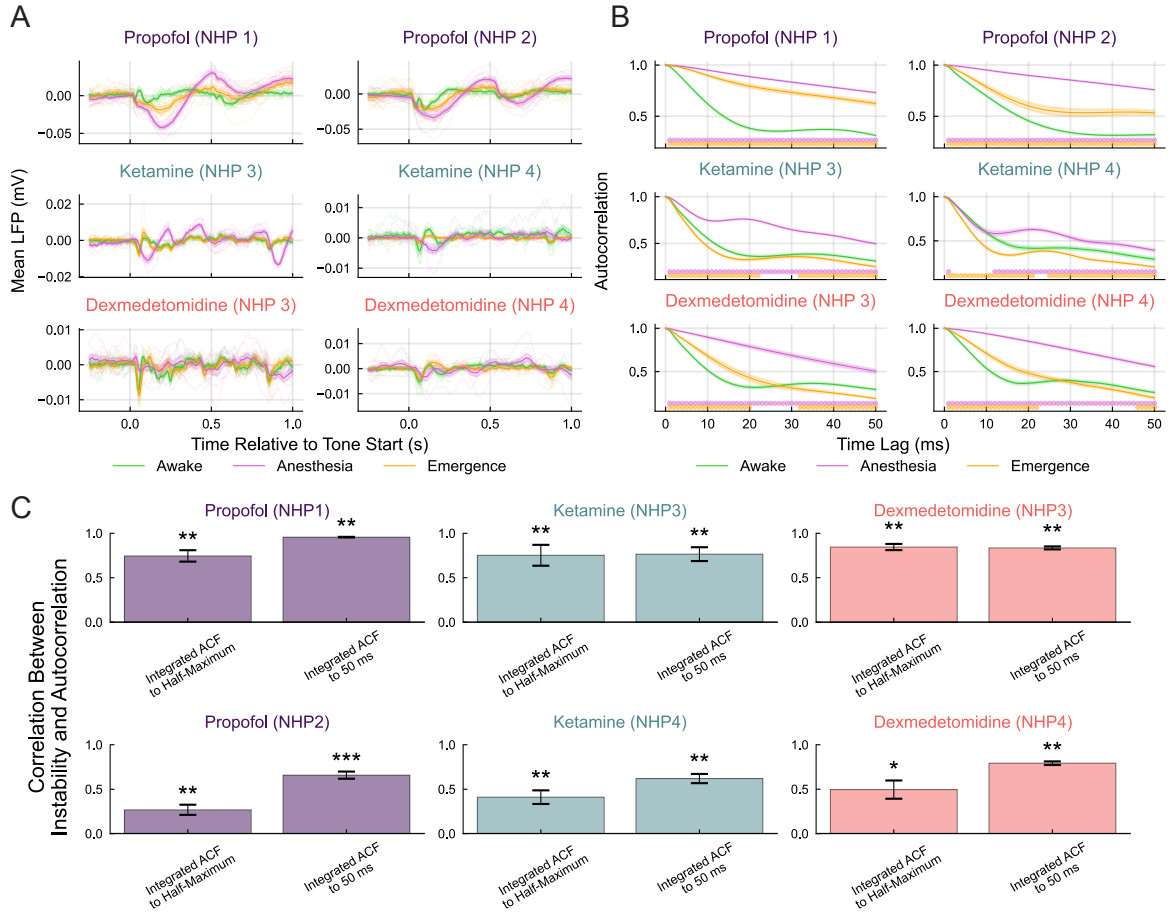


Figure 2.3: Destabilized dynamics explain sensory responses to tones during anesthetic-induced unconsciousness. (A) The LFP response to auditory stimuli for propofol (top), ketamine (middle), and dexmedetomidine (bottom). Response was averaged across electrodes and trials within a session, then the mean was computed across sessions. Translucent curves correspond to response averages from single sessions. Green curve is the Awake epoch, purple is the Anesthesia epoch, and orange is the Emergence epoch. Auditory stimuli were 500 milliseconds for propofol, and auditory oddball stimuli consisting of five 50 millisecond tones spaced by 150 milliseconds for ketamine and dexmedetomidine. Compared to the Awake epoch (green), the Anesthesia epoch (purple) shows a visibly slower recovery from the sensory perturbation, with a more prolonged LFP deflection. (B) Mean autocorrelation functions (ACF) computed from the 1-second period following tone onset for each epoch. The Anesthesia epoch (purple) shows a much slower-decaying ACF for all three anesthetics, quantifying the observation from (A) that the system takes longer to return to baseline after a perturbation. Stars on the bottom indicate whether Anesthesia is significantly different than Awake (purple stars) and whether Emergence is significantly different than Awake (orange stars). * indicates $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. $n = 10$ for propofol NHP 1, $n = 11$ for propofol NHP 2, and $n = 8$ for each NHP in ketamine and dexmedetomidine, using the two-sided Wilcoxon signed-rank test. (C) Correlation between instability and autocorrelation. All anesthetics show a significant positive correlation, confirming that decreased stability is associated with longer “memory” of perturbations. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. $n = 10$ for propofol NHP 1, $n = 11$ for propofol NHP 2, and $n = 8$ for each NHP in ketamine and dexmedetomidine, using the two-sided Wilcoxon signed-rank test. Data are represented as mean $\pm$ SEM, with results averaged over sessions for each NHP. See also Figures B.5A,B for zoomed-in responses for ketamine and dexmedetomidine and Table B.4 for full statistical results.

mobility of the underlying physical process.²⁰⁰ This revealed a strong and significant positive correlation across all three anesthetic agents and all NHPs (Figure 2.3C, Table S4). This result confirms that as dynamic instability increases, the "memory" of perturbations to the system (i.e., the time they linger in the neural activity) correspondingly increases, validating the observed slower recoveries (Figure 2.3A,B).

Autocorrelation values are directly related to the frequency of LFP dynamics. In the next section, we investigate the relationship between stability and spectral power.

2.3.3 Dynamic stability is reflected across lower frequencies

We established in the previous section that sensory responses during anesthetic-induced unconsciousness exhibit longer autocorrelation values, matching the slower recovery times found from the stability analysis. Due to the direct correspondence between the ACF and the power spectrum (in fact, the power spectrum is the Fourier transform of the ACF), a slower decaying ACF implies an increase in lower frequency power.²⁰¹ To analyze this relationship, for each session we computed the correlation between the time series of mean instability (DeLASE) values and the time series of average power in each frequency band. The frequency bands we considered were delta (0.1 - 4 Hz), theta (4 - 8 Hz), alpha (8 - 12 Hz), beta (12 - 30 Hz), and gamma (30 - 100 Hz).

Destabilization, for all three anesthetic agents, correlated positively with delta, theta, and alpha power (Figure 2.4, $p < 0.01$ for all combinations of NHP, agent, and band, two-sided Wilcoxon signed-rank test). Once again, there was a contrast between propofol and dexmedetomidine vs ketamine. For propofol and dexmedetomidine, gamma power was negatively correlated with destabilization ($p < 0.01$ for all combinations of NHP and agent, two-sided Wilcoxon signed-rank test). For ketamine it was positively correlated with destabilization ($p < 0.05$ for both NHPs, two-sided Wilcoxon signed-rank test), reflecting increases in gamma power during ketamine-induced unconsciousness.¹⁶⁷ Results were weaker and not consistent for beta for propofol and dexmedetomidine (Figure 2.4, left & right), while for ketamine destabilization was positively correlated with beta (Figure 2.4 middle, $p < 0.05$ for both NHPs, two-sided Wilcoxon signed-rank test). While a shift to delta in the cortex is a hallmark of loss of consciousness due to anesthesia (Figure B.5C,D),^{41,42,46,154,165–168,202} the destabilization results were qualitatively similar after removing the 0.1 – 4 Hz frequency range (Figure B.5E). This indicated that destabilization was reflected in a multitude of low-frequency responses.

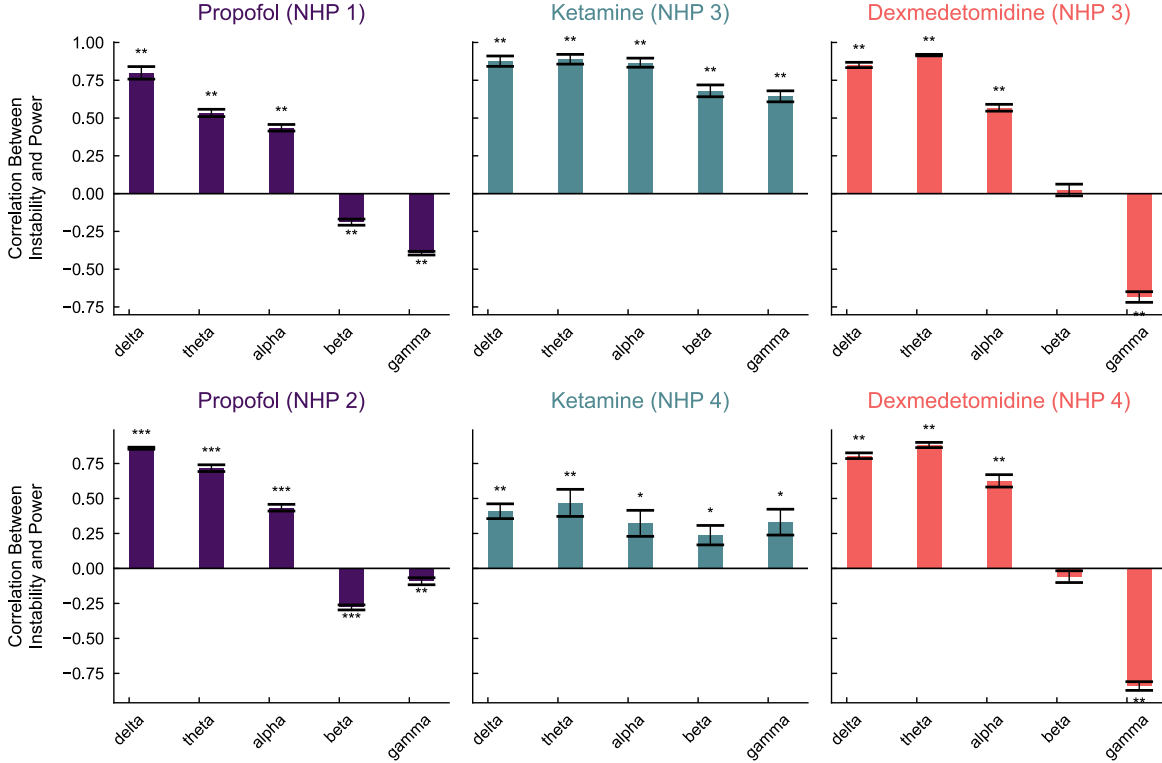


Figure 2.4: Changes in dynamic stability during anesthetic unconsciousness are reflected across the low frequency spectrum. Data are represented as mean $\pm$ SEM, with results averaged over sessions for each NHP. Mean correlation between instability values and band power for propofol (left), ketamine (middle), and dexmedetomidine (right). Correlation values are averaged across sessions. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. $n = 10$ for propofol NHP 1, $n = 11$ for propofol NHP 2, and $n = 8$ for each NHP in ketamine and dexmedetomidine, using two-sided Wilcoxon signed-rank test. See also Figures B.5C-E for power spectra comparisons and stability results with delta removed, as well as Table B.5 for full statistical results.

2.4 Discussion

Our results show that propofol, ketamine, and dexmedetomidine destabilize cortical dynamics to a similar degree. The destabilization was evident in neural responses to sensory inputs, which exhibited longer timescales during anesthesia. The effect of longer timescales on the electrophysiology would be increased low-frequency power. Indeed, cortical destabilization was correlated with increased power in the low-frequency spectrum (0.1 - 4 Hz delta band, 4 - 8 Hz theta band, 8 - 12 Hz alpha band). These results are consistent with previous work showing that, under anesthesia, cortical activity exhibits longer intrinsic timescales,^{203–207} a characteristic of destabilized systems approaching the boundary of chaos.^{25,208–210} Taken together, this suggests destabilization of neural dynamics as an explanation for the observed shift to lower frequencies that is shared between propofol, ketamine, and dexmedetomidine.^{41,42,46,154,165–168,202,211,212} Thus, despite their different actions on different molecular targets in different sets of neurons, there is a common signature of loss of consciousness (LOC)

among the three anesthetic agents at the more emergent level of neural population dynamics.

2.4.1 Possible mechanisms of destabilization

One hypothesis is that the destabilization is a result of the molecular action of the anesthetic agents on the dynamics of recurrent cortical networks. As we have previously demonstrated in computational models, the increased inhibition resulting from propofol's action at GABA receptors is sufficient to cause the dynamical destabilization observed here.³⁷ Notably, all the agents disrupt the balance between excitation and inhibition, which is crucial for maintaining dynamic stability.^{27,213}

A related hypothesis, not inconsistent with our results, is that the destabilization is related to the gating of bottom-up sensory inputs to cortex. Strong inputs, such as sensory stimuli, can stabilize otherwise unstable dynamics.^{214–217} Anesthesia may disrupt the relay of sensory input from subcortical to cortical regions, thus preventing their stabilizing effects.^{218–222} Specifically, propofol agonizes GABA receptors in thalamic relay neurons, and has been shown to disrupt the progression of inputs to higher level cortices.^{43,131,223–226} Ketamine has also been shown to disrupt the propagation of sensory information through the cortex.²²⁷ Dexmedetomidine, by activating the endogenous sleep pathway via α_2 -adrenergic receptors, is thought to reduce the coordination across brain areas.²²² Thus, it is possible that all three anesthetic agents, despite their distinct molecular targets, manage to disrupt the projection of sensory inputs. The observed destabilization may therefore reflect a lack of input-induced stabilization resulting from disrupted sensory inputs. This hypothesis is consistent with theories of consciousness that propose the reliable formation of meta ensembles that coordinate cortex on a large scale,^{15–19,48,181,228,229} and it is supported by the fact that deep-brain stimulation can induce arousal from anesthetic-induced unconsciousness.^{42,230–233} In fact, the relationship between stability and consciousness is not limited to anesthesia. It has been theorized that dynamic stability and consciousness are closely linked.^{37,49,52–54,61–65,180–190} When humans are presented with threshold-level sensory stimuli, neural representations are more stable when the stimuli were perceived versus not perceived.¹⁸⁹

2.4.2 Slow timescales and low-frequency power during anesthesia

Our finding that destabilization (longer characteristic timescales) correlates strongly with increased low-frequency power is supported by a fundamental mathematical principle. Mathematically, the Wiener-Khinchin theorem formalizes the relationship between a signal's time-domain structure and its frequency-domain content. It states that the power spectral density (the "power spectrum") is the Fourier transform of the autocorrelation function.²⁰¹ A system with a longer characteristic timescale, as revealed by our DeLASE analysis, will necessarily have a longer autocorrelation time. That is, perturbations "ring" or decay more slowly.²⁵ Per the theorem, a signal with a slowly decaying autocorrelation function will, by def-

initiation, have its power concentrated in the low-frequency bands. Thus, our finding of increased timescales in neural dynamics provides a direct explanation for the ubiquitous low-frequency dynamics observed in electrophysiology during anesthetic-induced unconsciousness.

While the intrinsic timescales of the frequency ranges we studied (0.1Hz – 100Hz) decreased, we note a complementary finding by Luppi et al. that in the infraslow oscillatory range (< 0.1 Hz), the intrinsic timescales of neural dynamics may actually increase.²³⁴ These changes in the dynamics, measured by fMRI and calcium imaging (0.5 s – 2.4 s sampling), rather than our high-speed electrophysiology approach, were found to be preserved across phylogeny and anesthetic agents.²³⁴ Together, our findings point to complementary aspects of the same fundamental mechanism: A collapse of the brain's rich, multi-scale temporal repertoire into a more stereotyped and less complex dynamical state.

2.4.3 Connection to other signatures of anesthetic-induced unconsciousness

Previous work has identified structure-function similarity as a robust correlate of anesthesia-induced loss of consciousness.^{55,174–178} Structure-function similarity is the similarity between resting state functional connectivity, as measured by fMRI, and anatomical connectivity. A natural link exists between these findings and our characterization of slower recoveries from perturbation during anesthesia. In the awake state, with greater dynamic stability, transients decay quickly. Thus, neural dynamics converge quickly to input-guided trajectories, and the brain can robustly respond to inputs. This leads to a rich repertoire of neural states, and thus low structure-function similarity. However, during anesthesia, transient perturbations linger, and there is a slow convergence to input-guided trajectories. Instead, the perturbed activity is propagated by the connectivity, leading to the observed increase in structure-function similarity. Thus, the increased structure-function similarity observed during anesthetic-induced unconsciousness is indicative of decreased dynamic stability, despite the difference in the timescales of fMRI and electrophysiology recordings.

A recent cross-phylogenetic comparison of anesthetic effects identified a reduction in interareal synchrony as a robust feature preserved across species and agents.²³⁴ This is in line with other studies of interareal synchrony during anesthesia, as well as the finding that ketamine and dexmedetomidine both reduce phase alignment between neighboring prefrontal regions within the same hemisphere.^{63,154,235} These misalignments may be a direct consequence of the decreased dynamic stability we report. A less stable system is less responsive to the inputs required to align and synchronize distinct cortical areas, leading to functional fragmentation. Interestingly, Bardon et al.¹⁵⁴ also found that these drugs increase phase alignment between homologous areas across hemispheres. This relates to the aforementioned finding of increased structure-function similarity across anesthetics. It could be a different manifestation of the same underlying process. As dynamics become less input-driven and more constrained by anatomy, they may be more influenced by strong, direct connections like the corpus callosum.

Another consequence of anesthetic-induced loss of consciousness is a change in cortical response complexity.^{236–242} A study found that the perturbational complexity index is decreased when there is a lack of conscious experience (propofol and xenon) but increased when there is conscious experience but lack of responsiveness (ketamine, administered at a dose approximately one fifth of that in the present study).²⁴³ Complexity thus corresponds to richer representations, which presumably depends on a stable cortex. However, too much can also be detrimental: LOC due to seizures is associated with a large increase in cortical stability, suggesting there is an optimal range of dynamic stability in cortex.⁶¹

2.4.4 Unique dynamics under ketamine

Though all drugs showed similar destabilization at the start of anesthesia, ketamine showed unique dynamics. At first, as with the other drugs, there was a destabilization shortly after the bolus infusion that induced LOC. As the drug began to wear off, however, the cortex was *more stable than during wakefulness, unlike the other two anesthetic agents*. This biphasic effect likely reflects ketamine's nonlinear physiological mechanism. The initial high-dose destabilization corresponds to widespread NMDA receptor blockade. In contrast, the late-session stability reflects a “low-dose” state where lower concentrations are thought to preferentially inhibit active interneurons.^{150,157–160} This resulting disinhibition of pyramidal cells is known to increase high-frequency (gamma) power and neural complexity.^{244–247} This mechanism directly aligns with our findings, as the stable period was characterized by decreased low-frequency (delta, 0.1-4 Hz) power, and increased high-frequency (gamma, 30-100 Hz) power.

Although we did not study ketamine's dose-dependent effects directly, the unique dynamics under ketamine reinforce the observation that lower doses of ketamine have distinct effects from anesthetic doses.^{160,170,243,248,249} A study using a ketamine dose approximately one fourth of ours found increased stability in cortex and supports this conclusion.¹⁹⁰ Also, higher doses of ketamine decrease the complexity of brain activity while lower doses increase complexity.^{246,250–258} Low-dose ketamine has shown some promise as a treatment for mood disorders.^{259–269} Further work will be needed to determine if these treatment effects are after-effects of the cortex in a higher frequency, relatively stable state.

2.4.5 Stability tracks fluctuating recovery under dexmedetomidine

Interestingly, towards the end of our dexmedetomidine sessions our NHPs fluctuated in and out of responsiveness. Correspondingly, cortex fluctuated between stability and destabilization. This likewise suggests that dynamic stability may serve as a marker of consciousness level. For example, it could be used, via EEG, to track level of consciousness during surgery.^{270,271} Given the tight, real-time coupling we observed between instability and behavioral responsiveness (Figure B.4), we believe that developing this measure into a robust index for clinical monitoring

is a promising and important future direction. Furthermore, it is important to note that the analysis of autocorrelations in the neural data in response to sensory stimuli provides a model-independent method for quantifying response timescales. It does not rely on the DeLASE framework and serves as a separate, robust validation of the observed destabilization. This may offer a broader applicability metric for assessing system recovery times and may present a possible utility in detecting changes in conscious state in clinical contexts.

2.4.6 Limitations of the study

The pattern of destabilization observed in all three anesthetic agents in this paper emerged when looking at 15-second windows of neural activity, with instability values averaged across sessions. However, it is important to note that anesthetic states often involve alternating patterns of brain activity, such as up and down states during propofol, gamma bursting during ketamine, and spindle oscillations during dexmedetomidine.^{41,42,131,166,167,171,173} Thus, while the relationship between instability and frequency content is clear when considering the data in aggregate, we note that a more nuanced understanding of how these alternating patterns of activity fit into the dynamic stability framework put forward here remains as a topic for future work.

We note that our analysis measures dynamic stability (recovery from perturbation) using local linear models, not functional stability (computational robustness). While these models focus on the recoveries from perturbation, we hypothesize that dynamic stability is a necessary precursor for robust function. A system that recovers more slowly from perturbations is likely less capable of maintaining the precise dynamic motifs required for reliable neural computation.^{35,36,54,56}

We demonstrated that our primary findings are robust to variations in window length, model hyperparameters, and the percent of characteristic roots used (Figure B.1). However, we note that because DeLASE relies on a linear approximation of nonlinear dynamics, the stability metric requires aggregating a distribution of roots rather than relying on a single dominant mode. While our results held across a broad range of thresholds (1% to 30%), reliance on a small subset of roots (e.g., $< 3\%$) can yield variable estimates. Thus, distributional averaging remains a necessary step when applying this framework.

Regarding statistical power, our study relies on many independent recording sessions across two subjects per anesthetic agent, consistent with standards in non-human primate neurophysiology.²⁷² A simulation-based post-hoc power analysis confirmed that the study was highly powered ($>99\%$) to detect the primary effect of destabilization for the majority of comparisons. However, the analysis also indicated that one specific comparison, the Awake-Anesthesia comparison in NHP 4 under ketamine, exhibited lower observed power ($\sim 63\%$). This suggests that while the destabilization effect is robust across the population, the magnitude of the effect relative to baseline variability can differ between subjects.

2.4.7 Future directions

Beyond anesthesia, the DeLASE framework offers a quantitative tool to explore other states of consciousness. Future studies could apply this method to neural data recorded during sleep to determine if NREM and REM sleep exhibit distinct stability profiles. Similarly, investigating neural dynamics during psychedelic experiences, or in patients with neuropsychiatric conditions or disorders of consciousness may reveal whether these states also represent deviations from the brain’s typical range of dynamic stability.

2.4.8 Conclusion

Leveraging high-resolution intracortical recordings and dynamical systems analysis, we provide evidence for destabilization as a convergent effect of different anesthetics. This suggests destabilization as a population-level signature of loss of consciousness potentially common among anesthetics. It furthermore provides an explanation for the low-frequency oscillations observed during anesthetic-induced unconsciousness. Our results point to an optimal range of dynamic stability as a prerequisite for consciousness.

Acknowledgments

The authors would like to acknowledge Leo Kozachkov for thoughtful and helpful discussions. E.K.M. acknowledges funding from ONR MURI N00014-23-1-2768, NIMH 1R01MH131715-01, NIMH R01MH11559, The Simons Center for the Social Brain, The Freedom Together Foundation, and The Picower Institute for Learning and Memory. I.R.F. acknowledges funding from The National Science Foundation Computer and Information Science and Engineering Directorate, The Simons Collaboration on the Global Brain, and The McGovern Institute at MIT. E.N.B. acknowledges funding from The Freedom Together Foundation, and The Picower Institute for Learning and Memory. Y.I. acknowledges funding from NIH 1R21AG077275-01A1. A.G.B. acknowledges funding from the U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing Research, Department of Energy Computational Science Graduate Fellowship under Award Number DE-SC0024386.

Chapter 3

Characterizing control between interacting subsystems with deep Jacobian estimation

Authors: Adam J. Eisen, Mitchell Ostrow, Sarthak Chandra, Leo Kozachkov, Earl K. Miller, Ila R. Fiete.

Statement of contribution and publication

*The work presented in this chapter was published at **NeurIPS** in 2025.*³⁹

Conceptualization, A.J.E.; methodology, A.J.E., M.O., S.C., L.K. and I.R.F.; software, A.J.E.; formal analysis, A.J.E.; writing – original draft, A.J.E.; writing – review & editing, A.J.E., M.O., S.C., L.K., and I.R.F.; visualization, A.J.E.; supervision, E.K.M. and I.R.F.; funding acquisition, E.K.M. and I.R.F.

3.1 Abstract

Biological function arises through the dynamical interactions of multiple subsystems, including those between brain areas, within gene regulatory networks, and more. A common approach to understanding these systems is to model the dynamics of each subsystem and characterize communication between them. An alternative approach is through the lens of control theory: how the subsystems control one another. This approach involves inferring the directionality, strength, and contextual modulation of control between subsystems. However, methods for understanding subsystem control are typically linear and cannot adequately describe the rich contextual effects enabled by nonlinear complex systems. To bridge this gap, we devise a data-driven nonlinear control-theoretic framework to characterize subsystem interactions via the Jacobian of the dynamics. We address the challenge of learning Jacobians from time-series data by proposing the JacobianODE, a deep learning method that leverages properties of the

Jacobian to directly estimate it for arbitrary dynamical systems from data alone. We show that JacobianODE models outperform existing Jacobian estimation methods on challenging systems, including high-dimensional chaos. Applying our approach to a multi-area recurrent neural network (RNN) trained on a working memory selection task, we show that the “sensory” area gains greater control over the “cognitive” area over learning. Furthermore, we leverage the JacobianODE to directly control the trained RNN, enabling precise manipulation of its behavior. Our work lays the foundation for a theoretically grounded and data-driven understanding of interactions among biological subsystems.

3.2 Introduction

Complex systems are ubiquitous in nature. These systems exhibit a wide range of behavior and function, in large part through the dynamic interaction of multiple component subsystems within them. One approach to understanding such complex systems is to build detailed models of their underlying dynamics. An alternative and simpler yet powerful approach is offered by control theory, focusing instead on how subsystems influence and regulate one another, and how they can be controlled.

Control theory thus offers a complementary approach to both understanding and manipulating biological systems. The theory describes how inputs must be coordinated with system dynamics to achieve desired behaviors, and can be applied across domains ranging from robotics to biology (Figure 3.1A). The brain coordinates neural activity across multiple interconnected brain areas, dynamically modulating which regions receive information from which others depending on need and context.^{273,274} Interareal interactions play central roles in cognition and consciousness,^{21,275–278} in selective attention,^{44,279–286} decision making,^{287,288} working memory,^{289,290} feature binding,^{291,292} motor control,^{293–297} and learning and memory.^{298–302}

A common approach to characterizing interareal interactions is to quantify *communication* between them, using methods such as reduced-rank regression to define “communication subspaces”. These subspaces determine low-dimensional projections that maximally align high-dimensional states of the input area with high-dimensional states of the target area.^{303–305} However, effective *control* not only involves alignment of high-variance input states with high-variance target states, but also appropriate alignment of the inputs with the *dynamics* of the target area. Given connected subsystems A and B, an identical signal from B will have dramatically different control effects on A, depending on whether the signal aligns with stable or unstable directions of A’s dynamics: projections onto more unstable eigenvectors can much more readily drive the system to novel states. (For more detail see Appendix C.3.1.)

Accordingly, recent work in neuroscience has espoused control-theoretic perspectives on interareal interactions (Figure 3.1B).^{182,294,306–311} The dominant approach has involved linear control.^{140,142,144,312–319} However, the inherently nonlinear dynamics of the brain enable richer contextual control than possible to fully model with linear systems, necessitating a nonlinear

control approach.

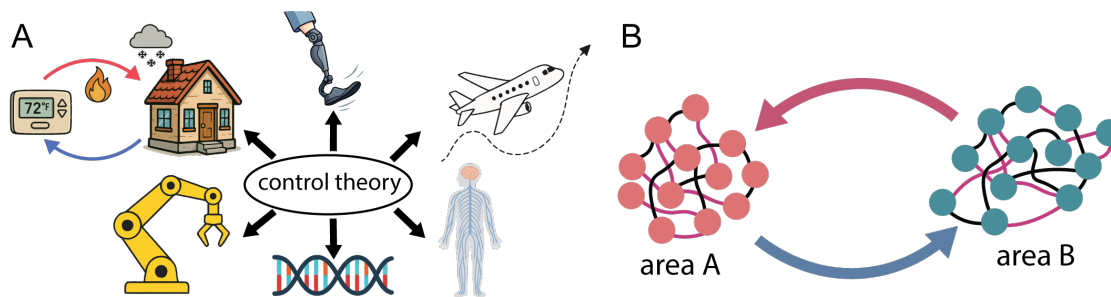


Figure 3.1: Schematic overview of control-theoretic framework applied to neural interactions. (A) Control theory generalizes across diverse systems. (B) Illustration of interareal control, highlighting how neural activity in one area directly influences dynamics in another.

One approach to extend control-theoretic analyses to nonlinear systems is by linearizing the nonlinear dynamics through Taylor expansion, which involves the Jacobian. This converts the nonlinear system into a linear state- and time-dependent one,^{182,320–322} allowing for simple control. Jacobian linearization for control is straightforward with access to analytical expressions for the non-linear system, but it becomes non-trivial in purely data-driven scenarios. Estimating the Jacobian involves conjunctively inferring both a function and its derivative, yet good function approximation need not yield good approximations of its derivatives (see Section 3.5 and Appendix C.3.2).

This paper introduces several key contributions:

- **Robust data-driven Jacobian estimation.** We present JacobianODE, a deep learning-based method for estimating Jacobians from noisy trajectories in high-dimensional dynamical systems. We demonstrate the validity of this approach in several sample systems that include high-dimensional chaos.
- **A framework to characterize control between interacting subsystems via data-driven Jacobian estimation.** Harnessing our Jacobian estimation method, we devise a rigorous, data-driven approach for nonlinear control-theoretic analysis of how paired interacting systems, including brain areas, drive and regulate each other across different contexts.
- **Data-driven inference of control dynamics in trained recurrent neural networks.** We apply our data-driven framework to a recurrent neural network (RNN) trained on a working memory task. We show that, purely from data, we can identify key control-theoretic interactions between the areas, and that these interactions crucially evolve over the course of learning to produce the desired behavior.
- **Demonstration of accurate control of rich interacting high-dimensional coupled dynamical subsystems.** We demonstrate high accuracy in a challenging high-dimensional data-driven nonlinear control task, enabling precise control of the behavior of the RNN.

Overall, our work lays the foundation for data-driven control-theoretic analyses in complex high-dimensional nonlinear systems, including the brain.

3.3 Related work

Interareal communication A wide range of tools have been developed and harnessed to study interareal communication in neural data.^{304,323} This includes, but is not limited to, methods based on reduced rank regression,^{303,305} recurrent neural network models of neural dynamics,^{182,324–329} Gaussian process factor analysis,^{330,331} canonical correlation analysis,^{332–334} convergent cross mapping,^{83,87,191} switching dynamical systems,^{335,336} granger causality,^{337,338} dynamic causal mapping,³³⁹ point process models,³⁴⁰ and machine learning methods.^{341,342}

Nonlinear controllability Classical results assess the controllability of nonlinear control systems via the Lie theory.^{179,343–346} Another approach to nonlinear network controllability is based on attractor strength.³⁴⁷ Liu, Slotine, and Barabási [320] and Parkes et al. [321] note that the large literature of linear controllability analyses could be extended locally to nonlinear systems with an appropriate linearization method.

Neural network controllability Linear structural network controllability has been implicated across a wide range of contexts, tasks, and neuropsychiatric conditions.^{140,144,306,311,313–316} This approach has been extended to functional brain networks.^{317–319} Recent work has characterized the subspaces of neural activity that are most feedback controllable as opposed to feedforward controllable using linear methods.³¹² Other approaches to neural control analyses consider the identification of structural driver nodes,³⁴⁸ and input novelty.³⁴⁹

Data-driven Jacobian estimation Besides approaches estimating Jacobians through weighted linear regressions,^{85,88} some methods have used direct parameterization via neural networks to learn Jacobians of general functions.^{350,351} These approaches inform our method but do not explicitly address dynamical systems. Applying path-integral-based Jacobian estimation to dynamical systems is challenging, as the target function (the system’s time derivative) is typically unobserved. Beik-Mohammadi et al. [352] utilized this idea in dynamical systems to learn contracting latent dynamics from demonstrations.

3.4 JacobianODE: learning Jacobians from data

3.4.1 Jacobian linearization

We consider nonlinear dynamical systems in $\mathbb{R}^n$, defined by $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t))$. The Jacobian of the dynamics is a matrix-valued function $\mathbf{J}_{\mathbf{f}} : \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$ (henceforth, $\mathbf{J}$) given by

$$\mathbf{J}(\mathbf{x}(t)) = \frac{\partial}{\partial \mathbf{x}} \mathbf{f}(\mathbf{x}(t)) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix}. \quad (3.1)$$

At each time t , the Jacobian defines a linear subspace relating input and output changes, capturing how perturbations to the system will propagate. This recasts nonlinear dynamics as linear time-varying dynamics in the tangent space locally along trajectories (formally, $\delta \dot{\mathbf{x}}(t) = \mathbf{J}_{\mathbf{f}}(\mathbf{x}(t)) \delta \mathbf{x}(t)$, see Figure 3.2 left, also see Lohmiller and Slotine [353] for a discussion in the context of virtual displacements).

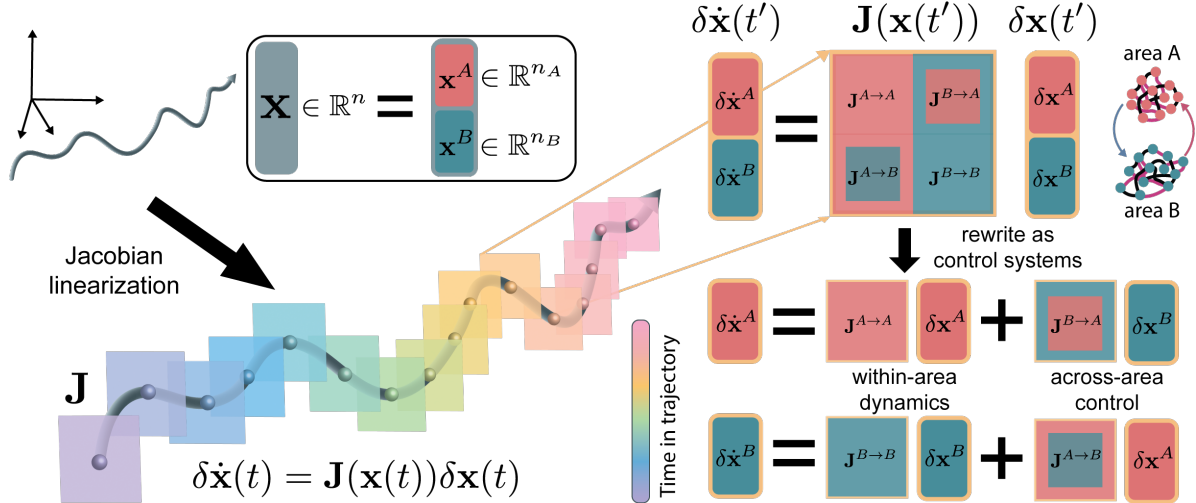


Figure 3.2: Analytical framework for pairwise interacting subsystem control. Trajectory dynamics (left, top) are locally linearized via Jacobians (left, bottom), explicitly separating within-area (diagonal blocks) and interareal (off-diagonal blocks) dynamics (right). These separated dynamics can be used to construct interaction-specific control systems.

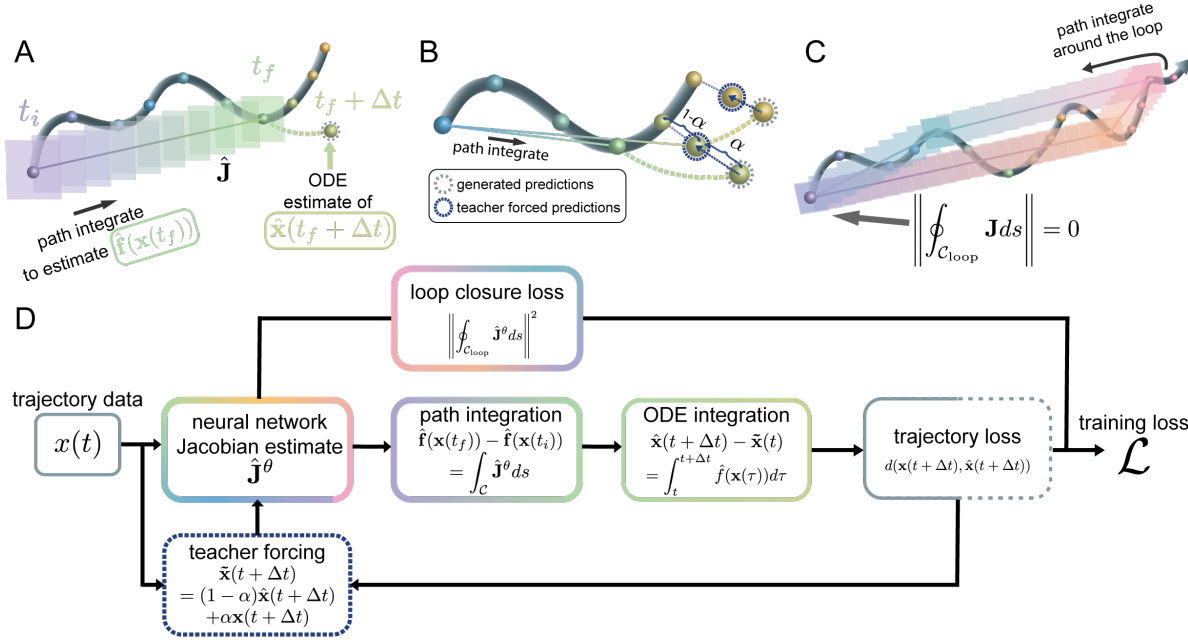


Figure 3.3: Jacobian estimation with JacobianODE models. (A) Path integration of the Jacobian predicts future states. (B) Generalized teacher forcing stabilizes trajectory predictions during training. (C) Loop-closure constraints enforce consistency of Jacobian estimates. (D) Training pipeline, combining neural Jacobian estimation, path integration, teacher forcing, and self-supervised loop-closure loss.

3.4.2 Parameterizing differential equations via the Jacobian

We now turn to the problem of how to estimate the Jacobian $\mathbf{J}$ from data. We assume that we have access only to observed trajectories of the system, of the form $\mathbf{x}^{(j)}(t_0^{(j)} + k\Delta t)$, $k = 1, 2, \dots$, $j = 1, 2, \dots$ where Δt is a fixed sampling interval, j indexes the trajectory, and $t_0^{(j)}$ is a trajectory-specific start time. Crucially, we do not assume access to the function $\mathbf{f}$. Our method estimates the Jacobian directly via a neural network. To do this, we parameterize a neural network function $\hat{\mathbf{J}}^\theta$ with learnable parameters θ that is then trained to approximate $\mathbf{J}$.

Path integration Following Lorraine and Hossain [351] and Beik-Mohammadi et al. [352], we exploit the fact that the path integral of the Jacobian is path independent in the construction of the loss function. This is because the rows of the Jacobian are conservative vector fields (i.e., they are the gradients of scalar functions). For an intuitive picture, consider the work done by the force of gravity as you climb a mountain. Regardless of the path you take to climb, the resulting work is dependent only on the start and end points of the path. Formally, for the time derivative function $\mathbf{f}$ we have that

$$\mathbf{f}(\mathbf{x}(t_f)) - \mathbf{f}(\mathbf{x}(t_i)) = \int_C \mathbf{J} ds = \int_{t_i}^{t_f} \mathbf{J}(\mathbf{c}(r)) \mathbf{c}'(r) dr, \quad (3.2)$$

where C is a piecewise smooth curve in $\mathbb{R}^n$ and $\mathbf{c} : [t_i, t_f] \rightarrow C$ is a parameterization of C with $\mathbf{c}(t_i) = \mathbf{x}(t_i)$ and $\mathbf{c}(t_f) = \mathbf{x}(t_f)$ (Figure 3.3A). Given estimates of $\mathbf{f}(\mathbf{x}(t_i))$ and the Jacobian $\mathbf{J}$,

we can then use Equation 3.2 to approximate $\mathbf{f}(\mathbf{x}(t_f))$ at any time t_f . For these integrals, we use a line between the endpoints as a simple choice of path (Figure 3.3A). Then, to generate an estimate $\hat{\mathbf{x}}(t_f + \Delta t)$ of the next step, we can use a standard ordinary differential equation (ODE) integrator (e.g., Euler, fourth-order Runge–Kutta, etc.) to integrate the estimated time derivative $\hat{\mathbf{f}}(\mathbf{x}(t))$ (see Appendix C.1 for more detail). To avoid the need to represent $\mathbf{f}$ directly (thereby enabling all gradients to backpropagate through the Jacobian network) we note that we parameterize an estimate of $\mathbf{f}(\mathbf{x}(t_i))$ in Equation 3.2 in terms of the Jacobian (see Appendix C.1.1).

3.4.3 Loss functions

Trajectory reconstruction loss Given an observed trajectory $\mathbf{x}(t_0 + k\Delta t)$, $k = 0, \dots, T - 1$ of length T , we can compute the trajectory reconstruction loss, $\mathcal{L}_{\text{traj}}(\theta; \mathbf{x})$, between the true trajectory and the estimated trajectory using an appropriate distance measure d (e.g., mean squared error, see Appendix C.1.2 for more detail on generating predictions and the trajectory prediction loss).

Generalized Teacher Forcing To avoid trajectory divergence in chaotic or noisy systems, we employ Generalized Teacher Forcing, generating recursive predictions partially guided by true states (Figure 3.3B, and Appendix C.4.8).³⁵⁴

Loop closure loss The Jacobian captures how perturbations to the system will propagate along *any* direction in state space. Estimating it purely from dynamics constrains only the direction of the flow, leaving the full solution underdetermined. To address this, we again exploit the fact that each row of the Jacobian is a conservative vector field. Specifically, we note that for any piecewise smooth loop $\mathcal{C}_{\text{loop}}$, we have $\left\| \oint_{\mathcal{C}_{\text{loop}}} \mathbf{J} ds \right\|_2 = 0$ (see Figure 3.3C). Thus, by integrating along loops that contain directions *orthogonal* to the system’s dynamics (and penalizing the deviation from zero), we encourage the estimated Jacobians to capture information about other directions in state space (see Appendices C.1.3 and C.4.8 for full technical details). To ensure broad coverage of tangent space directions, we form loops from concatenations of line integrals between randomly selected data points. This strategy samples diverse directions from the tangent space while remaining easy to compute. The resulting self-supervised loss term, $\mathcal{L}_{\text{loop}}(\theta; \mathbf{x})$, builds on the loss introduced by Iyer et al. [355]. It constrains $\hat{\mathbf{J}}^\theta$ to satisfy both the dynamics and conservativity. This improves Jacobian estimation accuracy significantly (see Appendix C.3.4 for ablation studies).

Training loss We therefore minimize the following loss function with respect to the parameters θ :

$$\mathcal{L}(\theta; \mathbf{x}) = \mathcal{L}_{\text{traj}}(\theta; \mathbf{x}) + \lambda_{\text{loop}} \mathcal{L}_{\text{loop}}(\theta; \mathbf{x}) \quad (3.3)$$

where λ_{loop} controls the relative weighting of the loop closure loss $\mathcal{L}_{\text{loop}}(\theta; \mathbf{x})$ compared to the trajectory prediction, and is a hyperparameter of the learning procedure (Figure 3.3D).

3.5 Jacobian estimation in dynamical systems

Data To evaluate the quality of the Jacobian estimation procedure, we apply our approach to several example systems for which the dynamics are known. For this analysis, we used the Van der Pol oscillator,¹²² Lorenz system,⁶⁰ and the Lorenz 96 system across three different system sizes (12, 32, and 64 dimensional).³⁵⁶ All systems were simulated using the `dysts` package, which samples dynamical systems with respect to the characteristic timescale τ of their Fourier spectrum.^{357,358} All training data consisted of 26 trajectories of 12 periods, sampled at 100 time steps per τ . All models were trained on a 10 time-step prediction task with teacher forcing.

To evaluate the performance of the methods in the presence of noise, we trained the models on data with 1%, 5%, and 10% Gaussian observation noise added i.i.d over time, where the percentage is defined via the ratio of the euclidean norm of the noise to the mean euclidean norm of the data.

JacobianODE model For the JacobianODE framework, loop closure loss weights were chosen via line search (Appendix C.4.8), where the neural network $\mathbf{J}^\theta$ was taken to be a four-layer multilayer perceptron (MLP) with hidden layer sizes 256, 1024, 2048 and 2048. Path integration was performed using the trapezoid method from the `torchquad` package, with each integral discretized into 20 steps.³⁵⁹ ODE integration was performed using the fourth-order Runge-Kutta (RK4) method from the `torchdiffeq` package.³⁶⁰ The JacobianODE models used 15 observed points to generate the initial estimate of $\mathbf{f}$ (see Section 3.4.2 and Appendices C.1.1 and C.1.2). All models were built in PyTorch.³⁶¹ Full implementation details are provided in Appendices C.1 and C.4.

Baselines We chose two different Jacobian estimation procedures for comparison. The first was a neural ordinary differential equation (NeuralODE) model trained to reproduce the dynamics.³⁶⁰ The NeuralODE was implemented as a four-layer MLP with hidden layers of the same size as the one used for the JacobianODE model. Jacobians were computed via automatic differentiation. NeuralODEs were regularized via a penalty on the Frobenius norm of the estimated Jacobians to prevent the model from learning unnecessarily large negative eigenvalues (see Appendix C.4.3).^{362–364} We also employed a baseline that estimates Jacobian via a weighted linear regression, which computes locally linear models at each point in the space^{85,88} (see Appendix C.4.4).

Performance We tested the approaches on Jacobian estimation on held-out trajectories without noise. The JacobianODE method outperforms the baseline methods in terms of mean Frobenius norm error for virtually all systems (Table 3.1). This was also true when considering the spectral matrix 2-norm (see Table C.3 in Appendix C.3.3).

We plot performance in Figure 3.4. While both the JacobianODEs and the NeuralODEs reproduce the observed dynamics (Figure 3.4A-D), the JacobianODE learns a more accurate

Table 3.1: Mean Frobenius norm error on Jacobian estimation, $\langle \|\mathbf{J} - \hat{\mathbf{J}}\|_F \rangle$, for each system and noise level. Errors are reported as mean $\pm$ standard deviation, with mean and standard deviation computed over five random initializations of the model architectures.

Project	Training noise	JacobianODE	NeuralODE	Weighted Linear
VanDerPol (2 dim)	1%	0.7 $\pm$ 0.1	1.0 $\pm$ 0.3	6.11
	5%	0.72 $\pm$ 0.05	0.72 $\pm$ 0.08	6.09
	10%	1.35 $\pm$ 0.06	2.2 $\pm$ 0.2	6.08
Lorenz (3 dim)	1%	3.3 $\pm$ 0.2	8.7 $\pm$ 0.3	21.94
	5%	5.1 $\pm$ 0.9	26.0 $\pm$ 1.5	21.90
	10%	6.4 $\pm$ 0.1	26.7 $\pm$ 0.9	21.84
Lorenz 96 (12 dim)	1%	1.2 $\pm$ 0.2	4.8 $\pm$ 0.2	28.67
	5%	2.7 $\pm$ 0.2	5.9 $\pm$ 0.2	28.64
	10%	4.6 $\pm$ 0.1	6.1 $\pm$ 0.1	28.56
Lorenz 96 (32 dim)	1%	8.7 $\pm$ 0.2	16.8 $\pm$ 0.4	47.13
	5%	7.8 $\pm$ 0.2	17.7 $\pm$ 0.6	46.96
	10%	13.45 $\pm$ 0.09	19.5 $\pm$ 0.4	47.03
Lorenz 96 (64 dim)	1%	30.9 $\pm$ 0.5	45.5 $\pm$ 0.1	66.39
	5%	34.0 $\pm$ 0.2	45.7 $\pm$ 0.1	66.26
	10%	36.0 $\pm$ 0.2	46.0 $\pm$ 0.2	66.29
Task-trained RNN	1%	188.5 $\pm$ 7.1	294.02 $\pm$ 0.03	301.46
	5%	166.8 $\pm$ 3.6	294.357 $\pm$ 0.003	297.63
	10%	180.4 $\pm$ 0.5	294.328 $\pm$ 0.004	296.63

estimate of the Jacobian (Figure 3.4 E,F). In particular, looking at Lyapunov spectra learned by the models, we note that the JacobianODE exceeds the other methods in estimating the full Lyapunov exponent spectrum, indicating a better overall representation of how perturbations along different directions will evolve (Figure 3.4G,H).

3.6 Control-theoretic analyses in a task-trained RNN

3.6.1 Characterizing control between subsystems with Jacobian linearization

Consider neural data recorded from two areas, A and B. Concatenating their data into a state vector $\mathbf{x} \in \mathbb{R}^n$, composed of $\mathbf{x}^A \in \mathbb{R}^{n_A}$ and $\mathbf{x}^B \in \mathbb{R}^{n_B}$, and assuming dynamics governed by $\mathbf{f}$, we linearize around a reference trajectory ($\delta \dot{\mathbf{x}}(t) = \mathbf{J}(\mathbf{x}(t)) \delta \mathbf{x}(t)$). Splitting the Jacobian into

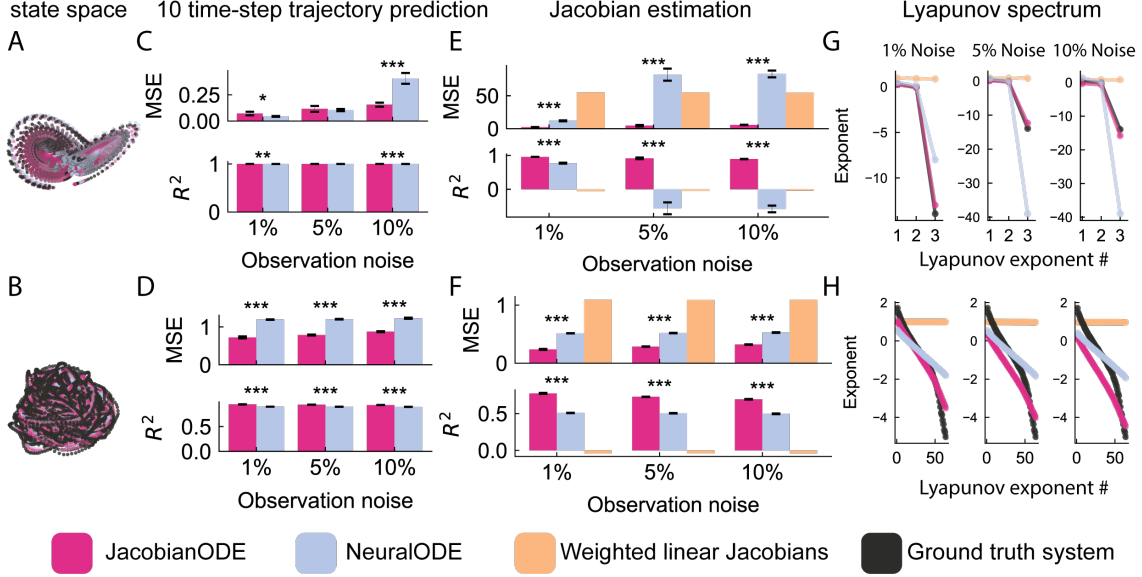


Figure 3.4: JacobianODE surpasses benchmark models on chaotic dynamical systems. Error bars indicate standard deviation, with statistics computed over five different model initializations. (A,B) State-space trajectories for (A) Lorenz and (B) 64-dimensional Lorenz 96 systems, with 10 time-step predictions. Spectral embedding was used to reduce the Lorenz 96 data to three dimensions. (C,D) Accuracy of 10 time-step trajectory predictions at varying noise levels. (E,F) Comparison of Jacobian estimation accuracy, quantified by mean squared error (MSE) and R^2 . (G,H) Estimated Lyapunov spectra averaged over initializations.

block matrices yields:

$$\begin{aligned}\delta\dot{\mathbf{x}}^A(t) &= \mathbf{J}^{A \rightarrow A}(\mathbf{x}(t)) \delta\mathbf{x}^A + \mathbf{J}^{B \rightarrow A}(\mathbf{x}(t)) \delta\mathbf{x}^B \\ \delta\dot{\mathbf{x}}^B(t) &= \mathbf{J}^{A \rightarrow B}(\mathbf{x}(t)) \delta\mathbf{x}^A + \mathbf{J}^{B \rightarrow B}(\mathbf{x}(t)) \delta\mathbf{x}^B\end{aligned}\tag{3.4}$$

where diagonal blocks $\mathbf{J}^{A \rightarrow A} \in \mathbb{R}^{n_A \times n_A}$ and $\mathbf{J}^{B \rightarrow B} \in \mathbb{R}^{n_B \times n_B}$ represent within-area dynamics, and off-diagonal blocks $\mathbf{J}^{B \rightarrow A} \in \mathbb{R}^{n_A \times n_B}$, $\mathbf{J}^{A \rightarrow B} \in \mathbb{R}^{n_B \times n_A}$ represent interareal interactions (Figure 3.2, right). Explicit separation of each area’s control dynamics quantifies the direct influence each exerts on the other, and readily generalizes beyond two areas (see Appendix C.2.3).

Since Jacobians are time-dependent, we obtain a linear time-varying representation of control dynamics along the trajectory. This enables computation of time-varying reachability ease, capturing how readily each area drives the other toward novel states.^{365–367} Reachability is quantified via the reachability Gramian, a matrix defining a local metric in tangent space. For the above control system capturing the influence of area B on area A, the time-varying reachability Gramian on the interval $[t_0, t_1]$ is defined as

$$\mathbf{W}_r(t_0, t_1) \triangleq \int_{t_0}^{t_1} \Phi(t_1, \tau) \mathbf{B}(\tau) \mathbf{B}^T(\tau) \Phi^T(t_1, \tau) d\tau\tag{3.5}$$

where Φ (computed from $\mathbf{J}^{A \rightarrow A}$) denotes the state-transition matrix of the intrinsic dynamics of subsystem A without any input (i.e., $\delta\mathbf{x}^A(t) = \Phi(t, t_0) \delta\mathbf{x}^A(t_0)$), and $\mathbf{B}(\tau) = \mathbf{J}^{B \rightarrow A}(\mathbf{x}(\tau))$

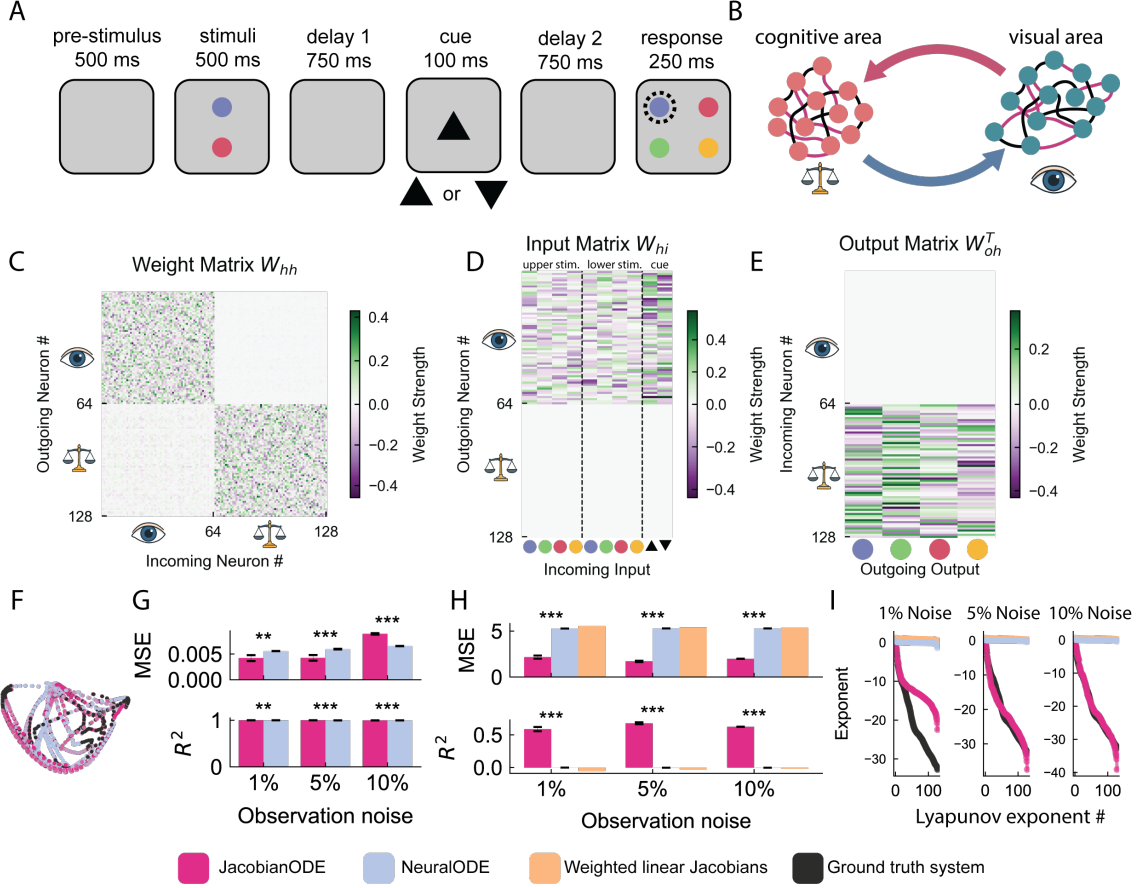


Figure 3.5: JacobianODE accurately infers trained RNN Jacobians. Error bars indicate standard deviation, with statistics computed over five different model initializations. (A) Task schematic, involving stimuli presentation, delays, cueing, and response. (B) RNN architecture with distinct visual and cognitive areas interacting. (C–E) Structure of trained weight matrices: (C) recurrent weights within RNN, (D) input connectivity pattern, and (E) output connectivity. (F) State-space trajectories from the trained RNN, with 10 time-step predictions. Spectral embedding was used to reduce the data to three dimensions. (G,H) JacobianODE performance evaluated against other models for trajectory prediction (G) and Jacobian estimation accuracy (H) at multiple noise levels. (I) Estimated Lyapunov spectra averaged over initializations.

(see Appendix C.2.1).³⁶⁵ The Gramian $\mathbf{W}_r(t_0, t_1)$ is symmetric and positive semidefinite for every $t_1 > t_0$.³⁶⁵ Each eigenvalue of the reachability Gramian quantifies how easily the target system can be driven along its corresponding eigenvector. Thus, the trace of the reachability Gramian reflects average ease of reaching new states, and its minimum eigenvalue reflects the ease along the most challenging direction of control.

3.6.2 Estimating the Jacobian of a task-trained RNN

Task To demonstrate how JacobianODE models could be used in neuroscience, we performed a control-theoretic analysis of a task-trained RNN. We used a working memory selection task from Panichello and Buschman [368] (Figure 3.5A). On each trial, the network is presented

with two of four possible "colors", denoted by one-hot vectors. After a delay, the network is presented with a cue indicating which of the colors to select. The network is then asked to reach a state of sustained activation that corresponds to the selected color.

RNN model To perform this task, we trained a 128-dimensional continuous-time RNN. The RNN had hidden dynamics defined by

$$\begin{aligned}\tau\dot{\mathbf{h}}(t) &= -\mathbf{h} + \mathbf{W}_{hh}\sigma(\mathbf{h}(t)) + \mathbf{W}_{hi}\mathbf{u}(t) + \mathbf{b} \\ \mathbf{o}(t) &= \mathbf{W}_{oh}\mathbf{h}(t)\end{aligned}\tag{3.6}$$

where $\mathbf{W}_{hh}$ defines the internal dynamics, $\mathbf{W}_{hi}$ maps the input into the hidden state, $\mathbf{W}_{oh}$ maps the hidden state to a four-dimensional output $\mathbf{o}(t)$, $\mathbf{b}$ is a bias term, and σ is the exponential linear unit activation. The RNN had two 64-neuron areas: a "visual" area (which received sensory input) and a "cognitive" (which output the RNN's color choice) (Figure 3.5B). To encourage multi-area structure, we initialized the within-area weights with greater connectivity strength than the across-area weights (Figure 3.5C). Since input comes only to the visual area and output only from the cognitive area, the two areas are forced to interact to solve the task (Figure 3.5D,E). The RNN solves this task with 100% accuracy.

Jacobian reconstruction quality We trained JacobianODEs on RNN trajectories from the post-cue delay and response portion of the trials. We used the same baselines as in Section 3.5. JacobianODEs are not only robust to noise, but can also benefit from it (for multiple systems in Table 3.1, 5% training noise improves estimation). Noise encourages the model to explore how perturbations around the observed trajectory evolve, which is crucial for learning accurate Jacobians in high-dimensional systems. Thus (for both the JacobianODE and NeuralODE) we add a small amount of additional noise to the data during learning. Although the models perform similarly on trajectory reconstruction (Figure 3.5F,G), on Jacobian estimation, JacobianODEs drastically outperform both baseline models (Figure 3.5H,I, and Table 3.1).

3.6.3 Reachability in the task-trained RNN across contexts

Next, we used the Jacobians learned by the JacobianODE (trained on 5% noise) to evaluate reachability control in the RNN, and compared it to evaluations using ground truth Jacobians. All analyses were performed using the 10 time-step reachability Gramian. We first found that it was on average easier for the visual area to drive the cognitive area, both when considering overall ease (Gramian trace) and worst-case ease (Gramian minimum eigenvalue) (Figure 3.6A, larger values indicate greater reachability). We next considered how reachability ease varied throughout the delay period. We found that the visual area could drive the cognitive area more easily at the beginning of the delay, with ease decreasing into the middle of the delay period (Figure 3.6B, bottom). The cognitive area was able to drive the visual area more easily slightly later in the delay period (Figure 3.6B, top). Finally, we considered whether reachability changes over the course of learning. We found that both directions of reachability

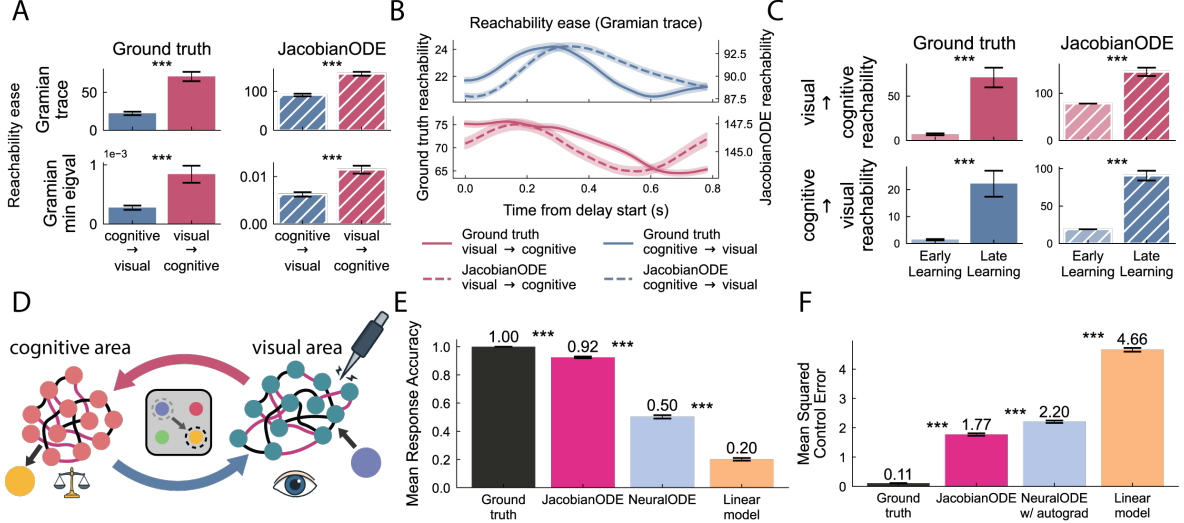


Figure 3.6: JacobianODE reveals differential interareal reachability. Error bars indicate standard deviation for (A) and (C), and standard error for (B), (E), and (F), with statistics computed over trajectories. (A) Comparison of interareal reachability between ground truth and JacobianODE estimates. (B) Temporal evolution of reachability (Gramian trace) throughout the delay period. (C) Comparison of reachability in early and late learning. (D) Schematic illustrating targeted control of cognitive-area via visual-area stimulation. (E) Mean accuracy of targeted responses using JacobianODE versus alternative methods. (F) MSE for multiple ILQR-based methods.

increased after learning (Figure 3.6C). The JacobianODE reproduced all results accurately (Figure 3.6A-C, comparison with ground truth). Our analysis reveals that reachability is crucial in the RNN’s ability to perform the working memory task, with the visual area’s ability to drive the cognitive area shortly after the cue being especially important.

3.6.4 Controlling the task-trained RNN

To further validate our approach, we used JacobianODE-learned Jacobians to control the task-trained RNN (Figure 3.6D). Given a trial in which the network was cued to respond with a particular color, we tested if we could induce a specific incorrect response by input to the visual area alone. To do so, we implemented Iterative Linear Quadratic Regulator (ILQR) control, which relies on knowledge of the Jacobian.^{369,370} The controller guided the network towards the mean hidden state of training trials corresponding to the desired incorrect color. We defined accuracy as the percentage of time points during which the RNN output the desired color. We found that the JacobianODE widely outperformed the baseline models on this task, achieving an accuracy nearing that of the ground truth system (Figure 3.6E). The JacobianODE was furthermore able to achieve the lowest mean squared error on the desired control trajectory (Figure 3.6F). This illustrates that while both the JacobianODE and the NeuralODE can learn the dynamics, only the JacobianODE learns a representation of the Jacobian that is sufficient for control.

3.7 Discussion

Extended Jacobian estimation A natural extension of JacobianODE models would add inductive biases for particular classes of dynamics. For example, one could parameterize a negative definite matrix and thus ensure contracting dynamics.³⁵² Other extensions could include an L1 penalty to encourage sparsity, as well as the inclusion of known dynamic structure (e.g., hierarchical structure, low-rank structure, etc.). In neuroscience, connectomic constraints could be incorporated to capture interactions within a neural circuit.

Limitations A future challenge for JacobianODE models is partially observed dynamics. Recent work has identified that it is possible to learn latent embeddings that approximately recover the true state from partial observation.^{89,193,371–373} Jacobian-based dynamics learning has been performed in a latent space,^{352,374} however it is unclear whether this translates to accurate Jacobian estimation, given the challenges related to automatic differentiation and Jacobian estimation presented here. We also note that reachability estimates depend sensitively on several factors: the alignment of cross-subsystem interactions and within-subsystem dynamics, the eigenvectors and eigenvalues of within-subsystem Jacobians, and the way activity propagates within each subsystem (see Appendix C.3.1). While our method reliably captures broad trends in reachability over time, fine-grained, time-point-specific comparisons should be interpreted with caution. Finally, although JacobianODE models scale well to moderately high-dimensional systems, their performance in systems that are orders of magnitude larger than those considered here (e.g., recordings of thousands of neurons via calcium imaging) remains to be tested. In many practical settings, this may not be a limitation: neural representations during tasks often exhibit intrinsic dimensionalities comparable to those studied here, enabling JacobianODE models to operate in reduced dimensionality (see Appendix C.3.7). Notably, most models converged in under 45 minutes on a single GPU, suggesting favorable scaling (see Table C.7). Future work should explore the viability of the method in higher dimensions, and assess whether dimensionality reduction strategies (such as latent state models or low-rank Jacobian approximations) can further improve scalability and accuracy.

Broader impact on dynamical systems We demonstrated in this work that it is possible to learn high accuracy Jacobians of arbitrary dynamical systems from data. While in this work we focus on control-theoretic analyses, we emphasize that JacobianODE models can be trained in any dynamical setting. Jacobians are widely applicable, and are essential in stability analysis,^{25,127,179,375–377} contraction theory,^{54,179,182,353,378,379} and nonlinear control,^{322,369,370,380–382} among many other settings.

Broader impact on control-theoretic analyses The control-theoretic analysis presented here is performed with respect to a particular reference trajectory. It therefore describes the local controllability properties along the reference trajectory, rather than global properties for the overall system.³²² This locality is desirable in biological settings due to the context-sensitive nature of the relevant processes. Applications of this approach could involve a

comparison of the functioning of the autonomic nervous system, or gene regulatory networks, between control and disease states, given the association of these processes with disease.^{383–391} Future work should investigate whether a more global Jacobian-based control analysis (e.g., contraction theory) is achievable in a data-driven manner via JacobianODE models. This would be useful especially in engineering settings where global controllability is desirable, such as robotics and prosthetic limbs.

Acknowledgments

The authors would like to thank Andrew Kirjner, Laureline Logiaco, Federico Claudi, Lakshmi Narasimhan Govindarajan, Jaedong Hwang, and Akhilan Boopathy for providing stimulating discussion about this work. The authors also acknowledge the MIT Office of Research Computing and Data for providing high-performance computing resources that have contributed to the research results reported within this paper. I.R.F. acknowledges funding from The National Science Foundation Computer and Information Science and Engineering Directorate, The Simons Collaboration on the Global Brain, and The McGovern Institute at MIT. E.K.M. acknowledges funding from ONR MURI N00014-23-1-2768, NIMH 1R01MH131715-01, NIMH R01MH11559, The Simons Center for the Social Brain, The Freedom Together Foundation, and The Picower Institute for Learning and Memory. L.K. acknowledges funding from the Goldstine Postdoctoral Fellowship at IBM Research. M.O. acknowledges funding from NSF GRFP 2141064.

Chapter 4

Disruption of interareal control during propofol anesthesia

Authors: Adam J. Eisen, André M. Bastos, Jacob A. Donoghue, Scott L. Brincat, Emery N. Brown, Ila R. Fiete^{**}, and Earl K. Miller^{**}.

Statement of contribution and publication

*The work presented in this chapter is available on bioRxiv as of 2026.*⁴⁰

^{**}*I.R.F. and E.K.M. are co-senior authors.*

4.1 Abstract

Anesthetic-induced unconsciousness may arise partly from a change in how brain areas can manipulate each other's activity. To quantify this change, we use control-theoretic tools that can precisely characterize how easily subsystems in complex networks can control each other. These tools rely on the Jacobian of the dynamics, an object that fully specifies how inputs to a function affect the outputs. We built on JacobianODE, a method for data-driven Jacobian learning, to enable its application to neural recordings. We developed a deep learning framework that recovers directional, nonlinear control from partially observed multi-area recordings by combining delay-coordinate embedding, a volume-preserving invertible encoder, and latent dynamics via JacobianODE. We validated our framework on the Lorenz system and a partially observed working-memory recurrent neural network. We then applied it to local field potential recordings from posterior parietal (PPC), superior temporal gyrus (STG), frontal eye fields (FEF), and ventrolateral prefrontal cortex (vLPFC) in two non-human

primates, comparing wakefulness with propofol anesthesia. Anesthesia pervasively reduced the ability of areas to control each other, both in terms of driving towards novel states and stabilizing along existing trajectories. This change in ease of control was driven by a decrease in the magnitude of interareal coupling. Directional ease of driving control from PPC to vLPFC and FEF to vLPFC was increased under anesthesia, providing a potential mechanism for paradoxical excitation observed during propofol infusion. Together, these results recast anesthetic unconsciousness as a directional breakdown of cortical control.

4.2 Introduction

Neural computation is supported by the dynamic, directional interactions between brain areas. Cross-area interaction underwrites a remarkably broad swath of neural function:^{277,278} it mediates the selective routing of attention,^{44,279–286} supports decision making,^{287,288} sustains working memory,^{289,290} binds features into coherent percepts,^{291,292} coordinates motor output,^{293–297} and underlies learning and memory.^{298–302} It is therefore unsurprising that interareal interactions are thought to underlie conscious experience itself.^{11,16,21,275,276,392–394}

General anesthetics produce a profound, reversible loss of consciousness. They thus present a promising setting for the study of consciousness in the brain.^{148–153} By comparing anesthetic-induced changes in neural dynamics between conscious and unconscious states, and identifying robustly altered interareal interactions, we aim to gain insight into the neural mechanisms underlying consciousness. Propofol, a GABA_A (γ -aminobutyric acid type A) receptor agonist, changes the balance between excitation and inhibition.^{41,42,45,46,155,156,395} It has been shown to produce robust disruptions to cortical connectivity and integration.^{41,55,131,149,154,176,177,234,241,396–404}

Control theory formalizes how external inputs can be coordinated with a system’s intrinsic dynamics in order to elicit desired behaviors, with applications spanning engineering, robotics, and biology (Figure 4.1A). Given the dynamic nature of cortical computation, control theory is a natural lens through which to interpret interactions in the brain (Figure 4.1B).^{182,294,306–311}

Two complementary quantities formalize measurements of ease of control (Figure 4.1C–E). Along a reference trajectory, *reachability* measures how easily the target system can be driven away from its trajectory, while *controllability* measures how easily the target system can be driven back toward its trajectory.³⁶⁷ Both trade off against the target’s dynamic stability (Figure 4.1C). Consider a passenger jet and a fighter jet attempting to fly in a straight line. The passenger jet is more dynamically stable, meaning it more easily returns back to its original heading after being perturbed by disturbances such as turbulence. When the goal is to stay on course (controllability, Figure 4.1C, top and Figure 4.1D, right), the passenger jet can more easily be guided back to the straight line path in the face of disturbances. However, when the planes must deftly navigate around storm clouds (reachability, Figure 4.1C, bottom and Figure 4.1D, left), the fighter jet can more easily be driven to deviate from its original

path and sharply steer around the obstacles.

Reachability and controllability thus have opposite relationships with the dynamic stability of the target system (Figure 4.1E, top left). As dynamic stability increases, controllability increases while reachability decreases. Coupling strength, on the other hand, is positively correlated with both reachability and controllability (Figure 4.1E, top right). Greater alignment of the coupling matrix with the leading eigenvector of the target area increases reachability (perturbations propagate along more unstable modes) but decreases controllability (harder to stabilize more unstable modes) (Figure 4.1E, bottom left). Accordingly, coupling intrinsic dimensionality can have varying effects, depending on whether the increased dimensionality distributes onto slower or faster modes of the target area (Figure 4.1E, bottom right). Thus reachability and controllability together depend intricately on a multitude of factors, including dynamic stability, coupling strength, as well as coupling dimensionality and the alignment of the inputs with the target system’s dynamics.

The lack of attentional control and awareness during anesthetic-induced unconsciousness suggests a dramatic change in control dynamics. Indeed, propofol and other general anesthetics have been shown using linear network control theory to increase the transition energy required to switch states from functional imaging data.⁴⁰⁵ While existing neural control analyses have leaned on linear formulations.^{140,142,144,312–319,406,407}, the brain’s dynamics are fundamentally nonlinear, and the context-dependent forms of control they afford cannot be captured in their entirety by linear models, motivating a nonlinear treatment of interareal control.^{39,320,321,408,409}

In this work, we leverage nonlinear control theory to characterize how interareal interactions reorganize under propofol anesthesia from intracortical depth electrodes recording local field potentials (LFPs). Our analysis uses the Jacobian of the dynamics, a matrix-valued function that captures the local interdependence of the dynamics of every component of the system. We reproduce findings from previous work on propofol anesthesia that found propofol induces destabilized dynamics.^{37,38} We then characterize how propofol dismantles both reachability and controllability in the brain, on top of demonstrating a reduction in the magnitude and dimensionality of local interareal coupling. Overall, our results suggest that propofol anesthesia disrupts consciousness through the disruption of interareal directional control dynamics.

4.3 Jacobian-based directional control estimation

We consider nonlinear dynamical systems in $\mathbb{R}^d$. The full system is defined by the dynamics $\dot{\mathbf{u}}(t) = \mathbf{f}(\mathbf{u}(t))$. We assume that within this full system, there are K coupled subsystems, each of which is defined by the dynamics $\dot{\mathbf{u}}^k(t) = \mathbf{f}^k(\mathbf{u}(t))$, $k = 1, \dots, K$. The subsystem dimensionalities are d_k , and $d = \sum_{k=1}^K d_k$. The system is then observed through a smooth map $\mathbf{x}(t) = \mathbf{h}(\mathbf{u}(t)) \in \mathbb{R}^{d_{\text{obs}}}$ with $\mathbf{h} : \mathbb{R}^d \rightarrow \mathbb{R}^{d_{\text{obs}}}$ that projects the full state vector $\mathbf{u}$ onto a lower-dimensional subspace. We will assume that our observation of each subsystem depends only on

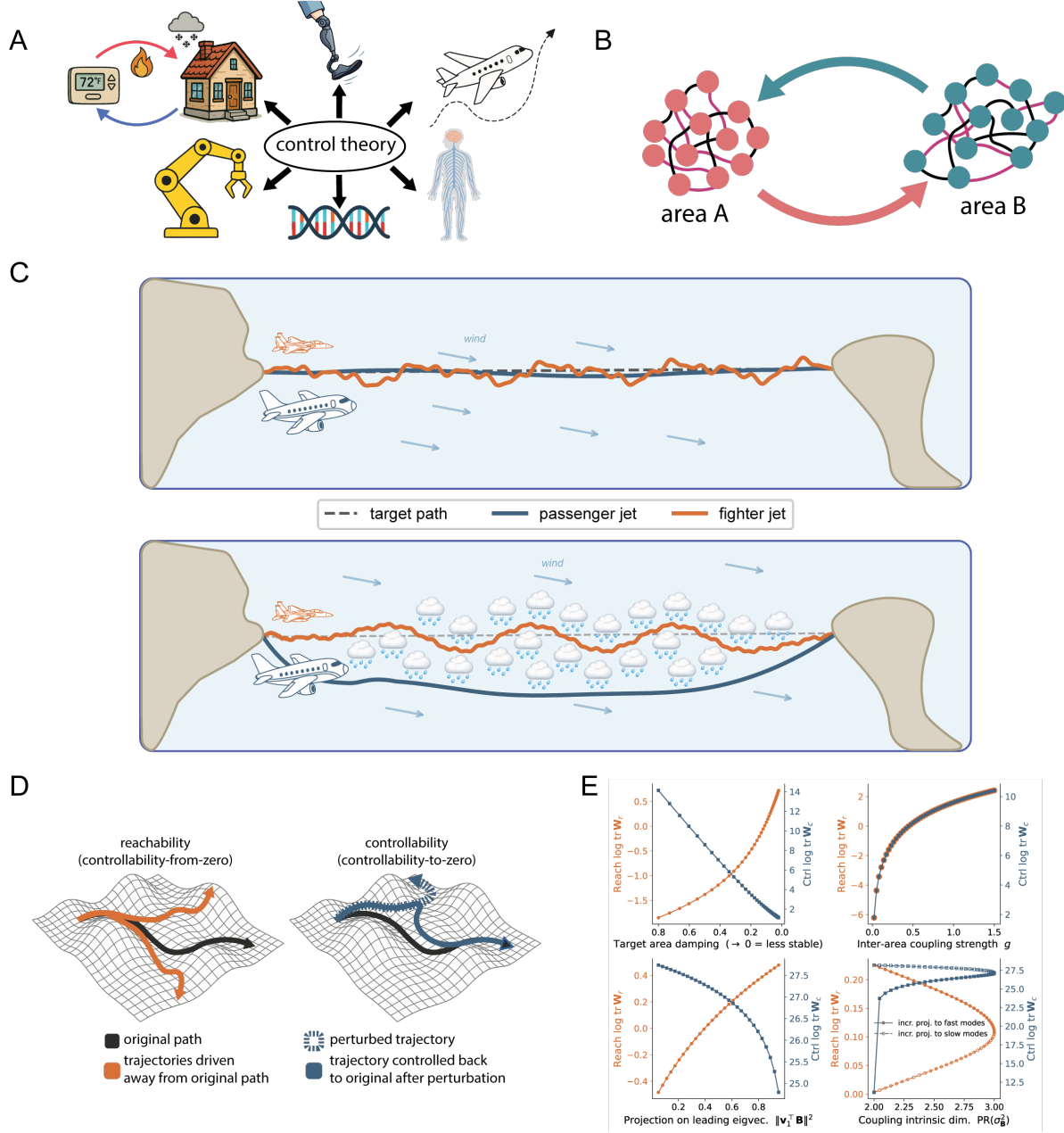


Figure 4.1: A control theoretic framing of interareal control. (A) Control problems share a common formal structure across engineering, biology, and brain networks. (B) Bidirectional control interactions between two areas. (C) Conceptual illustration of reachability and controllability interacting with dynamic stability. (D) Abstract illustration of reachability and controllability. (E) The dependence of reachability and controllability on dynamic stability, coupling magnitude, coupling alignment, and coupling dimensionality in a synthetic two-area linearization.

the state of the subsystem itself, and not on the state of the other subsystems. That is, we can decompose the observation map as $\mathbf{x}(t) = \mathbf{h}(\mathbf{u}(t)) = [\mathbf{h}^1(\mathbf{u}^1(t)), \mathbf{h}^2(\mathbf{u}^2(t)), \dots, \mathbf{h}^K(\mathbf{u}^K(t))]$.

4.3.1 Jacobian-based directional control estimation in fully observed dynamical systems

We begin with the simplest case discussed in the original JacobianODE paper³⁹, when $\mathbf{h}$ is the identity map, so $\mathbf{x}(t) = \mathbf{u}(t)$. The Jacobian of the dynamics is a matrix-valued function $\mathbf{J}_f : \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d}$ (henceforth, $\mathbf{J}$) given by

$$\mathbf{J}(\mathbf{x}(t)) = \frac{\partial}{\partial \mathbf{x}} \mathbf{f}(\mathbf{x}(t)) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_d} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_d}{\partial x_1} & \frac{\partial f_d}{\partial x_2} & \cdots & \frac{\partial f_d}{\partial x_d} \end{bmatrix}. \quad (4.1)$$

At each time t , the Jacobian captures how perturbations to the system will propagate. This recasts nonlinear dynamics as linear time-varying dynamics in the tangent space locally along trajectories (formally, $\delta \dot{\mathbf{x}}(t) = \mathbf{J}_f(\mathbf{x}(t))\delta \mathbf{x}(t)$) (Figure 4.2A, left).

To characterize control, consider the following formulation. As discussed in the original JacobianODE paper, the tangent space dynamics around the reference trajectory for subsystem $\alpha \in \{1, \dots, K\}$ are given by

$$\delta \dot{\mathbf{x}}^{(\alpha)}(t) = \sum_{k=1}^K \mathbf{J}^{k \rightarrow \alpha}(\mathbf{x}(t)) \delta \mathbf{x}^{(k)}(t)$$

where $\mathbf{J}^{k \rightarrow \alpha}(\mathbf{x}(t)) \in \mathbb{R}^{d_\alpha \times d_k}$ is the submatrix of $\mathbf{J}(\mathbf{x}(t))$ in which the columns correspond to subsystem k and the rows correspond to subsystem α . Now, for a given subsystem $\beta \in \{1, \dots, K\}$ with $\beta \neq \alpha$, we wish to analyze the ease with which β can control α locally around the reference trajectory, without intervention from other subsystems. Discounting the interventions from other subsystems equates to setting $\delta \mathbf{x}^{(k)}(t) = 0$ for $k \neq \alpha, \beta$, leaving the expression

$$\delta \dot{\mathbf{x}}^{(\alpha)}(t) = \mathbf{J}^{\alpha \rightarrow \alpha}(\mathbf{x}(t)) \delta \mathbf{x}^{(\alpha)}(t) + \mathbf{J}^{\beta \rightarrow \alpha}(\mathbf{x}(t)) \delta \mathbf{x}^{(\beta)}(t) \quad (4.2)$$

which quantifies the local dynamics of subsystem α with input from subsystem β , in the absence of perturbations from any other subsystem (Figure 4.2A, right).

To quantify the ease of control subsystem β can exert over subsystem α , we can use the reachability and controllability Gramians, defined in continuous time as

$$\mathbf{W}_r(t_0, t_1) \triangleq \int_{t_0}^{t_1} \Phi(t_1, \tau) \mathbf{B}(\tau) \mathbf{B}^T(\tau) \Phi^T(t_1, \tau) d\tau \quad (4.3)$$

and

$$\mathbf{W}_c(t_0, t_1) \triangleq \int_{t_0}^{t_1} \Phi(t_0, \tau) \mathbf{B}(\tau) \mathbf{B}^T(\tau) \Phi^T(t_0, \tau) d\tau \quad (4.4)$$

respectively. In these equations, Φ denotes the state-transition matrix of the intrinsic dynamics of subsystem α without any input (i.e., $\delta \mathbf{x}^{(\alpha)}(t) = \Phi(t, t_0) \delta \mathbf{x}^{(\alpha)}(t_0)$) and B denotes the coupling Jacobian $\mathbf{J}^{\beta \rightarrow \alpha}(\mathbf{x}(t))$.³⁶⁵

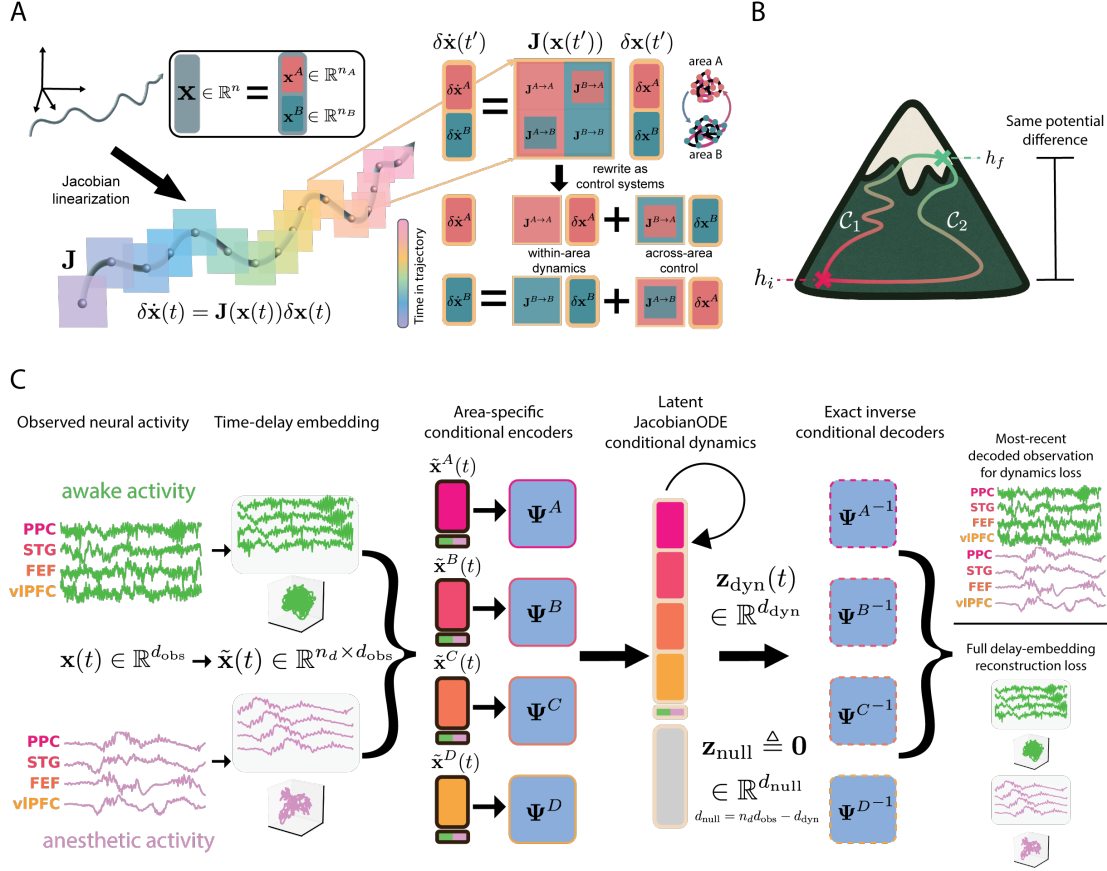


Figure 4.2: Latent JacobianODE pipeline for partially observed multi-area neural recordings. (A) Jacobians linearize dynamics along a reference trajectory, enabling the construction of pairwise interacting control systems. (B) Jacobians obey path independence, analogous to achieving the same potential energy difference taking two different paths up a mountain with the same endpoints. (C) The latent JacobianODE deep learning framework, demonstrated using the four-area neural recordings analyzed in this work.

4.3.2 Deep Jacobian estimation: learning Jacobians from data

Given only observed trajectories $\mathbf{x}^{(j)}(t_0^{(j)} + k\Delta t)$, $k = 0, 1, \dots$, and no access to the underlying dynamics $\mathbf{f}$, we estimate $\mathbf{J}$ directly by parameterizing it as a neural network $\hat{\mathbf{J}}^\theta$ with learnable parameters θ .³⁹ The core observation is that each row of $\mathbf{J}$ is a conservative vector field (the gradient of a scalar potential), so its path integral between any two points is path-independent^{351,352}: as in the work done climbing a mountain, the result depends only on the start and end points, not the route taken (Figure 4.2B). Formally,

$$\mathbf{f}(\mathbf{x}(t_f)) - \mathbf{f}(\mathbf{x}(t_i)) = \int_{\mathcal{C}} \hat{\mathbf{J}}^\theta ds, \quad (4.5)$$

where $\mathcal{C}$ is any piecewise-smooth curve joining $\mathbf{x}(t_i)$ to $\mathbf{x}(t_f)$. Together with a Jacobian-based parameterization of the reference value $\mathbf{f}(\mathbf{x}(t_i))$ so that all gradients backpropagate through θ (see Appendix D.2.1), Equation 4.5 yields an estimate of the time derivative everywhere along the trajectory, which a standard ODE integrator advances to predict $\hat{\mathbf{x}}(t_f + \Delta t)$.

Trajectory loss. Given an observed trajectory $\mathbf{x}(t_0 + k\Delta t)$, $k = 0, \dots, T - 1$, the trajectory reconstruction loss $\mathcal{L}_{\text{traj}}(\theta; \mathbf{x})$ penalizes deviation between predicted and true states under an appropriate distance (typically mean squared error, i.e. MSE). Recursive predictions in chaotic or noisy systems are stabilized by generalized teacher forcing, which partially anchors the rollout to ground-truth states.³⁵⁴

Loop-closure loss. Trajectory loss only constrains $\hat{\mathbf{J}}^\theta$ along the direction of the flow, leaving the orthogonal tangent-space directions underdetermined. Exploiting conservativity once more, the path integral of $\mathbf{J}$ around any piecewise-smooth closed loop $\mathcal{C}_{\text{loop}}$ must vanish (Figure 4.2C). The self-supervised loop-closure regularizer³⁵⁵,

$$\mathcal{L}_{\text{loop}}(\theta; \mathbf{x}) = \left\| \oint_{\mathcal{C}_{\text{loop}}} \hat{\mathbf{J}}^\theta ds \right\|_2^2, \quad (4.6)$$

penalizes the deviation of this integral from zero. In practice, we form from concatenations of line segments between randomly-sampled data points, sampling diverse tangent-space directions on the data manifold while remaining cheap to compute.

Training objective. The total loss combines the two terms,

$$\mathcal{L}(\theta; \mathbf{x}) = \mathcal{L}_{\text{traj}}(\theta; \mathbf{x}) + \lambda_{\text{loop}} \mathcal{L}_{\text{loop}}(\theta; \mathbf{x}), \quad (4.7)$$

with λ_{loop} a tunable hyperparameter controlling the weight of the loop-closure regularizer.

4.3.3 A latent JacobianODE pipeline for partially observed dynamical systems

Delay embedding. To recover the full system dynamics from partial observations, we can leverage Takens’ delay embedding theorem. The theorem guarantees that by looking backwards in time at the history of the observed data, we can reconstruct a manifold that is diffeomorphic to the full system state space. Specifically, we form delay-embedded vectors $\tilde{\mathbf{x}}(t) = [\mathbf{x}(t), \mathbf{x}(t - \tau\Delta t), \dots, \mathbf{x}(t - (n_d - 1)\tau\Delta t)] \in \mathbb{R}^{d'}$, with $d' = n_d d_{\text{obs}}$. Taking the intrinsic dimensionality of the system to be d^* , Takens’ theorem guarantees that for generic $\mathbf{h}$ and $n_d \geq 2d^* + 1$, the map $\tilde{\mathbf{x}} \mapsto \mathbf{u}$ is a diffeomorphism from the embedded manifold onto the full system state space. We will again assume subsystem-specificity of the observation map, so that $\tilde{\mathbf{x}}(t)$ can be written as $\tilde{\mathbf{x}}(t) = [\tilde{\mathbf{x}}^1(t); \tilde{\mathbf{x}}^2(t); \dots; \tilde{\mathbf{x}}^K(t)] \in \mathbb{R}^{d'}$, and the diffeomorphism is given by $\mathbf{u} = \psi(\tilde{\mathbf{x}}) = [\psi^1(\tilde{\mathbf{x}}^1(t)); \psi^2(\tilde{\mathbf{x}}^2(t)); \dots; \psi^K(\tilde{\mathbf{x}}^K(t))]$.

Diffeomorphic encoder. The latent version of the JacobianODE model builds on the insight from Takens’ theorem, and following previous work,⁴¹⁰ leverages diffeomorphic encoders to learn a latent representation that is diffeomorphic to the delay embedding. Specifically, we delay embed the observed data, and construct an encoder that consists of coupling layers interleaved with orthogonal mixing layers. This results in an encoder that is both diffeomorphic and exactly invertible (Figure 4.2C, left). In practice, we use additive coupling

layers for the coupling component, and Cayley-parameterized orthogonal mixers for the mixing component.^{411–413} Formally, an invertible additive coupling layer takes as input a vector $\mathbf{y} \in \mathbb{R}^{d'}$, partitions it into two halves $\mathbf{a} \in \mathbb{R}^{d'/2}$ and $\mathbf{b} \in \mathbb{R}^{d'/2}$, and applies

$$\mathbf{a}' = \mathbf{a}, \quad \mathbf{b}' = \mathbf{b} + \mathbf{m}_\vartheta(\mathbf{a}), \quad (4.8)$$

where $\mathbf{m}_\vartheta : \mathbb{R}^{d'/2} \rightarrow \mathbb{R}^{d'/2}$ is a (possibly non-invertible) neural network. The Jacobian of this map is

$$\frac{\partial(\mathbf{a}', \mathbf{b}')}{\partial(\mathbf{a}, \mathbf{b})} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \nabla_{\mathbf{a}} \mathbf{m}_\vartheta & \mathbf{I} \end{bmatrix}, \quad (4.9)$$

which has $\det = 1$ identically, regardless of $\mathbf{m}_\vartheta$. The inverse is $\mathbf{a} = \mathbf{a}'$, $\mathbf{b} = \mathbf{b}' - \mathbf{m}_\vartheta(\mathbf{a}')$.

To mix the dimensions in between coupling layers, we use a Cayley-parameterized orthogonal mixer. Formally, a Cayley-parameterized orthogonal mixer takes as input a vector $\mathbf{y} \in \mathbb{R}^{d'}$ and applies an orthogonal map given by $\mathbf{y}' = \mathbf{Q}\mathbf{y}$, where $\mathbf{Q} = (\mathbf{I} - \mathbf{A})(\mathbf{I} + \mathbf{A})^{-1}$ and $\mathbf{A}$ is a learnable skew-symmetric matrix. The determinant of $\mathbf{Q}$ is 1.

Thus the encoder as a whole is volume preserving (its determinant is 1), meaning it cannot arbitrarily reduce the geometric volume of the observed data to minimize noise. To handle multiple subsystems, we instantiate a separate encoder for each subsystem, so that the total encoder can be written as $\mathbf{z} = \Psi(\tilde{\mathbf{x}}(t)) = [\Psi^1(\tilde{\mathbf{x}}^1(t)); \Psi^2(\tilde{\mathbf{x}}^2(t)); \dots; \Psi^K(\tilde{\mathbf{x}}^K(t))]$.

A diffeomorphic function by construction must map to a space of the same dimensionality. However the intrinsic dimensionality of a system is in practice typically much smaller than the dimension of the delay embedding. We partition the latent space into two components: a dynamic latent $\mathbf{z}_{\text{dyn}} \in \mathbb{R}^{d_{\text{dyn}}}$ and an effectively-null latent $\mathbf{z}_{\text{null}} \in \mathbb{R}^{d_{\text{null}}} = \mathbb{R}^{n_d d_{\text{obs}} - d_{\text{dyn}}}$, with $\mathbf{z} = [\mathbf{z}_{\text{dyn}}; \mathbf{z}_{\text{null}}] \in \mathbb{R}^{n_d d_{\text{obs}}}$ (Figure 4.2C, center). The null subspace is zero-padded for all decoder calls (i.e., inverse encoder calls). Thus while the null subspace is allowed to freely vary, all information for both dynamics and reconstruction must live in the dynamic latent. The JacobianODE model then operates on $\mathbf{z}_{\text{dyn}}$ only. Note that since the encoder is volume-preserving, as mentioned, all information in the input must be preserved *somewhere* in the latent space. Thus the null subspace provides the opportunity to dump any noise or information not relevant to the dynamics or reconstruction. In practice, to pick the dimensionality of the dynamic subspace d_{dyn} , we perform a linear PCA on the delay-embedded training data and select the number of components that explain 99% of the variance. This is because linear dimensionality provides an upper bound on the intrinsic dimensionality of the system.

Reconstruction loss. The reconstruction loss $\mathcal{L}_{\text{rec}}(\theta_\Psi; \mathbf{x})$ penalizes the deviation between true and decoded latent states, where θ_Ψ are the parameters of the encoder (Figure 4.2C, right). We define a projection operator P_{dyn} that projects the latent into the dynamic subspace. Then, the reconstruction loss is given by

$$\mathcal{L}_{\text{rec}}(\theta_\Psi; \mathbf{x}) = \mathbb{E}_t [\|\Psi^{-1}([P_{\text{dyn}}\Psi(\tilde{\mathbf{x}}(t)); \mathbf{0}]) - \tilde{\mathbf{x}}(t)\|^2]. \quad (4.10)$$

Latent prediction loss. The latent prediction loss $\mathcal{L}_{\text{pred},\mathbf{z}}(\theta; \mathbf{x})$ penalizes the deviation between true and predicted latent dynamics, where $\theta = [\theta_\Psi; \theta_J]$ are the parameters of the encoder and JacobianODE model. Formally,

$$\mathcal{L}_{\text{pred},\mathbf{z}}(\theta; \mathbf{x}) = \mathbb{E}_t \left[\sum_{s=1}^T \left\| \hat{\mathbf{z}}(t + s\Delta t) - \mathbf{z}(t + s\Delta t) \right\|^2 \right], \quad (4.11)$$

where $\hat{\mathbf{z}}(t + s\Delta t)$ is the predicted latent state at time $t + s\Delta t$ and T is the number of prediction steps.

Decoded prediction loss. The decoded prediction loss $\mathcal{L}_{\text{pred},\mathbf{x}}(\theta; \mathbf{x})$ penalizes the deviation between true and predicted observed states. To minimize the amount of redundant information, we compute the decoded prediction loss only on the most recent slice of each predicted delay-embedded window (Figure 4.2C, right). Formally,

$$\mathcal{L}_{\text{pred},\mathbf{x}}(\theta; \mathbf{x}) = \mathbb{E}_t \left[\sum_{s=1}^T \left\| P_{\text{recent}} \Psi^{-1} ([P_{\text{dyn}} \hat{\mathbf{z}}(t + s\Delta t); \mathbf{0}]) - \mathbf{x}(t + s\Delta t) \right\|^2 \right], \quad (4.12)$$

where P_{recent} is a projection operator that projects the decoded latent into the most recent slice of the delay-embedded window, and $\mathbf{x}(t + s\Delta t)$ is the true observed state at time $t + s\Delta t$.

Training objective. The total loss is given by

$$\mathcal{L}(\theta; \mathbf{x}) = \mathcal{L}_{\text{rec}}(\theta_\Psi; \mathbf{x}) + \mathcal{L}_{\text{pred},\mathbf{z}}(\theta; \mathbf{x}) + \mathcal{L}_{\text{pred},\mathbf{x}}(\theta; \mathbf{x}) + \lambda_{\text{loop}} \mathcal{L}_{\text{loop}}(\theta; \mathbf{x}), \quad (4.13)$$

where λ_{loop} is a tunable hyperparameter controlling the weight of the loop-closure regularizer. Loop-closure loss is computed in latent space. The reconstruction loss and both prediction losses are given the same weight since the encoder is volume-preserving, and the relevant states should have approximately the same scale.

4.3.4 Conditional training

In the case where the data span multiple conditions, either discrete (e.g., awake vs. anesthesia) or continuous (e.g., different drug doses), we can train the JacobianODE model to be condition-aware. Specifically, we can add a conditional vector $\mathbf{c} \in \mathbb{R}^{d_c}$ to the input of both the encoder and the JacobianODE model, and train the model to predict the Jacobian of the dynamics conditioned on this variable. In practice, we take $c \in \{-1, +1\}$ to represent the difference between conditions.

4.4 Results

4.4.1 Validation on synthetic systems with ground-truth dynamics

We validated the pipeline on two simulated systems. In these systems, the true Jacobians are known, so we can compare against ground-truth invariants such as the Lyapunov spectrum, attractor topology, and ground-truth reachability and controllability gramians. The systems we used were the 3-D Lorenz attractor,⁶⁰ and a 128-D recurrent neural network (RNN) trained on a working memory selection task inspired by Panichello and Buschman [368]. For the task-trained RNN, on each trial, the network receives two cue inputs and, after a delay and a selection cue, must sustain activation corresponding to the chosen color. The RNN consists of two 64-neuron areas (visual input and cognitive output), with predominantly within-area connectivity to encourage multi-area structure. Inputs reach only the visual area, outputs are read from the cognitive area, and thus the two areas are forced to interact to solve the task. We trained Jacobians on activity from the autonomous delay epoch before the network provides a response.³⁹

Each system was trained in two variants: fully observed (the JacobianODE machinery receives all d state coordinates) and partially observed (the latent-space pipeline receives a $d_{\text{obs}} \ll d$ projection and recovers a coordinate system via delay embedding and the encoder). For the Lorenz system, only the x variable was observed. For the task-trained RNN, 16 coordinates were randomly chosen from each of the two areas. All models were trained on data with 5% observation noise.

We evaluated trajectory prediction via R^2 and decoder-corrected mean absolute standardized error (MASE). The decoder-corrected MASE normalizes the mean absolute error by the absolute error of the decoded persistence baseline ($\mathbb{E}_t [\|P_{\text{recent}} \Psi^{-1} ([P_{\text{dyn}} \Psi(\tilde{\mathbf{x}}(t)); \mathbf{0}]) - \mathbf{x}(t + \Delta t)\|^2]$), which separates the dynamic prediction quality from the decoder error. Trajectory-prediction quality is reported on held out test data (Figure 4.3A). In both the fully-observed and partially-observed cases, the models are able to reproduce trajectories, though this ability is predictably somewhat impaired for the partially-observed Lorenz system.

The Lyapunov spectrums are recovered by the latent JacobianODE models in the fully observed case (Figure 4.3B). For the partially-observed Lorenz system, the latent JacobianODE formulation recovers the top two exponents, while distributing the third exponent, corresponding to decay onto the attractor, into the remaining 12 latent dimensions (Figure 4.3C). Note that the partially observed latent spectrum has 14 entries due to the method of picking the latent dimension d_{dyn} by PCA on the delay-embedded training data at variance threshold 0.99. The results reflect that under partial observation, the encoded delay-embedded states do not contain the necessary resolution to resolve the third exponent exactly. However, the latent JacobianODE does recover the Kaplan-Yorke dimension of the Lorenz attractor, indicating that the attractor topology is preserved (Figure 4.3D,E). For the partially observed task-trained RNN, the latent JacobianODE model was trained conditioned on the condition

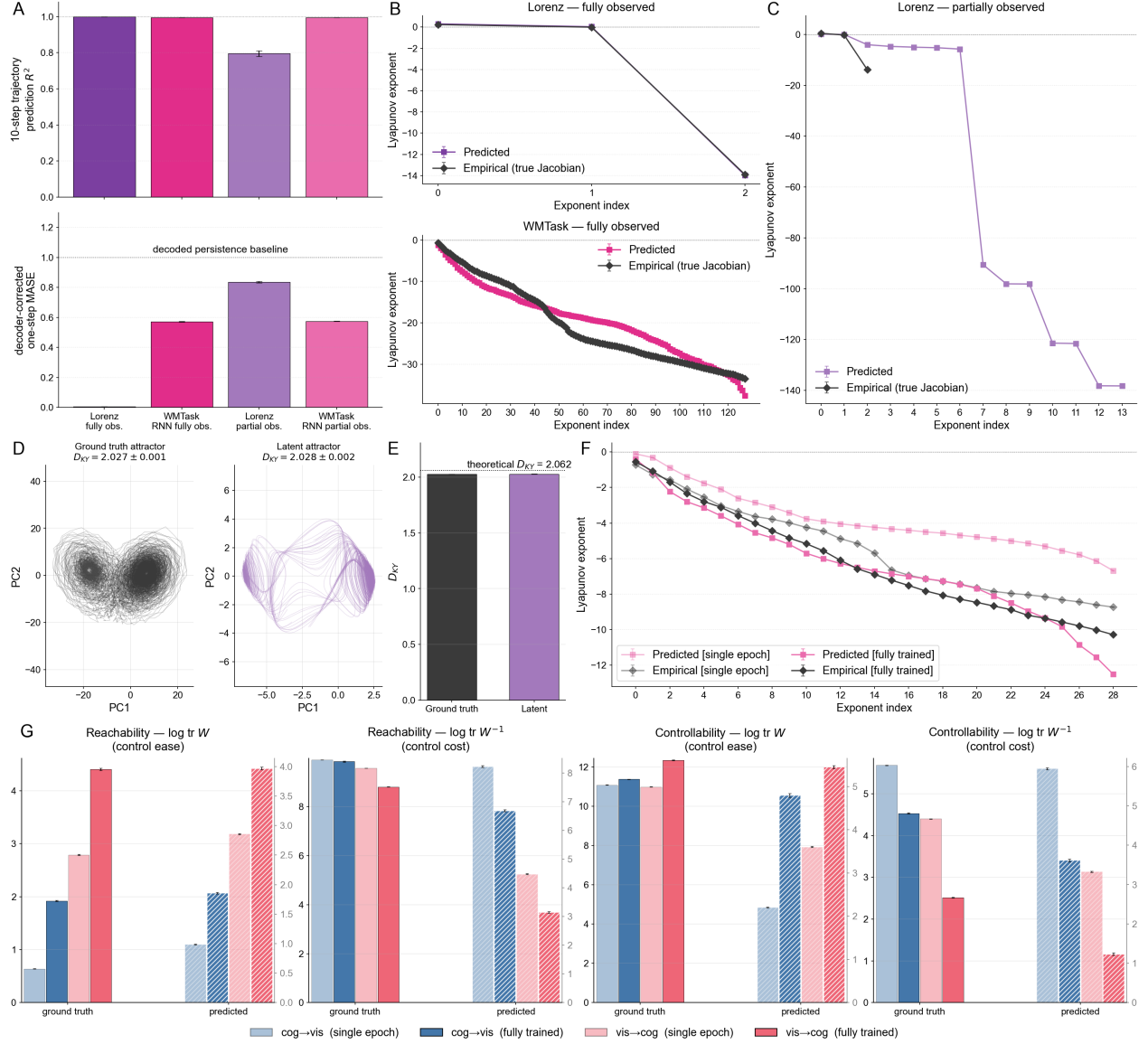


Figure 4.3: Validation on synthetic systems with ground-truth dynamics. (A) Ten-step free-running R^2 (top) and decoder-corrected one-step MASE (bottom) for the four runs. (B) Predicted vs. empirical Lyapunov spectra (sorted descending) for the two fully observed runs. (C) Same for Lorenz partially observed. (D–E) Ground-truth and latent-recovered Lorenz attractor in PC1–PC2 with Kaplan–Yorke dimension annotations. (F) Task-trained RNN single-epoch versus fully trained Lyapunov spectra, predicted (pink) vs. empirical (black). (G) Task-trained RNN partially observed directional reachability and controllability log-trace and inverse log-trace, ground-truth vs. predicted, per condition.

variable $c \in \{-1, +1\}$, indicating the first epoch or the final epoch of training. The latent JacobianODE model qualitatively recovers the leading exponents of the Lyapunov spectrum in each condition, as well as the relevant difference between them (Figure 4.3F).

The latent JacobianODE model also recovers the relative changes in reachability and controllability between the two conditions (Figure 4.3G). Reachability increases from the first epoch to the final epoch of training for both visual to cognitive control and cognitive to

visual control (Figure 4.3G, left and center-left). Additionally, visual to cognitive reachability is higher in both conditions, reflecting the need for the visual area to drive the cognitive area to make the correct decision. Controllability also increases from the first epoch to the final epoch of training for both visual to cognitive control and cognitive to visual control (Figure 4.3G, center-right and right).

4.4.2 JacobianODE models capture neural dynamics and show more nonlinear geometry in awake state relative to anesthesia

We applied the pipeline to multi-area LFP from two non-human primates (NHPs) recorded during propofol infusion. Each session simultaneously recorded LFP from posterior parietal cortex (PPC), superior temporal gyrus (STG, auditory cortex), frontal eye fields (FEF), and ventrolateral prefrontal cortex (vlPFC) at 1 kHz. We low-pass filtered the signals at 80 Hz. We used resting state trajectories for training the JacobianODE model, defined as starting at least 3 seconds after the onset of any external stimulus (either airpuff, or auditory tone). Data was taken from either the awake state or a maintenance dose anesthetic state, and the JacobianODE model was conditioned on the condition variable $c \in \{-1, +1\}$, indicating the state.

Free-running trajectory prediction and decoder-corrected one-step MASE are reported per condition, per NHP (Figure 4.4A). Trajectories are well captured over 10-step free-running prediction. Anesthetic state trajectories were generally easier to predict than awake state ($p = 9.5 \times 10^{-7}$, paired two-sided Wilcoxon signed-rank test, median difference anesthesia – awake $\Delta R^2 = 0.068$ [95% CI 0.060, 0.071]), though for both conditions the decoder-corrected one-step MASEs are below the decoder-corrected persistence baseline of 1.0.

The encoder leverages considerable transformations to encode the dynamics into the latent space. The angle of rotation rises through the depth of the encoder, and rotates the latent space into a more complex geometry (Figure 4.4B). Furthermore, the Riemannian metric induced on the latent space is farther from identity in the awake state than during anesthesia, indicating a more complex geometry in the awake state (Figure 4.4C; $p = 9.5 \times 10^{-7}$, paired two-sided Wilcoxon signed-rank test, median difference anesthesia – awake $\Delta = -0.059$ [95% CI $-0.067, -0.050$]). This suggests that the decreased prediction quality in the awake state relative to anesthesia may be related to the more nonlinear geometry of the data manifold.

4.4.3 Broad-spectrum destabilization during propofol anesthesia

Previous work has shown that neural dynamics are destabilized under propofol anesthesia.^{37,38} We confirmed this destabilization using the latent JacobianODE models. We computed Lyapunov exponents for all resting state trajectories within each epoch of each session and

then averaged across sessions. We report both the asymptotic Lyapunov spectrum as well as the finite time Lyapunov exponent (FTLE) spectrum, which was computed over 10 millisecond windows.

While the largest asymptotic Lyapunov exponents are similar across conditions, the broad spectrum of asymptotic Lyapunov exponents was significantly destabilized under anesthesia, as was previously found (Figure 4.4D,E; $p = 9.5 \times 10^{-7}$, paired two-sided Wilcoxon signed-rank test, median difference full spectrum mean anesthesia – awake $\Delta\langle\lambda\rangle = +5.61$ [95% CI +4.56, +6.78]). For the finite-time Lyapunov exponents, both the maximum Lyapunov exponent and full spectra show destabilizations during anesthesia (Figure 4.4D,E; max exponent $p = 0.038$, paired two-sided Wilcoxon signed-rank test, median paired difference, anesthesia – awake $\Delta\lambda = +3.82$ [95% CI +0.52, +6.97], full spectrum $p = 9.5 \times 10^{-7}$, $\langle\lambda\rangle \Delta = +5.61$ [95% CI +4.56, +6.78]). We note that the maximum Lyapunov exponents are not always faithfully reproduced in the sample systems and may be more sensitive to noise. They should therefore be interpreted with caution. The full spectrum however was robustly recovered by JacobianODE models, and more consistently depicts the destabilization across a multitude of directions.

Per-area asymptotic Lyapunov exponents qualitatively reproduce the shift found at the full system level (Figure 4.4F). Notably, PPC had the most unstable dynamics both in the awake and anesthetic states ($p \leq 3 \times 10^{-6}$, paired two-sided Wilcoxon signed-rank test).

4.4.4 Propofol anesthesia reduces directional reachability and controllability

We next sought to characterize how propofol anesthesia changes interareal reachability and controllability. We computed the directional reachability and controllability Gramians for each condition using the learned Jacobians. We report both the reachability and controllability Gramians computed over 10 millisecond horizons. Each gramian measures the influence of area β on area α through the time-dependent coupling, defined by $\mathbf{J}^{\beta \rightarrow \alpha}$. For $\beta = \alpha$, $\text{tr}(\mathbf{W}_r^{\alpha\alpha}) = \int_{t_0}^{t_1} \|\partial_\tau \Phi_\alpha(t_1, \tau)\|_F^2 d\tau$ and $\text{tr}(\mathbf{W}_c^{\alpha\alpha}) = \int_{t_0}^{t_1} \|\partial_\tau \Phi_\alpha(t_0, \tau)\|_F^2 d\tau$: the integrated squared Frobenius norms of how the within-area state-transition matrices evolve with the integration time. This can be interpreted roughly as the strength of the within-area recurrence, and provides a natural baseline for comparing the trace magnitudes.

Reachability was broadly decreased under anesthesia, both when considering the easiest directions of control (Figure 4.5A, left) and the most difficult directions of control (Figure 4.5A, right). Notably, the reachability along the most manipulable directions became *easier* under anesthesia for the interactions involving both PPC and FEF driving vIPFC. Controllability was also broadly decreased under anesthesia, both when considering the easiest directions of control (Figure 4.5B, left) and the most difficult directions of control (Figure 4.5B, right).

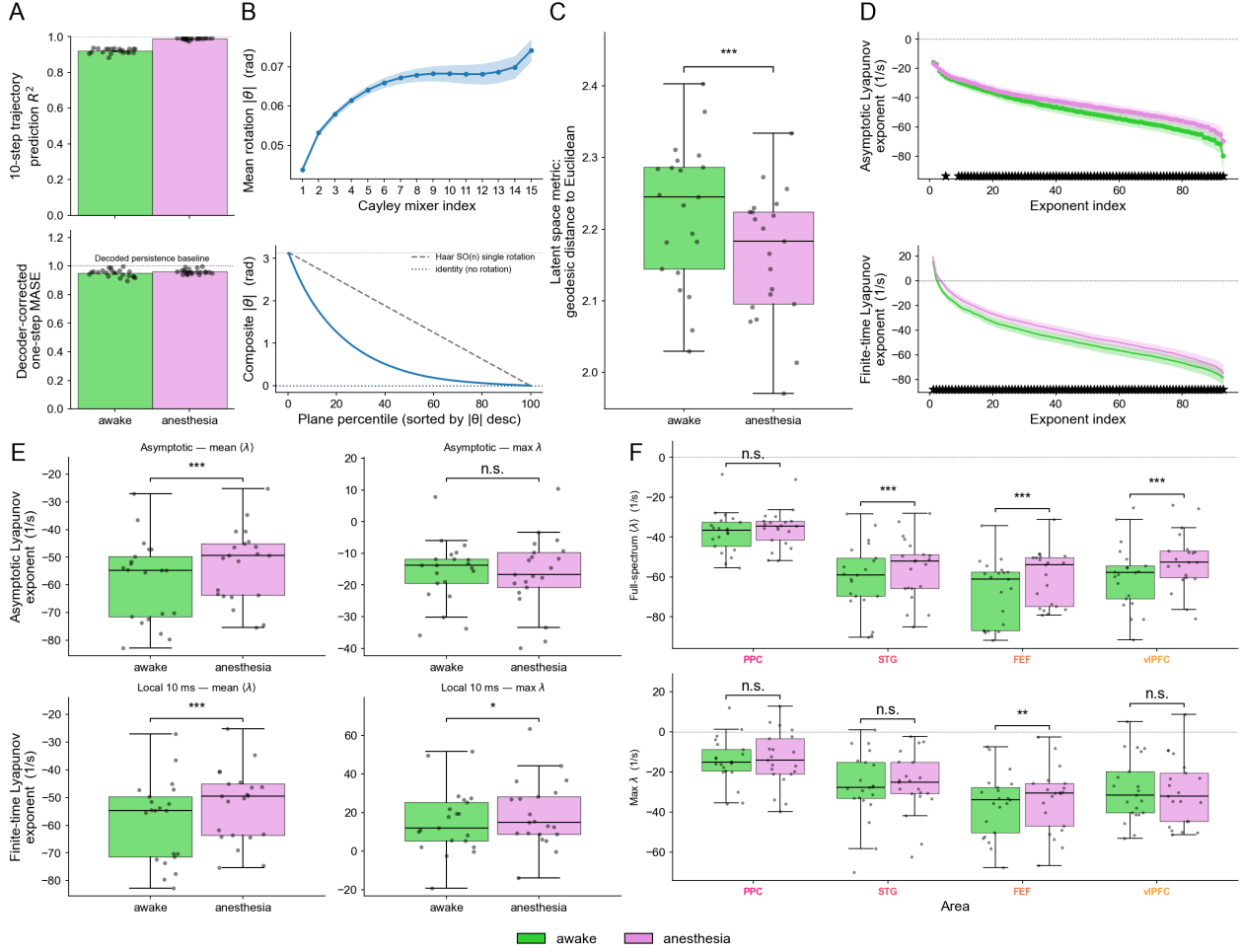


Figure 4.4: Trajectory prediction, encoder geometry, and Lyapunov spectra under propofol ($n = 21$ sessions). (A) Per-session 10-step R^2 and decoder-corrected one-step MASE, awake vs. anesthesia. (B) Mean per-Cayley-layer rotation angle through encoder depth, between identity ($\theta = 0$) and the Haar $SO(n)$ reference. (C) Geodesic distance from identity of the composite encoder Jacobian per session, awake (larger; more warped metric) vs. anesthesia ($p < 0.001$). (D) Full asymptotic (top) and 10-ms FTLE (bottom) Lyapunov spectra per session, sorted descending. (E) Boxplots of asymptotic mean and max exponents and of 10-ms FTLE mean and max exponents. (F) Per-area within-area asymptotic mean (top) and max (bottom) exponents per session for PPC, STG, FEF, vIPFC; sign-agree-gated significance annotated.

These disruptions in reachability and controllability are most evident at the short time horizon (10 milliseconds) presented here. At longer horizons, the intrinsic dynamics (more destabilized under anesthesia) increasingly dominate the Gramians over the interareal coupling. When the target system is less stable, reachability becomes easier and controllability becomes harder. Thus, over longer horizons, reachability becomes easier under anesthesia (particularly for projections to PPC and STG), while the controllability reduction instead deepens (Appendix D.1.2). This illustrates that a reduction in the ability to drive posterior areas manifests on shorter horizons (10, 20, and 100 milliseconds), while on longer timescales (1 second) the intrinsic instability of the target area dynamics leads to greater reachability.

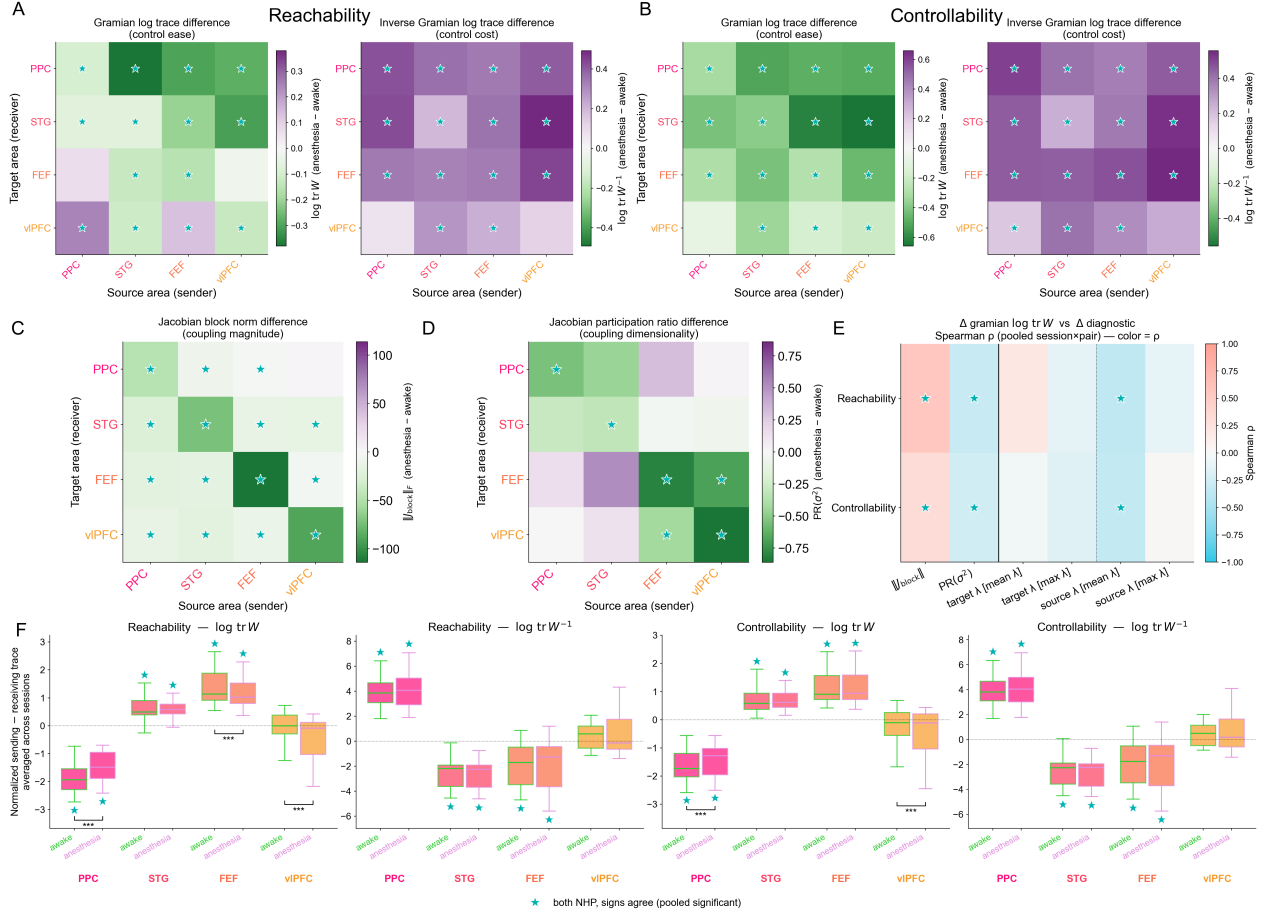


Figure 4.5: Directional reorganization of interareal control under propofol ($n = 21$ sessions). (A) 4×4 source–target heatmaps of pooled anesthesia–awake differences in reachability $\log \text{tr } \mathbf{W}$ (left) and $\log \text{tr } \mathbf{W}^{-1}$ (right); stars mark cells passing the sign-agree + BH-FDR gate at $q < 0.05$. (B) Same for controllability. (C) Jacobian block Frobenius-norm differences. (D) Block participation-ratio differences. (E) Spearman correlation between Gramian-change cells and block-norm/PR-change cells. (F) Per-area net send–receive axis normalized by diagonal within-area measure across all four metrics, placing PPC at the receiver end, FEF/STG at the sender end, vIPFC balanced.

4.4.5 Anesthesia reduces coupling strength and local coupling dimensionality

To probe the influence of anesthesia on coupling strength, we analyzed how anesthesia changes both the Frobenius norm of the cross-area and within-area Jacobian blocks ($\Delta \log \text{tr } \mathbf{W}$ and $\Delta \|\mathbf{J}_{\text{block}}\|$, the difference in Frobenius norm of the coupling blocks between states). We also analyzed the participation ratios of these blocks ($\Delta \text{PR}(\sigma^2)$, i.e. change in the participation ratio of the singular values of $\mathbf{J}_{\text{block}} \mathbf{J}_{\text{block}}^T$). We found that the Frobenius norm of the Jacobian blocks broadly decreased during anesthesia, indicative of reduced coupling strength between areas (Figure 4.5C). The participation ratios (dimensionality) of the Jacobian blocks showed more mixed results. Within frontal (FEF, vIPFC) and within posterior (PPC, STG) cortices anesthesia decreased coupling dimensionality (Figure 4.5D). All four within-frontal pairs

decreased significantly. For the frontal-to-posterior and posterior-to-frontal couplings, the results were not significant but tended to be higher during anesthesia. This suggests that anesthesia decouples more closely related areas but potentially creates more diffuse, albeit weakened, couplings across more distant areas. Indeed, within-cluster coupling was weakened more by anesthesia than between-cluster coupling, ($p = 9.5 \times 10^{-6}$, paired two-sided Wilcoxon signed-rank test, median difference within-cluster – between-cluster $\Delta[\Delta PR(\sigma^2)] = -0.69$ [95% CI $-0.94, -0.43$]).

4.4.6 Changes in reachability and controllability are driven by multi-factorial anesthesia-induced effects

We assessed how various factors explain changes in reachability and controllability. We correlated the changes in Gramian traces with changes in block-norm and participation-ratio, as well as with changes in target and source-area Lyapunov exponents across conditions (Figure 4.5E). The traces are positively correlated with block-norm changes, indicating that coupling magnitude is a significant contributor to control ease. Interestingly, they are negatively correlated with the participation ratio changes, indicating that more diffuse coupling might actually impede control along the most essential directions. Overall, the correlations were relatively weak, reflecting that directional control depends on a multitude of factors that include the coupling strength and dimensionality, as well as the dynamic stability of the source and target areas, and the alignment of the source projection with the directions of flow in the target area.

4.4.7 Anesthesia disrupts the hierarchy of cortical organization into senders and receivers

Finally, we evaluated whether anesthesia changes area-level patterns of sending and receiving. We first normalized each cross-area gramian measure by the within-area measure of the target area. We then computed the column-wise mean (corresponding to average sending capability) and row-wise mean (corresponding to average receiving capability). This gives us a per-area measure of the average sending capability relative to receiving capability within each condition (Figure 4.5F). We found that across both awake and anesthesia, PPC was consistently more of a receiving area, while STG and FEF were consistently more sending areas. vIPFC was balanced, indicative of its role as a hub. While the roles were preserved across conditions, anesthesia tended to push areas more toward a balanced regime (sending capability equivalent to receiving capability), thus disrupting cortical organization. PPC and FEF, the two areas with the most extreme awake-state roles, both shifted toward balance under anesthesia.

4.5 Discussion

Our results show that anesthesia decreases the ability of areas to control each other, both in terms of driving towards novel states and stabilizing along existing trajectories. We reproduced earlier results, showing that propofol anesthesia induces broad spectrum destabilization of neural dynamics. While the destabilization was most pronounced in PPC, it was present across all areas. Interestingly, despite decreased stability correlating with increased reachability ease in the absence of other changes, we found that reachability broadly decreased in anesthesia. This suggests that at the time horizon of 10 milliseconds considered in the chapter, the decreased stability in target areas could not compensate for the reduction in coupling strength also observed. This again highlights that changes in reachability and controllability were driven by multi-factorial anesthesia-induced effects, including changes in coupling strength and dimensionality, as well as the dynamic stability of the source and target areas.

Paradoxical excitation. Notably, the directional interactions from PPC to vLPFC and from FEF to vLPFC showed surprising increases in reachability under anesthesia. These pathways could potentially be interpreted as a mechanism of the paradoxical excitation observed during propofol anesthesia. Paradoxical excitation is characterized by an increase in abnormal movements as well as sensory hallucinations.^{414–416} Thus, it may be that the increased ease with which PPC and FEF can drive vLPFC under anesthesia may lead to the observed spurious movements and sensory hallucinations prior to the onset of loss of consciousness.

Reduced coupling magnitude during anesthesia. Our finding of reduced coupling magnitude during anesthesia aligns with previous findings of reduced synchrony and network integration during propofol infusion.^{41,55,131,149,154,176,177,234,241,396–404} Furthermore, the finding that coupling dimensionality decreases locally (within frontal and posterior areas) but increases more globally (across frontal and posterior areas) aligns with recent findings of reduced synchrony within hemisphere but increased synchrony across hemispheres.¹⁵⁴

Neuropsychiatric conditions. We also note that the latent JacobianODE pipeline leveraged in this work was agnostic to both the recording modality and the nature of the data (anesthetic infusion), and could be applied in a wide range of settings. Neuropsychiatric conditions, in particular, present a compelling setting in which to explore the application of nonlinear interareal control analyses. Neural control has been shown to be relevant in schizophrenia,^{141,313,319} and epilepsy,³¹⁴ and a lack of goal-directed control has been identified across multiple psychiatric conditions.⁴¹⁷ Additionally, dysregulation of the cortico-basal ganglia-thalamocortical loop has been associated with a wide range of neuropsychiatric conditions that includes addiction, attention-deficit/hyperactivity disorder, Parkinson’s, and many others.^{418–420} Notably, these conditions all involve a lack of control in some form, suggesting this might be a compelling setting in which to look for control theoretic biomarkers and treatment predictors.

Interventional control. Prior work has shown that JacobianODE models can be used to directly control synthetic neural models to produce desired behaviors.³⁹ This raises the possibility of using the latent JacobianODE pipeline for interventional control in both basic science and clinical settings, such as transcranial magnetic stimulation (TMS) and deep brain stimulation (DBS).^{421–425} Given an appropriate mapping from stimulus to brain activity, the control inputs could actually be provided as sensory stimuli, obviating the need for technologically complex manipulation.⁴²⁶

4.5.1 Future directions

We validated that the latent JacobianODE method is capable of capturing relative changes in reachability and controllability from partial observation in the task-trained RNN. However, it should be determined whether the model is capable of recapitulating these relative changes when there are more than two subsystems present.

A challenge of the latent Jacobian estimation pipeline is the need to pick the intrinsic dimensionality of the latent dynamics d_{dyn} . In the present work, we err on the side of caution and pick a dimensionality sufficient to explain 99% of the variance in principal component space in the data. However a nonlinear method of intrinsic dimensionality estimation that yields tighter bounds to the true intrinsic dimensionality of the system would both save compute (as higher dimensions require larger models) and likely yield improved Jacobian estimation. An outstanding possibility not explored in this work is to use the learned geometry of the encoder (the induced Riemannian metric on the latent space) to infer the true number of directions that are relevant for the latent manifold.

Furthermore, the models fit to neural data surpassed the decoder-corrected persistence baseline, but did not adequately reproduce the asymptotic dynamics of the system. To do so, a compelling alternative is to leverage the nonautonomous extension of the JacobianODE framework suggested in the initial presentation of the method.³⁹ Incorporating this change would vastly increase the expressivity of the model and potentially lead to more effective models of neural dynamics.

While we focus here on control, we note that a wide range of tools have been developed and harnessed to study interareal interactions in neural data.^{304,323} They include methods based on reduced rank regression,^{303,305} recurrent neural network models of neural dynamics,^{324–328} Gaussian-process factor analysis,³³⁰ canonical correlation analysis,^{332,333} convergent cross mapping,^{83,191} switching dynamical systems,^{335,427} granger causality,^{337,338} and point process models.³⁴⁰ A complete analysis should consider how these existing metrics vary with propofol anesthesia to provide a more comprehensive picture of changes in interareal interactions during anesthetic-induced unconsciousness.

Chapter 5

General discussion

Consciousness is supported by dynamical processes operating across many interacting cortical areas. This thesis approached the loss of those processes during anesthetic-induced unconsciousness through the lens of dynamical systems theory, focusing on two complementary properties of cortical activity: *dynamic stability*, the robustness of dynamics to perturbation, and *control*, the directional structure through which one area drives another. Each property motivated a distinct methodological contribution and a corresponding empirical finding, summarized below.

5.1 Main contributions

The main contributions of this thesis are as follows:

- **A data-driven framework for estimating dynamic stability of cortical activity.** We developed Delayed Linear Analysis for Stability Estimation (DeLASE; Chapter 1),³⁷ a delay-embedding-based method that infers the dynamic stability of complex systems from observed time series.
- **Cortical destabilization is a shared signature of anesthetic-induced unconsciousness.** Applying DeLASE across three anesthetics with distinct molecular mechanisms (propofol, ketamine, and dexmedetomidine) we found that all three similarly destabilize cortical dynamics (Chapter 2).³⁸
- **A deep learning framework for data-driven nonlinear control analysis.** We introduced JacobianODE (Chapter 3),³⁹ a method that estimates the time-varying Jacobian of nonlinear dynamics directly from observed trajectories using deep learning. Jacobian estimation enables a control-theoretic analysis of interacting subsystems.
- **A reduction in interareal control during propofol anesthesia.** We extended JacobianODE to partial observations via a latent dynamics formulation (Chapter 4).⁴⁰ Applied to four-area cortical recordings during propofol anesthesia, the latent Jacobian-

ODE revealed a reduction in the ease of intracortical control, both in terms of driving towards novel states and stabilizing along existing trajectories.

Together, these contributions develop dynamical systems theory as a quantitative lens for studying consciousness, and provide tools that can be applied broadly to other neural population datasets and questions in neuroscience.

5.2 Limitations and future directions

Mechanisms of shared signatures of unconsciousness. The stability and control findings of this thesis sit at different points along a single candidate causal chain underlying observed phenomena in anesthetic-induced unconsciousness. In Chapter 2, all three anesthetics similarly destabilized cortical dynamics, and we suggested that this destabilization may underlie the increase in low-frequency oscillations common to the three agents. Stronger low-frequency oscillations have in turn been shown to entrain spiking to their phase, concentrating activity into intermittent “up-states” and reducing overall firing rate.^{41,42,131} This local entrainment generates spike activity windows that are disjoint across time and space, inducing a state of cortical fragmentation.⁴¹ The fragmentation may then reduce the magnitude of interareal coupling, producing the broad-cohort reductions in reachability and controllability observed in Chapter 4.

A circuit-level model linking GABA_A agonism to destabilization was provided in Chapter 1, but analogous models for the NMDA-antagonist and α_2 -adrenergic-agonist mechanisms of ketamine and dexmedetomidine have not been constructed. A unified model that simultaneously accounts for destabilization, reduced spiking, low-frequency power, cortical fragmentation, and reduced directional control would be a valuable tool for understanding the convergent endpoint of mechanistically distinct anesthetics. We acknowledge the possibility of a shared upstream mechanism, not yet identified, that produces all of these changes.

Extending control analysis across multiple anesthetics. The latent JacobianODE was applied to propofol only. The cross-anesthetic generalization of destabilization documented in Chapter 2 has not yet been replicated for directional control. A natural extension is to apply the same pipeline to the ketamine and dexmedetomidine cohorts and ask whether the same reductions in reachability and controllability are shared across mechanistically distinct anesthetics or specific to propofol.

Causal intervention. The directional-control results of Chapter 4 are descriptive: they characterize how interareal control reorganizes under anesthesia, but do not perturb the directional modes they identify. The learned per-trajectory Jacobian field is, however, directly usable for active control: closed-loop nonlinear optimal control on the JacobianODE-learned Jacobians has been demonstrated to invert a task-trained recurrent neural network’s response (Chapter 3). The analogous experiment in neural circuits, involving designing optogenetic

or electrical stimulation protocols that reverse loss of consciousness, or maximally engage a target behavior or output, would convert the descriptive finding here into a prescriptive intervention target and provide a causal test of the directional-control framework.

From trajectory-local to global nonlinear control. The control-theoretic analyses in this thesis are local to the reference trajectory: they characterize the directional nonlinear control properties along a particular reference trajectory via the linearized Jacobian (which itself varies along the trajectory), rather than global properties of the full nonlinear system anywhere in state space. Trajectory locality is desirable in biological settings because the relevant computations are themselves context-sensitive, but a complementary global nonlinear treatment (grounded in tools like contraction theory or differential geometric control) would allow stronger statements about the cortical control landscape as a whole.

Clinical translation. The methods developed in this thesis are immediately applicable to several translational settings. Real-time dynamic-stability estimation could provide a continuous, mechanism-agnostic readout of depth of anesthesia and, with appropriate validation, support closed-loop anesthetic titration.^{270,271} The same dynamical-systems lens applies more broadly to neuropsychiatric conditions that may present biomarkers of disrupted stability or impaired control (depression, anxiety, addiction, schizophrenia).^{420,428–430} It also applies to the design of interventions such as transcranial magnetic stimulation and deep brain stimulation.

Brain-body connection. The dynamical analyses in this thesis consider only cortical recordings, but the connection between the brain and body has been argued to be a key component of consciousness, emotion, and cognition.^{431–435} A natural extension of the directional-control framework of Chapter 4 is to treat the autonomic nervous system as an additional subsystem in the block decomposition, jointly analyzing cortical LFPs with autonomic signals such as heart rate variability, respiration, and pupillometry. Anesthesia produces well-documented perturbations of autonomic regulation, and a unified cortical-autonomic dynamical model would let us ask how directional control reorganizes not just within cortex but between brain and body across the loss-of-consciousness transition. Beyond anesthesia, this approach would be particularly valuable for neuropsychiatric conditions such as anxiety, depression, and post-traumatic stress disorder, where emotion regulation depends on interoceptive integration and autonomic dysregulation is a robust biomarker.^{383,384}

5.3 Closing remarks

At the start of my PhD, I discussed with a friend my plans to study consciousness using math. He replied, lightheartedly, that to translate something as intimate as human experience into the language of math was a travesty. My rebuttal was to suggest that having a more formal understanding of our experience makes it more beautiful, not less. It formalizes our humanity, and therefore illuminates that we all share it. I do not believe I have come even close to

mathematically formalizing experience in this thesis. Nevertheless, there are times when I reflect on my own thinking processes and internal phenomenology, and have the sense that I recognize in myself some reflection of the dynamic processes that I have been studying. Moments of anxious rumination, for instance, when I notice I am failing to drive my own thoughts off the topic of my fixation. Perhaps the more we turn inward to understand our own minds, the more we recognize that no one is truly that different from ourselves. It has been a great privilege and an honor to undertake this doctoral work, and I am truly grateful for having had the opportunity to spend my time thinking about these topics.

Appendix A

Methods: Propofol destabilizes neural dynamics across cortex

A.1 Experimental Model Details

Multi-electrode neural activity was recorded from two rhesus macaques (*Macaca mulatta*), denoted NHP 1 and NHP 2 (standing for non-human primate). NHP 1 was female, aged 8 years, ~ 6.6 kg at the time of the study and NHP 2 was male, aged 14 years, ~ 13.0 -kg. Results were very similar between the male animal and female animal. Animals were pair-housed on 12 hr day/night cycles and maintained in a temperature-controlled environment (80 °F). The animals were not involved in previous procedures. Monkeys 1 and 2 were surgically implanted with a subcutaneous vascular access port (Model CP-6, Norfolk Access Technologies, Skokie, IL) at the cervicothoracic junction of the neck with the catheter tip reaching the termination of the superior vena cava via the external jugular vein. All procedures followed the guidelines of the Massachusetts Institute of Technology Committee on Animal Care and the National Institutes of Health.

The NHPs were implanted with 8×8 chronic Utah arrays, yielding 64 channels per multi-electrode array ('Utah arrays', MultiPort: 1.0 mm shank length, 400 mm spacing, Blackrock Microsystems, Salt Lake City, UT). The distance between electrodes ensured that we were not sampling spiking from the same neurons more than once. Electrodes were placed in four areas: ventrolateral prefrontal cortex (PFC), frontal eye fields (FEF), posterior parietal cortex (PPC) and auditory cortex (STG).⁴² In the present analysis, we use LFPs from each area. During Utah array recordings, areas FEF and PFC were ground and referenced to a common subdural site. Areas STG and PPC also shared a common subdural ground/reference channel. The LFPs were recorded at 30 kHz and filtered online via a lowpass 250 Hz software filter. The LFPs were then subsequently downsampled to 1 kHz. To ensure synchronization of signals recorded on the multiple data acquisition systems, we simultaneously recorded a synchronization test signal with locally unique temporal structure on one auxiliary analog channel of each system. Throughout the entire recording session, we regularly measured offline the relative timing of this test signal between each of the system's recorded datafiles. Measured timing offsets between datafiles were rectified by appropriately shifting event code

timestamps, and in addition by linearly interpolating analog signals to a common time base.

Note that for the stability analyses, since our method models population dynamics, all usable electrodes in each area were included in the analysis, and no averaging was performed. Each anesthetic infusion is referred to as a “session”. By “session”, we mean a continuous recording, from wakefulness through anesthetic infusion to recovery, in one animal. There are 21 anesthetic sessions across two non-human primate subjects (NHPs): 10 sessions with NHP 1, and 11 sessions with NHP 2. Sessions were always separated by at least 2 days, and were thus treated as independent samples. In each session, the NHP performed a non-demanding delayed saccade task pre-anesthesia. Then, following this task, the NHP experienced a passive airpuff/tone classical conditioning task. Specifically, on one third of trials the NHP was delivered a 500 ms ringing tone, which after a 500 ms blank delay, is followed by an airpuff (lasting about 10 ms) to the eye/face designed to elicit a blink response (notably, the airpuff also emitted a sound). In another third of trials there was an isolated airpuff with no tone, and on the last third of trials there was a distinct tone played that indicates that no airpuff will follow. After around 15-20 minutes of airpuff/tone classical conditioning the NHP was infused with propofol anesthesia for 60 minutes. 30 minutes at a higher loading dose (0.58 mg/kg/min for NHP 1 and 0.285 mg/kg/min for NHP 2) and 30 minutes at a lower maintenance dose (0.32 mg/kg/min for NHP 1 and approx. 0.075 mg/kg/min for NHP 2). The infusion was then stopped and the NHP gradually returned to a normal awake state. propofol was intravenously infused via a computer-controlled syringe pump (PHD ULTRA 4400, Harvard Apparatus, Holliston, MA).

All data analysis and simulation was done in Python using original code in addition to, Hydra, Matplotlib⁴³⁶, Pandas⁴³⁷, PyTorch³⁶¹, scikit-learn⁴³⁸, scipy⁴³⁹, and spynal.

A.2 Method Details

A.2.1 LFP Preprocessing

Recordings for each session were quite stable. Utah arrays have been shown to be stable on the order of days.⁴⁴⁰ But to further ensure instability across sessions did not affect results, we first removed the mean across the entire session from each LFP. Then, to handle noise sources, we removed line noise via temporally windowed sinusoid fits at 60 Hz and all harmonics, as well as empirically found line noise frequencies: 107.35, 214.7, 190.2, 196.8, 393.6. We then low pass filtered each LFP at 300 Hz (3rd order bidirectional Butterworth). It was not necessary to consider eye movement and muscle artifacts in the preprocessing as these are considered to be signals that are not cerebral in origin^{441,442}, and are more commonly considered in electroencephalography (EEG) and magnetoencephalography (MEG) studies due to their non-invasive nature.^{443,444} The invasive intracortical electrophysiology signals used in this study are fortunately not susceptible to the same signal contamination risks

as EEG and MEG. Furthermore, electrodes were not averaged at any point in the stability analysis pipeline. This is because our method models population dynamics and takes into account how the overall population representation transforms over time. That information is lost when electrodes are averaged.

A.2.2 Dynamics and stability estimation approach

Here we outline the DeLASE (Delayed Linear Analysis for Stability Estimation) approach to stability estimation.

Eigen-time-delay coordinate dynamics (HAVOK)

We consider observed data consisting of T observations of N dimensions (i.e. a matrix in $\mathbb{R}^{T \times N}$). Given a sampling interval of Δt , this corresponds to a window length of $T\Delta t$. Following HAVOK (Hankel Alternative View of Koopman), we construct a delay embedding, i.e. a matrix of the form

$$\mathbf{H} = \begin{bmatrix} \mathbf{x}_p & \mathbf{x}_{p+1} & \cdots & \mathbf{x}_T \\ \mathbf{x}_{p-1} & \mathbf{x}_p & \cdots & \mathbf{x}_{T-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_{T-p+1} \end{bmatrix},$$

where $\mathbf{x}_t \in \mathbb{R}^N$ are the state observations at time t (e.g. channel activities), and p is the number of lags in the delay embedding matrix. Thus $\mathbf{H} \in \mathbb{R}^{Np \times (T-p+1)}$. We can now perform SVD on $\mathbf{H}$ to obtain $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, where $\mathbf{U}$ are the delay embedding's spatial modes, $\mathbf{\Sigma}$ are the singular values, and $\mathbf{V}$ are the eigen-time-delay coordinates (temporal modes). We select the r coordinates with largest singular values to obtain a reduced rank matrix $\mathbf{V}_r \in \mathbb{R}^{T \times r}$ of temporal modes. The benefits of this SVD step are twofold. First, it avoids a somewhat ill-posed regression, as most of the terms in each column of the delay embedding matrix $\mathbf{H}$ are identical to the previous column (but shifted). Second, it also removes any extraneous information present in the delay embedding due to correlation dimensions and noise. The computation of eigen-time-delay coordinates is equivalent to performing PCA whitening on the delay embedding matrix.¹²⁰ PCA whitening projects a data matrix into a space in which coordinates are decorrelated and have equal variance. This is explored later in the STAR Methods in the section "Responses in eigen-time-delay coordinates". We then compute the dynamics matrix $\mathbf{A}_\mathbf{V} \in \mathbb{R}^{r \times r}$ as the matrix that solves the following least squares regression problem:

$$\mathbf{A}_\mathbf{V} = \arg \min_{\mathbf{A}} \left\| \mathbf{V}_r^{+T} - \mathbf{A} \mathbf{V}_r^{-T} \right\|_2,$$

where $\mathbf{V}_r^- \in \mathbb{R}^{(T-p) \times r}$ is the matrix consisting of the first r eigen-time-delay coordinates at times 1 through $T - p$ and $\mathbf{V}_r^+ \in \mathbb{R}^{(T-p) \times r}$ is the matrix consisting of the first r eigen-time-delay coordinates at times 2 through $T - p + 1$. We can therefore construct a discrete model of the delay-embedded neural dynamical system as

$$\mathbf{h}_{t+1} = \mathbf{U}_r \Sigma_r \mathbf{A}_\mathbf{V} \Sigma_r^\dagger \mathbf{U}_r^T \mathbf{h}_t,$$

where $\mathbf{h}_t$ is the column of the delay embedding matrix $\mathbf{H}$ corresponding to maximal time t (that is, including times $t - p + 1$ through t), $\mathbf{U}_r \in \mathbb{R}^{Np \times r}$ is the matrix corresponding to the first r columns of $\mathbf{U}$, Σ_r is the matrix corresponding to the first r singular values, and $\Sigma^\dagger$ is the pseudoinverse of Σ .

Estimating stability from delay dynamics

We now let $\mathbf{A} = \mathbf{U}_r \Sigma_r \mathbf{A}_\mathbf{V} \Sigma_r^\dagger \mathbf{U}_r^T$ so $\mathbf{A}$ is the linear representation of the dynamics in the delay-embedded neural space. We now note that the first N rows of the matrix $\mathbf{A}$ describe the dependence of $\mathbf{x}_t$ on the trajectory history:

$$\mathbf{x}_t = \sum_{k=1}^p \mathbf{A}_k \mathbf{x}_{t-1},$$

where $\mathbf{A}_k$ is the $N \times N$ matrix corresponding to columns kN through $(k+1)N$ of the first N rows of $\mathbf{A}$. We have thus arrived at a representation of the neural dynamics in the form of a (discrete) linear delay differential equation. To extract the stability from this representation of the dynamics, we turned to the field of delay differential equations. We first convert to continuous time (for purposes of interpretation) setting $\tilde{\mathbf{A}}_k = \frac{\mathbf{A}_k}{\Delta t}$ for $k > 1$ and $\tilde{\mathbf{A}}_k = \frac{\mathbf{A}_k - \mathbf{I}_N}{\Delta t}$ for $k = 1$ (this latter term is an approximation of the instantaneous term, and so it includes an identity matrix). This yields the delay differential equation $\dot{\mathbf{x}}(t) = \sum_{k=1}^p \tilde{\mathbf{A}}_k \mathbf{x}(t - k\Delta t)$. The stability of this delay differential equation is determined by the roots λ of its corresponding characteristic equation¹²³:

$$\det \left(\lambda \mathbf{I}_N - \sum_{k=1}^p \tilde{\mathbf{A}}_k e^{-\lambda k \Delta t} \right) = 0.$$

The intuition for why this equation is helpful in capturing the stability present in the system is because of the $e^{-\lambda_k \Delta t}$ term. This term discounts the impact of the matrix $\tilde{\mathbf{A}}_k$ based on how far it is from the current state (Figure 1.2E).

This characteristic equation has infinitely-many solutions.⁴⁴⁵ In fact, even for a simple delay differential equation with a single delay, the characteristic equation has infinite solutions. Intuitively, this is because in a continuous time setting there are infinite points in between the delay time and the current time - thus the system is infinite dimensional and must be specified with an infinite dimensional initial condition. Mathematically, the infinitely-many solutions are a result of a corollary to The Great Picard Theorem, namely that an entire non-polynomial function assumes every complex number infinitely many times (with one possible exception).⁴⁴⁶ Analytic investigation of the stability of even single delay equations can necessitate quite advanced mathematical machinery.⁴⁴⁷ There exist numerous methodologies to numerically approximate a finite portion of the roots of a given delay differential equation.^{123,124,448,449} These approaches typically discretize the delay period and are guaranteed to converge to the true characteristic roots as the discretization of the delay period becomes finer. For our analysis, we use the TRACE-DDE algorithm, which estimates the roots of the characteristic equation by constructing a discrete approximation of the infinitesimal generator.¹²⁴ We choose the number of discretization points in the delay period to be equal to the number of lags chosen for the delay embedding, with discretization points spaced out by Δt . This generates $N(p + 1)$ characteristic roots, of which for the present analyses we evaluate the top 10% (i.e. the 10% with greatest real part).

The characteristic roots are complex valued numbers. The real part of the root determines the rate at which perturbations to the system along a particular direction will decay. The complex part of the root determines the frequency at which such perturbations will decay. In a strictly linear delay differential equation, the root with largest real part determines the stability, as it determines the overall slowest rate at which perturbations will decay (if the real part is negative) or the fastest rate at which they will explode (if the real part is positive). For our analysis, since we are approximating complex nonlinear systems with linear delay differential equations, we look at the upper portion of the distribution of characteristic roots extracted to characterize stability. Further, while the convention is to report the real parts of characteristic roots as inverse timescales, in our analysis we represent these values both with inverse timescales and standard timescales where appropriate.

To handle instabilities introduced into the dynamics by sensory stimuli, inputs that are not accounted for, noise, and other artifacts, we filter the extracted characteristic roots in two ways. First, we enable filtering out all characteristic roots with a frequency component above a certain value. For our analysis, we filtered out characteristic roots with a frequency component greater than 500 Hz, since this is the maximum possible frequency in data sampled at 1 kHz. Next, we enable filtering out all unstable roots with a frequency component above a certain value. For our analysis, we filtered out unstable roots with a frequency component greater than 125 Hz. The intuition for this is that rapid, sharp changes due to sensory inputs and other artifacts will destabilize the dynamics with a high frequency component, due to the sharpness of the change. We also note that when visualizing the characteristic roots as

proper timescales in Figure 1.4B, we considered only the characteristic roots with negative real part. This was because the timescales associated with roots with negative real part are timescales of stable decay – in contrast to timescales associated with roots with positive real part, which are timescales of unstable explosion. Nearly all computed characteristic roots had a negative real part (see Figure 1.4C,E), and all statistics were performed on inverse timescales, which included any roots with positive real part.

A.2.3 Picking hyperparameters

To select (1) the minimum number of delay embedding coordinates and (2) the rank of the eigen-time-delay coordinates used to estimate the dynamics, a grid search was performed for each session (sample is shown in Figure A.1A). We use fixed window sizes for our analysis (10 seconds for simulated data and 15 seconds for neural data). These window sizes ensure a long enough time window to capture a wide range of dynamic motifs, but a short enough time window to preserve computational tractability. For the grid search on neural data, models were fit on 12 windows from each session, and parameters were chosen on a per-session and per-area basis. Windows were chosen to ensure that all sections of the session were included (i.e. awake, anesthetic induction, anesthetic unconsciousness, and recovery), with 4 windows being randomly selected from each section in each session. The metric used to assess model quality was Akaike Information Criterion (AIC). AIC quantifies the balance between high quality model prediction (as measured by one-step prediction error on test data) with the number of model parameters. Models are penalized for having larger numbers of parameters, such that the optimal model of minimal complexity can be obtained. While larger models tend to perform better, after a certain point the added parameters are not benefitting the prediction quality sufficiently to justify their inclusion (Figure A.1A, lower right corner of the grid). Because we compare stability between sections of each session, the hyperparameters minimizing AIC within a given session (and area) are chosen. The time interval between delays was held fixed at 1 time step (1 ms) for this analysis. Chosen hyperparameters varied across sessions but were typically around 750 delay coordinates (12 delays) with a rank of 750 for individual areas (usually ~ 64 channels per area), and approximately 1000 delay coordinates (4 delays) with a rank of 900 for all areas considered together (usually ~ 250 channels). All tested hyperparameter combinations preserve the core results of our analysis (see Figure A.1B).

A.2.4 Implementation of VAR

We implemented VAR(1) (1st-order vector autoregression) in the following way. We again consider observed data consisting of T observations from N dimensions (i.e. a matrix $\mathbf{X} \in \mathbb{R}^{T \times N}$). Given a sampling interval of Δt , this corresponds to a window length of $T\Delta t$. We then find the matrix $\mathbf{A}_{VAR} \in \mathbb{R}^{N \times N}$ that solves the following least squares regression problem:

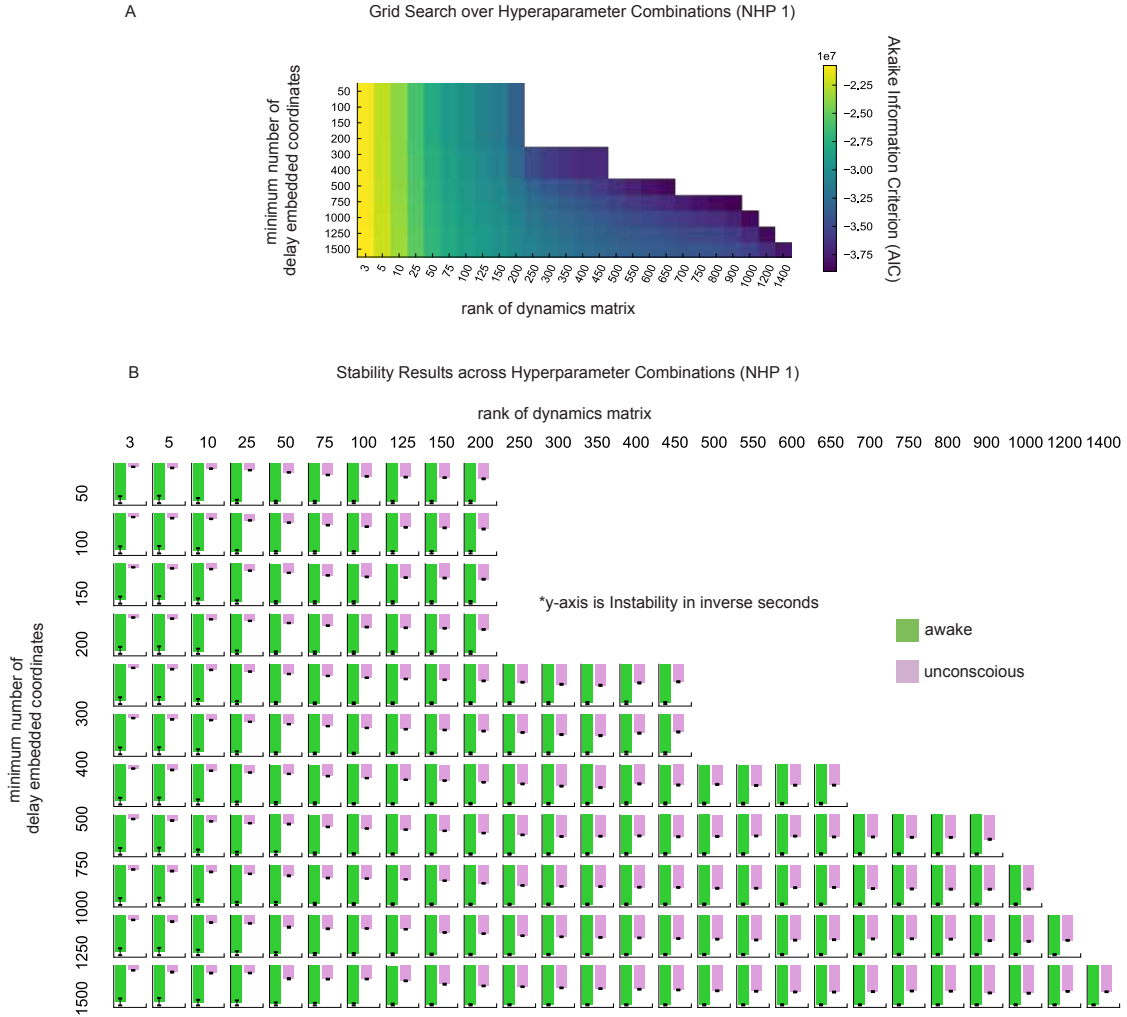


Figure A.1: Core results hold across hyperparameter combinations. (A) The mean AIC on test-set one-step prediction on 20 randomly chosen windows of awake data and anesthetic data (10 from each) from a single session. The y-axis is the minimum number of delay embedding coordinates, effectively determining the number of lags. The x-axis is the rank of the dynamics, or number of independent temporal modes, extracted through SVD on the delay embedding matrix that are used for fitting the dynamical model. Color indicates Akaike Information Criterion (AIC) normalized to the number of predicted values. AIC quantifies the balance between high quality model prediction (as measured by one-step prediction error on test data) with the number of model parameters. Models are penalized for having larger numbers of parameters, such that the optimal model of minimal complexity can be obtained. (B) The mean of the real part of the top 10% of characteristic roots (or the number of characteristic roots equivalent to the rank, whichever is smaller) from the 10 windows from each section of the session. The core results hold across all hyperparameter combinations. Data are reported as mean $\pm$ SEM, with results averaged over all considered characteristic roots from each section of the session.

$$\mathbf{A}_{VAR} = \arg \min_{\mathbf{A}} \left\| \mathbf{X}^{+T} - \mathbf{A} \mathbf{X}^{-T} \right\|_2,$$

where $\mathbf{X}^-$ is the $(T - 1) \times N$ matrix consisting of the N -dimensional state observations at times 1 through $T - 1$ and $\mathbf{X}^+$ is the $(T - 1) \times N$ matrix consisting of the N -dimensional state observations at times 2 through T .

To estimate stability, we compute the eigenvalues of $\mathbf{A}_{VAR}$. To convert to a continuous time representation, for each eigenvalue λ_i we convert to the continuous time instability measure $\hat{\lambda}_i = \frac{\log(|\lambda_i|)}{\Delta t}$ as in previous work.⁶⁴ To estimate stability, we considered the largest 10% of such instability measures extracted from VAR, mirroring our approach for DeLASE.

A.2.5 Estimating stability in simulated systems

To validate the DeLASE method, and to generate examples of dynamical systems to compare to neural data, we simulated sample systems. The systems included simple linear dynamics as well as randomly connected RNNs. Since we were interested in systems that are stable, we often simulated these systems in the stable regime. In this regime, however, these systems converge to fixed points. Thus, to avoid this trivial behavior, we simulated these systems with stochasticity injected into their dynamics. In this approach, the systems were simulated with a small amount of process noise, effectively perturbing the system a small amount at each time step. We simulated stochastic dynamical systems using the Euler-Maruyama method. Specifically, we simulated the systems using the update:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \mathbf{f}(\mathbf{x}_t) dt + \sigma d\mathbf{W}_t,$$

where $\mathbf{x}_t$ is the n -dimensional state of the system at time t , $\mathbf{f}$ is the system dynamics (i.e., $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$), dt is the time step of simulation, σ is a small scale parameter on the noise, and $d\mathbf{W}_t$ is the n -dimensional process noise at time t , with each dimension independently sampled from a normal distribution with mean 0 and standard deviation $\sqrt{dt}$.

A.2.6 Linear dynamics

We simulated n -dimensional linear systems of the base form $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) = \mathbf{A}\mathbf{x}$ where $\mathbf{A}$ is the $n \times n$ dynamics matrix. To set the maximal real part of the eigenvalues of this matrix, we first sample each matrix element from a normal distribution with mean 0 and standard deviation $\frac{1}{\sqrt{n}}$. Then we perform the update

$$\mathbf{A} = \mathbf{A} - \lambda_{\max} \mathbf{I}_n + \lambda_{\text{new}} \mathbf{I}_n,$$

where $\lambda_{\max}$ is the original maximum real part of the matrix eigenvalues and λ_{new} is what we would like to update it to. We simulated the systems in the stochastic framework with $\sigma = 1$. For the simulations used to generate Figure 1.3A, we simulated stochastic linear dynamics with maximal eigenvalues from -1 to -0.1 (inclusive) stepping by 0.1. We simulated each system 20 times. Each system had 100 dimensions and was simulated for 20,000 time steps. The time step used was 0.002ms. A transient of 2000 time steps was dropped from the simulated systems, and models were fit to a randomly chosen partial observation of 10 dimensions of the system over 10,000 time steps. A hyperparameter grid search was executed to choose the number of delays and the rank of the eigen-time-delay coordinates. Specifically, minimum delay embedding matrix sizes of 10, 20, 50, 100, 200, 300, 500, 750 and 1000 were tested, along with ranks of 3, 5, 10, 25, 50, 75, 100, 125, 150, 200, all ranks from 200 to 800 stepping by 50, and also ranks of 900 and 1000. For each of the 20 runs, the best performing parameter combinations across all linear systems were chosen (typically a delay embedding matrix size of 10 with a rank of 10) and the stability was estimated based on these models.

A.2.7 Randomly connected RNN dynamics

We simulated randomly connected RNNs according to Sompolinsky et al.²⁷. Specifically, we simulated the dynamics:

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) = \frac{1}{\tau} (-\mathbf{x} + g \mathbf{W} \tanh \mathbf{x}),$$

where $\mathbf{x}$ is the n -dimensional state of the system (effectively synaptic activity in the RNN), τ is the time constant of the dynamics, typically set to 100 ms (with a time step of simulation of 10 ms), $\mathbf{W}$ is the $n \times n$ weight matrix with each element drawn from a normal distribution of mean 0 standard deviation $\frac{1}{\sqrt{n}}$, and g is a gain parameter on the synaptic weights. It has been shown that as n increases, and as the gain is increased above 1, the network enters into increasingly chaotic regimes.²⁷

For the simulations presented in Figure 1.3B, we used a network size of 1024 neurons, and chose values for the gain parameter from 0.8 to 1.4 (inclusive), specifically: 0.8, 0.85, 0.9, 0.925, 0.95, 0.975, 1, 1.025, 1.05, 1.075, 1.1, 1.125, 1.15, 1.175, 1.2, 1.25, 1.3, 1.35, and 1.4. We used $\sigma = 0.05$. We sampled the matrix $\mathbf{W}$ 10 times, then for each sampled matrix simulated the dynamics with all possible values of the gain. The systems were simulated for 20,000 time steps. For stability analysis, a transient of 2000 time steps was dropped, and the models were fit to 10,000 time steps of data. We randomly observed 10 dimensions of the 1024-dimensional system to use for estimating stability. A hyperparameter grid search was

executed to choose the number of delays and the rank of the dynamics matrix. Specifically, minimum delay embedding matrix sizes of 10, 20, 30, 40, 50, 100, 150, 200, 250, 300, 350, and 400 were tested with ranks of 2, 3, 5, 10, 30, 40, 50, 75, 100, 125, 150, 175, 200, 225, 250, 275, 300, 325, 350, 375, and 400. For each of the 10 runs, the best performing parameter combinations across all trajectories were chosen (typically a delay embedding matrix size of 50 with a rank of 40, though there was some variation) and the stability was estimated based on these models.

To compute the Lyapunov exponents in the simulated networks, we applied a QR-based algorithm¹²⁷ to the Jacobian of the discrete dynamics, computed as

$$\mathbf{J} = \mathbf{I}_n + \frac{dt}{\tau} \left(-\mathbf{I}_n + g\mathbf{W}\text{diag} \left(\mathbf{1} - \tanh^2 \mathbf{x} \right) \right).$$

The exponents extracted from the QR algorithm are then converted into continuous time by dividing by the time step of simulation, dt .

A.2.8 Linear dynamics with inputs

One-dimensional linear sine wave filters

For Figure 1.5, we simulated a simple one-dimensional linear system with input according to the dynamics $\dot{x} = -\lambda x + u(t)$ where λ controls the instability of the filter. For the simple filter on a sine wave (Figure 1.5B), we found the analytical solution for the (deterministic) differential equation with $u(t) = \sin 2\pi\omega t$. The solution is $x(t) = \frac{1}{\lambda^2 + 1} (\lambda \sin 2\pi\omega t - \cos 2\pi\omega t)$. We computed this solution for values of λ from 0 to 4.5 (inclusive), stepping by 0.5, and $\omega = 1$ (Figure 1.5B). To simulate the dynamics in the presence of noise (Figure 1.5C), we simulated simple one-dimensional stochastic dynamics:

$$x_{t+1} = x_t + (-\lambda x_t + u(t)) dt + \sigma dW_t,$$

with λ set to 10 and 0.01, and σ set to 0.05. We used a time step of 1 ms to match the sampling rate of the neural data, and a simulation time of 3 seconds. 1 second into the simulation, we perturbed the above equation with a negative 2 Hz sine wave for 500 ms ($u(t) = \sin 4\pi t$). This matches the duration of the tone presentations in the experimental setup. We then provided no inputs for 500 ms before providing another negative 2 Hz sine wave input for 500 ms (this matches the spacing of tones and puffs in the experimental setup).

Two-dimensional linear system input responses

In the two-dimensional case, we simulated the two-dimensional stochastic dynamics given by:

$$\mathbf{x}_{t+1} = \mathbf{x}_t + (\mathbf{A}\mathbf{x}_t + u(t)\mathbf{I}_2) dt + \sigma dW_t.$$

To generate the dynamics $\mathbf{A}$, we took two parameters, a and b , as well as an orthogonal matrix $\mathbf{Q}$, and set

$$\mathbf{A} = \mathbf{Q} \begin{bmatrix} a & -b \\ b & a \end{bmatrix} \mathbf{Q}^{-1}.$$

Thus $\mathbf{A}$ has eigenvalues $a \pm bi$ and a controls the stability of the dynamics. For Figure 1.6B, we simulated the dynamics with both $a = -6$ and $a = -1$. We set $b = 2\pi$, and picked $\mathbf{Q}$ randomly. We used a time step of 0.25 ms and a simulation time of 7 seconds and set $\sigma = 0.005$. The first input was a constant function $u(t) = 0.6$ lasting 500 ms and starting at 2 seconds. The second input was a constant function $u(t) = 1$ lasting 150 ms and starting at 3 seconds. The structure of the inputs was chosen to mirror the tones and puffs in the experimental setup.

A.2.9 Neural trajectories of stimulus responses

To compute the neural trajectories depicted in Figure 1.5D,E and Figure 1.6C-H, we first compiled population responses from every session in each NHP for the trial type of interest (a 500 ms tone, followed by a 500 ms delay and then an air puff). To ensure enough resolution around the stimuli, we collected LFPs from 2 seconds before the first stimulus onset until 5 seconds after. For each NHP, this yielded a $K \times T \times N$ matrix, where K is the number of trials (aggregated across sessions), T is the number of time-points from the trial (as mentioned, we consider 7 seconds surrounding the stimulus onset), and N is the number of electrodes.

Average LFP responses

To compute the one-dimensional average LFP trial response (Figure 1.5D,E), for each NHP, we averaged the trial responses over trials, and then subsequently over electrodes.

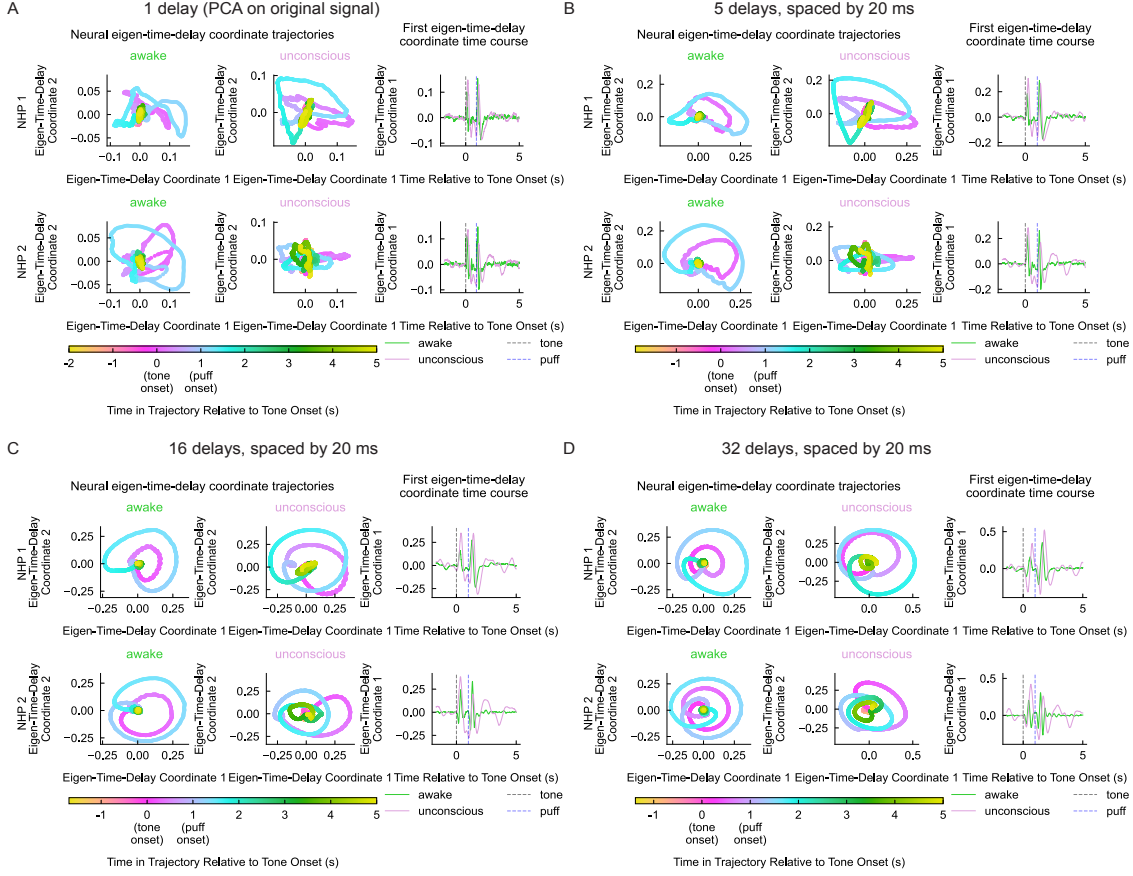


Figure A.2: Dimensionality reduction across delay embedding hyperparameters. Related to Figure 1.6. The core qualitative pattern is preserved across hyperparameters, though eigen-time-delay coordinates more clearly illuminate the state-space dynamics. (A) The original signal. The eigen-time-delay coordinates in this case reduce to PCA. (B) 5 delays, each spaced out by 20 ms. (C) 16 delays, each spaced out by 20 ms. (D) 32 delays, each spaced out by 20 ms.

Responses in eigen-time-delay coordinates

To compute the (scaled) eigen-time-delay coordinates of the average stimulus response (Figure 1.6C-H, Figure A.2), we first delay embedded each trial individually. Specifically, for a specific trial response, given that the neural population state at time t is $\mathbf{x}_t$, and given the delay embedding parameters (the delay interval τ and delays), we computed the $Np \times (T - (p - 1)\tau)$ matrix given by

$$\mathbf{H} = \begin{bmatrix} \mathbf{x}_{1+(p-1)\tau} & \mathbf{x}_{2+(p-1)\tau} & \cdots & \mathbf{x}_T \\ \mathbf{x}_{1+(p-2)\tau} & \mathbf{x}_{2+(p-1)\tau} & \cdots & \mathbf{x}_{T-\tau} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_{T-(p-1)\tau} \end{bmatrix}.$$

After performing a delay embedding for every trial, for each NHP we end up with a $K \times Np \times (T - (p - 1)\tau)$ matrix. We then average across trials to arrive at a $Np \times (T - (p - 1)\tau)$ matrix of the average delay embedding. Finally, we perform PCA on (the transpose of) this matrix to reduce the number of dimensions from Np to 2. Since each dimension of the delay embedding matrix contains observations from multiple time-points, we consider the “effective” time of the delay embedding coordinate to be the latest time-point in the dimension. That is, the effective time of the first column of $\mathbf{H}$ above is $1 + (p - 1)\tau$. These are the times used for plotting in Figure 1.6C-H.

Furthermore, we note again here that this process equates to computing scaled versions of the eigen-time-delay coordinates used by the HAVOK models. To see why this is the case, consider a delay-embedded matrix $\mathbf{X} \in \mathbb{R}^{N' \times T'}$ where N' is the number of coordinates in the delay embedding and T' is the number of delay embedded time-points. Thus, given the singular value decomposition $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, the first r eigen-time-delay coordinates used by HAVOK are the first r columns of $\mathbf{V}$. Note that by rearranging this equation we can show that the first r eigen-time-delay coordinates can be computed as $\mathbf{\Sigma}_r^{-1}\mathbf{U}_r^T\mathbf{X} = \mathbf{V}_r^T$, and equivalently $\mathbf{V}_r = \mathbf{X}^T\mathbf{U}_r\mathbf{\Sigma}_r^{-1}$ where the r in the subscript denotes the rank- r column truncation. Note here that $\mathbf{X}^T\mathbf{U}_r$ is the expression that computes the projection of the data onto the top r principal components in PCA (assuming the N' rows of $\mathbf{X}$ are mean-centered). Thus, in the eigen-time-delay coordinates, the diagonal matrix $\mathbf{\Sigma}_r^{-1}$ alters the PCA only in that it scales each component by the inverse of its corresponding singular value. This is known as PCA whitening and produces a scaled version of the data projected onto the principal components such that each component has unit variance.

A.2.10 Increasing inhibition in RNNs

To examine the effects of increasing inhibition (Figure 1.7), we again simulated randomly connected RNNs according to Sompolinsky et al.²⁷ following the dynamics

$$\tau \dot{\mathbf{x}} = -\mathbf{x} + g\mathbf{W} \tanh \mathbf{x},$$

where $\mathbf{x}$ is the n -dimensional state of the system (effectively synaptic activity in the RNN), τ is the time constant of the dynamics (10 ms, with a time step of simulation of 1 ms), $\mathbf{W}$ is the $n \times n$ weight matrix with each element drawn from a normal distribution of mean 0 standard deviation $\frac{1}{\sqrt{n}}$, and g is a gain parameter on the synaptic weights. We used a network size of 512 neurons. We simulated the dynamics with gains 0.8, 0.9, 1, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7. To scale up and down the inhibition in the network, we took all elements of $\mathbf{W}$ with negative weight, and scaled them by an inhibitory scaling factor, κ_I . For each gain, we simulated the network with $\kappa_I = 0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.1, 1.2, 1.3, 1.4$, and 1.5. We sampled the matrix $\mathbf{W}$ 10 times, then for each sampled matrix simulated the dynamics with all possible pairings of g and κ_I . The systems were simulated for 10,000 time steps. We computed the

Lyapunov exponents of the networks as described above in the section “Randomly connected RNN dynamics”.

A.2.11 Example dynamical systems

HAVOK models of the Van der Pol Oscillator

We simulated the Van der Pol Oscillator¹²² with a time step of 20 ms and a simulation time of 400 seconds (generating 20,000 time steps) using the governing equations

$$\dot{x} = y$$

$$\dot{y} = \mu (1 - x^2) y - x.$$

We set the parameter μ to 2. We chose an initial condition from a standard Gaussian distribution then dropped the first 5,000 time steps (100 seconds) as a transient. We then observed only the second (y) dimension of the dynamics. We fit a HAVOK model on 10,000 time steps (200 seconds) using 500 delays. Predictions were generated on the remaining 5,000 time steps by autonomously running the dynamics - that is, the initial state is fed through the model to get the first prediction, then that prediction is fed back through the model, and so forth. The results of this simulation are presented in Figure 1.2D.

Mass-spring simulation

For the cartoon shown in Figure 1.5A, a mass-spring system was simulated with governing dynamics $m\ddot{x} + R\dot{x} + kx = I$ where x is the vertical position of the mass, m, k, R are parameters governing the dynamics and I is the input. For our simulations, we set $m = 10, k = 1, R = 10$. We simulated using scipy’s `solve_ivp` function for 1,000 time steps with a time step of 10 ms.⁴⁴⁵ From time step 100 to 120, we perturbed the system with a constant input of value $I = 10$. The results of this simulation are in Figure 1.5A (right).

A.2.12 Pharmacokinetics analysis

Estimating integrated propofol dosage

To estimate the integrated (i.e., accumulated) propofol dosage up to a particular moment in the session, we began by estimating the context-sensitive half time (CSHT) of propofol. CSHTs describe the time it takes for the concentration of anesthetic in the system to halve, as a function of the duration of infusion (i.e., the “context”). Detailed modeling has estimated that the CSHT of propofol is around 3 minutes for a very short infusion and around 8 minutes after 60 minutes of infusion.⁴⁵⁰ We then linearly interpolated between these two values for infusion times in between 0 minutes and 60 minutes, and assumed the CSHT stays constant at 8 minutes after the 60 minute infusion ends. Given the CSHT, we can estimate the instantaneous time constant of decay of propofol by assuming that the quantity present in the system undergoes exponential decay in the absence of any input. We thus obtained, using the definition of the CSHT, the expression

$$0.5x_0 = e^{-\frac{CSHT(t)}{\tau}}x_0,$$

where x_0 is the initial quantity present in the system, τ is the instantaneous time constant of decay in minutes, and $CSHT(t)$ is the CSHT after t minutes of infusion. We can solve this equation to obtain that $\tau = \tau(t) = -\frac{CSHT(t)}{\log(0.5)}$ minutes. Now, we integrate, over the course of a full session, a first-order differential equation for the quantity of propofol present in the system:

$$\dot{x} = \alpha - \frac{1}{\tau(t)}x,$$

where x is the propofol quantity in mg/kg, $\dot{x}$ is the rate of change of propofol quantity in mg/kg/min, and α is the infusion rate in mg/kg/min (see Experimental Model Details). We solve this equation using scipy’s `solve_ivp` function. Given the complexity of propofol pharmacokinetics,⁴⁵¹ we stress that these estimates were intended only to be approximations for the purposes of investigating the relationship between propofol dosage and instability.

Predicting instability from propofol dose

Using the estimated integrated propofol dosage, we attempted to predict the estimated instability values. For each window in each session, we computed both the estimated integrated propofol dose up to that time, as well as the mean instability in that window.

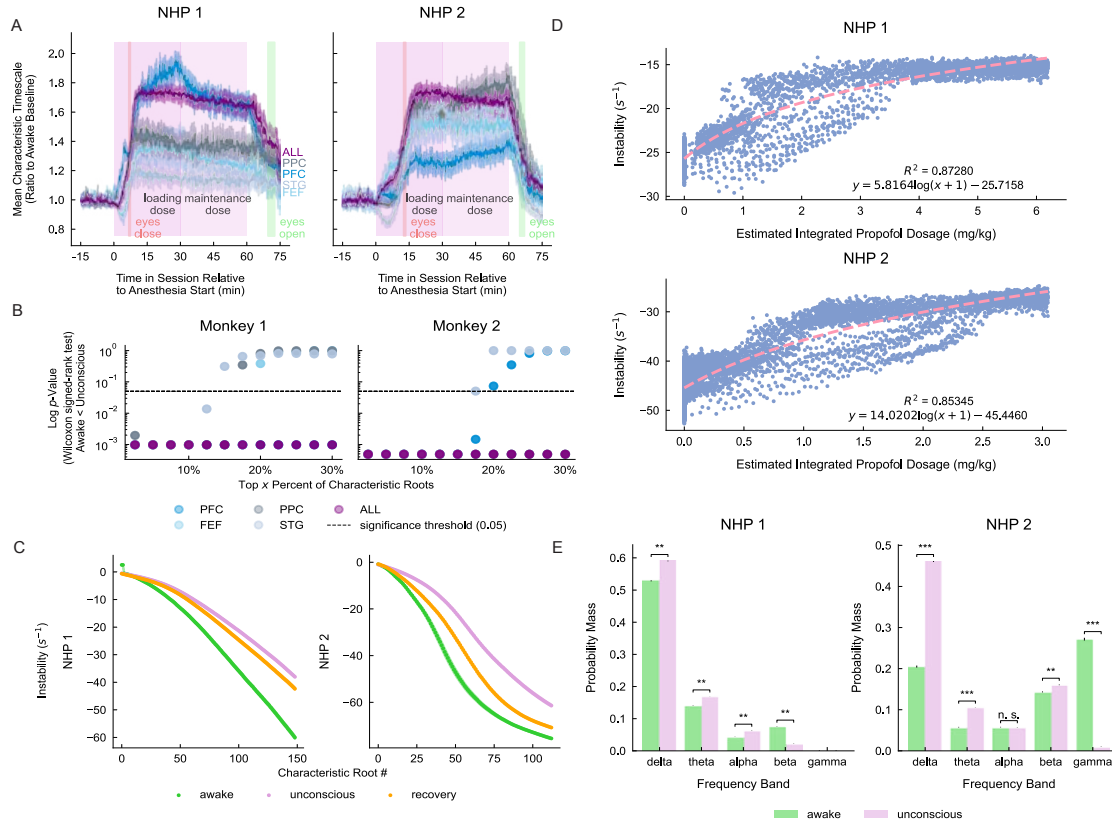


Figure A.3: Extended results demonstrating propofol destabilizes neural dynamics. Related to Figure 1.4. Data in (A), (C) and (E) are represented as mean $\pm$ SEM, with results averaged over sessions for each NHP. (A) The curves representing mean timescales normalized to the awake baseline for all areas from both NHPs. The same as Figure 1.4B, but showing results from all areas. The awake baseline is computed for each session by taking the geometric mean of the timescales associated with the characteristic roots across all windows in the awake section of the session. Then, for each window, the geometric mean of the ratio of the timescales to the awake baseline is computed. After accounting for the baseline awake stability, the degree of destabilization in the timescales in each NHP is very similar. (B) The p-value from the Wilcoxon signed-rank test testing whether the distribution of characteristic roots from the awake state is less than the distribution of characteristic roots from anesthesia. For each analysis, we consider the maximal x% of characteristic roots, where x is the x-axis value. For all areas, and for almost all portions of the distribution above 20%, the upper portion of the distribution of characteristic roots is more unstable in anesthesia. When considering all areas together, this is true in both NHPs for portions of the distribution up to 30%. (C) The spectrum of characteristic roots. Real parts are sorted then averaged across all sessions for a particular state in the session. In both NHPs, the spectrum of characteristic roots relaxed closer to the wakeful baseline during recovery. (D) Mean instability indices (maximal 10% of characteristic roots) for each window of every session plotted against the estimated integrated propofol dosage at the beginning of the window (see Appendix A.2.12). A linear-log regression was fit to predict the mean instability values from the estimated integrated propofol dose. (E) The probability mass function on which bands the frequency components of the (maximal 10%) of characteristic roots fall into. Anesthesia (purple) is compared to the awake state (green). There is more mass on lower frequency bands in anesthesia, and more mass on higher frequency bands during the awake state.

Then, for each NHP, we fitted a linear-log regression to predict instability from the estimated integrated propofol dose. The results of this analysis are in Figure A.3E.

A.3 Quantification and Statistical Analysis

For all data plots, “*” denotes a p-value less than 0.05, “**” denotes a p-value less than 0.01, and “***” denotes a p-value less than 0.001.

A.3.1 Prediction quality of dynamical models

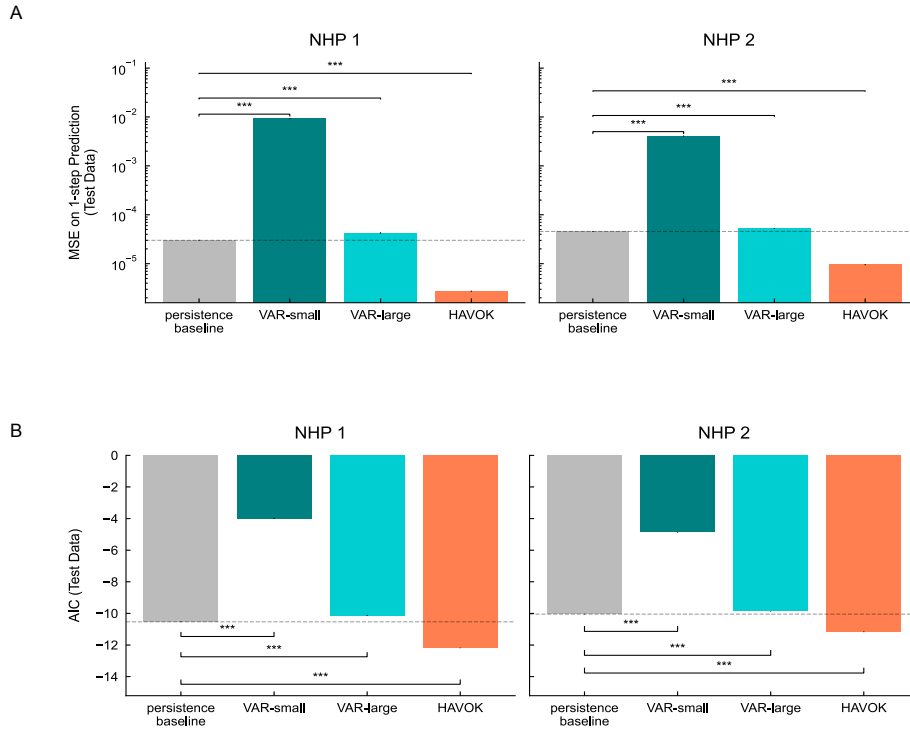


Figure A.4: Neural dynamics prediction, separated by NHP. Related to Figure 1.3. Data are represented as mean $\pm$ SEM, with results averaged over the number of simulations. (A) Comparison of different model’s one-step predictions on neural data - same as Figure 1.3, but split by NHP. The persistence baseline - predicting the next state solely with the previous state - is plotted in the gray bar and with the dotted line. The y-axis is plotted on a log scale. HAVOK models are the only model shown capable of beating the persistence baseline at this essential prediction task. (B) The same as (A), but with AIC, depicting that the superior performance of HAVOK is not due to having more parameters than the other models. The same as Figure 1.3, but split by NHP.

To quantify the prediction quality of the dynamical models considered in the manuscript (Figure 1.3 and section “Delayed linear dynamical systems models capture neural dynamics”), we first fit each model for every window of every session. Then, the performance metrics (MSE and AIC) were averaged across windows within a single session. To compute the statistics, each session was an independent sample. Since, for each session, we were comparing between two models, we used the one-sided Wilcoxon signed-rank test. Since each session was an independent sample, the sample size was $n = 10$. For Figure A.4, we split the analysis by NHP, yielding sample sizes of:

- NHP 1: $n = 10$
- NHP 2: $n = 11$

A.3.2 Changes in stability in neural dynamics

Statistical tests referenced in the section “Propofol anesthesia induces unstable cortical dynamics” were conducted as follows. In each NHP, the real parts (i.e., inverse timescales of response) of the top 10% of characteristic roots from every window in every session were grouped according to the section (i.e., awake, unconscious, recovery, early recovery, late recovery, loading dose, and maintenance dose) of the session. The mean of these inverse timescales was then taken to yield one value per section, per session.

To compute statistics, the sections were defined as follows. Awake windows were collected from the 15 minutes preceding the start of the anesthetic infusion. Induction windows were collected from the start of propofol infusion until 15 minutes following the start of propofol infusion. Unconscious windows were collected from 15 minutes following the start of propofol infusion until the end of propofol infusion (a period of 45 minutes). Recovery windows were collected from the end of propofol infusion until 15 minutes following the end of propofol infusion. Early recovery was collected from the end of propofol infusion until 8 minutes following the end of propofol infusion. Late recovery was assessed from 8 minutes following the end of propofol infusion until 15 minutes following the end of propofol infusion. Loading dose windows were collected from 15 minutes following the start of propofol infusion until 30 minutes following the start of propofol infusion (the final 15 minutes of the loading dose). Maintenance dose windows were collected from 45 minutes following the start of propofol infusion until the end of propofol infusion (the final 15 minutes of the maintenance dose). See Figure A.5 for a visual depiction of the section definitions.

To test whether one section of the session was more unstable than another, one-sided Wilcoxon signed-rank tests were performed on mean instability values (top 10% of the real parts of characteristic roots) and averaged within each section for each session. Thus for each section, for every area, the sample sizes are:

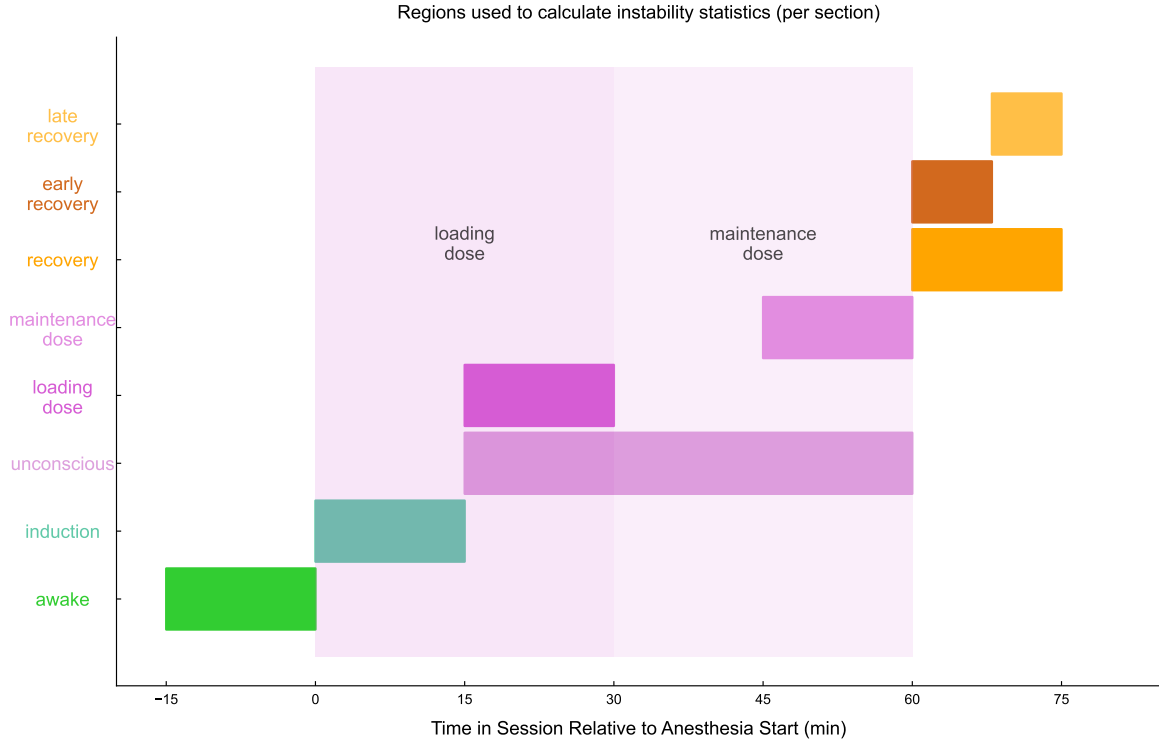


Figure A.5: Visual depiction of section definitions used for statistics computation. Related to Figure 1.4. A visual illustration of the definition of section times, all relative to the start of anesthesia. These section definitions were used to compute the between section statistical differences referenced throughout the paper. In each case, the value in question was aggregated within the section of each session, and statistics were computed using the number of sessions as the number of samples.

- NHP 1: $n = 10$
- NHP 2: $n = 11$

The results of all comparisons between sections are reported in the chapter in the section “Propofol anesthesia destabilizes cortical neural dynamics”. Most of these comparisons are also visually presented in Figure 1.4D.

A.3.3 Distances between stability distributions

To test whether the recovery distribution returned to awake levels of dynamic stability, we first constructed a test value that consisted of the awake-unconscious Cramér-von Mises criterion, subtracted from the awake-recovery Cramér-von Mises criterion. The Cramér-von Mises criterion was chosen as it constructs a non-parametric notion of distance between distributions that considers the whole distribution (as opposed to the widest gap, as is the case with Kolmogorov-Smirnov distance). To compute the Cramér-von Mises criterion, all

characteristic roots from a given section (i.e., awake, unconscious, etc.) of a single session were included to ensure the maximum resolution. This approach yielded one test value per session. With this construction, if the test value is negative, it suggests that the recovery distribution is closer to the awake than unconscious state, indicating a return to awake stability levels. We then performed a one sample Wilcoxon signed-rank test on these differences to test if they were significantly below zero (i.e., if the test statistic was indeed negative). The sample sizes are the same as above. The results of this analysis are reported in the chapter in the section “Propofol anesthesia destabilizes cortical neural dynamics”.

A.3.4 Changes to characteristic root frequencies

To analyze how the frequency (imaginary) components of the top 10% of characteristic roots changed in anesthesia, we first aggregated the top 10% of characteristic roots from all windows from all sessions in each NHP. We used the roots estimated from models fit to all areas considered together. Then, to test whether the proportion of roots in each band changed, we computed the proportion of characteristic roots that fell into that band during both the awake state and the unconscious state. This is equivalent to calculating the probability mass function for each band. To test whether the mass was significantly different between the awake and unconscious states, for each band, we performed a two-sided Wilcoxon signed-rank test on the computed probability mass for that frequency band. Sample sizes are the same as above. The results of this analysis are in Figure A.3D.

A.3.5 Destabilization in simulated networks

As mentioned above, for each sampling of the weight matrix $\mathbf{W}$, we simulated every possible pairing of the synaptic gain g and inhibitory scale κ_I . Thus, to test whether a given inhibitory scale κ_{I_1} induced more instability than another inhibitory scale κ_{I_2} , we compared the maximum Lyapunov exponents of the simulations from each gain and from each sampling of the weight matrix. Specifically, given a particular sample of $\mathbf{W}$ and value of g , we can consider the two networks with inhibitory scales κ_{I_1} and κ_{I_2} to be paired dependent samples. Thus, we can conduct a one-sided Wilcoxon signed-rank test using the 100 paired samples for (10 samples of $\mathbf{W}$ times 10 values of g) for each pair of inhibitory scales. The results of this analysis are in Figure A.6.

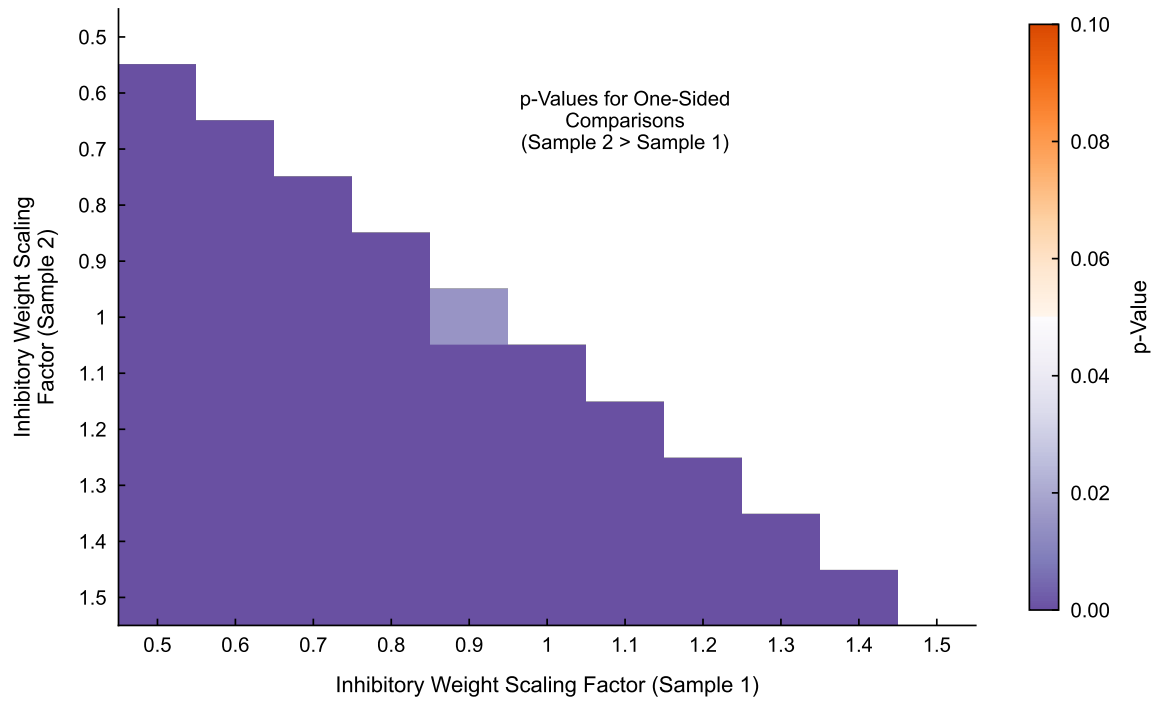


Figure A.6: p-values for statistical comparisons between the stability of RNNs with scaled inhibitory weights. Related to Figure 1.7. We compared the distributions of maximum Lyapunov exponents from all the simulations at varying values of the gain parameter, which controls the stability in the networks. For each inhibitory weight scaling factor, 10 simulations were analyzed. Comparisons are across inhibitory weight scaling factors. The statistical test used was the one-sided Wilcoxon signed-rank test.

Appendix B

Methods: Similar destabilization of neural dynamics

B.1 Experimental Model Details

Multi-electrode neural activity was recorded from four rhesus macaques (*Macaca mulatta*), referred to as non-human primates (NHPs). They were denoted NHP 1, NHP 2, NHP 3, and NHP 4 throughout the manuscript. NHP 1 was female, aged 8 years, ~6.6kg at the time of the study. NHP 2 was male, aged 14 years, ~13.0-kg. NHP 3 was a male, aged 11 years, ~10.2 kg. NHP 4 was a male, aged 14 years, ~13.3 kg, who had previously had headpost and chamber insertion procedures. The other animals were not involved in previous procedures. Propofol experiments were conducted with NHPs 1 and 2, while ketamine and dexmedetomidine experiments were conducted with NHPs 3 and 4. Animals were pair-housed on 12 hr day/night cycles and maintained in a temperature-controlled environment (80 °F). All surgical and animal care procedures were approved by the Massachusetts Institute of Technology (MIT)’s Committee on Animal Care and were conducted in accordance with the guidelines of the National Institute of Health and MIT’s Division of Comparative Medicine.

Each anesthetic infusion is referred to as a “session.” By “session,” we mean a continuous recording, from wakefulness through anesthetic infusion to emergence, in one animal.

B.1.1 Anesthetic protocols

For the propofol data, there were 21 anesthetic sessions across two non-human primate subjects: 10 sessions with NHP 1, and 11 sessions with NHP 2. In each session, the NHP performed a non-demanding delayed saccade task pre-anesthesia. Then, following this task, the NHP experienced a passive air puff/tone classical conditioning task. Specifically, on one third of trials the NHP was delivered a 500-millisecond ringing tone, which after a 500-millisecond blank delay, is followed by an air puff (lasting about 10 milliseconds) to the eye/face designed to elicit a blink response (notably, the air puff also emitted a sound). In another third of trials there was an isolated air puff with no tone, and on the last

third of trials there was a distinct tone played that indicates that no air puff will follow. After around 15-20 minutes of air puff/tone classical conditioning the NHP was infused with propofol anesthesia for 60 minutes, with 30 minutes at a higher loading dose (0.58 milligrams/kilogram/minute for NHP 1 and 0.285 milligrams/kilogram/minute for NHP 2) and 30 minutes at a lower maintenance dose (0.32 milligrams/kilogram/minute for NHP 1 and approx. 0.075 milligrams/kilogram/minute for NHP 2). The infusion was then stopped and the NHP gradually returned to a normal awake state. Propofol was intravenously infused via a computer-controlled syringe pump (PHD ULTRA 4400, Harvard Apparatus, Holliston, MA). Sessions were always separated by at least 2 days and were thus treated as independent samples. Loss of consciousness was determined subjectively by the experimenters as the time following drug onset when the subject’s eyes closed and remained closed for an extended period. Recovery of consciousness (ROC) was similarly estimated as the time when the subject’s eyes reopened following drug offset.

For the ketamine data, there were 12 anesthetic sessions across two non-human primate subjects: 6 sessions with NHP 3 and 6 sessions with NHP 4. For the dexmedetomidine data, there were 16 sessions across the two NHPs: 8 sessions with NHP 3 and 8 sessions with NHP 4. Throughout each session, the NHPs performed a lever task, in which a single (~220 milliseconds) high-pitched auditory tone (4 kHz sinusoid) was presented. This tone instructed the pressing of a lever (8-direction joystick) within 3.5 seconds to receive a juice reward. 10 minutes into the session, the NHPs began the first of three oddball tasks. During the oddball tasks, the NHPs were in each trial presented with a 5-item auditory sequence composed of two distinct tones (‘A’=1300 Hz and ‘B’=500 Hz, each 50 milliseconds in duration with 150 millisecond silent intervals between [850 milliseconds total trial], with a 1 second inter-trial interval). All oddball tasks were 16 minutes in length. This oddball task was repeated 10 minutes into the anesthetic infusion, as well as at the end of the session, when the NHP had awakened from the anesthetic infusion. The start of the final oddball session ranged from 216 minutes after the start of anesthetic infusion to 326 minutes after the start of anesthetic infusion. 10 minutes following the conclusion of the first auditory oddball task, NHPs were administered a bolus of either ketamine at 10 mg/kg or dexmedetomidine at 20 μ g/kg. LOC was determined by a consistent failure to respond over an extended period; ROC was determined by a return of the lever press response (see section “Identifying unresponsive and responsive windows” below).

All data analysis and simulation was done in Python using original code in addition to NumPy,⁴⁵² Hydra,⁴⁵³ Matplotlib,⁴³⁶ Pandas,^{437,454} PyTorch,³⁶¹ scikit-learn,⁴³⁸ statsmodels,⁴⁵⁵ and SciPy.⁴³⁹

B.2 Method Details

B.2.1 LFP recording and preprocessing

NHPs 1 and 2 were implanted with 8×8 chronic Utah arrays, yielding 64 channels per multi-electrode array. The electrodes had a depth of 1.0 mm and 400 μm spacing. Electrodes were placed in four areas: ventrolateral prefrontal cortex (PFC), frontal eye fields (FEF), posterior parietal cortex (PPC) and auditory cortex (STG). During Utah array recordings, areas FEF and PFC were ground and referenced to a common subdural site. Areas STG and PPC also shared a common subdural ground/reference channel. The LFPs were recorded at 30 kHz by a Blackrock Cerebus Digital Acquisition systems and filtered online via a lowpass 250 Hz software filter. The LFPs were then subsequently downsampled to 1 kHz.

NHPs 3 and 4 were implanted with 6x6 chronic Utah arrays, which excluding the corner locations yielded 32 channels per electrode array. The electrodes had a depth of 1.0 mm and 400 μm spacing. Electrodes were placed in four areas: left and right ventrolateral prefrontal cortex (vlPFC-L and vlPFC-R) and left and right dorsolateral prefrontal cortex (dlPFC-L and dlPFC-R). During Utah array recordings, all signals were referenced to a stripped low-impedance silver wire implanted over or under the dura (but not in the brain) either on the left or right hemisphere. The LFPs were recorded at 1 kHz using the Ripple Grapevine system.

For all datasets, behavioral data was aligned to neural recordings by having the behavioral control system send event codes timestamping key task/behavioral events (stimulus onsets/offsets, behavioral response times, reward delivery, etc.) to the neurophysiological recording system. These timestamps were saved alongside the neural data and could thus be used to realign the neural data to task/behavioral events post hoc.

To preprocess the LFPs, we first removed the mean across the entire session from each LFP. Then, to handle noise sources, for NHP1 and NHP2 we removed line noise via temporally windowed sinusoid fits at 60 Hz and all harmonics, as well as empirically found line noise frequencies: 107.35, 214.7, 190.2, 196.8, 393.6. As a filtering step, we low pass filtered each LFP at 300 Hz (3rd order bidirectional Butterworth). For sessions with NHP 4 (excluding three ketamine sessions), noise was removed at 60, 120, and 180 Hz with online Butterworth notch filtering. For NHP 3 (and the three NHP 4 sessions), filtering was performed offline, after data collection. We estimated the line noise at the 1st through 8th harmonics of 60 Hz (i.e., 60–480 Hz in steps of 60) by fitting the raw data on each channel with a temporally adaptive sine function at each given frequency. The fitted sinusoids were then subtracted from the data.

It was not necessary to consider eye movement and muscle artifacts in the preprocessing as these are considered to be signals that are not cerebral in origin, and are more commonly considered in electroencephalography (EEG) and magnetoencephalography (MEG) studies

due to their non-invasive nature.^{441–444} The invasive intracortical electrophysiology signals used in this study are fortunately not susceptible to the same signal contamination risks as EEG and MEG. Furthermore, electrodes were not averaged at any point in the stability analysis pipeline. This is because our method models population dynamics and considers how the overall population representation transforms over time. That information is lost when electrodes are averaged.

To perform artifact removal in the LFP data, we first split each session up into the length of window used for analysis (e.g., 15 seconds). In each window, we find the maximum amplitude of any electrode over the entire time interval. If the maximum amplitude of any two electrodes is greater than 1 mV, we exclude that window from all subsequent analysis. We furthermore exclude any offending electrode from any subsequent analysis in that session. To ensure data quality, we exclude any windows immediately preceding a window in which noise was identified. For the propofol sessions, we excluded between 4 and 23 electrodes per session, and between 6% and 44% of windows. For the ketamine sessions, we excluded between 14 and 30 electrodes per session, and between 4% and 57% of windows. For the dexmedetomidine sessions, we excluded between 11 and 28 electrodes per session, and between 1% and 52% of windows.

B.2.2 Dynamic stability analysis: DeLASE framework

To quantify the dynamic stability of cortical population activity, we applied the Delayed Linear Analysis for Stability Estimation (DeLASE) method, as introduced in Eisen et al. (2024).³⁷ This method allows us to build a predictive model of the high-dimensional, nonlinear neural dynamics and use that model to estimate the system's stability. The analysis was performed on 15-second windows of preprocessed LFP data.

The first step in the framework is state-space reconstruction. Because we are only observing a small subset (LFP electrodes) of a vast, high-dimensional system (the brain), we apply Takens' Theorem to reconstruct the essential dynamics by using the history of the observed signals. We implemented this by creating a time-delay embedding (or Hankel matrix), $\mathbf{H}$, which stacks the LFP data (from all N channels) from the current time and $p-1$ previous time steps (lags). This matrix $\mathbf{H}$ represents the system's trajectory through a high-dimensional state space. To be specific, if $\mathbf{x}_k \in \mathbb{R}^N$ is the state observation at time $k\Delta t$ (e.g., channel activities), then

$$\mathbf{H} = \begin{bmatrix} \mathbf{x}_p & \mathbf{x}_{p+1} & \cdots & \mathbf{x}_T \\ \mathbf{x}_{p-1} & \mathbf{x}_p & \cdots & \mathbf{x}_{T-1} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_{T-p+1} \end{bmatrix} \quad (\text{B.1})$$

Next, we built a linear model of the system's nonlinear dynamics. While neural dynamics are highly nonlinear, Koopman operator theory suggests that they can be represented by a linear model in a suitable high-dimensional space, such as the one provided by our time-delay embedding. We used the Hankel Alternative View of Koopman (HAVOK) approach to find the best-fit linear model.⁶⁷ This involves using Singular Value Decomposition (SVD) on the embedding matrix $\mathbf{H}$ ($\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$) to identify the most important temporal patterns in the data, called eigen-time-delay coordinates (the columns of $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$). The computation of eigen-time-delay coordinates is equivalent to performing PCA whitening on the delay embedding matrix.¹²⁰ We kept the top r most significant patterns (where r is the model rank) and built a simple linear model, $\mathbf{A}_\mathbf{V} \in \mathbb{R}^{r \times r}$, that best predicts how these core patterns (the first r columns of $\mathbf{V}$) evolve from one time step to the next. Following HAVOK and dynamic mode decomposition, we fit the linear model as

$$\mathbf{A}_\mathbf{V} = \underset{\mathbf{A}}{\operatorname{argmin}} \left\| \mathbf{V}_r^{+T} - \mathbf{A}\mathbf{V}_r^{-T} \right\|_2$$

This model $\mathbf{A}_\mathbf{V}$ was then projected back into the full delay-embedded space to create the final discrete-time dynamics matrix, $\mathbf{A}$ ($\mathbf{A} = \mathbf{U}_r \mathbf{\Sigma}_r \mathbf{A}_\mathbf{V} \mathbf{\Sigma}_r^\dagger \mathbf{U}_r^T$) which linearly approximates the brain's complex nonlinear dynamics within that 15-second window.

The final step was to estimate the system's stability from this model. The stability of a dynamical system is determined by how it responds to perturbations. The discrete-time model $\mathbf{A}$ was converted into its continuous-time equivalent, a linear delay differential equation (DDE). We then numerically solved for the characteristic roots λ of this DDE, which determine how the system will respond to perturbation. These roots are complex numbers ($\lambda = a + bi$) where the real part (a) determines the stability along an associated direction. If $a < 0$, the response to perturbation along the corresponding direction is stable and will return to baseline. The value of a is an inverse timescale (units of s^{-1}), where a large negative value signifies a fast recovery (high stability) and a small negative value signifies a slow recovery (low stability, i.e., destabilized). If $a > 0$, the response to perturbation along the corresponding direction is unstable and will diverge away from the baseline state. We defined the Mean Instability in units of inverse seconds for each window as the mean of the real parts of the top 10% of characteristic roots (i.e., the 10% with the largest real parts, corresponding to the slowest-decaying modes). We note that this choice of 10% was validated in previous work,³⁷ but our results also hold for a wide range of thresholds. We performed a robustness analysis by varying this threshold between 1% and 30% (Figure B.1B,C), finding that the significant destabilization observed during Anesthesia was consistent across this range. The Mean Instability estimates the dynamic stability of the cortical population for each 15-second window.

To precisely specify how characteristic roots are computed, we first convert to continuous time (for purposes of interpretation) setting $\tilde{\mathbf{A}}_k = \frac{\mathbf{A}_k}{\Delta t}$ for $k > 0$ and $\tilde{\mathbf{A}}_k = \frac{\mathbf{A}_k - \mathbf{I}_N}{\Delta t}$ for $k = 0$ (this latter term is an approximation of the instantaneous term, and so it includes an identity matrix). This yields the delay differential equation $\dot{\mathbf{x}}(t) = \sum_{k=0}^{p-1} \tilde{\mathbf{A}}_k \mathbf{x}(t - k\Delta t)$. The stability of this delay differential equation is determined by the roots λ of its corresponding

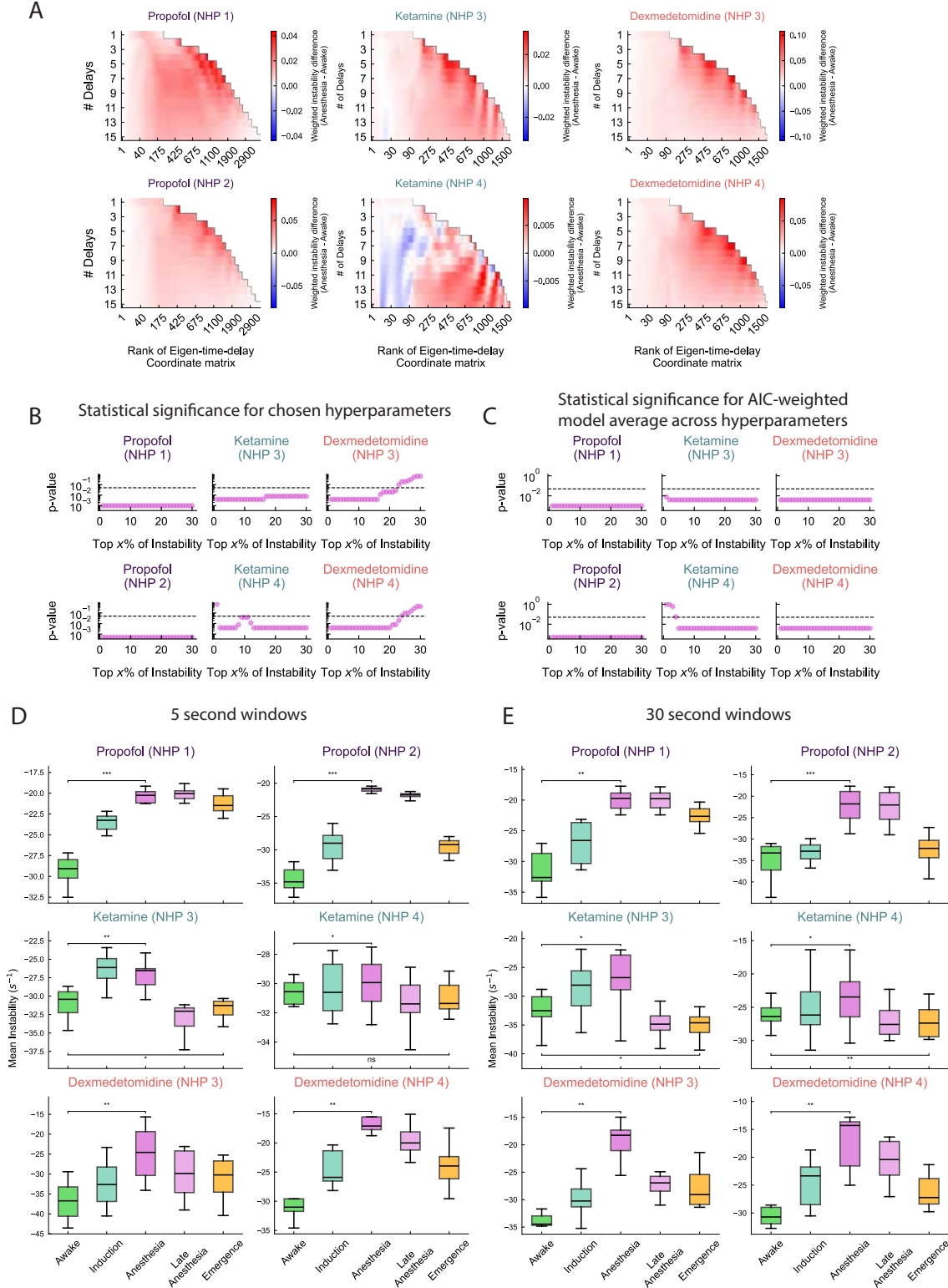


Figure B.1: DeLASE stability estimates are robust to hyperparameter choices. Related to Figure 2.2. (A) AIC-weighted average of stability differences across the full hyperparameter space (number of delays vs. rank of the model). There is generally a dominant destabilization (red) across model parameters and agents. (B) Robustness to the 'top x% of roots' threshold for the chosen best-fit models (from Figure 2.2). p-values for the Anesthesia - Awake comparisons are plotted. The findings generally remain highly significant for a wide range of threshold between 1% and 30%, validating our choice of 10% as a representative...

Figure B.1: (Continued from previous page) value. $n=10$ for propofol NHP 1, $n=11$ for propofol NHP 2, and $n=8$ for each NHP in ketamine and dexmedetomidine, using one-sided Wilcoxon signed-rank test. (C) Robustness to the 'top $x\%$ of roots' threshold for the AIC-weighted model average (from panel A). This panel confirms the findings from (B) are also robust when averaging across all model parameters rather than using a single best-fit model. $n=10$ for propofol NHP 1, $n=11$ for propofol NHP 2, and $n=8$ for each NHP in ketamine and dexmedetomidine, using one-sided Wilcoxon signed-rank test. (D, E) The primary stability analysis from Figure 2, recomputed using (D) 5-second windows and (E) 30-second windows. The central finding of significant destabilization during the Anesthesia epoch is clearly and statistically reproduced at both shorter and longer timescales, demonstrating robustness to the choice of window size. * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$, $n=10$ for propofol NHP 1, $n=11$ for propofol NHP 2, and $n=8$ for each NHP in ketamine and dexmedetomidine, using one-sided Wilcoxon signed-rank test.

characteristic equation:

$$\det \left(\lambda \mathbf{I}_N - \sum_{k=0}^{p-1} \tilde{\mathbf{A}}_k e^{-\lambda k \Delta t} \right) = 0$$

which can be identified using algorithms such as TRACE-DDE, which we implement here. [124,445,447,449](#)

To reduce artifacts from noise or sharp sensory inputs, we filtered the computed characteristic roots. We excluded roots with frequency components > 300 Hz and any unstable roots (real part > 0) with frequencies > 125 Hz, as these high-frequency instabilities typically reflect non-cerebral artifacts. For timescale visualizations (Figure 2.2), we used only stable roots (real part < 0) to compute decay timescales. However, all reported statistics were performed on the "Mean Instability" metric (inverse timescales), which included the full distribution of roots. The vast majority of roots were stable (for the main analyses, 98.93% of propofol roots, 99.81% of ketamine roots, and 99.71% of dexmedetomidine roots were stable, and see Figure B.2D for a depiction of the distribution).

To compute the Mean Characteristic Timescale normalized to the awake baseline (Figure 2.2A-C), we first established a baseline for each session by calculating the geometric mean of the timescales from all Awake windows. The timescales in every other window were then expressed as a ratio to this Awake baseline. The Mean Characteristic Timescale was computed as the geometric mean of the slowest 10% of timescales as a ratio to the Awake baseline. For plotting, these ratio time-series were aligned across sessions by linearly interpolating them onto a common 15-second grid.

Picking hyperparameters

To select (1) the number of delay embedding coordinates and (2) the rank of the eigen-time-delay coordinates used to estimate the dynamics, a grid search was performed for each

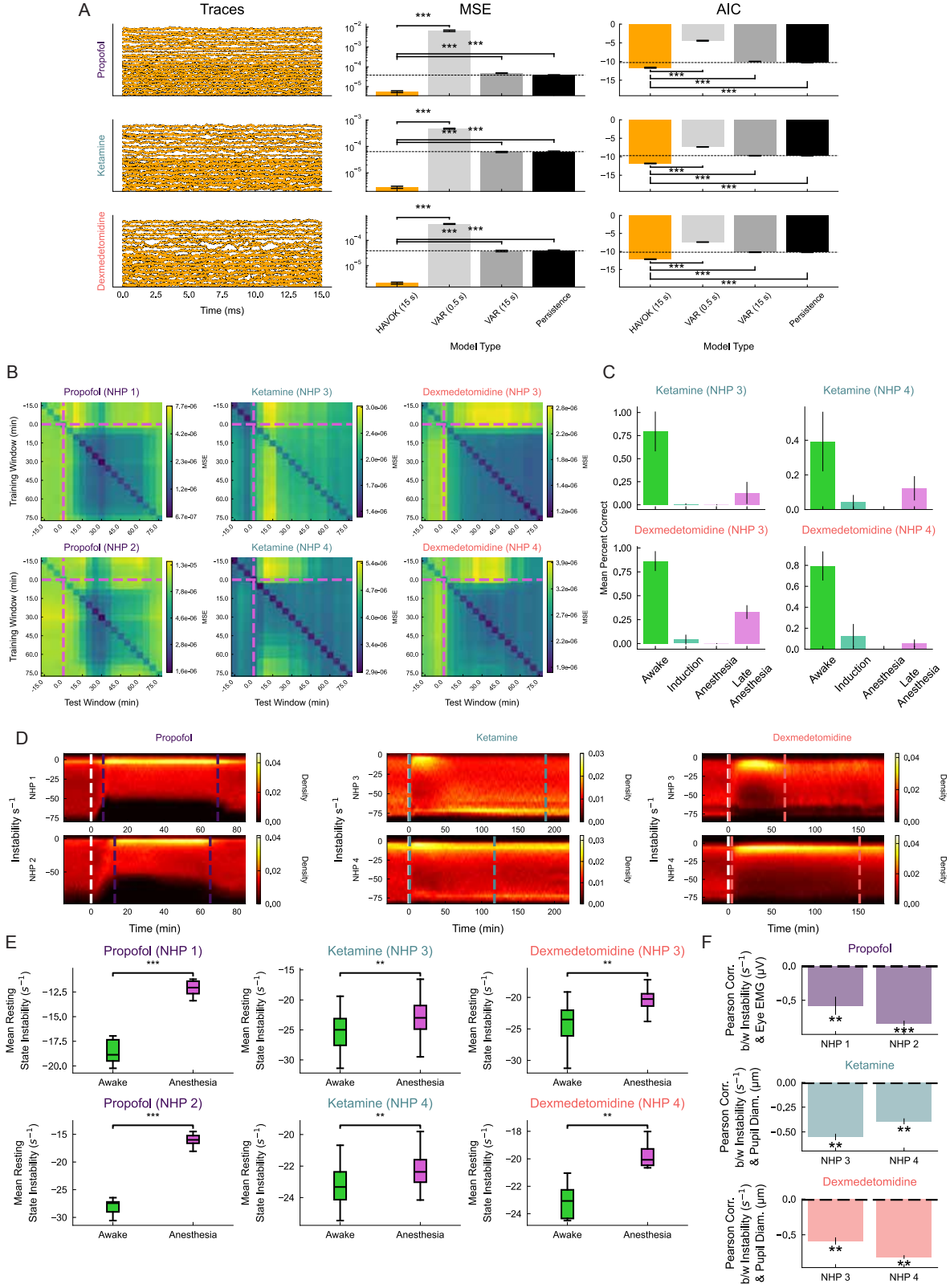


Figure B.2: Model validation, behavioral confirmation, and supplementary stability analyses. Related to Figure 2.2. (A) DeLASE (HAVOK) model validation and comparison. Bar plots show Mean Squared Error (MSE) and Akaike Information Criterion (AIC) for one-step-ahead prediction on held-out test data, averaged across all windows and sessions for each anesthetic. The DeLASE (HAVOK) model, using a 15 s training window, consistently outperforms standard Vector Autoregressive (VAR) models (using 0.5 s ...

Figure B.2: (Continued from previous page) or 15 s training windows) and a baseline persistence model. Lower MSE and AIC validate DeLASE as providing a more accurate and parsimonious fit. Sample spike raster traces (15 s) are shown at left, with the ground truth in black and the HAVOK one-step predictions in orange. *** indicates $p < 0.001$, $n=10$ for propofol NHP 1, $n=11$ for propofol NHP 2, and $n=8$ for each NHP in ketamine and dexmedetomidine, using one-sided Wilcoxon signed-rank test. (B) Cross-temporal generalization analysis. Heatmaps show the MSE when a model trained on a specific 15-second window (y-axis) is used to predict data from other 15-second windows in the session (x-axis). The high off-diagonal error demonstrates non-stationary dynamics, justifying the use of tiled windows over a single global model. (C) Behavioral confirmation of anesthetic-induced unresponsiveness. Plots show the mean percent correct on the lever-press task during ketamine and dexmedetomidine sessions, grouped by epoch. Performance drops to near-zero levels during the Anesthesia epoch, confirming that the NHPs were behaviorally unresponsive. While performance was low during the Late Anesthesia epoch, behavioral responsiveness varied. Propofol is not shown as the NHP was unresponsive from the moment in which eyes closed until eyes opened in all sessions. (D) Time-resolved histograms of the top 10% of characteristic roots from DeLASE stability analysis. Histograms were computed over characteristic roots computed from all sessions with a particular NHP. The white vertical dotted line indicates the start of anesthetic infusion. The first colored vertical dotted line indicates the mean time of loss of consciousness, and the second indicates the mean time of return of consciousness. (E) Comparison of instability between Awake and Anesthesia epochs using only resting state data. Resting state windows were identified by isolating times in which at least 3 s had passed since an external stimulus was presented. In all sessions, these windows were limited to be approximately 3 s. In all NHPs and for all three anesthetics, the mean instability (real part of the characteristic root, s^{-1}) was significantly closer to zero (less negative) during the Anesthesia epoch compared to the Awake epoch. This demonstrates that the destabilization of cortical dynamics during anesthetic-induced unconsciousness is robust even when considering resting state dynamics. During resting state, the cortex is presumably perturbed by endogenous inputs. ** indicates $p < 0.01$ and *** indicates $p < 0.001$, $n=10$ for propofol NHP 1, $n=11$ for propofol NHP 2, and $n=8$ for each NHP in ketamine and dexmedetomidine, using one-sided Wilcoxon signed-rank test. (F) Pearson correlation between instability and physiological variables for propofol (top), ketamine (middle) and dexmedetomidine (bottom). Instability was correlated with the magnitude of eye EMG activity for propofol and with pupil diameter for ketamine and dexmedetomidine. ** indicates $p < 0.01$ and *** indicates $p < 0.001$, $n=10$ for propofol NHP 1, $n=11$ for propofol NHP 2, and $n=8$ for each NHP in ketamine and dexmedetomidine, using one-sided Wilcoxon signed-rank test.

session. Our hyperparameter search was extensive (16,000–20,000 model fits *per session*) to find the best predictive model for each session, ensuring independence. For the main analysis, we used fixed window sizes (15 seconds of neural data). These window sizes ensure a long enough time window to capture a wide range of dynamic motifs, but a short enough time window to preserve computational tractability. For the grid search, models were fit on 4 windows from each subsection of the session, and parameters were chosen on a per-session basis. Windows were chosen to ensure that all sections of the session were included (see below). Since certain windows were excluded based on estimated electrical noise (see above), if there was not enough valid data in a subsection to constitute the 4 windows needed for the hyperparameter search, we used as many windows as were available. The metric used to assess model quality was Akaike Information Criterion (AIC). AIC quantifies the balance between high quality model prediction (as measured by one-step prediction error on test data) with the number of model parameters. Models are penalized for having larger numbers of parameters, such that the optimal model of minimal complexity can be obtained. For each session, we computed the AIC on a window of test data of equal size to the training window (15 seconds) and immediately following the training window in the session. We averaged the AIC across all windows from all subsections. While larger models tend to perform better, after

a certain point the added parameters are not benefitting the prediction quality sufficiently to justify their inclusion. Because we compare stability between sections of each session, the hyperparameters minimizing AIC within a given session (and area) are chosen. The time interval between delays was held fixed at 1 time step (1 millisecond) for this analysis. Chosen hyperparameters varied across sessions. All hyperparameters for models used in the primary analysis are included in Table B.1.

Table B.1: DeLASE model hyperparameters selected for each recording session. This table lists the optimal hyperparameters used for the stability analysis in each independent session, determined via a grid search minimizing the Akaike Information Criterion (AIC). “# of delays” corresponds to the number of time-delay embeddings used to reconstruct the state space. “rank” corresponds to the rank of the eigen-time-delay coordinates used to approximate the linear dynamics in the HAVOK model.

Agent	Animal	Session #	# of delays	Rank
Propofol	NHP 1	1	5	1150
Propofol	NHP 1	2	5	1100
Propofol	NHP 1	3	5	1150
Propofol	NHP 1	4	5	1150
Propofol	NHP 1	5	5	1100
Propofol	NHP 1	6	5	1100
Propofol	NHP 1	7	5	1100
Propofol	NHP 1	8	5	1100
Propofol	NHP 1	9	5	1100
Propofol	NHP 1	10	5	1100
Propofol	NHP 2	1	5	1000
Propofol	NHP 2	2	5	1000
Propofol	NHP 2	3	5	1000
Propofol	NHP 2	4	5	1050
Propofol	NHP 2	5	5	1050
Propofol	NHP 2	6	5	1050
Propofol	NHP 2	7	5	1000
Propofol	NHP 2	8	5	1050
Propofol	NHP 2	9	5	1050
Propofol	NHP 2	10	5	1050
Propofol	NHP 2	11	5	1000
Ketamine	NHP 3	1	8	900
Ketamine	NHP 3	2	9	850
Ketamine	NHP 3	3	9	900
Ketamine	NHP 3	4	7	750
Ketamine	NHP 3	5	7	725
Ketamine	NHP 3	6	9	900
Ketamine	NHP 3	7	9	950
Ketamine	NHP 3	8	9	950
Ketamine	NHP 4	1	9	900

Continued on next page

Table B.1 – continued from previous page

Agent	Animal	Session #	# of delays	Rank
Ketamine	NHP 4	2	9	950
Ketamine	NHP 4	3	9	900
Ketamine	NHP 4	4	9	900
Ketamine	NHP 4	5	9	1000
Ketamine	NHP 4	6	9	900
Ketamine	NHP 4	7	9	1000
Ketamine	NHP 4	8	9	950
Dexmedetomidine	NHP 3	1	9	900
Dexmedetomidine	NHP 3	2	7	800
Dexmedetomidine	NHP 3	3	7	725
Dexmedetomidine	NHP 3	4	9	950
Dexmedetomidine	NHP 3	5	8	900
Dexmedetomidine	NHP 3	6	8	850
Dexmedetomidine	NHP 3	7	9	950
Dexmedetomidine	NHP 3	8	9	1050
Dexmedetomidine	NHP 4	1	9	900
Dexmedetomidine	NHP 4	2	9	900
Dexmedetomidine	NHP 4	3	9	1000
Dexmedetomidine	NHP 4	4	8	850
Dexmedetomidine	NHP 4	5	9	1000
Dexmedetomidine	NHP 4	6	9	950
Dexmedetomidine	NHP 4	7	9	950
Dexmedetomidine	NHP 4	8	9	900

Grid search section definitions

For propofol sessions, the relevant subsections were awake (the 15 minutes preceding the onset of infusion), anesthetic induction (the 15 minutes following the onset of infusion), loading dose (from 15 minutes to 30 minutes following infusion), maintenance dose (from 30 minutes to 60 minutes following infusion), and recovery (from 60 minutes to 75 minutes following infusion). For ketamine and dexmedetomidine sessions, the relevant sections are the awake oddball, the awake lever press sections (the 10 minutes preceding and following the awake oddball, respectively), induction (the 10 minutes following infusion), anesthesia oddball, the early unconscious period (from the end of the oddball until 45 minutes post-infusion), the late unconscious period (from 45 minutes post infusion until the recovery oddball), and the recovery oddball. 4 windows were randomly selected from each subsection in each session.

B.2.3 Session epochs

We divided each session into epochs for the purposes of plotting and comparison. These epochs were the Awake epoch, the Anesthesia epoch, and the Emergence epoch. For the propofol sessions, the Awake epoch was defined as the 15 minutes preceding the start of anesthetic infusion. The Anesthesia epoch was defined as the 45-minute period starting 15 minutes after the onset of anesthetic infusion and ending 60 minutes after the onset of infusion. The Emergence epoch was defined as the 15 minutes starting from 60 minutes after the onset of infusion (i.e., the end of propofol infusion) and ending 75 minutes after the onset of infusion. The Induction epoch was defined as the 15 minutes following the onset of infusion. The Late Anesthesia epoch was defined as the 30-minute period starting from 30 minutes after the start of infusion until the end of infusion (60 minutes after the onset of infusion). For the ketamine and dexmedetomidine sessions, the Awake epoch was defined as the approximately 36-minute period leading up to the start of anesthetic infusion, which included the awake oddball subsection as well as both awake lever-press task subsections. The Anesthesia epoch was defined as the approximately 35-minute period beginning with the anesthesia oddball subsection and ending 45 minutes after the onset of infusion. The Emergence epoch was taken to be the final oddball subsection, which was administered after the NHPs had awakened from the anesthetic infusion. The Induction epoch was defined as the 10 minutes following the start of anesthetic infusion, which leads up to the anesthetic oddball subsection. The Late Anesthesia epoch was defined from 45 minutes after the onset of infusion until the final oddball subsection.

B.2.4 VAR and baseline comparisons

To validate the stability estimates derived from DeLASE and to benchmark its predictive performance, we implemented a classical Vector Auto-Regressive (VAR) model and a persistence baseline.

We fit a first-order VAR(1) model to the same 15-second LFP windows used for the DeLASE analysis, as well as 500 millisecond windows as in previous work (Figure B.2A and Figure B.3A,B).^{63,190} For a given window of multichannel LFP data $\mathbf{X} \in \mathbb{R}^{T \times N}$ (where T is time points and N is the number of channels), the VAR(1) transition matrix $\mathbf{A}_{VAR}$ was estimated using ordinary least squares:

$$\mathbf{A}_{VAR} = \underset{\mathbf{A}}{\operatorname{argmin}} \left\| \mathbf{X}_{t+1}^T - \mathbf{A} \mathbf{X}_t^T \right\|_F^2$$

where $\mathbf{X}_t$ and $\mathbf{X}_{t+1}$ are the data matrices time-shifted by one sample.

We evaluated the model fit and predictive accuracy of DeLASE against two baselines: the VAR(1) model described above and a "persistence baseline" which assumes the system state does not change (i.e., predicting the next state as the same as the current state).

For each analysis window, models were fit on the first portion of the data (training set) and evaluated on a subsequent held-out test window of equal size (i.e., either 15 seconds or 500 milliseconds, depending on the training window). Prediction error was quantified via Akaike Information Criterion

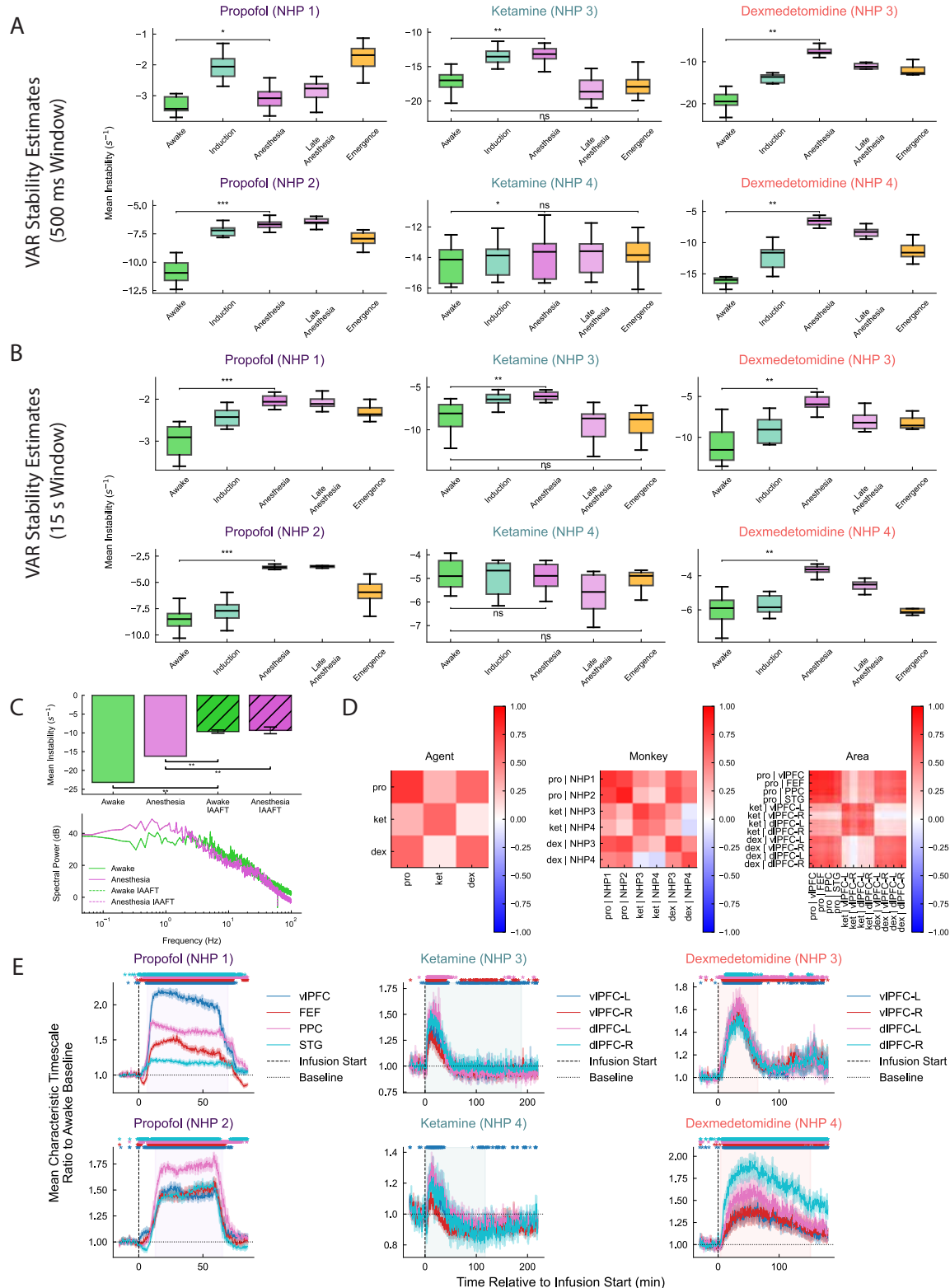


Figure B.3: Validation of DeLASE, robustness checks, and regional destabilization results. Related to Figure 2.2. (A, B) Stability analysis using classical Vector Autoregressive (VAR) models. Stability was estimated from the eigenvalues of first-order VAR models fit to either (A) 500 ms windows, as in previous work, or (B) 15 s windows. While the VAR models are outperformed by DeLASE in predictive accuracy (see Figure B.2), these more classical methods qualitatively reproduce the central finding: a significant shift toward instability (less negative eigenvalues) during the Anesthesia epoch compared to...

Figure B.3: (Continued from previous page) the Awake epoch. * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$, $n=10$ for propofol NHP 1, $n=11$ for propofol NHP 2, and $n=8$ for each NHP in ketamine and dexmedetomidine, using one-sided Wilcoxon signed-rank test. (C) Surrogate data analysis confirms DeLASE captures dynamics beyond spectral power. (Top) Stability estimates for a sample Awake and Anesthesia window (solid bars) compared to Iterative Amplitude Adjusted Fourier Transform (IAAFT) surrogates (hatched bars). (Bottom) The IAAFT method generates surrogate data that perfectly matches the power spectrum of the original data. Despite having identical power spectra, the surrogate data has significantly different stability estimates from the real data. This confirms that the DeLASE measurement is not trivially driven by changes in spectral power and depends on the complex, cross-channel temporal structure of the data. (D) Similarity analysis of instability time-series. Pearson correlation matrices show the similarity between all recording sessions and areas, grouped by Agent (left), Agent and Monkey (middle), and Agent and Area (right). The clear block-diagonal structure, dominated by the anesthetic agent, demonstrates that the instability profile is determined by the drug administered, not by subject-specific (animal) or region-specific (area) differences. (E) Destabilization is consistent across all recorded cortical areas. The Mean Characteristic Timescale (from Figure 2.2A) is shown but computed per each individual recording area. For all three anesthetics, the destabilization effect (an increase in the timescale ratio) is observed robustly and concurrently across all measured brain regions (e.g., vIPFC, FEF, 7b, CPB for propofol; bilateral vIPFC and dlPFC for ketamine and dexmedetomidine). * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$, $n=10$ for propofol NHP 1, $n=11$ for propofol NHP 2, and $n=8$ for each NHP in ketamine and dexmedetomidine, using one-sided Wilcoxon signed-rank test.

(AIC) and mean squared error (MSE) on one-step-ahead prediction (Figure B.2A). As described above, AIC is a measure of model quality penalizing for complexity (number of parameters).

These metrics were computed for the DeLASE (HAVOK) model, the VAR(1) model on both 15 second and 500 millisecond time windows, and the persistence baseline to demonstrate the superior predictive capability of the HAVOK models.

To estimate stability comparable to DeLASE, we computed the eigenvalues λ_i of the fitted $\mathbf{A}_{VAR}$ matrix. These discrete-time eigenvalues were converted to continuous-time timescales τ_i using the transformation $\tau_i = \frac{\ln(|\lambda_i|)}{\Delta t}$ as in previous work,¹⁹⁰ where Δt is the sampling interval. The maximum 10% of the real part of these exponents was taken as the stability metric for the window, for comparison with DeLASE (Figure B.3A,B).

B.2.5 Cross-window prediction quality analysis

To evaluate the stationarity of neural dynamics and the validity of our windowed modeling approach, we performed a cross-temporal generalization analysis (Figure B.2B). For each recording session, we partitioned the LFP data into the same 15-second windows used for the primary stability analysis. We trained a separate HAVOK model on each 15-second window (the "training window"). We then evaluated the predictive performance of each trained model on every other 15-second window in the session (the "test windows"). Predictive performance was quantified using the MSE of the one-step-ahead prediction on the held-out test data. This procedure yielded a cross-temporal generalization matrix $\mathbf{M}$, where the entry $M_{i,j}$ represents the MSE of the model trained on window i when predicting the dynamics of window j . This matrix allows us to visualize how well dynamical models generalize across different epochs (e.g., Awake vs. Anesthesia). High off-diagonal errors

indicate that the dynamical rules governing the system change over time, necessitating the use of time-resolved, local models rather than a single global model for the entire session.

B.2.6 Identifying unresponsive windows

For the analysis in Figure B.2C and Figure B.4, to identify windows in which the NHPs were unresponsive, we considered the percent of correct lever trials (i.e., the trials in which the NHP successfully pressed the lever in response to the indicating tone). For each 15 second window used for stability analysis, we computed the percent of correct lever trials in a 120 second window centered on the start of the 15 second stability window. Any window in which the NHP responded to fewer than 10% of lever presses correctly was labeled an unresponsive window. To estimate times for which the NHPs were unresponsive (shaded regions in Figure 2.2A), we performed the same analysis for all ketamine and dexmedetomidine sessions. The mean time of the first window for which fewer than 10% of lever presses were correct was taken to be the mean time of loss of consciousness. The mean time of the first window following this time for which more than 10% of lever presses were correct was taken to be the mean time of return of consciousness. For propofol, the moments at which the NHP eyes closed and the moment at which they reopened, were judged and recorded by the experimenter. Thus, the mean time of loss of consciousness was taken to be the mean time of eyes closing, and the mean time of return of consciousness was taken to be the mean time of eyes opening.

B.2.7 Resting state stability analysis

To confirm that the observed destabilization was not driven by sensory-evoked responses, we performed a control analysis restricted to resting-state data (Figure B.2E). Resting-state segments were identified by isolating periods in the recording where at least 3 seconds had elapsed since the presentation of any external stimulus (e.g., auditory tones or air puffs). These segments were strictly limited to a duration of 3 to 3.5 seconds. This restriction was imposed to strictly ensure the absence of any late-component evoked responses or stimulus artifacts, as finding longer contiguous blocks of stimulus-free data was constrained by the frequent trial structure of the experimental tasks. We applied the DeLASE method to these 3 second windows. To ensure comparability with the main results, we used the same model hyperparameters (embedding delays and rank) as were selected for the primary 15 second window analysis. Instability was quantified as the mean of the top 10% of the real parts of the characteristic roots. We then compared the mean instability of these resting-state windows between the Awake and Anesthesia epochs.

B.2.8 Correlation with physiological markers

To validate that the estimated dynamic instability tracked independent physiological measures of anesthetic depth, we correlated the time-resolved stability metric with concurrently recorded physiological signals (Figure S1F). For propofol, we analyzed the electromyography (EMG) signal

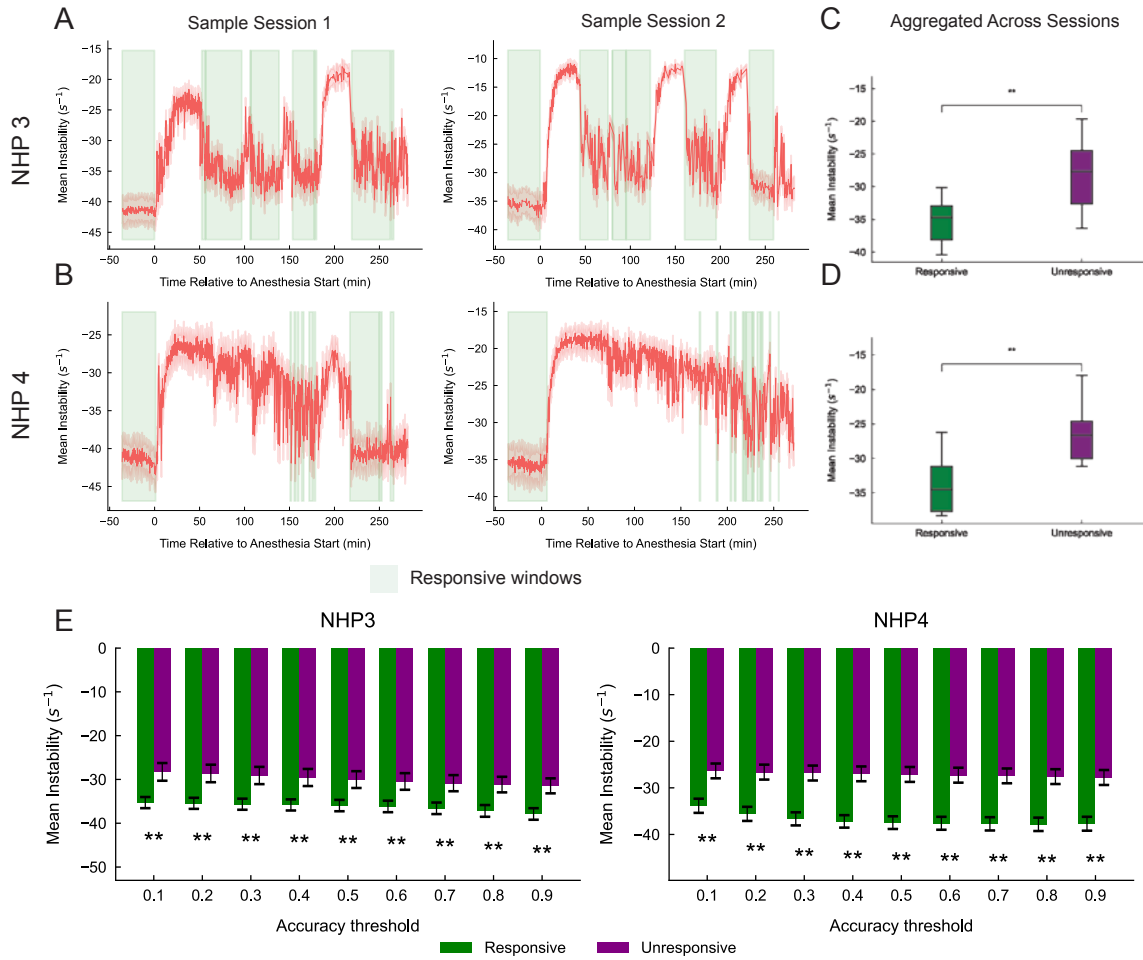


Figure B.4: Dynamic instability tracks real-time responsiveness. Related to Figure 2.2. (A, B) Time-resolved instability (red line) from two sample dexmedetomidine sessions for NHP 3 (A) and NHP 4 (B). Periods in which the NHP was identified as behaviorally responsive (based on a lever-press task) are shaded in green. Throughout the sessions, NHPs were presented with tones indicating that they should press a lever. To identify the periods in which the NHPs were behaviorally responsive, we computed the percentage of correct lever presses within two-minute windows. For this analysis, responsive windows were taken to be those in which the NHP pressed the lever in response to more than 10% of the tones. Instability visibly decreases (becomes more stable) during brief moments of responsiveness and increases (becomes less stable) during periods of unresponsiveness. (C, D) Mean instability aggregated across all sessions for NHP 3 (C) and NHP 4 (D). This analysis uses the 10% lever-press accuracy threshold. Unresponsive periods (purple) were significantly more unstable (less negative) than responsive periods (green) in both NHPs. ** indicates $p < 0.01$, $n=8$ for each NHP, using one-sided Wilcoxon signed-rank test. (E) Robustness analysis for the behavioral threshold. The analysis from (C) and (D) was repeated using accuracy thresholds ranging from 10% (0.1) to 90% (0.9). The significant difference in stability between responsive (green) and unresponsive (purple) windows was robustly maintained regardless of the specific threshold chosen. ** indicates $p < 0.01$, $n=8$ for each NHP, using one-sided Wilcoxon signed-rank test.

recorded from electrodes placed adjacent to the eyes. The magnitude of this signal serves as a proxy for muscle tone and eye movement, which are known to decrease during anesthesia.^{456–458} For ketamine, and dexmedetomidine, we analyzed pupil diameter, which was continuously tracked throughout the sessions. Pupil diameter is a well-established autonomic correlate of arousal and anesthetic depth.^{459,460} For each session, we computed the Pearson correlation between the time series of mean instability (inverse timescale, s^{-1}) and the corresponding physiological time series (mean eye EMG magnitude or pupil diameter) calculated in matching 15-second windows. To assess the consistency of this relationship across sessions, we performed a one-sided Wilcoxon signed-rank test on the distribution of correlation coefficients obtained for each subject. The null hypothesis was that the median correlation is zero, while the alternative hypothesis was that the correlation is less than zero (indicating that as the cortex becomes more unstable/less stable, the physiological marker of arousal decreases).

B.2.9 Surrogate data generation and analysis

To validate that the estimated stability metrics captured complex temporal dynamics rather than simple spectral properties (Figure B.3C), we generated surrogate data using the Iterative Amplitude Adjusted Fourier Transform (IAAFT) method.^{461–463} For a given LFP window, the IAAFT algorithm iteratively randomizes the phases of the signal's Fourier transform while enforcing the original signal's power spectrum and amplitude distribution. This process yields surrogate signals that possess identical spectral power and amplitude distributions to the original neural data but lack the specific nonlinear temporal correlations. We used a public implementation of the IAAFT.⁴⁶⁴ As a proof by contradiction, we picked a sample window from the Awake epoch and from the Anesthesia epoch and intended to demonstrate that the stability is not trivially determined by the power spectrum. We then generated 10 surrogates for each window and performed stability analysis on these surrogates using the same DeLASE model hyperparameters applied to the original data.

B.2.10 Single-area stability and similarity

To determine if the observed destabilization was driven by specific brain regions or was a global cortical phenomenon, we performed a region-specific stability analysis (Figure B.3E) and a time-profile similarity analysis (Figure B.3D).

We applied the DeLASE framework independently to data from each recorded cortical area. For the propofol sessions, the areas were ventrolateral prefrontal cortex (PFC), frontal eye fields (FEF), posterior parietal cortex (PPC) and auditory cortex (STG). For ketamine and dexmedetomidine sessions, we analyzed the left and right ventrolateral and dorsolateral prefrontal cortices (vlPFC-L/R, dlPFC-L/R). We conducted a full hyperparameter grid search for each area, in each session. For each area, we computed the mean instability (inverse timescale) in 15-second windows across the session. We then visualized the mean characteristic timescale ratio relative to the awake baseline for each area to confirm that destabilization occurred consistently across all recorded regions (Figure B.3E).

To quantify the similarity of instability dynamics across different anesthetic agents, subjects, and

brain areas, we computed the Pearson correlation between their time-courses (Figure B.3D). First, the instability time-series for each session (or specific area within a session) were linearly interpolated onto a common temporal grid with a resolution of 15 seconds to ensure alignment across experiments with slightly varying timings. We then constructed a correlation matrix by calculating the Pearson correlation coefficient between every pair of interpolated time-series (clipped to from 15 minutes preceding anesthetic infusion to 85 minutes following anesthetic infusion). These correlations were grouped by agent, subject, and area to assess whether the dynamic profile of instability was primarily determined by the anesthetic agent or by other factors. High correlations within agent groups (even across different subjects and areas) indicated that the temporal progression of destabilization is a robust signature of the specific drug's action.

B.2.11 AIC-weighted stability estimates

To ensure that our findings were not dependent on the specific choice of model hyperparameters (embedding delay and rank), we performed a robustness analysis using model averaging (Figure B.1A, C). We recomputed the stability estimates across a comprehensive grid of hyperparameters for every session. Due to the computational cost of fitting thousands of models per session, this analysis was performed with a temporal stride of 120 seconds (instead of 15 seconds) between analysis windows.

For each combination of hyperparameters (rank r and number of delays p), we computed the mean Akaike Information Criterion (AIC) of the fitted HAVOK models across the session. We then calculated Akaike weights $w_{r,p}$ for each model specification using the formula:

$$w_{r,p} = \frac{\exp\left(-\frac{1}{2}\Delta AIC_{r,p}\right)}{\sum_{i,j} \exp\left(-\frac{1}{2}\Delta AIC_{i,j}\right)}$$

where $\Delta AIC_{r,p} = AIC_{r,p} - AIC_{\min}$ and $AIC_{\min}$ is the minimum mean AIC value observed across the hyperparameter grid for that session.⁴⁶⁵

We then computed an AIC-weighted estimate of the difference in instability (inverse timescales) for each animal/agent combination by taking the weighted sum of the instability difference values produced by each model. That is, for a given animal and agent combination, let there be K sessions. Then, for each session let ΔS_k be the change in mean instability (top 10%) from the Awake to Anesthesia epoch (i.e. Anesthesia – Awake instability) for $k \in \{1, \dots, K\}$. Then, the model weighted changed in instability for parameters rank r and number of delays p is

$$\Delta S_{r,p} = \sum_{k=1}^K w_{r,p}^k \Delta S_k$$

where $w_{r,p}^k$ corresponding to session k (note that AIC weights are computed on a *per session* basis). These values are plotted on a per hyperparameter basis in Figure B.1A. In this Figure, we compared these weighted stability estimates to the null hypothesis of zero-difference using one-sided Wilcoxon

signed-rank test to confirm that the destabilization effect was robust to model specification. We performed further model averaging to identify the AIC-weighted overall estimate of the change in instability as

$$\Delta S_{AIC} = \sum_{r,p} \Delta S_{r,p}$$

We then used these weighted stability estimates to conduct multiple Wilcoxon signed-rank tests with various top % of characteristic roots in Figure B.1C.

B.2.12 Robustness across % of characteristic roots

To verify that our stability estimates were not dependent on the specific choice of using the top 10% of characteristic roots (the threshold used in the primary analysis), we performed a robustness check across a wide range of threshold values (Figure B.1B,C). We recomputed the mean instability metric using thresholds ranging from the top 1% to the top 30% of characteristic roots, in increments of 1%. This range was chosen to encompass both very conservative estimates (relying only on the few most dominant modes) and broader estimates (averaging over a larger portion of the dynamical spectrum). This analysis was performed in two ways to ensure robustness. (1) Best-Fit Models (Figure B.1B): We applied the varying thresholds to the single best-fitting HAVOK model (lowest AIC) identified for each session. (2) AIC-Weighted Average (Figure B.1C): We applied the varying thresholds to the AIC-weighted average stability estimate (calculated as described in the "AIC-weighted stability estimates" section) to account for model uncertainty. For each threshold value in this range (1% to 30%), we compared the mean instability between the Awake and Anesthesia epochs (and Awake vs. Emergence) using a one-sided Wilcoxon signed-rank test. We plotted the resulting p-values as a function of the threshold percentage to demonstrate that the significant destabilization effect was maintained across large portions of the tested range.

B.2.13 Stability analysis with different windows

To ensure that our findings were not artifacts of the specific temporal resolution chosen for the primary analysis, we tested the robustness of our results to variations in the analysis window length (Figure B.1D,E). We repeated the complete DeLASE stability analysis using two alternative window sizes: 5 seconds and 30 seconds. For each window size, we partitioned the recording sessions into non-overlapping segments (stride equal to window size). We then conducted full per-session grid searches to identify the optimal hyperparameters, and fit HAVOK models to these windows. We estimated the dynamic instability using the same pipeline as the primary 15-second analysis. This included (1) constructing time-delay embeddings and fitting linear dynamics models for every 5-second or 30-second window, (2) computing the mean real part of the top 10% of characteristic roots for each window, and (3) comparing the mean instability between the Awake, Anesthesia, and Emergence epochs using a one-sided Wilcoxon signed-rank test. The results from these analyses demonstrated that the significant destabilization observed during anesthesia was consistent across

timescales of training windows, remaining robust whether the dynamics were estimated from short (5s) or long (30s) snapshots of neural activity.

B.2.14 Responsiveness-tracking in dexmedetomidine

To investigate whether dynamic instability tracked real-time fluctuations in consciousness during dexmedetomidine sedation, we classified 15-second analysis windows as "responsive" or "unresponsive" based on behavioral performance (Figure B.4). Dexmedetomidine is in fact known to produce an arousable sedation state, in which subjects can move easily between wakefulness and unconsciousness.^{466–468} For each 15-second window used in the stability analysis, we calculated the percentage of correct lever presses in a 120-second window centered on the start of the analysis window (as discussed above in section “Identifying unresponsive and responsive windows”). Windows in which the subject responded correctly to fewer than 10% of the auditory cues were classified as unresponsive, while those exceeding this threshold were classified as responsive. For each session, we then computed the mean instability (inverse timescales) separately for all responsive and unresponsive windows (Figure B.4A,B). We then compared these session-averaged stability values using a one-sided Wilcoxon signed-rank test to determine if unresponsive periods were significantly more unstable than responsive periods (Figure B.4C,D). To ensure that our results were not dependent on the specific choice of the 10% accuracy threshold, we repeated this analysis using a range of thresholds from 10% to 90% (in 10% increments). We confirmed that the significant difference in stability between responsive and unresponsive states was maintained across this range (Figure B.4E).

B.2.15 Sensory response and autocorrelation analysis

We computed sensory responses to auditory stimuli for Figure 2.3 and Figure B.5. For the propofol sessions, sensory responses to the isolated tones were analyzed. For the ketamine and dexmedetomidine sessions, sensory responses to the auditory oddball task were analyzed. This task was chosen as opposed to the tones that indicated lever presses to avoid the presence of a motor signal in the neural data. Sensory responses were averaged over all trials and electrodes within a single session.

To quantify the temporal characteristics of these responses, we computed their autocorrelation functions (ACFs) (Figure 2.3B). For each trial, the LFP time series was collected in a 1-second window immediately following the onset of the first auditory stimulus. Given the 1 kHz sampling rate, this resulted in a time series consisting of $T = 1000$ samples. Let M be the number of trials in a given epoch (Awake, Anesthesia, Emergence) of a given session. Let $\mathbf{x}_t^{(j)} \in \mathbb{R}^N$ be the vector of LFP activity at time $t \in \{1, \dots, T\}$ and for trial $j \in \{1, \dots, M\}$. Let $x_{i,t}^{(j)}$ be the activity of the i -th electrode on trial j at time t . To compute the autocorrelation function of this univariate time series, one first computes mean $\mu_i^{(j)} = \frac{1}{T} \sum_{t=1}^T x_{i,t}^{(j)}$ as well as the variance $\left(\sigma_i^{(j)}\right)^2 = \frac{1}{T} \sum_{t=1}^T \left(x_{i,t}^{(j)} - \mu_i^{(j)}\right)^2$. The ACF $R(\tau)$ measures the correlation of the signal with a delayed copy of itself as a function of the time lag τ . For a given lag τ , one computes the autocovariance function

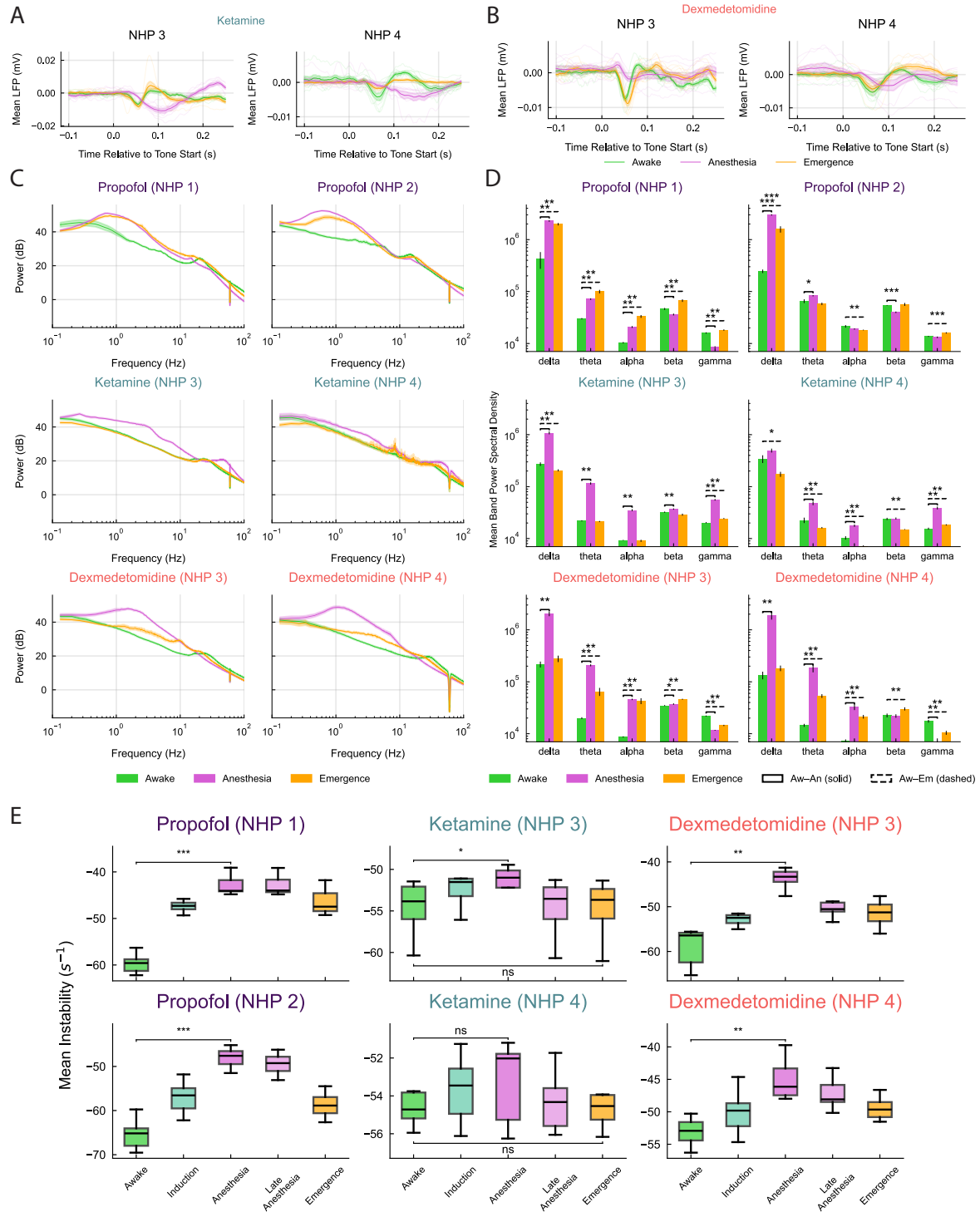


Figure B.5: Supplementary sensory and spectral analyses. Related to Figure 2.2-2.4. (A, B) Temporally zoomed-in LFP responses to auditory stimuli from Figure 3A for (A) ketamine and (B) dexmedetomidine. The plots show the initial ~250 ms of the response, highlighting the early-latency components of the evoked potential. (C) Mean power spectral densities (PSDs) for each epoch, averaged across all sessions for propofol (top), ketamine (middle), and dexmedetomidine (bottom). These plots visualize the well-known shift to low-frequency power during the Anesthesia epoch (purple) compared to the Awake (green) and Emergence (orange) epochs. (D) Comparison of mean power in specific frequency bands (delta, theta, alpha, beta, gamma) across epochs for propofol (top), ketamine (middle), and dexmedetomidine (bottom). This quantifies the spectral shifts seen in the PSDs, confirming the significant increase in low-frequency (e.g., delta)...

Figure B.5: (Continued from previous page) power that is a hallmark of anesthetic-induced unconsciousness. * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$, $n=10$ for propofol NHP 1, $n=11$ for propofol NHP 2, and $n=8$ for each NHP in ketamine and dexmedetomidine, using two-sided Wilcoxon signed-rank test. (E) Dynamic stability analysis repeated after removing the delta (0.1–4 Hz) frequency range. The analysis from Figure 2B was re-run on data that was band-stop filtered to remove the primary delta band. The results show that the significant destabilization during the Anesthesia epoch (purple) persists even after this dominant low-frequency component is removed, indicating that the effect is not merely an artifact of delta oscillations but is present in a multitude of low-frequency responses. * indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$, $n=10$ for propofol NHP 1, $n=11$ for propofol NHP 2, and $n=8$ for each NHP in ketamine and dexmedetomidine, using one-sided Wilcoxon signed-rank test.

$$c_i^{(j)}(\tau) = \frac{1}{T} \sum_{t=1}^{T-\tau} \left(x_{i,t}^{(j)} - \mu_i^{(j)} \right) \left(x_{i,t+\tau}^{(j)} - \mu_i^{(j)} \right)$$

The autocorrelation function for electrode i on trial j at time lag τ is then obtained as

$$R_i^{(j)}(\tau) = \frac{c_i^{(j)}(\tau)}{\left(\sigma_i^{(j)} \right)^2}$$

This computation yields $R_i^{(j)}(0) = 1$ by definition. For each session and epoch, we compute the mean across trials and electrodes as

$$R(\tau) = \frac{1}{MN} \sum_{j=1}^M \sum_{i=1}^N R_i^{(j)}(\tau)$$

We can then use these trial- and electrode- averaged values within each epoch to compute the mean across sessions. For propofol, only the maintenance dose of anesthesia was used for sensory response analysis. Statistical analyses were performed across sessions as well. In our implementation, autocorrelations were computed using the statsmodels package.⁴⁵⁵

To quantify the relationship between dynamic stability and the "memory" of the system's response to perturbation, we computed a time-resolved metric of autocorrelation. For each 15 second window used in the stability analysis, we computed the ACF of each electrode up to a lag of 5 seconds (i.e., 5000 milliseconds, or 5000 lags). We then compute the integral of the ACF (for each electrode's ACF) in two ways. First, we integrate it up to 50 milliseconds. Second, we integrate it up to half-maximum. That is, we integrate up to the lag value at which the ACF crosses 0.5. If the ACF does not cross 0.5 in the 5 seconds analyzed, we integrate up to a time lag of 5 seconds. We then average these values across electrodes in each window. Then, for each session, we correlate the sequence of Mean Instability values (in units of inverse seconds) with the sequence of mean integrated ACF values.

B.2.16 Power-spectral analyses

For the analysis in Figure 2.4 and Figure B.5C,D, within each 15 second window, spectral analysis was conducted using NumPy’s rfft function.⁴⁵² We computed the mean spectral power in each frequency band. The frequency bands we considered were delta (0.1 - 4 Hz), theta (4 - 8 Hz), alpha (8 - 12 Hz), beta (12 - 30 Hz), and gamma (30 - 100 Hz). For Figure 2.4, we then correlated the time-series of mean spectral band power at each window (and in each band) with the time series of mean inverse timescales in each window computed using DeLASE. We used SciPy’s pearsonr function to compute the Pearson correlation.⁴³⁹ To filter out the delta frequency range for Figure B.5E, we used a bi-directional Butterworth band-stop filter of order 2 implemented using SciPy.⁴³⁹

B.3 Quantification and Statistical Analysis

For all data plots, “*” denotes a p-value less than 0.05, “**” denotes a p-value less than 0.01, and “***” denotes a p-value less than 0.001. We note that experiments for each anesthetic were performed on two subjects, as is standard in the field.²⁷² Our statistical power is primarily derived not from the number of animals, but from the large number of independent recording sessions collected from each animal, which provides a strong basis for within-subject analysis. Statistical tests were performed on a per-animal basis, and in all cases n was equal to the # of sessions. There were 10 sessions for NHP 1 under propofol, 11 sessions for NHP 2 under propofol, and 8 sessions for both NHP 3 and NHP 4 under both ketamine and dexmedetomidine. Errors are reported as standard error of the mean (SEM).

B.3.1 Testing session independence

To rigorously test the assumption that sessions are independent, we fit a linear mixed-effects (LME) model to test if there was any systematic effect of session order on our stability measurements. Specifically, we tested whether adding the 'sequential session number' as a predictor improved the model's ability to predict instability, beyond the effect of the state (Awake vs. Anesthesia vs. Emergence) and a random intercept at the level of the session (to account for baseline). The results of this analysis support our independence assumption. For all NHP/anesthetic agent combinations, a model including the session number did not provide a statistically better fit than a model without it (as assessed by a likelihood ratio test). Even when evaluating the model including the session number information, the coefficient for the session number predictor was not significant in all cases, indicating no systematic drift in changes in stability from one session to the next. Results are reported in Table B.2.

Table B.2: Linear mixed-effects model results assessing instability changes and session independence. This table presents the fixed-effects coefficients and p-values from a linear mixed-effects (LME) model predicting mean instability (s^{-1}). The Anesthesia and Emergence columns show the estimated change in instability relative to the Awake baseline for each subject and agent; positive coefficients indicate destabilization. The Session Order columns show the effect of the sequential session number included as a predictor. The non-significant p-values ($p > 0.05$) for both the Session Order coefficient and the Session Order Model relative to a model with no session order (via log-likelihood ratio test) across all subjects and agents confirm that there was no systematic drift in stability measurements over time. This supports the assumption that recording sessions were statistically independent.

Agent	Propofol		Ketamine		Dexmedetomidine	
Subject	NHP 1	NHP 2	NHP 3	NHP 4	NHP 3	NHP 4
Anesthesia Coeff.	12.70	14.63	5.43	1.60	15.34	14.30
Anesthesia p-value	4.38e-237	4.32e-113	1.91e-13	0.002	9.11e-18	1.13e-62
Emergence Coeff.	10.23	6.09	-1.82	-1.58	5.63	5.06
Emergence p-value	1.95e-154	4.83e-21	0.013	0.002	0.002	3.4e-09
Session Order Coeff.	0.10	0.07	-0.48	-0.37	-0.82	0.04
Session Order p-value	0.605	0.780	0.433	0.516	0.148	0.964
Session Order Model p-value	0.566	0.757	0.377	0.461	0.122	0.958

B.3.2 Changes in stability in neural dynamics

In each NHP, we computed the real parts (i.e. inverse timescales of response) of the top 10% of characteristic roots from every window using DeLASE. For Figure 2.2A the data was interpolated onto a grid of time points spaced by 30 seconds for the purposes of aligning the data across sessions. We performed a two-sided Wilcoxon signed-rank test comparing the mean inverse timescales at that time point to the per-session mean inverse timescales from the awake state. For Figure 2.2B, Figure B.3 and Figure B.1, and for the in-text statistics in the section “All three anesthetic agents destabilize neural dynamics,” we computed the mean inverse timescale within each epoch of session, with the epochs defined as above. Statistical tests to compare the inverse timescales between epochs were then performed using a one-sided Wilcoxon signed-rank test, with n again being the number of sessions for each NHP/agent. For Figure B.4C-E, we computed the mean inverse timescale in responsive and unresponsive windows, as defined above, and then performed a statistical test using a one-sided Wilcoxon signed-rank test, with n again being the number of sessions for each NHP/agent.

B.3.3 Effect sizes for changes in stability

To quantify the magnitude of the change in dynamic stability between epochs (e.g., Anesthesia vs. Awake), we computed a robust estimate of the location shift for paired data. For each subject and anesthetic condition, we calculated the paired differences in mean instability between the two epochs across for each session. That is, for a given animal and agent combination, let there be K

sessions. Then, for each session let ΔS_k be the change in mean instability (top 10%) from the Awake to Anesthesia epoch (i.e. Anesthesia – Awake instability) for $k \in \{1, \dots, K\}$. We estimated the effect size as the median of these paired differences, which is the effectively the Hodges-Lehmann Estimator for paired data.^{469,470}

B.3.4 Estimating confidence intervals

To estimate the precision of our stability difference estimates, we computed 95% confidence intervals (CIs) using a non-parametric bootstrap procedure. For each pairwise comparison, we generated 10,000 bootstrap samples by resampling the observed set of paired differences with replacement. For each resampled dataset, we re-calculated the median difference. The 95% confidence interval was then derived using the percentile method, defined as the interval between the 2.5th and 97.5th percentiles of the bootstrap distribution.

B.3.5 Post-hoc statistical power analysis

To verify the statistical power of our results given the sample size (number of sessions), we performed a simulation-based post-hoc statistical power analysis. We treated the observed paired differences in instability as the empirical population distribution. We then ran 1,000 simulations where, in each iteration, we generated a new synthetic dataset of the same size (K sessions) by resampling with replacement from the observed differences. We performed a Wilcoxon signed-rank test on each synthetic dataset with significance level 0.05. The observed power was calculated as the proportion of these 1,000 simulations that yielded a statistically significant result. Results are reported in Table B.3.

Table B.3: Post-hoc statistical power analysis results. This table displays the estimated statistical power for the primary stability comparisons (Anesthesia vs. Awake and Emergence vs. Awake) for each subject and anesthetic agent. Power estimates were calculated using a simulation-based approach with 1,000 bootstrap simulations per comparison, treating the observed session-level differences as the empirical population.

Agent	NHP	Comparison	Power
Propofol	NHP 1	Anesthesia vs. Awake	1.000
Propofol	NHP 2	Anesthesia vs. Awake	1.000
Ketamine	NHP 3	Anesthesia vs. Awake	1.000
Ketamine	NHP 3	Emergence vs. Awake	0.899
Ketamine	NHP 4	Anesthesia vs. Awake	0.630
Ketamine	NHP 4	Emergence vs. Awake	1.000
Dexmedetomidine	NHP 3	Anesthesia vs. Awake	1.000
Dexmedetomidine	NHP 4	Anesthesia vs. Awake	1.000

B.3.6 Comparing autocorrelation values

For the statistical tests in Figure 2.3B, we computed the autocorrelation of the stimulus-induced response in each electrode and averaged across all electrodes and trials within each epoch of each session, as described above. The autocorrelation values at each lag were then compared using a two-sided Wilcoxon signed-rank test, with n again being the number of sessions for each NHP/agent.

B.3.7 Autocorrelation-instability correlations

As mentioned, for each session, we computed two time-series: (1) the mean instability (inverse timescale) derived from DeLASE, and (2) the integrated autocorrelation function (ACF) of the sensory response (integrated either to 50ms or to its half-maximum). We then calculated the Pearson correlation coefficient between these two time-series for each session. This procedure yielded a distribution of correlation coefficients for each subject and anesthetic agent (one coefficient per session). To test whether instability was significantly correlated with the system's memory, we performed a two-sided Wilcoxon signed-rank test on these coefficients against a null hypothesis of zero correlation. We report the median correlation and the 95% confidence interval of the median, estimated via bootstrapping, alongside the p-values in Table B.4.

Table B.4: Statistical results for the correlation between dynamic instability and sensory response autocorrelation. This table presents the detailed statistics for the correlations summarized in Figure 2.3C. It reports the relationship between the time-resolved instability estimates (from DeLASE) and two metrics of the sensory response autocorrelation function (ACF): the integrated ACF up to its half-maximum and the integrated ACF up to 50 ms. Median Correlation represents the median Pearson correlation coefficient computed across sessions. 95% CI denotes the 95% confidence interval of the correlation coefficients. p-values were computed using a two-sided Wilcoxon signed-rank test against a null hypothesis of zero correlation. The consistently positive correlations confirm that periods of high instability are associated with slower decay (longer memory) in the sensory response.

Agent	NHP	Metric	p-value	Med. Corr.	95% CI
Propofol	NHP 1	Integrated ACF to Half-Max	0.001953	0.81	[0.55, 0.92]
Propofol	NHP 1	Integrated ACF to 50 ms	0.001953	0.95	[0.95, 0.97]
Propofol	NHP 2	Integrated ACF to Half-Max	0.001953	0.24	[0.15, 0.36]
Propofol	NHP 2	Integrated ACF to 50 ms	0.000977	0.68	[0.52, 0.77]
Ketamine	NHP 3	Integrated ACF to Half-Max	0.007812	0.92	[0.29, 0.95]
Ketamine	NHP 3	Integrated ACF to 50 ms	0.007812	0.85	[0.72, 0.89]
Ketamine	NHP 4	Integrated ACF to Half-Max	0.007812	0.38	[0.22, 0.66]
Ketamine	NHP 4	Integrated ACF to 50 ms	0.007812	0.64	[0.44, 0.77]
Dexmed.	NHP 3	Integrated ACF to Half-Max	0.007812	0.87	[0.82, 0.92]
Dexmed.	NHP 3	Integrated ACF to 50 ms	0.007812	0.83	[0.79, 0.90]
Dexmed.	NHP 4	Integrated ACF to Half-Max	0.015625	0.53	[0.22, 0.81]
Dexmed.	NHP 4	Integrated ACF to 50 ms	0.007812	0.79	[0.74, 0.87]

B.3.8 Band power-instability correlations

For each session, we computed the mean inverse timescale (as described above) in 15 second windows that were tiled across the session. Then, across each session we computed the Pearson correlation of the time series of spectral band power and mean inverse timescales as described above. Then, for each band, we performed a two-sided Wilcoxon signed-rank test comparing the set of Pearson correlations obtained from each session for each NHP to the null-hypothesis of a correlation value of zero correlation. We report the median correlation and the 95% confidence interval of the median, estimated via bootstrapping, alongside the p-values in Table B.5.

Table B.5: Statistical results for the correlation between dynamic instability and spectral power. This table presents the detailed statistics for the correlations summarized in Figure 2.4. It reports the relationship between the time-resolved instability estimates (from DeLASE) and the mean spectral power in five frequency bands: delta (0.1–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–30 Hz), and gamma (30–100 Hz). Median Correlation represents the median Pearson correlation coefficient computed across sessions between the instability time-series and the band power time-series. “95% CI” denotes the 95% confidence interval of these correlation coefficients. p-values were computed using a two-sided Wilcoxon signed-rank test against a null hypothesis of zero correlation. The results quantitatively confirm that destabilization is positively correlated with increased power in the low-frequency (delta, theta, alpha) bands across all three anesthetics.

Agent	NHP	Dose	Band	p-value	Med. Corr.	95% CI
Propofol	NHP1	High	Delta	0.001953	0.8471	[0.7111, 0.8951]
		High	Theta	0.001953	0.5301	[0.4759, 0.6101]
		High	Alpha	0.001953	0.4170	[0.3719, 0.4999]
		High	Beta	0.001953	-0.1710	[-0.2395, -0.1347]
		High	Gamma	0.001953	-0.3839	[-0.4264, -0.3598]
	NHP2	High	Delta	0.000977	0.8592	[0.8452, 0.8823]
		High	Theta	0.000977	0.6979	[0.6483, 0.7946]
		High	Alpha	0.000977	0.4254	[0.3746, 0.4892]
		High	Beta	0.000977	-0.2869	[-0.3209, -0.2204]
		High	Gamma	0.002930	-0.0611	[-0.1570, -0.0479]
Ketamine	NHP3	High	Delta	0.007812	0.9172	[0.8813, 0.9263]
		High	Theta	0.007812	0.9208	[0.8809, 0.9459]
		High	Alpha	0.007812	0.8950	[0.8505, 0.9196]
		High	Beta	0.007812	0.7221	[0.5975, 0.7512]
		High	Gamma	0.007812	0.6870	[0.5099, 0.7355]
	NHP4	High	Delta	0.007812	0.3491	[0.3061, 0.5826]
		High	Theta	0.007812	0.5065	[0.0606, 0.7333]
		High	Alpha	0.023438	0.3794	[0.0400, 0.6273]
		High	Beta	0.039062	0.3063	[-0.0081, 0.4385]
		High	Gamma	0.023438	0.3404	[0.0414, 0.5899]
Dexmed.	NHP3	High	Delta	0.007812	0.8462	[0.8403, 0.8926]

Continued on next page

Table B.5 – continued from previous page

Agent	NHP	Dose	Band	p-value	Med. Corr.	95% CI
		High	Theta	0.007812	0.9185	[0.9015, 0.9277]
		High	Alpha	0.007812	0.5557	[0.5155, 0.6455]
		High	Beta	0.546875	0.0469	[-0.0927, 0.0839]
		High	Gamma	0.007812	-0.7224	[-0.7637, -0.5499]
	NHP4	High	Delta	0.007812	0.8006	[0.7396, 0.8723]
		High	Theta	0.007812	0.9067	[0.8005, 0.9238]
		High	Alpha	0.007812	0.6759	[0.4644, 0.7296]
		High	Beta	0.312500	-0.0497	[-0.1031, 0.0440]
		High	Gamma	0.007812	-0.8589	[-0.9389, -0.7354]

B.3.9 Multiple comparison in time-resolved analyses

To statistically test for changes in stability over time (Figure 2.2A) and across areas (Figure B.3E), we compared the instability metrics during anesthesia to those during the awake state at multiple time points. For the time-resolved analysis in Figure 2.2A and Figure B.3E, instability values were interpolated onto a common grid of time points spaced by 30 seconds. At each time point, we performed a two-sided Wilcoxon signed-rank test comparing the distribution of instability values across sessions to the session-specific awake baseline. Because this procedure involved repeated statistical testing across many time points, we corrected for multiple comparisons to control the False Discovery Rate (FDR). We applied the Benjamini-Hochberg procedure to the set of p-values obtained from the time-resolved tests. We used a false discovery rate (FDR) threshold of 0.05. This correction ensures that the reported periods of significant destabilization are robust and not due to the inflated Type I error rate associated with multiple sequential hypothesis tests.

B.3.10 Comparing mean band power

For the analysis in Figure B.5, we computed the mean band power as described above and averaged it within each epoch of each session (Awake, Anesthesia, Emergence). We then compared the epoch-wise band powers between sessions using a two-sided Wilcoxon signed-rank test, with n again being the number of sessions for each NHP/agent.

Appendix C

Methods: Characterizing control between interacting subsystems

C.1 JacobianODE technical details

C.1.1 Parameterizing the time derivative via the Jacobian

Recall that, for the time derivative function $\mathbf{f}$, we have that the path integral is independent of the choice of path:

$$\mathbf{f}(\mathbf{x}(t_f)) - \mathbf{f}(\mathbf{x}(t_i)) = \int_{\mathcal{C}} \mathbf{J} ds = \int_{t_i}^{t_f} \mathbf{J}(\mathbf{c}(r)) \mathbf{c}'(r) dr, \quad (\text{C.1})$$

where $\mathcal{C}$ is a piecewise smooth curve in $\mathbb{R}^n$ and $\mathbf{c} : [t_i, t_f] \rightarrow \mathcal{C}$ is a parameterization of $\mathcal{C}$ with $\mathbf{c}(t_i) = \mathbf{x}(t_i)$ and $\mathbf{c}(t_f) = \mathbf{x}(t_f)$ (also see equation 3.2). Path integration notably provides only differences between $\mathbf{f}$ at distinct points. Thus, knowing the value of $\mathbf{f}$ at one point is needed for trajectory reconstruction, which we do not assume. To address this, we note it is possible to represent $\mathbf{f}$ solely through the Jacobian at a point $\mathbf{x}(t)$ as

$$\mathbf{f}(\mathbf{x}(t)) = \frac{G(t_0, t; \mathbf{J}, \mathbf{c}) + \mathbf{x}(t) - \mathbf{x}(t_0)}{t - t_0}, \quad (\text{C.2})$$

where

$$G(t_0, t; \mathbf{J}, \mathbf{c}) = \int_{t_0}^t \int_s^t \mathbf{J}(\mathbf{c}_{s,t}(r)) \mathbf{c}'_{s,t}(r) dr ds \quad (\text{C.3})$$

and we have abbreviated $\mathbf{c}_{s,t}(r) = \mathbf{c}(r; s, t, \mathbf{x}(s), \mathbf{x}(t))$, a piecewise smooth curve on $[s, t]$ parameterized by r and beginning and ending at $\mathbf{x}(s)$ and $\mathbf{x}(t)$ respectively. Intuitively, we can circumvent the need to know $\mathbf{f}$ by recognizing that integrating $\mathbf{f}$ between time points will produce the difference in the system states at those time points, which are known to us. With this formalism, we avoid the need to represent $\mathbf{f}$ directly, thereby enabling all gradients to backpropagate through the Jacobian network. The proof of this formalism is presented below.

Proposition 1 (Jacobian-parameterized ODEs). *Let $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t))$ and let $\mathbf{J}_{\mathbf{f}}(\mathbf{x}(t)) = \mathbf{J}(\mathbf{x}(t)) = \frac{\partial}{\partial \mathbf{x}} \mathbf{f}(\mathbf{x}(t))$. Then given times t_0, t we can express $\mathbf{f}$ parameterized by the Jacobian as*

$$\mathbf{f}(\mathbf{x}(t)) = \frac{G(t_0, t; \mathbf{J}, \mathbf{c}) + \mathbf{x}(t) - \mathbf{x}(t_0)}{t - t_0},$$

where

$$G(t_0, t; \mathbf{J}, \mathbf{c}) = \int_{t_0}^t \int_s^t \mathbf{J}(\mathbf{c}_{s,t}(r)) \mathbf{c}'_{s,t}(r) dr ds$$

and we have abbreviated $\mathbf{c}_{s,t}(r) = \mathbf{c}(r; s, t, \mathbf{x}(s), \mathbf{x}(t))$, a piecewise smooth curve on $[s, t]$ parameterized by r and beginning and ending at $\mathbf{x}(s)$ and $\mathbf{x}(t)$ respectively.

Proof. For given times t , t_0 , and s , from the fundamental theorem of calculus, we have that

$$\mathbf{x}(t) - \mathbf{x}(t_0) = \int_{t_0}^t \mathbf{f}(\mathbf{x}(s)) ds$$

and

$$\mathbf{f}(\mathbf{x}(t)) - \mathbf{f}(\mathbf{x}(s)) = \int_s^t \mathbf{J}(\mathbf{c}_{s,t}(r)) \mathbf{c}'_{s,t}(r) dr.$$

Letting

$$H(s, t) = \int_s^t \mathbf{J}(\mathbf{c}_{s,t}(r)) \mathbf{c}'_{s,t}(r) dr = \mathbf{f}(\mathbf{x}(t)) - \mathbf{f}(\mathbf{x}(s)),$$

we then have that

$$\begin{aligned} \int_{t_0}^t H(s, t) ds &= \int_{t_0}^t \mathbf{f}(\mathbf{x}(t)) - \mathbf{f}(\mathbf{x}(s)) ds \\ &= \int_{t_0}^t \mathbf{f}(\mathbf{x}(t)) ds - \int_{t_0}^t \mathbf{f}(\mathbf{x}(s)) ds \\ &= (t - t_0) \mathbf{f}(\mathbf{x}(t)) - (\mathbf{x}(t) - \mathbf{x}(t_0)). \end{aligned}$$

Letting now

$$G(t_0, t; \mathbf{J}, \mathbf{c}) = \int_{t_0}^t H(s, t) ds = \int_{t_0}^t \int_s^t \mathbf{J}(\mathbf{c}_{s,t}(r)) \mathbf{c}'_{s,t}(r) dr ds,$$

we can see that

$$G(t_0, t; \mathbf{J}, \mathbf{c}) = (t - t_0) \mathbf{f}(\mathbf{x}(t)) - (\mathbf{x}(t) - \mathbf{x}(t_0)),$$

and thus

$$\mathbf{f}(\mathbf{x}(t)) = \frac{G(t_0, t; \mathbf{J}, \mathbf{c}) + \mathbf{x}(t) - \mathbf{x}(t_0)}{t - t_0}.$$

□

C.1.2 Path integrating to generate predictions

Given an initial observed trajectory $\mathbf{x}(t_0 + k\Delta t)$, $k = 0, \dots, b$ of length at least two ($b \geq 1$), we compute an estimate of $\hat{\mathbf{f}}(\mathbf{x}(t_0 + b\Delta t))$ of $\mathbf{f}(\mathbf{x}(t_0 + b\Delta t))$ by replacing $\mathbf{J}$ with $\hat{\mathbf{J}}^\theta$ in equation C.2. Specifically, we compute an estimate of $\hat{\mathbf{f}}(\mathbf{x}(t_b))$ of $\mathbf{f}(\mathbf{x}(t_b))$ as

$$\hat{\mathbf{f}}(\mathbf{x}(t_b)) = \frac{G(t_0, t_b; \hat{\mathbf{J}}^\theta, \mathbf{c}) + \mathbf{x}(t_b) - \mathbf{x}(t_0)}{t_b - t_0}, \quad (\text{C.4})$$

In practice, we compute this estimate by constructing a cubic spline on the initial observed trajectory using 15 points (i.e., $b = 14$). Given that the computation of G involves a double integral - one over states and over time - using a spline is computationally advantageous. This is because the path integral (and all intermediate steps) can be quickly computed along the full spline, with the result of each intermediate step along the path then being summed to approximate the time integral. The integral is computed by interpolating 4 points for every gap between observed points, resulting in a discretization of 58 points along the spline. Integrals are computed using the trapezoid method from `torchquad`.³⁵⁹

Once we have constructed our estimate of $\hat{\mathbf{f}}(\mathbf{x}(t_b))$ we can estimate $\mathbf{f}$ at any other point $\mathbf{x}(t)$ as

$$\hat{\mathbf{f}}(\mathbf{x}(t)) = \begin{cases} H(t_b, t) + \hat{\mathbf{f}}(\mathbf{x}(t_b)), & \text{if } t_b < t \\ \hat{\mathbf{f}}(\mathbf{x}(t_b)) - H(t, t_b), & \text{if } t_b > t \\ \mathbf{f}(\mathbf{x}(t_b)) & \text{if } t_b = t \end{cases}$$

where by convention we integrate forwards in time and H is the path integral defined above, with $\hat{\mathbf{J}}$ in place of $\mathbf{J}$. In practice, for the integration path $\mathbf{c}(r; s, t, \mathbf{x}(s), \mathbf{x}(t))$, we construct a line from $\mathbf{x}(s)$ to $\mathbf{x}(t)$ as

$$\mathbf{c}(r; s, t, \mathbf{x}(s), \mathbf{x}(t)) = \left(1 - \frac{r-s}{t-s}\right) \mathbf{x}(s) + \frac{r-s}{t-s} \mathbf{x}(t)$$

to maintain the interpretability of having r in the range $[s, t]$ however it can be easily seen that setting $r' = \frac{r-s}{t-s}$ we recover the familiar line

$$\mathbf{c}(r') = (1 - r')\mathbf{x}(s) + r'\mathbf{x}(t)$$

with r' taking values on $[0, 1]$. Line integrals are computed with 20 discretization steps using the trapezoid method from `torchquad`.³⁵⁹

Using $\hat{\mathbf{f}}(\mathbf{x}(t))$, we can then generate predictions as

$$\hat{\mathbf{x}}(t + \Delta t) = \mathbf{x}(t) + \int_t^{t+\Delta t} \hat{\mathbf{f}}(\mathbf{x}(\tau)) d\tau \quad (\text{C.5})$$

where the integral can be computed by a standard ODE solver (we used the RK4 method from `torchdiffeq` with default values for the relative and absolute tolerance).

We can then compute the trajectory reconstruction loss as

$$\mathcal{L}_{\text{traj}}(\theta; \mathbf{x}) = \sum_{k=b+1}^{T-1} d(\mathbf{x}(t_0 + k\Delta t), \hat{\mathbf{x}}(t_0 + k\Delta t)) \quad (\text{C.6})$$

where d is a distance measure between trajectories (e.g. mean squared error).

C.1.3 Loop closure loss

While the space of Jacobian functions $\hat{\mathbf{J}}^\theta$ that solve the dynamics problem may be quite large, we are interested only in functions $\hat{\mathbf{J}}^\theta$ that *both* solve the dynamics problem and generate matrices with rows that are conservative vector fields. To effectively restrict our optimization to this space, we employ a self-supervised loss building on the loss introduced by Iyer et al. [355]. This loss imposes a conservative constraint on the rows of the Jacobian matrix. Specifically, for any piecewise smooth curve $\mathcal{C}_{\text{loop}}$ starting and ending at the same state $\mathbf{x}_0$, we have that $\left\| \int_{\mathcal{C}_{\text{loop}}} \mathbf{J} ds \right\|_2 = 0$.

Thus, given n_{loops} sets of any L (not necessarily sequential) points $\{\mathbf{x}^{(l)}(t_1), \mathbf{x}^{(l)}(t_2), \dots, \mathbf{x}^{(l)}(t_L)\}$ we can form a loop $\mathcal{C}_{\text{loop}}^{(l)}$ consisting of the sequence lines from $\mathbf{x}^{(l)}(t_i)$ to $\mathbf{x}^{(l)}(t_{i+1})$, $i = 1, \dots, L - 1$, followed by a line from $\mathbf{x}^{(l)}(t_L)$ to $\mathbf{x}^{(l)}(t_1)$. We then define the following regularization loss to enforce this conservation constraint:

$$\mathcal{L}_{\text{loop}}(\theta; \mathbf{x}) = \frac{1}{n_{\text{loops}}} \sum_{l=1}^{n_{\text{loops}}} \frac{1}{n} \left\| \int_{\mathcal{C}_{\text{loop}}^{(l)}} \hat{\mathbf{J}}^\theta ds \right\|_2^2 \quad (\text{C.7})$$

We note that while other choices for loops would have been possible (for instance, forward and backward passes along observed trajectories, or arbitrary circles beginning and ending at the same point), the approach presented here has several advantages (as discussed in Section 3.4.3). Namely, it balances computational tractability with uniform sampling of the directions in the tangent space. Enforcing that all the points that comprise the loop are on the data manifold ensures that the loops are integrated in directions that are as informative as possible for estimating the Jacobian of the given dynamical system.

C.2 Control-theoretic analysis details

C.2.1 Gramian computation

We begin with equation 3.4 from Section 3.6, in which we separate the locally-linearized dynamics in the tangent space into separate pairwise inter-subsystem control interactions. These pairwise control interactions are in general of the form

$$\delta \dot{\mathbf{x}}^A(t) = \mathbf{A}(t)\delta \mathbf{x}^A(t) + \mathbf{B}(t)\delta \mathbf{x}^B(t)$$

where $\mathbf{A}$ is the within-subsystem Jacobian, $\mathbf{B}$ is the across subsystem Jacobian and $\mathbf{x}^A$, $\mathbf{x}^B$ are the states of subsystems A and B, respectively. For a given control system, the time-varying reachability Gramian on the interval $[t_0, t_1]$ is defined as

$$\mathbf{W}_r(t_0, t_1) \triangleq \int_{t_0}^{t_1} \Phi(t_1, \tau) \mathbf{B}(\tau) \mathbf{B}^T(\tau) \Phi^T(t_1, \tau) d\tau$$

where Φ denotes the state-transition matrix of the intrinsic dynamics of subsystem A without any input (i.e., $\delta \mathbf{x}^A(t) = \Phi(t, t_0) \delta \mathbf{x}^A(t_0)$).³⁶⁵ Differentiating with respect to t , we obtain $\delta \dot{\mathbf{x}}^A(t) = \frac{\partial}{\partial t} \Phi(t, t_0) \delta \mathbf{x}^A(t_0)$. Noting also that, in the absence of input from subsystem B, we have

$$\delta \dot{\mathbf{x}}^A(t) = \mathbf{A}(t) \delta \mathbf{x}^A(t) = \mathbf{A}(t) \Phi(t, t_0) \delta \mathbf{x}^A(t_0)$$

Thus setting the equations equal to each other and canceling $\delta \mathbf{x}^A(t_0)$ from both sides we obtain

$$\frac{\partial}{\partial t} \Phi(t, t_0) = \mathbf{A}(t) \Phi(t, t_0)$$

Note that $\Phi(t_0, t_0) = \mathbf{I}$. Now, letting $\Gamma(t, \tau) = \Phi(t, \tau) \mathbf{B}(\tau) \mathbf{B}^T(\tau) \Phi^T(t, \tau)$ and differentiating the Gramian expression with respect to the second argument (and using the Leibniz integral rule) yields

$$\begin{aligned} \frac{\partial}{\partial t} \mathbf{W}_r(t_0, t) &= \frac{\partial}{\partial t} \int_{t_0}^t \Gamma(t, \tau) d\tau \\ &= \Gamma(t, t) \frac{\partial}{\partial t}(t) - \Gamma(t, t_0) \frac{\partial}{\partial t}(t_0) + \int_{t_0}^t \frac{\partial}{\partial t} \Gamma(t, \tau) d\tau \\ &= \mathbf{I} \mathbf{B}(t) \mathbf{B}^T(t) \mathbf{I}(1) - 0 + \int_{t_0}^t \frac{\partial}{\partial t} \Gamma(t, \tau) d\tau \end{aligned}$$

and observing

$$\begin{aligned} \frac{\partial}{\partial t} \Gamma(t, \tau) &= \left(\frac{\partial}{\partial t} \Phi(t, t_0) \right) \mathbf{B}(\tau) \mathbf{B}^T(\tau) \Phi^T(t, \tau) + \Phi(t, t_0) \mathbf{B}(\tau) \mathbf{B}^T(\tau) \left(\frac{\partial}{\partial t} \Phi^T(t, \tau) \right)^T \\ &= \mathbf{A}(t) \Phi(t, t_0) \mathbf{B}(\tau) \mathbf{B}^T(\tau) \Phi^T(t, \tau) + \Phi(t, t_0) \mathbf{B}(\tau) \mathbf{B}^T(\tau) \Phi^T(t, \tau) \mathbf{A}^T(t) \\ &= \mathbf{A}(t) \Gamma(t, \tau) + \Gamma(t, \tau) \mathbf{A}^T(t) \end{aligned}$$

we can continue

$$\begin{aligned} \frac{\partial}{\partial t} \mathbf{W}_r(t_0, t) &= \mathbf{B}(t) \mathbf{B}^T(t) + \mathbf{A}(t) \int_{t_0}^t \Gamma(t, \tau) d\tau + \left(\int_{t_0}^t \Gamma(t, \tau) d\tau \right) \mathbf{A}^T(t) \\ &= \mathbf{B}(t) \mathbf{B}^T(t) + \mathbf{A}(t) \mathbf{W}_r(t_0, t) + \mathbf{W}_r(t_0, t) \mathbf{A}^T(t) \end{aligned}$$

which illustrates that the reachability Gramian can be solved for using an ODE integrator (with initial condition $\mathbf{W}_r(t_0, t_0) = \mathbf{0}$).³²² In practice, to compute the reachability Gramians using the trained JacobianODE models, we fit a cubic spline $\mathbf{c}(t)$ to the reference trajectory $\mathbf{x}(t)$ and compute $\mathbf{J}(t) = \mathbf{J}(\mathbf{c}(t))$. We can then parse the Jacobian matrix into its component submatrices and compute the Gramians accordingly.

The reachability Gramian is a symmetric positive semidefinite matrix.³⁶⁵ The optimal cost of driving the system from state $\delta\mathbf{x}_0$ to state $\delta\mathbf{x}_1$ on time interval $[t_0, t_1]$ can be computed as

$$(\delta\mathbf{x}_1 - \Phi(t_1, t_0)\delta\mathbf{x}_0)^T \mathbf{W}_r^{-1}(t_0, t_1) (\delta\mathbf{x}_1 - \Phi(t_1, t_0)\delta\mathbf{x}_0)$$

Thus each eigenvalue of $\mathbf{W}_r$ reflects the ease of control along the corresponding eigenvector.^{365,367} Eigenvectors with larger corresponding eigenvalues will have smaller inverse eigenvalues, and thus scale the cost down along those directions when computing the cost as above.

C.2.2 Extension to non-autonomous dynamics

While we deal with autonomous systems in this work, we note that this construction can easily be extended to include non-autonomous dynamical systems, of the form $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}, t)$ by constructing an augmented state $\tilde{\mathbf{x}} \in \mathbb{R}^{n+1}$ which is simply the concatenation of $\mathbf{x}$ with t . This yields the autonomous dynamics $\dot{\tilde{\mathbf{x}}} = \tilde{\mathbf{f}}(\tilde{\mathbf{x}})$, where $\tilde{\mathbf{f}} : \mathbb{R}^{n+1} \rightarrow \mathbb{R}^{n+1}$ is the concatenation of $\mathbf{f}$ with a function that maps all states to 1.

C.2.3 Extension to more than two subsystems

Consider a nonlinear dynamical system in $\mathbb{R}^n$, defined by $\dot{\mathbf{x}}(t) = \mathbf{f}(\mathbf{x}(t))$. Suppose now that the system is composed of K subsystems, where the time evolution of subsystem k is given by $\mathbf{x}^{(k)}(t) \in \mathbb{R}^{n_k}$. $\mathbf{x}(t)$ is then comprised of a concatenation of the $\mathbf{x}^{(k)}(t)$, with $\sum_{k=1}^K n_k = n$. Given a particular reference trajectory, $\mathbf{x}(t)$, with $\mathbf{J}$ as the Jacobian of $\mathbf{f}$, then the tangent space dynamics around the reference trajectory for subsystem $\alpha \in \{1, \dots, K\}$ are given by

$$\delta\dot{\mathbf{x}}^{(\alpha)}(t) = \sum_{k=1}^K \mathbf{J}^{k \rightarrow \alpha}(\mathbf{x}(t)) \delta\mathbf{x}^{(k)}(t)$$

where $\mathbf{J}^{k \rightarrow \alpha}(\mathbf{x}(t)) \in \mathbb{R}^{n_\alpha \times n_k}$ is the submatrix of $\mathbf{J}(\mathbf{x}(t))$ in which the columns correspond to subsystem k and the rows correspond to subsystem α . Now, for a given subsystem $\beta \in \{1, \dots, K\}$ with $\beta \neq \alpha$, we wish to analyze the ease with which β can control α locally around the reference trajectory, without intervention from other subsystems. Discounting the interventions from other subsystems equates to setting $\delta\mathbf{x}^{(k)}(t) = 0$ for $k \neq \alpha, \beta$, leaving the expression

$$\delta\dot{\mathbf{x}}^{(\alpha)}(t) = \mathbf{J}^{\alpha \rightarrow \alpha}(\mathbf{x}(t)) \delta\mathbf{x}^{(\alpha)}(t) + \mathbf{J}^{\beta \rightarrow \alpha}(\mathbf{x}(t)) \delta\mathbf{x}^{(\beta)}(t)$$

which quantifies the influence β can exert over α in the absence of perturbations from any other subsystem. Using this representation, the reachability Gramian can be computed as described above. Performing this procedure for all pairs of subsystems $\alpha, \beta \in \{1, \dots, K\}$ thus characterizes all pairwise control relationships between subsystems along the reference trajectory.

C.3 Supplementary results

C.3.1 Control and communication capture different phenomena

In the Introduction (Section 3.2) we note that measuring communication between two systems is different than measuring control. To illustrate this, consider two brain areas, A and B, whose dynamics are internally linear. The areas are connected only through a linear feedforward interaction term from area B to area A (Figure C.1A). Concretely, we consider the dynamics:

$$\begin{aligned}\dot{\mathbf{x}}^A(t) &= \mathbf{J}^{A \rightarrow A} \mathbf{x}^A(t) + \mathbf{J}^{B \rightarrow A} \mathbf{x}^B(t) \\ \dot{\mathbf{x}}^B(t) &= \mathbf{J}^{B \rightarrow B} \mathbf{x}^B(t)\end{aligned}$$

Here, the $\mathbf{J}$ matrices are time-invariant. When considering communication between brain areas, one might aim to find the subspaces in which communication between B and A occurs, as well as the messages passed.^{303,336} In this construction, the subspace in which area B communicates with area A is explicitly given by $\mathbf{J}^{B \rightarrow A}$, and thus the message, or input, from area B to area A at any time t is simply given by $\mathbf{J}^{B \rightarrow A} \mathbf{x}(t)$. We pick the dimensions of each region to be $n_A = n_B = 4$ and let the eigenvectors of the (negative definite matrix) $\mathbf{J}^{A \rightarrow A}$ be given by $\mathbf{v}_i, i = 1, 2, 3, 4$. The eigenvectors are numbered in order of decreasing real part of their corresponding eigenvalue. Now, suppose we construct the interaction matrix $\mathbf{J}^{B \rightarrow A}$ in two different ways: If we let (1) $\mathbf{J}_1^{B \rightarrow A} = \mathbf{v}_1 \mathbf{1}^T$ (Figure C.1B, left), then the signal from B is projected onto the most stable mode of region A, whereas if we let (2) $\mathbf{J}_2^{B \rightarrow A} = \mathbf{v}_4 \mathbf{1}^T$ (Figure C.1B, right), the signal is projected onto the least stable mode. While interaction 1 and interaction 2 communicated messages with identical magnitudes, interaction 2 led to much lower cost reachability control (Figure C.1C). This illustrates that control depends not only on the directions along which the areas can communicate, but also on how *aligned* the communication is with the target area's dynamics.

C.3.2 Derivative estimation is not implied by function estimation

An alternative approach to directly estimating the Jacobians would be to learn an approximation $\hat{\mathbf{f}}$ of the function $\mathbf{f}$, and then approximate the Jacobian via automatic differentiation (i.e., an estimate $\hat{\mathbf{J}} = \frac{\partial}{\partial x} \hat{\mathbf{f}}$, as with the NeuralODEs). While this can be effective in certain scenarios, it is not generally the case that approximating a function well will yield a good approximation of its derivative. To illustrate this, we recall an example from Latrémollière, Narayanappa, and Vojtěchovský [350], in which functions $f_n = \frac{1}{n} \cos(nx)$ are used to approximate the function $f(x) = 0$ (Figure C.1D). While these approximations improve with increasing n (i.e., $\lim_{n \rightarrow \infty} f_n = f$), this is not the case for the derivative ($\lim_{n \rightarrow \infty} f'_n = -\sin(nx) \neq f'$). This demonstrates the necessity of learning $\mathbf{J}$ directly (rather than first approximating $\mathbf{f}$), which we also demonstrate empirically. In the context of machine learning, this setting could be interpreted as overfitting.³⁵⁰ As long as function estimates match at the specific points in the training set, how the function fluctuates between these points is not constrained to match the true function. For this reason, we included the Frobenius norm Jacobian

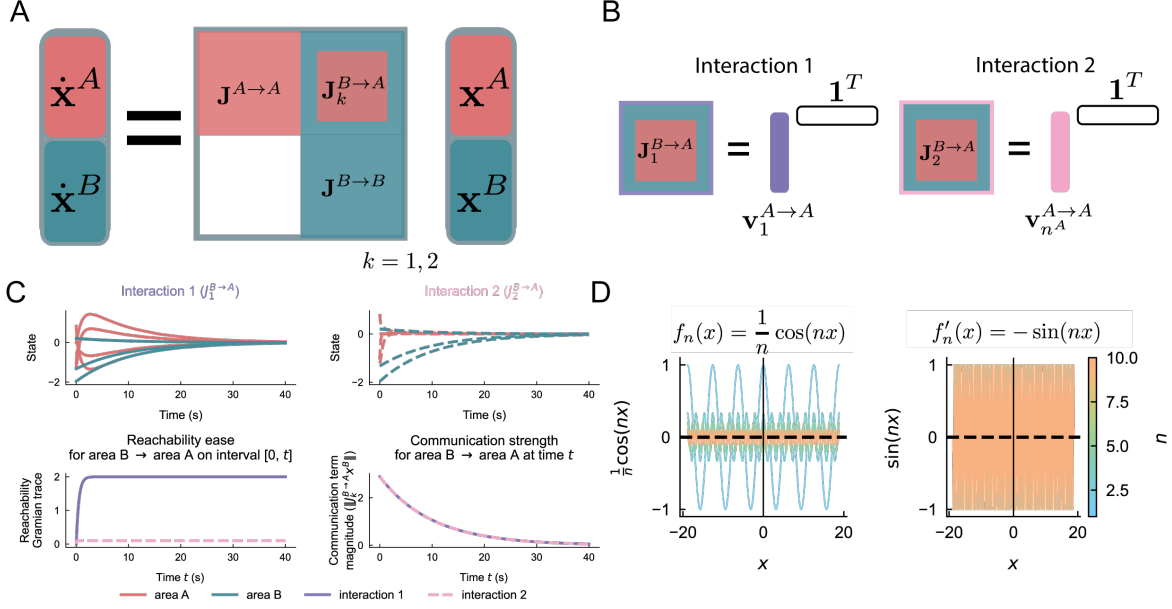


Figure C.1: Communication versus control in linear systems and the challenge of Jacobian estimation. (A) Setup of two linearly connected brain areas, A and B. (B) Interaction matrices projecting signals from area B onto either the most stable (Interaction 1) or least stable (Interaction 2) eigenvectors of area A. (C) Although both interactions communicate identical signal magnitudes, Interaction 2 provides significantly enhanced reachability due to alignment with unstable modes of area A dynamics. (D) An illustrative example demonstrating that accurate approximation of a function (left panel) does not guarantee accurate approximation of its derivative (right panel), emphasizing the necessity of directly estimating the Jacobian.

regularization in our implementation of the NeuralODEs (Appendix C.4.3).

C.3.3 Full benchmark dynamical systems results

We here present the full results for all considered example dynamical systems. 10 time-step trajectory predictions along with Jacobian estimation and estimated Lyapunov spectra are displayed in Figure C.2, along with MSE and R^2 in Table C.1 (trajectory prediction) and Table C.2 (Jacobian estimation). Jacobian estimation errors using the 2-norm are presented in Table C.3.

C.3.4 Ablation studies

To determine the value of the different components of the JacobianODE learning framework, we performed several ablation studies. We chose to evaluate on the Lorenz system and the task-trained RNNs, as these together provide two common settings of chaos and stability, as well as low- and high-dimensional dynamics.

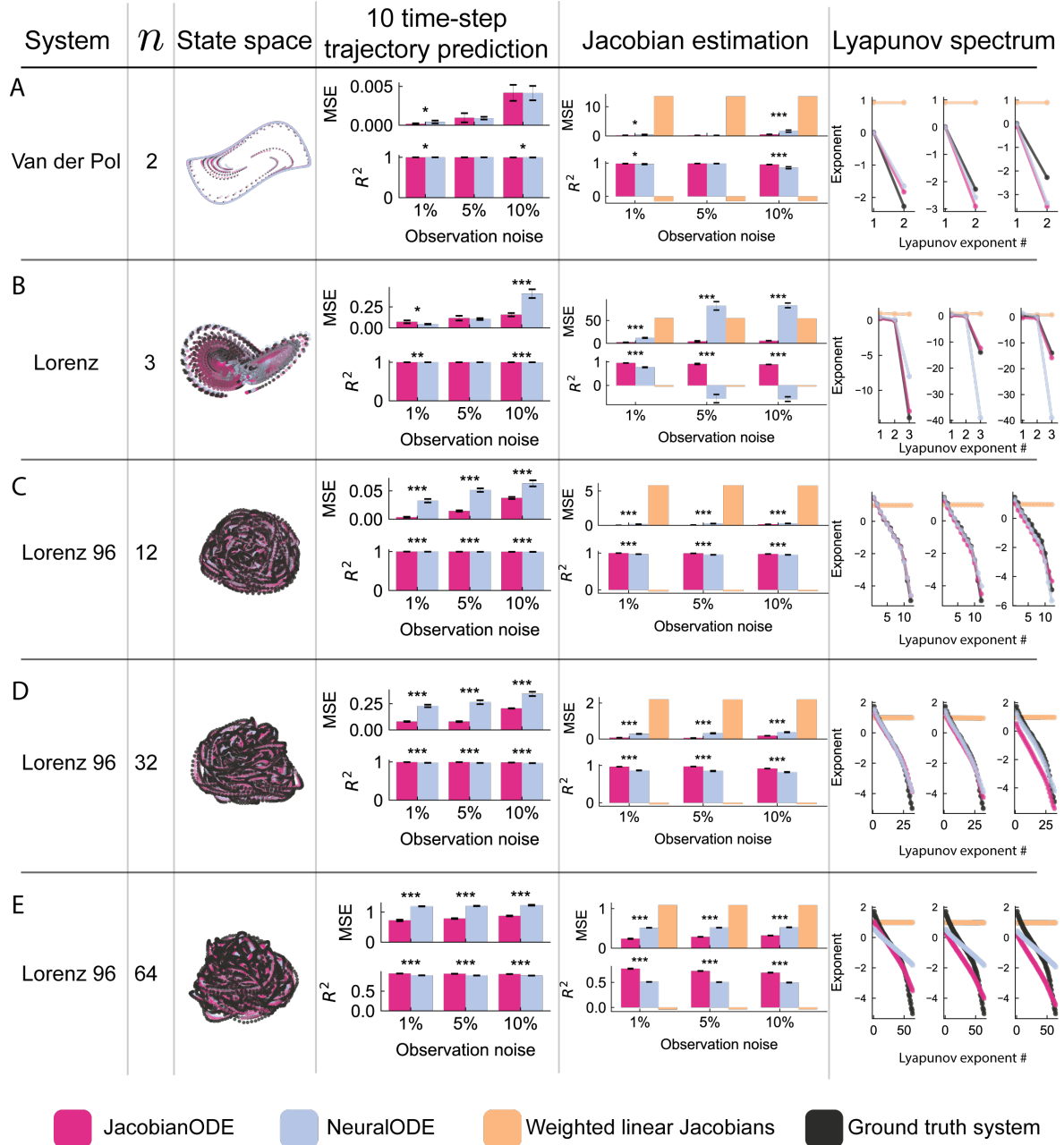


Figure C.2: Full dynamical systems prediction results. State space representations (using spectral embedding for systems that have more than 3 dimensions), 10 time-step trajectory predictions, Jacobian estimation, and Lyapunov spectrum estimates for each of (A) the Van der Pol oscillator, (B) the Lorenz system, and the Lorenz 96 system with (C) 12, (D) 32, and (E) 64 dimensions.

Ablating the Jacobian-parameterized ODEs The JacobianODE framework constructs an estimate of the initial time derivative $\mathbf{f}$ via a double integral of the Jacobian as described in Section 3.4.2. To briefly recall, Jacobian path integration is described as

Table C.1: Trajectory prediction metrics (MSE and R^2) for each system and noise level. Errors are mean $\pm$ standard deviation across five random initializations of the model architectures.

Project	Training noise	MSE		R^2	
		JacobianODE	NeuralODE	JacobianODE	NeuralODE
VanDerPol (2 dim)	1%	0.00017 $\pm$ 0.00007	0.0004 $\pm$ 0.0002	0.99987 $\pm$ 0.00005	0.9996 $\pm$ 0.0002
	5%	0.0010 $\pm$ 0.0006	0.0009 $\pm$ 0.0002	0.9995 $\pm$ 0.0003	0.99949 $\pm$ 0.00009
	10%	0.004 $\pm$ 0.001	0.0041 $\pm$ 0.0009	0.9979 $\pm$ 0.0005	0.9969 $\pm$ 0.0003
Lorenz (3 dim)	1%	0.07 $\pm$ 0.02	0.045 $\pm$ 0.004	0.99979 $\pm$ 0.00004	0.99986 $\pm$ 0.00001
	5%	0.12 $\pm$ 0.03	0.10 $\pm$ 0.01	0.9996 $\pm$ 0.0001	0.99958 $\pm$ 0.00004
	10%	0.16 $\pm$ 0.02	0.41 $\pm$ 0.05	0.99928 $\pm$ 0.00008	0.9981 $\pm$ 0.0003
Lorenz 96 (12 dim)	1%	0.0035 $\pm$ 0.0007	0.033 $\pm$ 0.003	0.99969 $\pm$ 0.00007	0.9969 $\pm$ 0.0003
	5%	0.014 $\pm$ 0.001	0.051 $\pm$ 0.003	0.9985 $\pm$ 0.0001	0.9955 $\pm$ 0.0003
	10%	0.038 $\pm$ 0.002	0.063 $\pm$ 0.005	0.9960 $\pm$ 0.0002	0.9940 $\pm$ 0.0007
Lorenz 96 (32 dim)	1%	0.079 $\pm$ 0.004	0.23 $\pm$ 0.01	0.9939 $\pm$ 0.0003	0.977 $\pm$ 0.002
	5%	0.079 $\pm$ 0.004	0.26 $\pm$ 0.02	0.9939 $\pm$ 0.0003	0.974 $\pm$ 0.002
	10%	0.205 $\pm$ 0.001	0.34 $\pm$ 0.02	0.9837 $\pm$ 0.0001	0.968 $\pm$ 0.002
Lorenz 96 (64 dim)	1%	0.72 $\pm$ 0.03	1.194 $\pm$ 0.005	0.946 $\pm$ 0.002	0.8997 $\pm$ 0.0007
	5%	0.79 $\pm$ 0.01	1.205 $\pm$ 0.007	0.9400 $\pm$ 0.0010	0.8986 $\pm$ 0.0010
	10%	0.87 $\pm$ 0.01	1.23 $\pm$ 0.01	0.9332 $\pm$ 0.0009	0.896 $\pm$ 0.001
Task-trained RNN	1%	0.0042 $\pm$ 0.0006	0.00558 $\pm$ 0.00002	0.9981 $\pm$ 0.0003	0.99762 $\pm$ 0.00001
	5%	0.0043 $\pm$ 0.0006	0.00596 $\pm$ 0.00005	0.9981 $\pm$ 0.0002	0.99747 $\pm$ 0.00002
	10%	0.0089 $\pm$ 0.0001	0.00655 $\pm$ 0.00003	0.99591 $\pm$ 0.00004	0.99721 $\pm$ 0.00001

$$\mathbf{f}(\mathbf{x}(t_f)) - \mathbf{f}(\mathbf{x}(t_i)) = \int_C \mathbf{J} ds = \int_{t_i}^{t_f} \mathbf{J}(\mathbf{c}(r)) \mathbf{c}'(r) dr, \quad (\text{C.8})$$

When we do not have access to the initial derivative and base point $\mathbf{f}(\mathbf{x}(t_i)), \mathbf{x}(t_i)$, we use the following formulation

$$\mathbf{f}(\mathbf{x}(t)) = \frac{G(t_0, t; \mathbf{J}, \mathbf{c}) + \mathbf{x}(t) - \mathbf{x}(t_0)}{t - t_0}, \quad (\text{C.9})$$

where

$$G(t_0, t; \mathbf{J}, \mathbf{c}) = \int_{t_0}^t \int_s^t \mathbf{J}(\mathbf{c}_{s,t}(r)) \mathbf{c}'_{s,t}(r) dr ds \quad (\text{C.10})$$

Alternatively, one could use Equation C.8 and learn $\mathbf{f}(\mathbf{x}(t_i))$ and $\mathbf{x}(t_i)$ as learnable parameters ($\mathbf{x}(t_i)$ is needed as the first point of the path, i.e. $\mathbf{c}(t_i) = \mathbf{x}(t_i)$ inside the integral).³⁵² Here, the path integral between $\mathbf{x}(t_i)$ and $\mathbf{x}(t_f)$ is a linear interpolation, as before. These ablated models were trained on 10 time-step prediction, as with the original JacobianODEs. For these models, we completed a full sweep over the loop closure loss weight λ_{loop} in order to determine the best hyperparameter.

Ablating teacher forcing We trained models without any teacher-forcing. That is, models were able to generate only one-step predictions, without any recursive predictions. Again we did a full hyperparameter sweep to pick λ_{loop} .

Table C.2: Jacobian estimation metrics (MSE and R^2) for each system and noise level. Errors are mean $\pm$ standard deviation across five random initializations of the model architectures.

Project	Training noise	MSE		R^2	
		JacobianODE	NeuralODE	JacobianODE	NeuralODE
VanDerPol (2 dim)	1%	0.18 $\pm$ 0.06	0.4 $\pm$ 0.1	0.985 $\pm$ 0.005	0.97 $\pm$ 0.01
	5%	0.16 $\pm$ 0.02	0.17 $\pm$ 0.03	0.987 $\pm$ 0.002	0.985 $\pm$ 0.002
	10%	0.50 $\pm$ 0.04	1.6 $\pm$ 0.3	0.958 $\pm$ 0.003	0.86 $\pm$ 0.03
Lorenz (3 dim)	1%	2.4 $\pm$ 0.1	12.0 $\pm$ 0.7	0.954 $\pm$ 0.002	0.77 $\pm$ 0.01
	5%	4.6 $\pm$ 1.2	82.0 $\pm$ 9.2	0.91 $\pm$ 0.02	-0.6 $\pm$ 0.2
	10%	5.7 $\pm$ 0.2	83.0 $\pm$ 5.2	0.891 $\pm$ 0.004	-0.58 $\pm$ 0.10
Lorenz 96 (12 dim)	1%	0.012 $\pm$ 0.003	0.18 $\pm$ 0.01	0.9979 $\pm$ 0.0006	0.969 $\pm$ 0.002
	5%	0.053 $\pm$ 0.006	0.27 $\pm$ 0.02	0.991 $\pm$ 0.001	0.953 $\pm$ 0.003
	10%	0.153 $\pm$ 0.008	0.27 $\pm$ 0.01	0.973 $\pm$ 0.001	0.951 $\pm$ 0.002
Lorenz 96 (32 dim)	1%	0.075 $\pm$ 0.004	0.28 $\pm$ 0.01	0.965 $\pm$ 0.002	0.866 $\pm$ 0.006
	5%	0.060 $\pm$ 0.003	0.32 $\pm$ 0.02	0.972 $\pm$ 0.001	0.850 $\pm$ 0.009
	10%	0.180 $\pm$ 0.003	0.38 $\pm$ 0.02	0.915 $\pm$ 0.001	0.820 $\pm$ 0.008
Lorenz 96 (64 dim)	1%	0.238 $\pm$ 0.007	0.515 $\pm$ 0.002	0.773 $\pm$ 0.007	0.509 $\pm$ 0.002
	5%	0.286 $\pm$ 0.003	0.519 $\pm$ 0.003	0.727 $\pm$ 0.003	0.505 $\pm$ 0.003
	10%	0.321 $\pm$ 0.004	0.527 $\pm$ 0.005	0.694 $\pm$ 0.003	0.497 $\pm$ 0.005
Task-trained RNN	1%	2.2 $\pm$ 0.2	5.2777 $\pm$ 0.0009	0.59 $\pm$ 0.03	-0.0008 $\pm$ 0.0002
	5%	1.70 $\pm$ 0.08	5.2900 $\pm$ 0.0001	0.68 $\pm$ 0.01	-0.00315 $\pm$ 0.00002
	10%	1.99 $\pm$ 0.01	5.2889 $\pm$ 0.0001	0.623 $\pm$ 0.002	-0.00295 $\pm$ 0.00002

Ablating loop closure loss We ablated the loop closure loss in two ways. The first was to set $\lambda_{\text{loop}} = 0$ to illustrate what would happen if there were no constraints placed on the learned Jacobians. The second was to instead use the Jacobian Frobenius norm regularization that was used for the NeuralODEs (details are in Appendix C.4.3). We did a full sweep to pick λ_{jac} , the Frobenius norm regularization weight.

Ablation results The performance of the ablated models on Jacobian estimation in the Lorenz system are presented in Table C.4. The original JacobianODE outperforms all ablated models, indicating that all components of the JacobianODE training framework improve the model’s performance in this setting. Ablating the Jacobian-parameterized initial derivative estimate resulted in a slight decrease in the estimation loss. This is potentially because the network could offload some of the responsibility for generating correct trajectory predictions onto the estimated base point $\mathbf{x}(t_i)$ and derivative estimate $\hat{\mathbf{f}}(\mathbf{x}(t_i))$, slightly reducing the necessity of estimating correct Jacobians. Ablating the teacher forcing annealing predictably led to worse Jacobian estimation, as the network no longer has to consider how errors will propagate along the trajectory. The most dramatic increase in error was with the ablation of the loop closure loss. Without this important regularization, the learned Jacobians reproduced the dynamics but were not constrained to be conservative, resulting in poor Jacobian estimation. The inclusion of the Frobenius penalty on the Jacobians mitigated this, although it did not encourage accurate Jacobian estimation to the same degree as the loop closure loss.

We then tested the ablated the models on Jacobian estimation in the task-trained RNN, with results presented in Table C.5. Again, ablating the Jacobian-parameterized derivative estimates, teacher forcing, and loop closure resulted in worse Jacobian estimation. Interestingly, in this setting, the

Table C.3: Mean 2-norm error on Jacobian estimation, $\langle \|\mathbf{J} - \hat{\mathbf{J}}\|_2 \rangle$, for each system and noise level. Errors are reported as mean $\pm$ standard deviation, with mean and standard deviation computed over five random initializations of the model architectures.

Project	Training noise	JacobianODE	NeuralODE	Weighted Linear
VanDerPol (2 dim)	1%	0.7 $\pm$ 0.1	1.0 $\pm$ 0.3	6.05
	5%	0.71 $\pm$ 0.05	0.71 $\pm$ 0.08	6.03
	10%	1.31 $\pm$ 0.05	2.2 $\pm$ 0.2	6.02
Lorenz (3 dim)	1%	3.2 $\pm$ 0.2	8.6 $\pm$ 0.3	17.06
	5%	4.9 $\pm$ 0.9	25.9 $\pm$ 1.5	17.02
	10%	6.0 $\pm$ 0.1	26.4 $\pm$ 0.9	16.95
Lorenz 96 (12 dim)	1%	0.8 $\pm$ 0.1	3.5 $\pm$ 0.2	14.94
	5%	1.75 $\pm$ 0.09	4.1 $\pm$ 0.1	14.93
	10%	2.91 $\pm$ 0.09	4.2 $\pm$ 0.1	14.92
Lorenz 96 (32 dim)	1%	3.75 $\pm$ 0.08	7.8 $\pm$ 0.1	16.29
	5%	3.10 $\pm$ 0.09	8.1 $\pm$ 0.2	16.26
	10%	4.89 $\pm$ 0.05	8.6 $\pm$ 0.1	16.28
Lorenz 96 (64 dim)	1%	8.4 $\pm$ 0.1	12.63 $\pm$ 0.02	16.84
	5%	9.04 $\pm$ 0.07	12.66 $\pm$ 0.03	16.81
	10%	9.42 $\pm$ 0.06	12.72 $\pm$ 0.05	16.82
Task-trained RNN	1%	30.4 $\pm$ 0.4	38.565 $\pm$ 0.006	39.29
	5%	29.4 $\pm$ 0.9	38.5611 $\pm$ 0.0004	38.91
	10%	36.0 $\pm$ 0.1	38.5600 $\pm$ 0.0004	38.70

Table C.4: Mean Frobenius norm error $\langle \|\mathbf{J} - \hat{\mathbf{J}}\|_F \rangle$ for different model ablations on the Lorenz system with 10% observation noise. Errors are reported as mean $\pm$ standard deviation, with statistics computed over 8 test trajectories, each consisting of 1200 points.

Model variant	Frobenius norm error
JacobianODE (original)	6.46 $\pm$ 1.50
With learned base derivative point	7.87 $\pm$ 1.14
No teacher forcing	11.59 $\pm$ 1.30
No loop closure	58.22 $\pm$ 2.00
With Jacobian penalty instead of loop closure	9.59 $\pm$ 2.45

inclusion of a penalty on the Frobenius norm of the Jacobians outperformed the use of the loop closure loss. This could potentially be because the loop closure loss is more difficult to drive to zero in high dimensional systems, or because the loop closure loss is more important in chaotic systems like the Lorenz system considered above. Future work should consider in what contexts each kind of regularization is most beneficial to JacobianODE models.

Table C.5: Mean Frobenius norm error $\langle \|\mathbf{J} - \hat{\mathbf{J}}\|_F \rangle$ for different model ablations on the task-trained RNN with 10% observation noise. Errors are reported as mean $\pm$ standard deviation, with statistics computed over 409 test trajectories, each consisting of the 49 points from the second delay and response epochs.

Model variant	Frobenius norm error
JacobianODE (original)	180.13 ± 1.55
With learned base derivative point	186.28 ± 1.17
No teacher forcing	187.72 ± 1.63
No loop closure	313.31 ± 29.88
With Jacobian penalty instead of loop closure	163.69 ± 4.22

C.3.5 NeuralODEs achieve improved performance at the cost of increased inference time

In the main paper, we implemented both the JacobianODEs and the NeuralODEs as four-layer MLPs, with the four layers having sizes of 256, 1024, 2048, and 2048 respectively. This was done for the fairest architectural comparison between the models, to ensure that both models had the same representational capacity when generating their respective outputs. However, there are many architectural changes that we could make to this setup that impact performance. We hypothesized based on the discussion in Appendix C.3.2 that increasing the hidden layer size of the NeuralODEs would improve Jacobian estimation, as larger models have been known to learn smoother representations. Furthermore, we wondered whether including residual blocks in place of the standard MLP implementation would improve Jacobian estimation.

To test this, we implemented the NeuralODEs as four-layer residual networks and tested three different sizes of hidden layer: 1024, 2048, and 4096. Results are in Table C.6. For nearly all models, these changes yielded only marginal improvements over the original NeuralODE model. Only the model with 4096-dimensional hidden layers under 10% training noise achieves a performance near the *original* JacobianODEs. However, the performance limit is still below that of the JacobianODEs, even with a large increase in the representational capacity of the model. It is furthermore of note that the NeuralODEs were able to significantly improve performance only in the high noise setting. This suggests that high noise is necessary for the model to be forced to learn the response of the system to perturbation. In contrast, JacobianODEs perform similarly across all noise levels, indicating a stronger inductive bias to learn the response of the system to perturbation.

Table C.6: Mean Frobenius norm error $\langle \|\mathbf{J} - \hat{\mathbf{J}}\|_F \rangle$ for different model types on the task-trained RNN data across observation noise levels. Errors are reported as mean $\pm$ standard deviation, with statistics computed over 409 test trajectories.

Model type	Training noise 5%	Training noise 10%	Learnable parameters
JacobianODE (original)	174.1 ± 2.4	180.1 ± 1.5	4.02e+07
NeuralODE (original)	294.0 ± 2.4	293.9 ± 2.4	6.85e+06
NeuralODE (1024 dim. hidden layers)	291.6 ± 2.3	289.8 ± 2.3	3.41e+06
NeuralODE (2048 dim. hidden layers)	288.5 ± 2.3	275.0 ± 2.7	1.31e+07
NeuralODE (4096 dim. hidden layers)	275.8 ± 2.4	184.2 ± 5.6	5.14e+07

While the increased hidden layer size seems to induce smoother hidden representations (as expected), it comes with increasing computational cost. Recall that the NeuralODE models directly parameterize the dynamics as $\dot{\mathbf{x}} = \hat{\mathbf{f}}^\theta(\mathbf{x})$. The Jacobian at state $\mathbf{x}$ is computed by directly backpropagating through the Neural ODE $\hat{\mathbf{f}}^\theta$, thereby significantly increasing compute. On the other hand, the JacobianODEs only require a forward pass. We evaluated the Jacobian inference time of JacobianODE and NeuralODE models with four hidden layers of the same size on an H100 GPU. Each model was timed on inferring the Jacobians of 100 batches of the 128-dimensional task-trained RNN, with each batch containing 16 sequences of length 25. Timings were repeated ten times for each model. Results are plotted in Figure C.3.

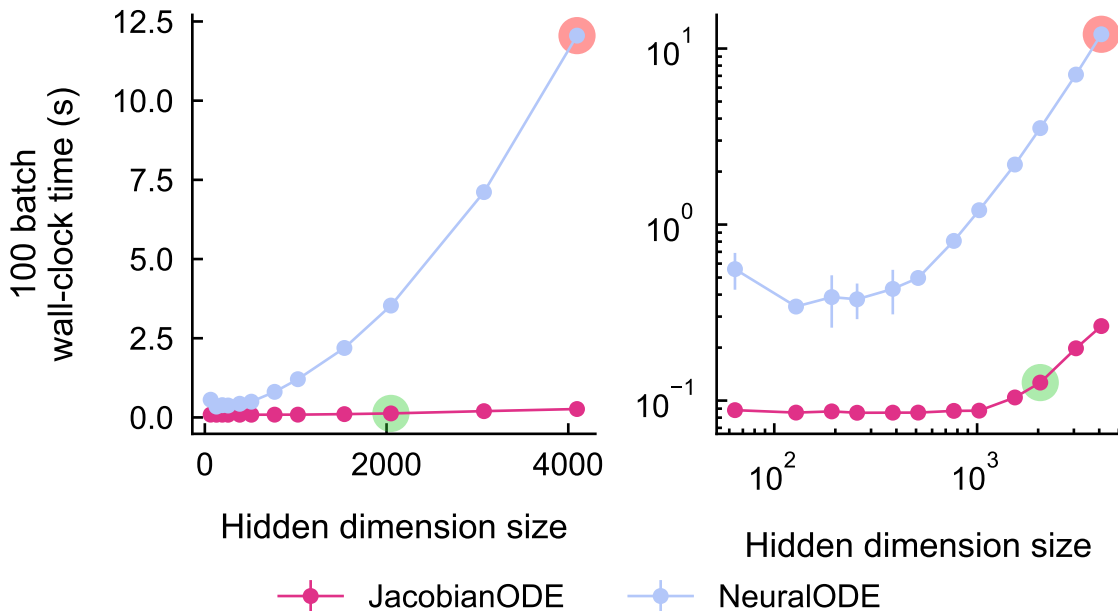


Figure C.3: JacobianODEs achieve highly efficient Jacobian inference. Jacobian inference times computed over ten repetitions of 100 batches (error bars indicate mean $\pm$ standard deviation). Red circles indicate the inference time corresponding to the largest hidden layer dimension of the highest performing NeuralODE model in Table C.6. Green circles indicate the inference time corresponding to the largest hidden layer dimension of the highest performing JacobianODE model in Table C.6. Each plot illustrates the same data but with different x and y scaling.

The JacobianODE achieves much faster inference times than the NeuralODE – approximately two orders of magnitude faster at large hidden dimension sizes. Furthermore, as shown in Table C.6, the JacobianODE with a maximum hidden layer size of 2048 outperforms the NeuralODE with a maximum hidden layer size of 4096 on Jacobian estimation, and does so with orders of magnitude faster inference (Figure C.3, green and red circles, Table C.6). This suggests that while both architectures appear to have inference times that scale approximately exponentially, the JacobianODE achieves more favorable scaling across every hidden layer size we tested. Our analysis therefore illustrates that it is possible to improve the NeuralODEs’ Jacobian estimation with larger models, but the inference time scaling renders these models ill-equipped for important settings such as real-time control and the analysis of high-volume neural data.

C.3.6 Comparison to echo state networks

Echo state networks (ESNs, or reservoir networks) are tools for time-series forecasting and analysis. ESNs are typically implemented as RNNs with fixed, sparse, hidden connectivity. The output weights of the ESNs can then be trained via linear regression to reproduce specific temporal patterns. ESNs have been shown to be able to forecast the Lorenz 63 system for a particular choice of hyperparameters.⁴⁷¹ We note that Pathak et al. [471] implements the same method as Pathak et al. [472]. We implemented the ESNs exact hyperparameters reported in Pathak et al. 2017 for the Lorenz system, which involved a reservoir of 300 nodes. Our implementation of the Lorenz system was simulated with a sampling interval of $\tilde{0}.015$ s, very similar to the 0.02 s used in the paper. Thus, for consistency of comparison with JacobianODEs, we fit ESN models to data sampled every 0.015 s.

To enable Jacobian estimation, we changed one small implementation detail from Pathak et al. [471]. Specifically, we trained the network output weights to predict the third variable z using only the network state $\mathbf{r}(t)$, as opposed to the vector $\tilde{\mathbf{r}}(t)$ for which the node state is squared ($\tilde{r}_i(t) = r_i^2(t)$) for half of the nodes i . This is because squaring the reservoir state complicates Jacobian computation, and was stated by Pathak et al. [471] and Lu et al. [473] to only be necessary for symmetry breaking if only the z variable of the Lorenz system were provided. We trained the ESNs on the same volume of data as the JacobianODEs and NeuralODEs, with 1% observation noise. We swept the regularization parameter β to ensure a good fit. We furthermore generated predictions using a 150 time-step lead-up to ensure the ESNs converged to a good internal representation. The ESNs predicted the Lorenz system well, achieving an MSE of 0.03 on the 10-step prediction task from our paper (on test data). This was very similar to the reported 10-step prediction MSEs of 0.07 and 0.04 for the JacobianODEs and NeuralODEs, respectively.

In Pathak et al. [471], the Lyapunov exponents of the system were computed using the Jacobian of the *interior* dynamics of the ESN, as the recursion necessary to compute the Jacobian can be formulated uniquely only in regard to this state representation. Thus, each of these Jacobians would be a 300×300 matrix. While this suffices to compute the Lyapunov exponents, it does not suffice for our purposes of obtaining Jacobian estimates of the dynamics of the *original* system. For this purpose, it has been shown that it is possible to obtain an estimate of the Jacobian of the original dynamical system (i.e., not the echo state space) using ESNs.⁴⁷⁴ However, as Banerjee et al. [474] noted, it is necessary to invert the non-square ESN output matrix in order to obtain these Jacobians, thus they are non-uniquely determined, and are used in the paper "only for causality estimation purposes". Nevertheless, we computed ESN-estimated Jacobians on the 1% observation noise test data. We used a large (7,200 time-step) lead-up and ensured coherence to the trajectory by feeding in the true state as input. The Frobenius norm error over five random seeds was 85.8 ± 3.1 , much larger than the reported errors of 3.3 ± 0.2 and 8.7 ± 0.3 for the JacobianODEs and NeuralODEs respectively. Thus, similar to the NeuralODE, while ESNs can achieve comparable prediction on chaotic systems, they do so without enabling an accurate reconstruction of the Jacobian of the original dynamical system.

C.3.7 Analyzing the dimensionality of Jacobian manifolds

RNN activity is often confined to low-dimensional manifolds, potentially simplifying Jacobian estimation. To address the extent to which this influences our results, we performed the following analyses.

Jacobian rank: We first note that both the RNN hidden weight matrix and nearly all Jacobians had full rank (128; $\sim 0.1\%$ had rank 127). To assess the *effective* rank, we computed the participation ratio (PR), a singular value-based measure of intrinsic dimensionality. The RNN weight matrix had a PR of 123, and the Jacobians had a mean PR of ~ 106 (std 0.7).

Intrinsic-data dimensionality: Another possibility is that the data itself may be low-dimensional. Since the Jacobian is a function of the trajectory data, the Jacobian manifold’s intrinsic dimension is at most that of the trajectory manifold. To estimate the intrinsic dimensions, we pooled time points from a large sample of trajectories. For RNN state vectors, we computed PR directly; for Jacobians, we flattened them into n^2 -dimensional vectors before computing PR. We found the RNN trajectory PR to be ~ 13.5 , and the Jacobian PR to be ~ 6 . Given the 128D extrinsic space, this confirms both RNN states and Jacobians lie on lower-dimensional manifolds.

Notably, prior work estimated the dimensionality of prefrontal cortex delay activity at 24 using ~ 4000 neurons, falling to ~ 6 when subsampled to 100 neurons.⁴⁷⁵ These values were interpreted as high-dimensional. Thus, the dimensionalities we observe are consistent with biological systems performing complex tasks.

Furthermore, in the 64-dimensional Lorenz 96 system, the trajectory PR was ~ 47.5 , and the Jacobians had a PR of ~ 50 , which we believe to be high-dimensional enough to demonstrate the effectiveness of our method on complex systems.

We note that in larger systems, approximating the Jacobian as low rank may be more efficient. Rather than computing each of the outputs directly, we can constrain the Jacobian to be low-rank by designing the output as a low-rank combination of vectors (e.g., it learns to output a low-rank singular value decomposition).

C.4 Experimental details

C.4.1 Dynamical systems data

We used the following dynamical systems for the testing and validation of Jacobian estimation.

Van der Pol oscillator We implement the classic Van der Pol oscillator.¹²² The system is governed by the following equations

$$\begin{aligned}\dot{x} &= y \\ \dot{y} &= \mu(1 - x^2)y - x\end{aligned}$$

We pick $\mu = 2$ in our implementation.

Lorenz We implement the Lorenz system introduced by Lorenz [60] as

$$\begin{aligned}\dot{x} &= \sigma(y - x) \\ \dot{y} &= x(\rho - z) - y \\ \dot{z} &= xy - \beta z\end{aligned}$$

with the typical choices for parameters ($\sigma = 10, \rho = 28, \beta = 8/3$).

Lorenz 96 We implement the Lorenz 96 system introduced by Lorenz [356] and defined by

$$\dot{x}_i = (x_{(i+1) \pmod N} - x_{(i-2) \pmod N})x_{(i-1) \pmod N} - x_i \pmod N + F$$

with $F = 8$ and $N \in \{12, 32, 64\}$.

Simulation We use the `dysts` package to simulate all dynamical systems.^{357,358} The characteristic timescale of their Fourier spectrum τ is selected and the systems are sampled with respect to τ . For all systems, the training data consisted of 26 trajectories of 12 periods, sampled at 100 time steps per τ . The validation data consisted of 6 trajectories of 12 periods sampled at 100 time steps per τ . The test data consisted of 8 trajectories of 12 periods sampled at 100 time steps per τ . Trajectories were initialized using a random normal distribution with standard deviation 0.2. The simulation algorithm used was Radau. Batches were constructed by moving a sliding window along the signal. The sequence length was selected such that the generated predictions would generate 10 novel time points (i.e., 11 time steps for the NeuralODE, and 25 time steps for the JacobianODE, due to the 15 time steps used to estimate the initial time derivative $\mathbf{f}$).

Noise We define $P\%$ observation noise with $P = 100p$ in the following way. Let $A_{\text{signal}} = \mathbb{E} [\|x(t)\|_2^2]$ be the expected squared norm of the signal with $\mathbf{x}(t) \in \mathbb{R}^n$. Then consider a noise signal $\eta(t) \in \mathbb{R}^n$ where each component $\eta_i(t) \sim \mathcal{N}(0, \frac{1}{\sqrt{n}}p\sqrt{A_{\text{signal}}})$. Then

$$\mathbb{E} [\|\eta(t)\|_2^2] = \mathbb{E} \left[\sum_{i=1}^n \eta_i^2(t) \right] = \sum_{i=1}^n \mathbb{E} [\eta_i^2(t)] = \sum_{i=1}^n \frac{1}{n} p^2 A_{\text{signal}} = p^2 A_{\text{signal}}$$

and thus the noise percent is

$$\sqrt{\frac{\mathbb{E} [\|\eta(t)\|_2^2]}{\mathbb{E} [\|x(t)\|_2^2]}} = \sqrt{\frac{p^2 A_{\text{signal}}}{A_{\text{signal}}}} = p$$

C.4.2 Task-trained RNN

The task used to train the RNN was exactly as defined in Section 3.6. The hidden dimensionality of the RNN was 128, and the input dimensionality was 10, where the first four dimensions represented the one-hot encoded "upper" color, the second four dimensions represented the one-hot encoded "lower" color, and the last two dimensions represented the one-hot encoded cue. The RNN used for the task had hidden dynamics defined by

$$\begin{aligned}\tau \dot{\mathbf{h}}(t) &= -\mathbf{h} + \mathbf{W}_{hh}\sigma(\mathbf{h}(t)) + \mathbf{W}_{hi}\mathbf{u}(t) + \mathbf{b} \\ \mathbf{o}(t) &= \mathbf{W}_{oh}\mathbf{h}(t)\end{aligned}$$

with $\tau = 50$ ms, which for the purposes of training was discretized with Euler integration with a time step of $\Delta t = 20$ ms. $\mathbf{W}_{hh}$ is the 128x128 dimensional matrix that defines the internal dynamics, $\mathbf{W}_{hi}$ is the 128 x 10 dimensional matrix that maps the input into the hidden state, $\mathbf{W}_{oh}$ is the 4x128 dimensional output matrix that maps the hidden state to a four-dimensional output $\mathbf{o}(t)$, and $\mathbf{b}$ is a static bias term. σ was taken to be an exponential linear unit activation with $\alpha = 1$. The RNN hidden state was split into two "areas" each with 64 dimensions. The input matrix $\mathbf{W}_{hi}$ was masked during training so that inputs could only flow into the first 64 dimensions – the "visual" area. The same procedure was performed for the output matrix $\mathbf{W}_{oh}$, except the mask was such that outputs could stem only from the second group of 64 dimensions – the "cognitive" area. The within-area subblocks of the matrix $\mathbf{W}_{hh}$ were first initialized such that the real part of the eigenvalues were randomly distributed on the interval $[-0.1, 0]$ and the imaginary part of the eigenvalues were randomly distributed on the interval $[0, 2\pi]$. The eigenvectors were random orthonormal matrices. We then computed the matrix exponential of this matrix. The across-area weights were first initialized to be random normal, then divided by the 2-norm of the resulting matrix and multiplied by 0.05. The input and output matrices were initialized as random normal and then scaled by the 2-norm of the resulting matrix. The static bias $\mathbf{b}$ was initialized at $\mathbf{0}$. After initialization, all weights could be altered unimpeded (except for the masks). Notably, the inputs $\mathbf{u}(t)$ were present only during the stimulus and cue presentation epochs – otherwise the network evolved autonomously. The loss was computed via cross entropy loss on the RNN outputs during the response period (the final 250 ms of the trial).

For the training data, we generated 4096 random trials, and used 80% for training and the remainder for validation. The batch size used was 32. Training was performed for 40 epochs. The learning rate was 0.0005. For use with the Jacobian estimation models, data was batched and used for training exactly as was done with the other dynamical systems data (see Appendix C.4.1). Observation noise was also computed in the same way.

C.4.3 NeuralODE details

NeuralODE models directly estimate the time derivative $\mathbf{f}$ with a neural-network parameterized function $\hat{\mathbf{f}}^\theta$. Then the Jacobians can be computed as $\hat{\mathbf{J}} = \frac{\partial}{\partial \mathbf{x}} \hat{\mathbf{f}}^\theta$.

The NeuralODEs were implemented as described in Section 3.5 and ODE integration was done exactly as for the JacobianODE using the `torchdiffeq` package with the RK4 method.³⁶⁰ To regularize the NeuralODE we implemented a Frobenius norm penalty on the estimated Jacobians, i.e.

$$\mathcal{L}_{\text{jac}} = \lambda_{\text{jac}} \langle \|\hat{\mathbf{J}}(\mathbf{x}(t))\|_F \rangle$$

where $\hat{\mathbf{J}}$ is the estimated Jacobian computed via automatic differentiation and λ_{jac} is a hyperparameter that controls the relative weighting of the Jacobian penalty.^{362–364} As mentioned in the chapter, this penalty prevents the model from learning unnecessarily large eigenvalues and encourages better Jacobian estimation.

C.4.4 Weighted linear Jacobian details

We implemented a baseline Jacobian estimation method using weighted linear regression models as described in Deyle et al. [85]. Given a reference point $\mathbf{x}(t^*)$ at which the (discrete) Jacobian will be computed, all other points are weighted according to

$$w_k = \exp \frac{-\theta \|\mathbf{x}(t_k) - \mathbf{x}(t^*)\|}{\bar{d}}$$

where

$$\bar{d} = \sum_{i=1}^T \|\mathbf{x}(t_i) - \mathbf{x}(t^*)\|$$

is the average distance from $\mathbf{x}(t^*)$ to all other points. We then perform a linear regression using the weighted points (and a bias term), the result of which is an estimate of the discrete Jacobian at $\mathbf{x}(t^*)$, which can be converted to continuous time by subtracting the identity matrix and dividing by the sampling time step (i.e., $\hat{\mathbf{J}} = \frac{\hat{\mathbf{J}}_{\text{discrete}} - \mathbf{I}}{\Delta t}$, where $\hat{\mathbf{J}}_{\text{discrete}}$ is the discrete Jacobian). The parameter θ tunes how strongly the regression is weighted towards local points. To pick θ , we sweep over values range from 0 to 10, and pick the value that yields the best one-step prediction according to

$$\mathbf{x}(t^* + 2\Delta t) = \mathbf{x}(t^* + \Delta t) + e^{\hat{\mathbf{J}}\Delta t}(\mathbf{x}(t^* + \Delta t) - \mathbf{x}(t^*))$$

where $\hat{\mathbf{J}}$ is the estimated Jacobian. This form of prediction has been previously been used to learn Jacobians in machine learning settings.³⁵⁰ To test the method, we pick θ based on data with observation noise at a particular noise level, then add in the denoised data to the data pool in order to compute regressions and estimate Jacobians at the true points.

C.4.5 Model details

All models were implemented as four-layer MLPS, with the four layers having sizes of 256, 1024, 2048, and 2048 respectively. All models used a sigmoid linear unit activation. JacobianODE models output to the dimension n^2 which was then reshaped into the matrix of the appropriate dimension. NeuralODEs output to the dimension n .

C.4.6 Lyapunov spectrum computation

To compute the Lyapunov spectrum, we employ a QR based algorithm.^{127,376} We discretize the Jacobians using the matrix exponential (i.e., $\hat{\mathbf{J}}_{\text{discrete}} = e^{\mathbf{J}\Delta t}$) and then propagate a bundle of small vectors through the Jacobians, using QR to ensure the perturbations remain bounded.

C.4.7 Iterative Linear Quadratic Regulator (ILQR)

We implement the standard algorithm for ILQR, the details of which can be found in Li and Todorov [369] and Tassa, Erez, and Todorov [370]. In brief, the ILQR algorithm linearizes the system dynamics around a nominal trajectory using the Jacobian, and then iteratively optimizes the control sequence using forward and backward passes to minimize the total control cost. The state cost matrix $\mathbf{Q}$ was a diagonal matrix with 1.0 along the diagonal. The final state cost matrix $\mathbf{Q}_f$ was a diagonal matrix with 1.0 along the diagonal. The control cost matrix $\mathbf{R}$ was a diagonal matrix with 0.01 along the diagonal. The control matrix was a 128×128 matrix in which the 64×64 block corresponding to the first 64 neurons (the "visual" area) was the 64-dimensional identity matrix. The control algorithm was seeded with only the initial state of the test trajectory with 5% noise. The control sequence was initialized random normal with standard deviation 0.001 and mean 0. The ILQR algorithm was run for a max of 100 iterations. The regularization was initialized at 1.0, with a minimum of 1×10^{-6} and a maximum of 1×10^{10} . Δ_0 was set to 2, as in Tassa, Erez, and Todorov [370]. If the backward pass failed 20 times in a row, the optimization was stopped. The list of values for the line search parameter α was 1.1^{-k^2} for $k \in 0, \dots, 9$ (see Tassa, Erez, and Todorov [370]). The linear model used for the linear baseline was computed via linear regression.

C.4.8 Training details

All models were implemented in PyTorch. The batch size used was 16. Gradients were accumulated for 4 batches. Training epochs were limited to 500 shuffled batches. Validation epochs were limited to 100 randomly chosen batches. Testing used all testing data. Training was run for a maximum of 1000 epochs, 3 hours, or until the early stopping was activated (see Appendix C.4.8), whichever came first.

Generalized Teacher Forcing

The Jacobian can be best learned when training predictions are generated recursively (i.e., replacing $\mathbf{x}(t)$ by $\hat{\mathbf{x}}(t)$). However, in chaotic systems, and/or systems with measurement noise (as considered here), this could lead to catastrophic divergence of the predicted trajectory from the true trajectory during training. We therefore employ Generalized Teacher Forcing when training all models.³⁵⁴ Generalized Teacher Forcing prevents catastrophic divergence by forcing the generated predictions along the line from the prediction to the true state. Specifically, for a given predicted state $\hat{\mathbf{x}}(t)$ and

true state $\mathbf{x}(t)$, the teacher forced state is

$$\tilde{\mathbf{x}}(t) = (1 - \alpha)\hat{\mathbf{x}}(t) + \alpha\mathbf{x}(t)$$

with $\alpha \in [0, 1]$. This effectively forces the predictions along a line from the prediction to the true state, by an amount with proportion α . $\alpha = 1$ corresponds to fully-forced prediction (i.e., one-step prediction) and $\alpha = 0$ corresponds to completely unforced prediction (i.e., autonomous prediction). Hess et al. [354] suggested that a good estimate of α is

$$\alpha = \max \left(\max_p \left[1 - \frac{1}{\|\mathcal{G}(\mathbf{J}_{T:2}^{(p)})\|} \right], 0 \right) \quad (\text{C.11})$$

where $\mathbf{J}_{T:2}^{(p)}$ are the Jacobians of the modeled dynamics computed at data-constrained states, p indicates the batch or sequence index, and

$$\|\mathcal{G}(\mathbf{J}_{T:2}^{(p)})\| = \left\| \left(\prod_{k=0}^{T-2} \mathbf{J}_{T-k} \right)^{\frac{1}{T-1}} \right\|$$

effectively computes the discrete maximum Lyapunov exponent. In our implementation, we compute $\|\mathcal{G}(\mathbf{J}_{T:2}^{(p)})\|$ using a QR-decomposition-based Lyapunov computation algorithm.¹²⁷ As the Jacobian of the dynamics is necessary to compute this quantity, the JacobianODEs enjoy an advantage over other models in that the Jacobians are directly output by the model, and do not have to be computed via differentiating the model itself.

We furthermore employ a slightly modified version of the suggested annealing process in Hess et al. [354], which sets $\alpha_0 = 1$ and updates α_n as

$$\alpha_n = \gamma\alpha_{n-1} + (1 - \gamma)\alpha$$

where α is computed according to equation C.11. Following the suggested hyperparameters, we set $\gamma = 0.999$ and update α_n every 5 batches. Once the teacher forced state $\tilde{\mathbf{x}}(t)$ is computed, it can simply replace $\mathbf{x}(t)$ in equation C.5 to generate predictions.

Loop closure loss

We implemented a loop closure loss as discussed in Section 3.4.3 and Appendix C.1.3. For each loop, we used 20 randomly chosen points from the batch. For each batch, we constructed the same number of loops as there were batches. Path integrals were discretized in 20 steps and computed using the trapezoid method from `torchquad`.³⁵⁹

Validation loss

All models were validated on 10 time-step prediction task with teacher forcing parameter $\alpha = 0$ (i.e., autonomous prediction).

Learning rate scheduling

For all models, the learning rate was annealed in accordance with teacher forcing annealing. Given an initial and final learning rates η_i and η_f we compute the effective learning rate as

$$\eta = \eta_f + \sigma(\alpha_n)(\eta_i - \eta_f)$$

where α_n is the current value of the teacher forcing parameter and

$$\sigma(\alpha_n) = \frac{\alpha_n}{\alpha_n + (1 - \alpha_n)e^{-k\alpha_n}}$$

$\sigma(\alpha_n)$ is a scaling function with $\sigma(1) = 1$ and $\sigma(0) = 0$ and for which the shape of the scaling is controlled by the parameter k . For positive values of k , the scaling is super-linear, and for negative values of k it is sub-linear. We use $k = 1$, ensuring that the learning rate does not decrease too quickly at the start of learning. We set $\eta_i = 10^{-4}$ and $\eta_f = 10^{-6}$ for all models.

Optimizer and weight decay

All models were trained with PyTorch’s AdamW optimizer with the learning rate as described above, and weight decay parameter 10^{-4} . All other parameters were default ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$). We also used gradient norm clipping, with a clipping value of 1.0.

Early stopping

For all models, we implemented an early stopping scheme that halted the training if the validation loss improved by less than 1% for two epochs in a row.

Added noise during learning

For the models trained on the task-trained RNN dynamics, we added 5% Gaussian i.i.d. noise (defined relative to the norm of the training data with observation noise already added). Noise was sampled for each batch and added prior to the trajectory generation step of the learning process. Additional noise was not added for the loop closure computation.

Hyperparameter selection

For the JacobianODEs, the primary hyperparameter to select is the loop closure loss λ_{loop} . To select this hyperparameter, we trained JacobianODE models with

$$\lambda_{\text{loop}} \in [0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10].$$

For each run, the epoch with the lowest trajectory validation loss ($\mathcal{L}_{\text{traj}}$) is kept. Then, for this model, we compute the one-step prediction error on validation data, the validation loop closure loss ($\mathcal{L}_{\text{loop}}$), and the percentage of Jacobian eigenvalues on all validation data that have a decay rate faster than the sampling rate $\frac{1}{\Delta t}$. We exclude any models that meet any of the following criteria:

1. **One-step prediction error greater than the persistence baseline.** The persistence baseline is computed as the mean error between each time step $k\Delta t$ and the subsequent time step $(k+1)\Delta t$ across the dataset, and constitutes a sanity check for whether a model is capturing meaningful information about the dynamics.
2. **Loop closure loss greater than $\sqrt{n}$,** where n is the system dimension (see Appendix C.4.11 for the derivation of this bound). As discussed in the chapter, we are interested in Jacobians that not only solve the trajectory prediction problem, but that also are constructed so that the rows of the matrix are approximately conservative vector fields.
3. **More than 0.1% of the Jacobian eigenvalues have a decay rate faster than the sampling rate $\frac{1}{\Delta t}$.** Since large negative eigenvalues do not impact trajectory prediction, the models may erroneously learn Jacobians with large negative eigenvalues. If the decay rate of these eigenvalues is faster than the sampling rate, we can infer that the eigenvalues are not aligned with the observed data.

If none of the models that meet criterion (2) meet criterion (1), we discount criterion (2), as this suggests that a loop closure loss below $\sqrt{n}$ bound is too strict to obtain good prediction on this system. Additionally, if none of the models that meet criterion (3) meet criterion (1), we discount criterion (1), as this suggests that noise is very high in the data, which leads to both high one-step prediction error, and large negative eigenvalues to compensate for the perturbations introduced by the noise. Of the remaining models, we select the one with the lowest trajectory validation loss $\mathcal{L}_{\text{traj}}$.

For the NeuralODEs, we needed to select the hyperparameter λ_{jac} , which regularized the mean frobenius norm of the Jacobians computed through automatic differentiation. Again, to select this hyperparameter, we trained JacobianODE models with

$$\lambda_{\text{jac}} \in [0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10].$$

We followed exactly the above procedure with the exception of criterion (2), which was deemed unnecessary, as computing the Jacobians implicitly via a gradient of the model (using automatic differentiation) ensures that the rows of the matrix are conservative.

All other hyperparameters (model size, learning rate, length of initial trajectory, number of discretization steps, etc.) were fixed for all systems. Given the wide range of systems and behaviors and dimensionalities that the JacobianODEs are capable of capturing, this indicates that the method is robust given a reasonable choice of these hyperparameters.

Model hyperparameters and training details

Below are presented the details of all Jacobian estimation models considered in the main paper.

Table C.7: Hyperparameters used and model details for each system and noise level. Training time is reported in seconds.

System	Noise	Model	Loop closure weight	Jacobian penalty	Training time (s)	Final epoch	Learning rate	Min learning rate	Weight decay	Learnable parameters
VanDerPol (2 dim)	1%	JacobianODE	0.0010	0	882.139	15	0.0001	1.00e-06	0.0001	6.568e+06
	1%	NeuralODE	0	1.00e-06	949.217	21	0.0001	1.00e-06	0.0001	6.564e+06
	5%	JacobianODE	0.0001	0	692.980	11	0.0001	1.00e-06	0.0001	6.568e+06
	5%	NeuralODE	0	0	249.451	6	0.0001	1.00e-06	0.0001	6.564e+06
	10%	JacobianODE	0.010	0	678.794	10	0.0001	1.00e-06	0.0001	6.568e+06
	10%	NeuralODE	0	0.0010	713.221	16	0.0001	1.00e-06	0.0001	6.564e+06
Lorenz (3 dim)	1%	JacobianODE	0.0010	0	972.268	16	0.0001	1.00e-06	0.0001	6.578e+06
	1%	NeuralODE	0	0.0010	889.819	18	0.0001	1.00e-06	0.0001	6.566e+06
	5%	JacobianODE	0.010	0	1.147e+03	18	0.0001	1.00e-06	0.0001	6.578e+06
	5%	NeuralODE	0	0.0010	680.651	15	0.0001	1.00e-06	0.0001	6.566e+06
	10%	JacobianODE	0.010	0	388.346	6	0.0001	1.00e-06	0.0001	6.578e+06
	10%	NeuralODE	0	0.010	421.488	8	0.0001	1.00e-06	0.0001	6.566e+06
Lorenz 96 (12 dim)	1%	JacobianODE	0.0010	0	1.817e+03	31	0.0001	1.00e-06	0.0001	6.857e+06
	1%	NeuralODE	0	1.00e-06	1.892e+03	32	0.0001	1.00e-06	0.0001	6.587e+06
	5%	JacobianODE	0.0010	0	716.408	12	0.0001	1.00e-06	0.0001	6.857e+06
	5%	NeuralODE	0	0.0010	1.147e+03	19	0.0001	1.00e-06	0.0001	6.587e+06
	10%	JacobianODE	0.010	0	1.213e+03	18	0.0001	1.00e-06	0.0001	6.857e+06
	10%	NeuralODE	0	0.0001	851.803	14	0.0001	1.00e-06	0.0001	6.587e+06
Lorenz 96 (32 dim)	1%	JacobianODE	0.010	0	5.160e+03	85	0.0001	1.00e-06	0.0001	8.665e+06
	1%	NeuralODE	0	0.0010	4.204e+03	43	0.0001	1.00e-06	0.0001	6.633e+06
	5%	JacobianODE	0.0010	0	1.341e+03	22	0.0001	1.00e-06	0.0001	8.665e+06
	5%	NeuralODE	0	0.0010	3.855e+03	39	0.0001	1.00e-06	0.0001	6.633e+06
	10%	JacobianODE	0.0001	0	868.262	14	0.0001	1.00e-06	0.0001	8.665e+06
	10%	NeuralODE	0	0.0010	2.695e+03	28	0.0001	1.00e-06	0.0001	6.633e+06
Lorenz 96 (64 dim)	1%	JacobianODE	0.0010	0	2.749e+03	40	0.0001	1.00e-06	0.0001	1.497e+07
	1%	NeuralODE	0	0.010	4.801e+03	30	0.0001	1.00e-06	0.0001	6.706e+06
	5%	JacobianODE	0.0001	0	1.748e+03	27	0.0001	1.00e-06	0.0001	1.497e+07
	5%	NeuralODE	0	0.010	4.879e+03	30	0.0001	1.00e-06	0.0001	6.706e+06
	10%	JacobianODE	0.0010	0	1.701e+03	25	0.0001	1.00e-06	0.0001	1.497e+07
	10%	NeuralODE	0	0.010	4.237e+03	26	0.0001	1.00e-06	0.0001	6.706e+06
Task-trained RNN	1%	JacobianODE	0.0001	0	2.496e+03	32	0.0001	1.00e-06	0.0001	4.016e+07
	1%	NeuralODE	0	0	9.777e+03	38	0.0001	1.00e-06	0.0001	6.854e+06
	5%	JacobianODE	0.0001	0	2.352e+03	29	0.0001	1.00e-06	0.0001	4.016e+07
	5%	NeuralODE	0	1.00e-05	7.770e+03	27	0.0001	1.00e-06	0.0001	6.854e+06
	10%	JacobianODE	0.010	0	2.172e+03	27	0.0001	1.00e-06	0.0001	4.016e+07
	10%	NeuralODE	0	1.00e-05	5.260e+03	18	0.0001	1.00e-06	0.0001	6.854e+06

C.4.9 Information about computing resources and efficiency

All models were able to be trained on a single H100 GPU, with 80 GB of memory.

Jacobian inference times Jacobian inference times for the JacobianODE and NeuralODE models are discussed in Appendix C.3.5. As discussed, models were implemented with four hidden layers of the same size, and tested on 100 batches of the 128-dimensional task-trained RNN data, with each batch consisting of 16 sequences of length 25. Timings were repeated ten times for each model (see Figure C.3 for details).

Training time Total training times for each of the chosen models are presented in Table C.7. Furthermore, we include the training time (including backward pass) for 100 batches (with 16 sequences per batch), using 10 time-step prediction, in Table C.8.

C.4.10 Statistical details

All statistics were computed using `scipy`. For the comparison between JacobianODE and NeuralODE trajectory and Jacobian predictions, as well as the comparison of Gramian traces and minimum eigenvalues, we used a two-sample t-test. For the comparison of ILQR control accuracies and errors, we used a Wilcoxon signed-rank test.

Table C.8: Trajectory training time (seconds) for each system and noise level. Training used 100 batches, with 16 sequences per batch, as well as 10 time-step prediction.

Model	Lorenz (3 dim)	VanDerPol (2 dim)	Lorenz 96 (12 dim)	Lorenz 96 (32 dim)	Lorenz 96 (64 dim)	Task-trained RNN
1% noise						
JacobianODE	15.772	13.737	11.604	16.469	16.214	23.855
NeuralODE	8.215	8.403	11.852	12.075	18.192	14.955
5% noise						
JacobianODE	12.422	16.772	10.655	13.121	12.949	23.197
NeuralODE	6.191	5.841	13.501	8.021	21.307	30.209
10% noise						
JacobianODE	12.590	15.719	18.447	10.506	18.900	22.482
NeuralODE	11.083	9.875	15.269	8.008	17.826	33.074

C.4.11 Derivation of loop closure loss bound

We consider the loop closure loss as defined in Section 3.4.3. We are interested in estimating a bound on the error

$$\mathbb{E} \left[\frac{1}{n} \left\| \int_{\mathcal{C}_{\text{loop}}^{(l)}} \mathbf{J} ds \right\|_2^2 \right]$$

where n is the system dimension. While in theory this quantity should be 0, in practice due to numerical estimation error, it will not be. First recall, that

$$\int_{\mathcal{C}_{\text{loop}}^{(l)}} \mathbf{J} ds = \sum_{i=1}^L \int_{\mathbf{x}(t_i \pmod{L})}^{\mathbf{x}(t_{i+1} \pmod{L})} \mathbf{J}(\mathbf{c}(r)) \mathbf{c}'(r) dr$$

where L is the number of loop points and $\mathbf{c}$ is a line from $\mathbf{x}(t_i \pmod{L})$ to $\mathbf{x}(t_{i+1} \pmod{L})$. We assume that

$$\mathbb{E} \left[\left\| \int_{\mathbf{x}(t_i \pmod{L})}^{\mathbf{x}(t_{i+1} \pmod{L})} \mathbf{J}(\mathbf{c}(r)) \mathbf{c}'(r) dr \right\|_2^2 \right] = \mathbb{E} \left[\left\| \int_{\mathbf{x}(t_j \pmod{L})}^{\mathbf{x}(t_{j+1} \pmod{L})} \mathbf{J}(\mathbf{c}(r)) \mathbf{c}'(r) dr \right\|_2^2 \right]$$

$\forall i, j \in [1, \dots, L]$, which is justified as the numerical error accrued will likely be similar along different lines for the same system. Thus

$$\begin{aligned}
\mathbb{E} \left[\frac{1}{n} \left\| \int_{\mathcal{C}_{\text{loop}}^{(l)}} \mathbf{J} ds \right\|_2^2 \right] &= \mathbb{E} \left[\frac{1}{n} \left\| \sum_{i=1}^L \int_{\mathbf{x}(t_i \pmod{L})}^{\mathbf{x}(t_{i+1} \pmod{L})} \mathbf{J}(\mathbf{c}(r)) \mathbf{c}'(r) dr \right\|_2^2 \right] \\
&\leq \frac{1}{n} \mathbb{E} \left[\sum_{i=1}^L \left\| \int_{\mathbf{x}(t_i \pmod{L})}^{\mathbf{x}(t_{i+1} \pmod{L})} \mathbf{J}(\mathbf{c}(r)) \mathbf{c}'(r) dr \right\|_2^2 \right] \\
&= \frac{1}{n} \sum_{i=1}^L \mathbb{E} \left[\left\| \int_{\mathbf{x}(t_i \pmod{L})}^{\mathbf{x}(t_{i+1} \pmod{L})} \mathbf{J}(\mathbf{c}(r)) \mathbf{c}'(r) dr \right\|_2^2 \right] \\
&= \frac{1}{n} \sum_{i=1}^L \mathbb{E} \left[\left\| \int_{\mathbf{x}(t_0)}^{\mathbf{x}(t_1)} \mathbf{J}(\mathbf{c}(r)) \mathbf{c}'(r) dr \right\|_2^2 \right] \\
&= \frac{L}{n} \mathbb{E} \left[\left\| \int_{\mathbf{x}(t_0)}^{\mathbf{x}(t_1)} \mathbf{J}(\mathbf{c}(r)) \mathbf{c}'(r) dr \right\|_2^2 \right]
\end{aligned}$$

where we have used the fact that the errors are assumed to be equivalent for each of the line segments comprising the overall loop path. Recall now that for the trapezoid integration rule, the error E in integrating $\int_a^b f(x)dx$ can be computed as $E = -\frac{(b-a)^3}{12M^2} f''(u)$ for some $u \in [a, b]$. Thus the squared error is bounded as

$$E^2 \leq \left\| \max_{u \in [a, b]} \frac{(b-a)^3}{12M^2} f''(u) \right\|_2^2$$

In our case, when path integrating lines, we are effectively integrating from $a = 0$ to $b = 1$. Furthermore, let $\mathbf{g}(r) = \mathbf{J}(\mathbf{c}(r))\mathbf{c}'(r)$, the integrand of our line integrations. We assume that $\mathbb{E} \left[\|\max_r \mathbf{g}''(r)\|_2^2 \right] \propto n \cdot \sqrt{n}$ where the first n comes from the fact that the norm involves summing over the n components of the vector $\mathbf{g}''(r)$ and the second $\sqrt{n}$ involves an assumption that loop closure loss will be more difficult to compute accurately in higher dimensions, though this will be more pronounced as dimensionality initially starts increasing. Thus $\mathbb{E} \left[\|\max_r \mathbf{g}''(r)\|_2^2 \right] = kn \cdot \sqrt{n}$ for some $k \in \mathbb{R}^+$. Now, continuing on,

$$\begin{aligned}
\mathbb{E} \left[\frac{1}{n} \left\| \int_{\mathcal{C}_{\text{loop}}^{(l)}} \mathbf{J} ds \right\|_2^2 \right] &\leq \frac{L}{n} \mathbb{E} \left[\left\| \int_{\mathbf{x}(t_0)}^{\mathbf{x}(t_1)} \mathbf{J}(\mathbf{c}(r)) \mathbf{c}'(r) dr \right\|_2^2 \right] \\
&\leq \frac{L}{n} \mathbb{E} \left[\left\| \frac{1^3}{12M^2} \max_r \mathbf{g}''(r) \right\|_2^2 \right] \\
&= \frac{L}{12^2 n M^4} \mathbb{E} \left[\left\| \max_r \mathbf{g}''(r) \right\|_2^2 \right] \\
&= \frac{Lkn \cdot \sqrt{n}}{12^2 n M^4} \\
&= \frac{Lk\sqrt{n}}{12^2 M^4}
\end{aligned}$$

Finally, assuming that the number of discretization steps M was chosen to be large enough such

that $M^4 \geq \frac{1}{12^2} Lk$ we finally obtain

$$\mathbb{E} \left[\frac{1}{n} \left\| \int_{\mathcal{C}_{\text{loop}}^{(l)}} \mathbf{J} ds \right\|_2^2 \right] \leq \sqrt{n}$$

Appendix D

Methods: Disruption of interareal control during propofol anesthesia

D.1 Supplementary results

D.1.1 Linearity of the learned dynamics and linear baseline analysis

A natural question raised by the conditional JacobianODE model is whether the dynamics it learns from the neural data are genuinely nonlinear, or whether a strictly linear time-invariant (LTI) description of the same data would suffice. We address this question in two ways (Figure D.1). We first investigate how strongly the learned Jacobian vary along the trajectory, and how that variation trades off against predictive accuracy (Figure D.1A,B). Second, we retrain the dynamics with a linear delay-embedded model and ask whether the same anesthesia-related Gramian reorganization is retained (Figure D.1C,D).

The learned Jacobian is approximately linear, despite targeted anti-LTI interventions. We compared the baseline JacobianODE model against five variants designed to push the Jacobian off the constant-matrix solution: separate per-condition dynamics, a per-batch latent exponential moving average normalization, a null-space penalty, a stop-gradient applied to the loop-closure target, and a smaller encoder. As a non-LTI reference we include the partially observed working-memory selection RNN of §D.4.1 (a 128-D task-trained network whose ground-truth dynamics are genuinely nonlinear). For each model we report three diagnostics on the validation set: the per-condition Jacobian temporal coefficient of variation, $CV_c = \langle \sigma_t(\hat{\mathbf{J}}_{ij}^\theta(\mathbf{z}(t), c)) / |\langle \hat{\mathbf{J}}_{ij}^\theta(\mathbf{z}(t), c) \rangle_t| \rangle_{i,j,\text{traj}}$, the loop-closure loss $\mathcal{L}_{\text{loop}}$, and the free-running trajectory validation loss (Figure D.1A). The baseline model had a relatively low temporal CV, indicating approximately LTI latent dynamics. Across all five interventions, temporal CV stays within a factor of two of the baseline ($CV \approx 10^{-3.5} - 10^{-3}$), three orders of magnitude below the task-trained RNN reference ($CV \approx 3$). Thus the variation of $\hat{\mathbf{J}}^\theta$ along the trajectory is small even when the architecture is explicitly biased against the constant-matrix solution. None of the interventions yields a meaningful reduction in $\mathcal{L}_{\text{loop}}$ or trajectory loss either. The model is therefore capable of representing strongly nonlinear Jacobians when the data demand it (the task-trained RNN reference confirms this), but on the propofol cohort it settles into the approximately-LTI regime. This may have to do with the construction of the dynamics as

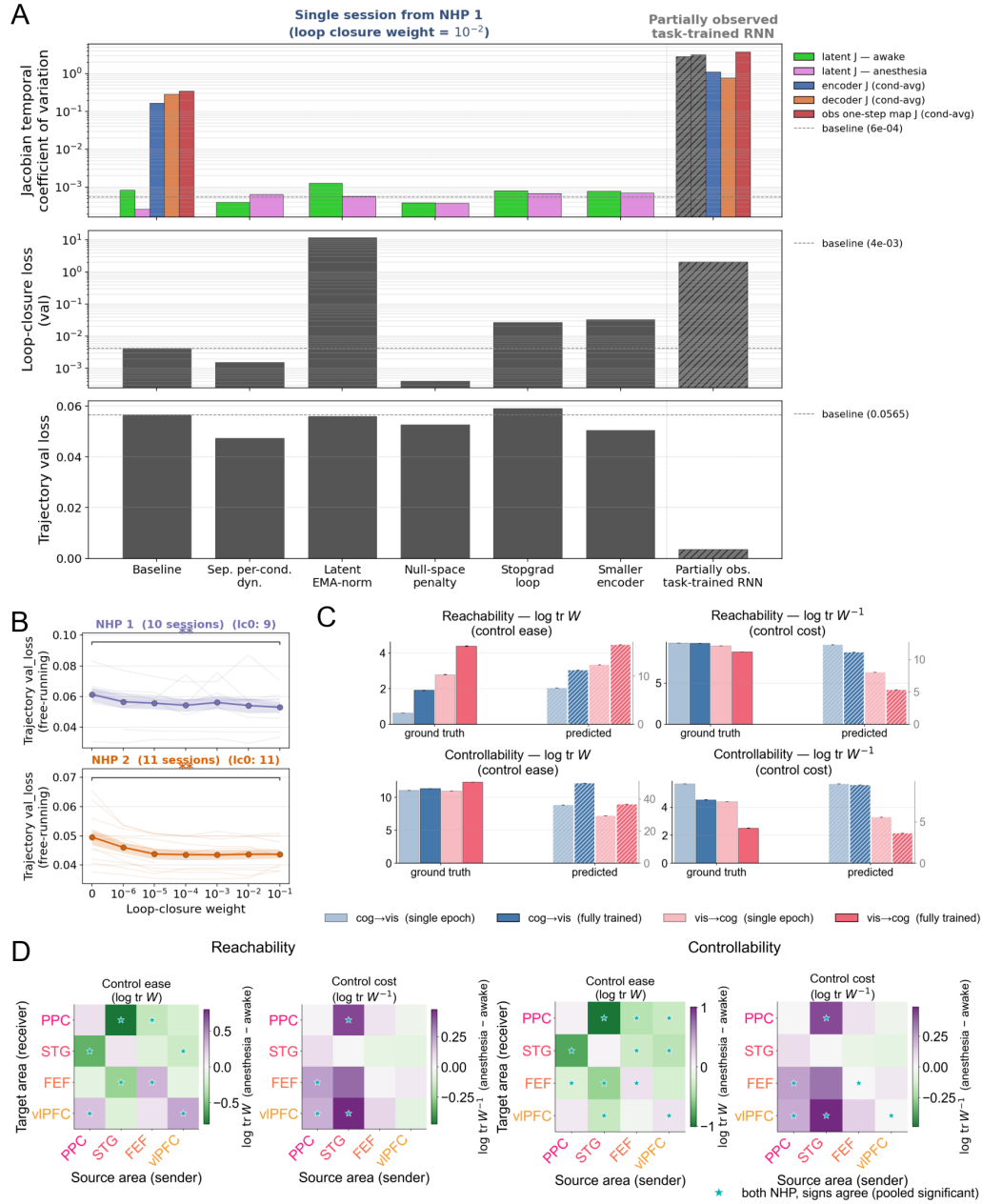


Figure D.1: Linearity diagnostics and linear delay-embedded baseline. (A) Baseline with five anti-LTI interventions on a single session from NHP 1 (using loop-closure weight 10^{-2}), with a partially observed task-trained RNN as nonlinear reference. (A, top) Per-condition latent Jacobian temporal coefficient of variation (CV), used as a measure of nonlinearity. The condition-averaged encoder (J_{Ψ}), decoder ($J_{\Psi^{-1}}$), and round-trip Jacobian CV area overlaid for the baseline and the RNN reference. (A, middle) Validation loop-closure loss. (A, bottom) Free-running trajectory validation loss. The NHP 1 variants cluster near the baseline on all three latent-Jacobian diagnostics. The encoder/decoder Jacobians, by contrast, are two to three orders of magnitude more state-dependent (Eq. D.1), and the reference shows the genuinely nonlinear regime the architecture is capable of. (B) Best free-running trajectory validation loss vs. loop-closure weight λ_{loop} , per NHP, across the cohort's chosen-cell sweep grid (per-session lines + mean $\pm$ SEM). There is a monotone-non-decreasing trend in both NHPs, indicating improved fits with models that obey loop closure laws. (C) Per-area linear delay-embedded baseline (Eqs. D.2–D.3) applied to the partially observed task-trained RNN. Reachability and controllability gramian and inverse gramian log traces ($\log \text{tr } W$ and $\log \text{tr } W^{-1}$) for the visual $\rightarrow$ cognitive and cognitive $\rightarrow$ visual blocks at single-epoch and fully-trained checkpoints. Ground truth (left of each pair) vs. linear baseline (right). (D) The same linear baseline on the propofol cohort, with the four sign-agree pooled-Wilcoxon 4×4 Gramian Δ -heatmaps that mirror Figure 4.5A,B of the chapter. The qualitative reorganization is preserved, with smaller effect sizes and more variability.

autonomous, or with the short prediction horizon (10 milliseconds) used for training.

The overall observation-space dynamics remain nonlinear through the encoder and decoder. The near-LTI behaviour of $\hat{\mathbf{J}}^\theta$ is a statement about the *latent* dynamics alone. The dynamics in the observed delay-embedded space need not be linear, because they are read out through the nonlinear diffeomorphic encoder. The model advances a delay-embedded observation by encoding it, evolving the latent state by encoding, integrating $\hat{\mathbf{J}}^\theta$, then decoding ($\tilde{\mathbf{x}}(t + \Delta t) = \Psi^{-1}(\varphi_\theta(\Psi(\tilde{\mathbf{x}}(t)), c))$). By the chain rule the observation-space Jacobian therefore factorizes as

$$\mathbf{J}_{\text{obs}}(\tilde{\mathbf{x}}) = \mathbf{J}_{\Psi^{-1}}(\mathbf{z}') \mathbf{J}_\varphi(\mathbf{z}) \mathbf{J}_\Psi(\tilde{\mathbf{x}}), \quad \mathbf{z} = \Psi(\tilde{\mathbf{x}}), \quad \mathbf{z}' = \varphi_\theta(\mathbf{z}, c), \quad (\text{D.1})$$

which is the product of the encoder Jacobian, the latent flow Jacobian, and the decoder Jacobian. Even when the central factor $\mathbf{J}_\varphi$ is nearly state-independent, the encoder and decoder Jacobians $\mathbf{J}_\Psi$ and $\mathbf{J}_{\Psi^{-1}} = \mathbf{J}_\Psi^{-1}$ vary with the state, so the composite observation-space map is genuinely nonlinear. To quantify this we computed the same per-condition temporal CV for the encoder Jacobian and the (exactly invertible) decoder Jacobian on the held-out validation set (Figure D.1A, top row; for NHP 1 the decoder Jacobian is obtained as $\mathbf{J}_\Psi^{-1}$, which matches a direct automatic-differentiation Jacobian of Ψ^{-1} to a relative error of $\sim 10^{-14}$). For NHP 1 the encoder and decoder Jacobians vary substantially along the trajectory. The near-constancy of $\hat{\mathbf{J}}^\theta$ thus reflects the encoder absorbing the nonlinearity of the observed dynamics into a coordinate system in which the latent flow is approximately linear, not the observed neural dynamics being linear.

This view connects the latent JacobianODE to Koopman operator theory, in which a nonlinear system is rendered linear by lifting it through a map of observables^{100,101}, and in particular to deep-learning approaches that learn such linearizing coordinates with autoencoders^{112,113}. The diffeomorphic encoder differs in being an exactly invertible, volume-preserving change of coordinates of the same dimension, rather than a (generally higher- or infinite-dimensional) lift to a space of observables; correspondingly the latent flow is only *approximately* linear, since an invertible, dimension-preserving change of coordinates cannot in general render a nonlinear flow exactly linear (for instance, when it possesses multiple isolated fixed points).

Loop-closure regularization trades nonlinearity against predictive accuracy. We found that as the loop closure loss weight λ_{loop} was varied, the best free-running trajectory validation loss was monotonically non-decreasing for both NHPs. Increasing λ_{loop} enforces row-conservativity of $\hat{\mathbf{J}}^\theta$, an integrability constraint that the constant-matrix solution satisfies trivially. Together with panel A, this is consistent with the interpretation that, for the propofol data, a near-LTI Jacobian is the most parsimonious model: gains in nonlinearity (lower $\mathcal{L}_{\text{loop}}$) come at the cost of a worse fit.

Linear delay-embedded baseline: per-area eigen-time-delay coordinates with joint DMD. Given the near-LTI behaviour of the JacobianODE model on the cohort, we asked whether substituting a strictly linear delay-embedded model for the learned Jacobian preserves the per-pair Gramian readouts reported in the chapter. The baseline is a per-area extension of the Hankel Alternative View Of Koopman (HAVOK) construction^{67,107}, fit as follows. For each area $k \in \{1, \dots, K\}$ and each time window (§D.4.3), we form the per-area Hankel matrix at depth n_d^k

from the $n_d^k d_{\text{obs}}^k$ stacked delay coordinates of §D.3.1, stack the windows of one condition, and take its reduced SVD to obtain the per-area eigen-time-delay basis:

$$\mathbf{H}^k = \mathbf{U}^k \mathbf{\Sigma}^k (\mathbf{V}^k)^\top, \quad \mathbf{P}^k \triangleq [\mathbf{V}^k]_{:,1:r^k} \in \mathbb{R}^{n_d^k d_{\text{obs}}^k \times r^k}, \quad (\text{D.2})$$

so that the per-area reduced state is $\mathbf{z}^k(t) = (\mathbf{P}^k)^\top \tilde{\mathbf{x}}^k(t) \in \mathbb{R}^{r^k}$. The pair (n_d^k, r^k) is selected per area by an AIC sweep over $n_d^k \in \{1, 3, 5, 10\}$ and $r^k \in \{4, 8, 16, \dots, 600\}$ on held-out within-condition windows (80/20 window-level split, per-area observation-space prediction error). Reduced states from the K areas are concatenated to a joint state $\mathbf{z}(t) = [\mathbf{z}^1(t); \dots; \mathbf{z}^K(t)] \in \mathbb{R}^R$ with $R = \sum_k r^k$ (aligning all windows to the maximum $n_d^k - 1$ leading offset so the per-area snapshot indices align), and a joint discrete-time linear model is fit by least-squares dynamic mode decomposition,

$$\mathbf{A}^* = \arg \min_{\mathbf{A} \in \mathbb{R}^{R \times R}} \sum_t \|\mathbf{z}(t + \Delta t) - \mathbf{A} \mathbf{z}(t)\|_2^2 = \mathbf{Z}_+ \mathbf{Z}_-^+, \quad (\text{D.3})$$

where $\mathbf{Z}_- = [\mathbf{z}(t_1), \dots, \mathbf{z}(t_{N-1})]$ and $\mathbf{Z}_+ = [\mathbf{z}(t_2), \dots, \mathbf{z}(t_N)]$ are the stacked snapshot pairs and $(\cdot)^+$ denotes the Moore–Penrose pseudoinverse. The construction of the K -area joint state imposes a natural block decomposition of $\mathbf{A}^*$: writing $\mathbf{A}^* = [\mathbf{A}^{*,\beta \rightarrow \alpha}]_{\alpha,\beta=1}^K$ with $\mathbf{A}^{*,\beta \rightarrow \alpha} \in \mathbb{R}^{r^\alpha \times r^\beta}$ acting as the linear analogue of $\mathbf{J}^{\beta \rightarrow \alpha}$ used in the chapter, the discrete-time reachability and controllability Gramians at horizon T are computed by the square-root QR iteration of §D.3.4 with $(\mathbf{A}_k, \mathbf{B}_k, \mathbf{C}_k) = (\mathbf{A}^{*,\alpha \rightarrow \alpha}, \mathbf{A}^{*,\beta \rightarrow \alpha}, \mathbf{A}^{*,\alpha \rightarrow \beta})$ time-invariant (the same numerical pipeline, with the constant LTV reference replaced by the fitted $\mathbf{A}^*$). All Gramians reported below use horizon $T = 10\Delta t$, matching the chapter’s JacobianODE horizon.

The linear baseline reproduces the JacobianODE ordering on the task-trained RNN. Applied to the partially observed task-trained working-memory selection RNN of §D.4.1, the per-area linear HAVOK pipeline qualitatively reproduces the ground-truth pattern of reach and controllability Gramian magnitudes for the visual $\rightarrow$ cognitive and cognitive $\rightarrow$ visual blocks, both at the single-epoch and fully-trained checkpoints used in the chapter (Figure D.1C). While the directional ordering between conditions is not perfectly preserved on every panel, the linear baseline is informative on a nonlinear test bed, and is therefore an appropriate sanity check to run on the neural data.

The linear baseline reproduces the main propofol findings, with reduced precision.

We ran the same per-area linear pipeline across all 21 sessions of the propofol cohort and computed the four sign-agree pooled-Wilcoxon $\log \text{tr } \mathbf{W}$ and $\log \text{tr } \mathbf{W}^{-1}$ heatmaps that constitute Figure 4.5A,B of the chapter (Figure D.1D, 10 ms integration horizon). The qualitative organization survives: anesthesia decreases reachability and controllability ease (more negative $\Delta \log \text{tr } \mathbf{W}$) and increases their associated cost ($\Delta \log \text{tr } \mathbf{W}^{-1}$) on most pairs, with the same notable exception that anesthesia *increases* reach ease from PPC and FEF into vIPFC. Compared to the JacobianODE composite, however, the linear baseline shows more variability and smaller effect sizes. This is the expected signature of a less precise model: a single time-invariant block matrix is a coarser representation of the data than a per-trajectory conditional Jacobian, but the directional reorganization the chapter reports is robust to this substitution.

D.1.2 Interareal control reorganization across integration horizons

The directional control summaries reported in the chapter use a $T = 10$ ms integration horizon (§D.3.4). The gramian integrand combines a cross-area input term (the coupling block $\mathbf{B} = \mathbf{J}^{\beta \rightarrow \alpha}$) with the within-target state-transition propagator Φ^α (Eqs. D.52–D.53), thus lengthening the horizon progressively up-weights the intrinsic within-area dynamics relative to the interareal coupling. To ask whether the anesthesia-driven reorganization is a property of the coupling or of the intrinsic dynamics, we recomputed the full set of summaries (reachability and controllability $\log \text{tr} \mathbf{W}$ and $\log \text{tr} \mathbf{W}^{-1}$, and their diagnostic correlations) at integration horizons of 10, 20, 100 and 1000 ms (Figure D.2).

At the control-relevant 10 ms horizon, as reported in the chapter, anesthesia shrinks reachability between areas (Figure D.2, top row). This reachability reduction progressively dissolves as the horizon lengthens: the number of pairs with a significant reachability $\log \text{tr} \mathbf{W}$ decrease falls from 12 (10 ms) to 9 (20 ms) to 7 (100 ms) to 3 (1 s), and by 1 s the pooled mean difference has reversed sign (pooled mean $\Delta = +0.22$, vs. -0.11 at 10 ms). Controllability behaves oppositely: the trace decrease not only persists but amplifies monotonically with horizon, as the longer integration window accumulates more of the within-target propagator relative to the bounded cross-area input.

The diagnostic correlations tell the same story from the explanatory side. At 10–20 ms, the change in reachability tracks the change in cross-area coupling magnitude ($\Delta \|\mathbf{J}_{\text{block}}\|$, $\rho = +0.56$), coupling dimensionality ($\Delta \text{PR}(\sigma^2)$, $\rho = -0.31$), and source-area stability ($\Delta \text{source-area } \lambda$, $\rho = -0.36$). As the horizon lengthens, the source-area- λ and participation-ratio associations vanish, and the only correlate that remains stable across all four horizons is coupling magnitude ($\rho = +0.56, +0.57, +0.48, +0.47$). A target-area- λ association emerges at the longer horizons but does not reach significance ($\rho = +0.45$ at 100 ms, $p = 0.08$). The explanatory weight thus migrates from the input coupling toward the target area’s own dynamics as the integration window grows.

Together, these trends indicate that the anesthesia-driven reachability reorganization reported in the chapter is a short-horizon, coupling-governed effect. Over longer windows, the within-area state-transition propagator $\Phi^\alpha \sim \exp(\mathbf{J}^{\alpha \rightarrow \alpha} t)$ (the intrinsic dynamics whose leading Lyapunov exponents are characterized in §D.7) increasingly dominates the Gramian over the cross-area input term, reshaping the apparent reachability landscape (the sign reversal at 1 s). At the same time, the destabilized anesthetic state drives the controllability ease even further down over long horizons. Crucially, the increased reachability capability under anesthesia is past the timescale on which one area can steer another before the target’s own dynamics carry its state away: by the point at which intrinsic dynamics set the reachability landscape, the window for directed interareal control has likely effectively closed. The reachability disruption (reduced reachability under anesthesia, set by the interareal coupling) is the one visible at the short, functionally relevant horizon.

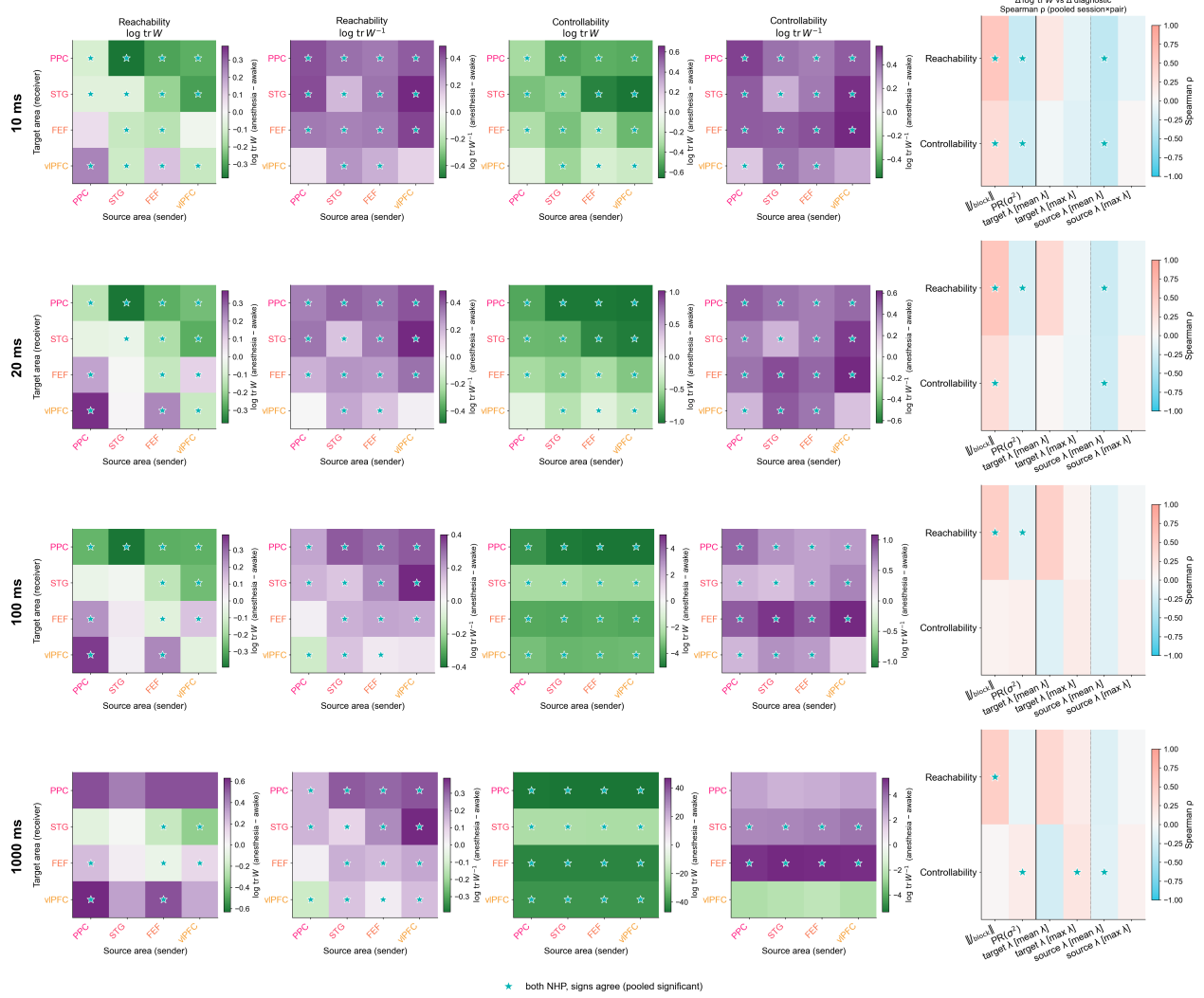


Figure D.2: Interareal control summaries across integration horizons ($n = 21$ sessions). Rows: gramian integration horizon (10, 20, 100, 1000 ms). Columns 1–4: 4×4 source–target heatmaps of pooled anesthesia–awake differences in reachability $\log \text{tr } \mathbf{W}$ and $\log \text{tr } \mathbf{W}^{-1}$ and controllability $\log \text{tr } \mathbf{W}$ and $\log \text{tr } \mathbf{W}^{-1}$; teal stars mark cells passing the sign-agree + BH-FDR gate at $q < 0.05$ (same convention as Fig. 4.5A,B). Column 5: Spearman correlation between the per-pair change in Gramian $\log \text{tr } \mathbf{W}$ and the change in block-norm, participation ratio, and source- and target-area Lyapunov exponents, with session-clustered permutation testing (same construction as Fig. 4.5E), computed at that horizon. The 10 ms row reproduces the main-text summaries.

D.2 Proofs

D.2.1 Jacobian-parameterized time derivative

The trick that makes JacobianODE estimation backpropagate cleanly through $\hat{\mathbf{J}}^\theta$ rather than through a separately-parameterized $\mathbf{f}$ is to express the instantaneous time derivative as a functional of the

Jacobian alone, evaluated along a smooth path between two observed states. By the fundamental theorem of calculus along the trajectory,

$$\mathbf{x}(t) - \mathbf{x}(t_0) = \int_{t_0}^t \mathbf{f}(\mathbf{x}(s)) ds. \quad (\text{D.4})$$

The conservativity-induced path-integral identity (Eq. 4.5) lets us replace each integrand $\mathbf{f}(\mathbf{x}(s))$ by $\mathbf{f}(\mathbf{x}(t))$ minus the path integral of $\mathbf{J}$ from $\mathbf{x}(s)$ to $\mathbf{x}(t)$, and solving the resulting linear relation for $\mathbf{f}(\mathbf{x}(t))$ gives the central proposition used throughout JacobianODE training.

Proposition 2 (Jacobian-parameterized $\mathbf{f}$). *Given observed states $\mathbf{x}(t_0)$ and $\mathbf{x}(t)$ along a trajectory of $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ with continuous Jacobian $\mathbf{J} = \nabla \mathbf{f}$, there exists a functional $\mathcal{F}$ of $\mathbf{J}$ alone such that*

$$\mathcal{F}[\mathbf{J}; \mathbf{x}(t_0), \mathbf{x}(t), t_0, t] = \mathbf{f}(\mathbf{x}(t)). \quad (\text{D.5})$$

Explicitly,

$$\boxed{\mathcal{F}[\mathbf{J}; \mathbf{x}(t_0), \mathbf{x}(t), t_0, t] = \frac{\mathbf{x}(t) - \mathbf{x}(t_0) + G(t_0, t; \mathbf{J})}{t - t_0}} \quad (\text{D.6})$$

where

$$G(t_0, t; \mathbf{J}) = \int_{t_0}^t \int_s^t \mathbf{J}(\mathbf{c}_{s,t}(r)) \mathbf{c}'_{s,t}(r) dr ds, \quad (\text{D.7})$$

and $\mathbf{c}_{s,t} : [s, t] \rightarrow \mathbb{R}^d$ is any smooth curve with $\mathbf{c}_{s,t}(s) = \mathbf{x}(s)$ and $\mathbf{c}_{s,t}(t) = \mathbf{x}(t)$.

Proof. By the fundamental theorem of calculus along the trajectory,

$$\mathbf{x}(t) - \mathbf{x}(t_0) = \int_{t_0}^t \mathbf{f}(\mathbf{x}(s)) ds. \quad (\text{D.8})$$

By the row-conservativity path-integral identity, for each $s \in [t_0, t]$,

$$\mathbf{f}(\mathbf{x}(s)) = \mathbf{f}(\mathbf{x}(t)) - \int_s^t \mathbf{J}(\mathbf{c}_{s,t}(r)) \mathbf{c}'_{s,t}(r) dr. \quad (\text{D.9})$$

Substituting Eq. D.9 into Eq. D.8 and pulling $\mathbf{f}(\mathbf{x}(t))$ outside the now-trivial s -integral yields

$$\mathbf{x}(t) - \mathbf{x}(t_0) = (t - t_0) \mathbf{f}(\mathbf{x}(t)) - \underbrace{\int_{t_0}^t \int_s^t \mathbf{J}(\mathbf{c}_{s,t}(r)) \mathbf{c}'_{s,t}(r) dr ds}_{G(t_0, t; \mathbf{J})}. \quad (\text{D.10})$$

Rearranging gives $\mathbf{f}(\mathbf{x}(t)) = [\mathbf{x}(t) - \mathbf{x}(t_0) + G(t_0, t; \mathbf{J})] / (t - t_0)$. $\square$

Remark. Equation D.6 is implemented in `JacobianODE/jacobians/jacobianODE.py`: the double integral G uses a straight-line path $\mathbf{c}_{s,t}(r) = \mathbf{x}(s) + (r - s)(\mathbf{x}(t) - \mathbf{x}(s)) / (t - s)$ with the trapezoid rule from `torchquad` (N=20 subdivisions); the resulting estimate of $\mathbf{f}(\mathbf{x}(t))$ is then advanced one step by a standard ODE integrator (default: RK4 from `torchdiffeq`). Because $\mathcal{F}$ depends only on $\mathbf{J}$, all gradients backpropagate through $\hat{\mathbf{J}}^\theta$, and no separate $\mathbf{f}$ -network is trained.

D.2.2 Invariance of the pairwise reachability Gramian trace under isometric direct-sum diffeomorphism

The pairwise reachability Gramian quantifies how strongly source area β can drive target area α along the reference trajectory. We show that when the encoder respects the subsystem product structure—that is, each subsystem’s latent coordinates depend only on that subsystem’s observed coordinates—and is an isometry on each subsystem factor, the pairwise reachability Gramian computed in the latent is conjugate by an orthogonal map to the Gramian one would compute on the (unknown) data side. Hence the trace and the full eigenvalue spectrum are exactly preserved.

Setup. Let $\mathcal{M} = \mathcal{M}^1 \times \dots \times \mathcal{M}^K$ and $\widetilde{\mathcal{M}} = \widetilde{\mathcal{M}}^1 \times \dots \times \widetilde{\mathcal{M}}^K$ be Riemannian product manifolds with product metrics $g_{\mathcal{M}} = \bigoplus_k g_{\mathcal{M}}^k$ and $g_{\widetilde{\mathcal{M}}} = \bigoplus_k g_{\widetilde{\mathcal{M}}}^k$. Assume the encoder is a *direct sum of diffeomorphisms*, with each area’s encoded coordinates depending only on that area’s observed coordinates—this is precisely the per-subsystem encoder construction of the Diffeomorphic-encoder subsection of the chapter Methods section, $\mathbf{z} = \Psi(\tilde{\mathbf{x}}) = [\Psi^1(\tilde{\mathbf{x}}^1); \dots; \Psi^K(\tilde{\mathbf{x}}^K)]$,

$$\Psi(\tilde{\mathbf{x}}^1, \dots, \tilde{\mathbf{x}}^K) = (\Psi^1(\tilde{\mathbf{x}}^1), \dots, \Psi^K(\tilde{\mathbf{x}}^K)), \quad (\text{D.11})$$

with each Ψ^k an isometric diffeomorphism. Then Ψ is itself an isometry, and its differential is block-diagonal:

$$D\Psi(\tilde{\mathbf{x}}) = D\Psi^1(\tilde{\mathbf{x}}^1) \oplus \dots \oplus D\Psi^K(\tilde{\mathbf{x}}^K). \quad (\text{D.12})$$

We will write $D\Psi_t^\alpha \triangleq D\Psi^\alpha(\tilde{\mathbf{x}}^\alpha(t))$ for short, and similarly for the other subsystems. All adjoints $(\cdot)^*$ below are taken with respect to the metric on each side. We write $\mathbf{J}^{\beta \rightarrow \alpha}(\tilde{\mathbf{x}})$ for the data-side cross-area Jacobian block (rows of area α , columns of area β), $\widetilde{\mathbf{J}}^{\beta \rightarrow \alpha}(\mathbf{z})$ for its latent-side counterpart, $\Phi^\alpha(t_1, \tau)$ for the within- α state-transition operator under $\mathbf{J}^{\alpha \rightarrow \alpha}$, and $\widetilde{\Phi}^\alpha(t_1, \tau)$ for its latent counterpart. Following the pairwise tangent-space equation in §4.3,

$$\mathbf{W}_r^{\beta \rightarrow \alpha}(t_0, t_1) = \int_{t_0}^{t_1} \Phi^\alpha(t_1, \tau) \mathbf{J}^{\beta \rightarrow \alpha}(\tilde{\mathbf{x}}(\tau)) [\mathbf{J}^{\beta \rightarrow \alpha}(\tilde{\mathbf{x}}(\tau))]^* [\Phi^\alpha(t_1, \tau)]^* d\tau, \quad (\text{D.13})$$

and analogously $\widetilde{\mathbf{W}}_r^{\beta \rightarrow \alpha}(t_0, t_1)$ with the tilde objects.

Theorem 1 (Spectral invariance of the pairwise reachability Gramian). *Under the setup above, for any pair $(\alpha, \beta) \in \{1, \dots, K\}^2$ with $\alpha \neq \beta$, the pairwise reachability Gramian on the latent is conjugate to the pairwise reachability Gramian on the data by an orthogonal map at the final time:*

$$\widetilde{\mathbf{W}}_r^{\beta \rightarrow \alpha}(t_0, t_1) = D\Psi_{t_1}^\alpha \mathbf{W}_r^{\beta \rightarrow \alpha}(t_0, t_1) (D\Psi_{t_1}^\alpha)^{-1}. \quad (\text{D.14})$$

In particular,

$$\text{tr } \widetilde{\mathbf{W}}_r^{\beta \rightarrow \alpha}(t_0, t_1) = \text{tr } \mathbf{W}_r^{\beta \rightarrow \alpha}(t_0, t_1), \quad (\text{D.15})$$

and the full eigenvalue spectrum—including $\lambda_{\min}$ —is preserved.

Proof. Each Ψ^k is an isometry, so $D\Psi^k$ is orthogonal:

$$(D\Psi^k)^* = (D\Psi^k)^{-1} \quad \text{for all } k \in \{1, \dots, K\}. \quad (\text{D.16})$$

The block-diagonal structure of Eq. D.12 means that differentiating $D\Psi^\alpha(\tilde{\mathbf{x}}^\alpha)$ with respect to $\tilde{\mathbf{x}}^\beta$ for $\beta \neq \alpha$ gives zero, so the chain rule applied to the cross-area block of the Jacobian yields a clean conjugation *between* the two subsystem factors, with no $D^2\Psi$ correction term:

$$\tilde{\mathbf{J}}^{\beta \rightarrow \alpha}(\mathbf{z}) = D\Psi^\alpha(\tilde{\mathbf{x}}^\alpha) \mathbf{J}^{\beta \rightarrow \alpha}(\tilde{\mathbf{x}}) (D\Psi^\beta(\tilde{\mathbf{x}}^\beta))^{-1}. \quad (\text{D.17})$$

Taking the adjoint of Eq. D.17 and using orthogonality,

$$[\tilde{\mathbf{J}}^{\beta \rightarrow \alpha}(\mathbf{z})]^* = D\Psi^\beta(\tilde{\mathbf{x}}^\beta) [\mathbf{J}^{\beta \rightarrow \alpha}(\tilde{\mathbf{x}})]^* (D\Psi^\alpha(\tilde{\mathbf{x}}^\alpha))^{-1}. \quad (\text{D.18})$$

So the β -side factors collapse in the symmetric product:

$$\tilde{\mathbf{J}}^{\beta \rightarrow \alpha} [\tilde{\mathbf{J}}^{\beta \rightarrow \alpha}]^* = D\Psi^\alpha \mathbf{J}^{\beta \rightarrow \alpha} [\mathbf{J}^{\beta \rightarrow \alpha}]^* (D\Psi^\alpha)^{-1}. \quad (\text{D.19})$$

The within- α state-transition operator transforms similarly: from $\delta \mathbf{z}^\alpha(t) = D\Psi_t^\alpha \delta \tilde{\mathbf{x}}^\alpha(t)$ and the definition of Φ^α ,

$$\tilde{\Phi}^\alpha(t_1, \tau) = D\Psi_{t_1}^\alpha \Phi^\alpha(t_1, \tau) (D\Psi_\tau^\alpha)^{-1}, \quad (\text{D.20})$$

and analogously for the adjoint $[\tilde{\Phi}^\alpha(t_1, \tau)]^* = D\Psi_\tau^\alpha [\Phi^\alpha(t_1, \tau)]^* (D\Psi_{t_1}^\alpha)^{-1}$. Substituting Eq. D.19 and Eq. D.20 into the integrand of $\tilde{\mathbf{W}}_r^{\beta \rightarrow \alpha}$,

$$\begin{aligned} & \tilde{\Phi}^\alpha(t_1, \tau) \tilde{\mathbf{J}}^{\beta \rightarrow \alpha} [\tilde{\mathbf{J}}^{\beta \rightarrow \alpha}]^* [\tilde{\Phi}^\alpha(t_1, \tau)]^* \\ &= D\Psi_{t_1}^\alpha \left[\Phi^\alpha(t_1, \tau) \mathbf{J}^{\beta \rightarrow \alpha} [\mathbf{J}^{\beta \rightarrow \alpha}]^* [\Phi^\alpha(t_1, \tau)]^* \right] (D\Psi_{t_1}^\alpha)^{-1}, \end{aligned} \quad (\text{D.21})$$

where every intermediate $D\Psi_\tau^\alpha$ pair cancels: each $D\Psi_\tau^\alpha$ appears next to its inverse, and likewise for the $D\Psi_\tau^\beta$ pair that hid inside the symmetric product Eq. D.19. Integrating over τ and pulling the τ -independent outer factors out gives Eq. D.14. Trace and eigenvalues are invariant under conjugation by an invertible operator. $\square$

Remark. The proof extends pointwise to any number of subsystems $K \geq 2$, with the only requirement being that the encoder respect the subsystem product structure (each Ψ^k acts only on $\tilde{\mathbf{x}}^k$). It also extends to the controllability Gramian by replacing $\Phi^\alpha(t_1, \tau)$ with $\Phi^\alpha(t_0, \tau)$ throughout. The hypothesis that each Ψ^k is an isometry is stronger than what the NICE–Cayley encoder of §D.3.1 guarantees by construction (the encoder is volume-preserving but not in general isometric); we therefore present the reachability and controllability log-trace summaries reported in the chapter as invariants up to encoder-induced similarity, which is what their trace summaries measure. The metric-adjusted Gramian construction of §D.3.4 provides one route to mitigate the residual mismatch by re-weighting the input matrix by the inverse encoder Jacobian on the source-area input space.

D.2.3 Bounded error under generalized teacher forcing

Generalized teacher forcing (GTF)³⁵⁴ stabilizes recursive prediction by partially anchoring the rollout to the true trajectory. The strength of anchoring is a scalar $\alpha \in [0, 1]$: the input to the next-step network at time t is the convex combination

$$\tilde{\mathbf{x}}_t = \alpha \mathbf{x}_t + (1 - \alpha) \hat{\mathbf{x}}_t, \quad \tilde{\mathbf{x}}_{t+1} = \hat{F}(\tilde{\mathbf{x}}_t), \quad (\text{D.22})$$

where $\mathbf{x}_t$ is the true state, $\hat{\mathbf{x}}_t$ is the (potentially drifting) recursive prediction, and $\hat{F}$ is the learned one-step update. $\alpha = 1$ recovers full teacher forcing; $\alpha = 0$ recovers free-running prediction.

Lemma 1 (Asymptotic bound on the product of one-step generators). *For every $\epsilon > 0$ there exists $C \in \mathbb{R}_+$ such that for all $T \geq 1$,*

$$\left\| \prod_{t=0}^{T-1} e^{\mathbf{J}_{T-t}\Delta t} \right\|_2 \leq C (e^{\lambda_{\max} + \epsilon})^T, \quad (\text{D.23})$$

where $\lambda_{\max} = \lim_{T \rightarrow \infty} \frac{1}{T} \log \|\mathcal{G}_T\|_2$ is the maximum Lyapunov exponent of the linearized system along the trajectory and $\mathcal{G}_T = \prod_{t=0}^{T-1} e^{\mathbf{J}_{T-t}\Delta t}$.

Proof. By the definition of the maximum Lyapunov exponent, for every $\epsilon > 0$ there exists N such that for all $T > N$, $\frac{1}{T} \log \|\mathcal{G}_T\|_2 < \lambda_{\max} + \epsilon$, which rearranges to

$$\|\mathcal{G}_T\|_2 < e^{T(\lambda_{\max} + \epsilon)} = (e^{\lambda_{\max} + \epsilon})^T. \quad (\text{D.24})$$

For the pre-asymptotic regime $1 \leq T \leq N$, the sequence consists of finitely many bounded matrix products, so there exists $C \in \mathbb{R}_+$ such that $\|\mathcal{G}_T\|_2 \leq C(e^{\lambda_{\max} + \epsilon})^T$ for all $1 \leq T \leq N$. Taking the larger of the two constants establishes the claim. $\square$

Proposition 3 (Bounded error under GTF; cf. ³⁵⁴). *Let $\hat{F}$ be the learned one-step update with bounded one-step approximation error, $\|\hat{F}(\mathbf{x}_t) - F(\mathbf{x}_t)\|_2 \leq L$ for some $L \in \mathbb{R}_+$, and let the GTF strength satisfy*

$$\alpha > 1 - e^{-\lambda_{\max}}. \quad (\text{D.25})$$

Then the prediction error $\mathbf{e}_T = \hat{\mathbf{x}}_T - \mathbf{x}_T$ remains bounded for all T .

Proof. Define $\delta_t = \hat{F}(\mathbf{x}_t) - F(\mathbf{x}_t)$ (one-step model error). Note that

$$\tilde{\mathbf{x}}_t = \alpha \mathbf{x}_t + (1 - \alpha) \hat{\mathbf{x}}_t = \alpha \mathbf{x}_t + (1 - \alpha)(\mathbf{x}_t + \mathbf{e}_t) = \mathbf{x}_t + (1 - \alpha) \mathbf{e}_t. \quad (\text{D.26})$$

The error propagates as

$$\mathbf{e}_{t+1} = \hat{F}(\tilde{\mathbf{x}}_t) - F(\mathbf{x}_t) \quad (\text{D.27})$$

$$= [\hat{F}(\tilde{\mathbf{x}}_t) - \hat{F}(\mathbf{x}_t)] + [\hat{F}(\mathbf{x}_t) - F(\mathbf{x}_t)] \quad (\text{D.28})$$

$$= [\hat{F}(\mathbf{x}_t + (1 - \alpha)\mathbf{e}_t) - \hat{F}(\mathbf{x}_t)] + \delta_t. \quad (\text{D.29})$$

Linearizing $\hat{F}$ about $\mathbf{x}_t$ (which is valid for sufficiently small $\mathbf{e}_t$),

$$\mathbf{e}_{t+1} \approx e^{\mathbf{J}_t \Delta t} (1 - \alpha) \mathbf{e}_t + \delta_t, \quad (\text{D.30})$$

where $\mathbf{J}_t = \nabla \hat{F}(\mathbf{x}_t)/\Delta t$ in the continuous-time interpretation (or simply the one-step linearization in the discrete formulation). This recursion has solution

$$\mathbf{e}_T = \sum_{k=0}^{T-1} (1 - \alpha)^{T-1-k} \left(\prod_{t=k+1}^{T-1} e^{\mathbf{J}_{T-t}\Delta t} \right) \delta_k, \quad (\text{D.31})$$

with norm bound

$$\|\mathbf{e}_T\|_2 \leq \sum_{k=0}^{T-1} (1-\alpha)^{T-1-k} \left\| \prod_{t=k+1}^{T-1} e^{\mathbf{J}_{T-t}\Delta t} \right\|_2 \|\delta_k\|_2. \quad (\text{D.32})$$

Using $\|\delta_k\|_2 \leq L$ and Lemma 1,

$$\|\mathbf{e}_T\|_2 \leq LC \sum_{k=0}^{T-1} [(1-\alpha)e^{\lambda_{\max}}]^{T-1-k} = LC \sum_{j=0}^{T-1} [(1-\alpha)e^{\lambda_{\max}}]^j. \quad (\text{D.33})$$

The geometric series on the right converges as $T \rightarrow \infty$ iff $(1-\alpha)e^{\lambda_{\max}} < 1$, which is equivalent to Eq. D.25. Under this condition, $\|\mathbf{e}_T\|_2 \leq LC/[1 - (1-\alpha)e^{\lambda_{\max}}]$ uniformly in T . $\square$

Remark. The bound is tight in the sense that for $\alpha < 1 - e^{-\lambda_{\max}}$ the geometric series diverges and the error grows exponentially; the threshold $\alpha^* = 1 - e^{-\lambda_{\max}}$ is therefore the minimum GTF strength that ensures bounded recursive prediction in the linearized regime. In practice we use $\alpha_{\text{teacher forcing}} \in [\alpha_{\min}, 1]$ annealed over training (§D.3.2); both the per-rollout maximum effective $\lambda_{\max}$ (via the learned Jacobians along the reference trajectory) and the geometric series prefactor enter directly into the validity of the bound.

D.3 Modeling details

D.3.1 Diffeomorphic encoder

Delay embedding

For each subsystem $k \in \{1, \dots, K\}$, the per-subsystem observation $\mathbf{x}^k(t) \in \mathbb{R}^{d_{\text{obs}}^k}$ is stacked with n_d delayed copies at unit stride ($\tau\Delta t = \Delta t$):

$$\tilde{\mathbf{x}}^k(t) = [\mathbf{x}^k(t); \mathbf{x}^k(t - \Delta t); \dots; \mathbf{x}^k(t - (n_d - 1)\Delta t)] \in \mathbb{R}^{n_d d_{\text{obs}}^k}. \quad (\text{D.34})$$

The full delay-embedded observation is the concatenation $\tilde{\mathbf{x}}(t) = [\tilde{\mathbf{x}}^1(t); \dots; \tilde{\mathbf{x}}^K(t)] \in \mathbb{R}^{d'}$ with $d' = n_d \sum_k d_{\text{obs}}^k$, ordered most-recent-first within each per-subsystem block. The choice of n_d is a per-session sweep variable for the neural data. In practice, we take $\tau\Delta t = \Delta t$ which avoids the dimensionality blow-up of larger strides. For the propofol cohort, $n_d \in \{5, 10, 15\}$ was searched per session and per area (the per-area block size is $n_d \cdot 64$ for the typical 64-channel block).

Additive coupling layers

Each additive coupling layer⁴¹¹ acts on $\mathbf{y} \in \mathbb{R}^n$ by an even/odd channel split $\mathbf{y} = (\mathbf{a}, \mathbf{b})$ with $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{n/2}$ and applies

$$\mathbf{a}' = \mathbf{a}, \quad \mathbf{b}' = \mathbf{b} + \mathbf{m}_{\vartheta}(\mathbf{a}, \mathbf{c}), \quad (\text{D.35})$$

where $\mathbf{m}_\vartheta : \mathbb{R}^{n/2} \times \mathbb{R}^{d_c} \rightarrow \mathbb{R}^{n/2}$ is a conditioner MLP and $\mathbf{c}$ is a (per-trial, per-batch) condition vector (binary $\{-1, +1\}$ for awake-vs-anesthesia). The layer Jacobian is

$$\frac{\partial(\mathbf{a}', \mathbf{b}')}{\partial(\mathbf{a}, \mathbf{b})} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \nabla_{\mathbf{a}} \mathbf{m}_\vartheta & \mathbf{I} \end{bmatrix}, \quad (\text{D.36})$$

which is unit-triangular: $\det = 1$ identically, regardless of $\mathbf{m}_\vartheta$, $\mathbf{a}$, or $\mathbf{c}$. The layer is exactly invertible by $\mathbf{a} = \mathbf{a}'$, $\mathbf{b} = \mathbf{b}' - \mathbf{m}_\vartheta(\mathbf{a}', \mathbf{c})$ in one forward pass.

Cayley orthogonal mixers

Between consecutive coupling layers we insert a learnable rotation $\mathbf{Q} \in \text{SO}(n)$ parameterized by the Cayley transform of a learnable skew-symmetric matrix $\mathbf{A} = -\mathbf{A}^\top \in \mathbb{R}^{n \times n}$:

$$\mathbf{Q} = (\mathbf{I} - \mathbf{A})(\mathbf{I} + \mathbf{A})^{-1}. \quad (\text{D.37})$$

Lemma 2. *For any skew-symmetric $\mathbf{A}$ with $-1 \notin \text{spec}(\mathbf{A})$, $\mathbf{Q}$ is orthogonal with $\det \mathbf{Q} = +1$.*

Proof. $\mathbf{Q}^\top \mathbf{Q} = (\mathbf{I} + \mathbf{A})^{-\top} (\mathbf{I} - \mathbf{A})^\top (\mathbf{I} - \mathbf{A})(\mathbf{I} + \mathbf{A})^{-1} = (\mathbf{I} - \mathbf{A})^{-1} (\mathbf{I} + \mathbf{A})(\mathbf{I} - \mathbf{A})(\mathbf{I} + \mathbf{A})^{-1} = \mathbf{I}$ since $(\mathbf{I} - \mathbf{A})$ and $(\mathbf{I} + \mathbf{A})$ share an eigenbasis (commute) and cancel pairwise. The eigenvalues of $\mathbf{A}$ are purely imaginary in conjugate pairs $\pm i\omega_k$, so the eigenvalues of $\mathbf{Q}$ are $(1 - i\omega_k)/(1 + i\omega_k)$ —complex of unit modulus in conjugate pairs—and their product is $+1$. $\square$

Implementation uses `torch.linalg.solve(I-A, I+A)` rather than an explicit inverse for stability. The skew parameter is built from the strict upper triangle of a free parameter $\mathbf{U} \in \mathbb{R}^{n \times n}$ via $\mathbf{A} = \text{striu}(\mathbf{U}) - \text{striu}(\mathbf{U})^\top$ (the lower triangle and diagonal of $\mathbf{U}$ are unused, and the operation projects to the skew subspace automatically).

Volume-preserving composition and identity-at-initialization

Let $\Psi^k = \mathbf{Q}_L^k \circ \mathcal{C}_L^k \circ \dots \circ \mathbf{Q}_1^k \circ \mathcal{C}_1^k$ denote the per-subsystem encoder (composition of L additive coupling layers $\mathcal{C}_\ell^k$ with Cayley mixers $\mathbf{Q}_\ell^k$). By the chain rule and Lemma 2,

$$\det \nabla \Psi^k(\mathbf{y}) = \prod_{\ell=1}^L \det \nabla \mathbf{Q}_\ell^k \cdot \det \nabla \mathcal{C}_\ell^k(\mathbf{z}_\ell^k) = \prod_{\ell=1}^L (+1) \cdot (+1) = +1, \quad (\text{D.38})$$

so $|\det \nabla \Psi^k| \equiv 1$ everywhere on its domain, and each Ψ^k is a C^∞ diffeomorphism on $\mathbb{R}^{n_d d_{\text{obs}}^k}$. The full encoder $\Psi = (\Psi^1, \dots, \Psi^K)$ inherits these properties on $\mathbb{R}^{d'}$ and respects the subsystem product structure required by Theorem 1.

Identity-at-initialization. We zero-initialize the last linear layer of every conditioner MLP, which makes $\mathbf{m}_\vartheta(\cdot, \cdot) \equiv \mathbf{0}$ at step 0; every coupling layer is then the identity map. We zero-initialize every

skew parameter $\mathbf{U} = \mathbf{0}$, which gives $\mathbf{Q} = \mathbf{I}$. A final fixed permutation is appended whose composition with the random per-layer permutations equals the identity, so the entire encoder reduces to $\Psi = \text{id}$ at step 0. Training therefore departs from the trivial identity-reconstruction solution only as the trajectory and loop-closure losses begin to drive learning, which empirically gives the LR schedule a clean reconstruction-floor starting point.

Zero-padded reconstruction loss

The encoder is dimension-preserving on $\mathbb{R}^{d'}$, but the intrinsic dimensionality of the dynamics is much smaller. The latent is partitioned as $\mathbf{z} = [\mathbf{z}_{\text{dyn}}; \mathbf{z}_{\text{null}}] \in \mathbb{R}^{d_{\text{dyn}}} \oplus \mathbb{R}^{d_{\text{null}}}$ with $d_{\text{null}} = d' - d_{\text{dyn}}$. The dynamic-subspace dimension d_{dyn} is fixed per session by a PCA on the delay-embedded training data at variance threshold 0.99 (per-area block selection; the resulting d_{dyn}^k values are concatenated to define the projection $P_{\text{dyn}} = \text{diag}(\mathbf{I}_{d_{\text{dyn}}}, \mathbf{0}_{d_{\text{null}}})$). All decoder calls in the model use the zero-padded latent $[P_{\text{dyn}}\mathbf{z}; \mathbf{0}]$ rather than $\mathbf{z}$ itself, so any information that lives in $\mathbf{z}_{\text{null}}$ is ignored on the round-trip. The reconstruction loss (Eq. 4.10) therefore takes the form

$$\mathcal{L}_{\text{rec}}(\theta_{\Psi}; \mathbf{x}) = \mathbb{E}_t \left[\left\| \Psi^{-1}([P_{\text{dyn}}\Psi(\tilde{\mathbf{x}}(t)); \mathbf{0}]) - \tilde{\mathbf{x}}(t) \right\|_2^2 \right], \quad (\text{D.39})$$

penalizing the round-trip error under the zero-padded forward map. Since Ψ is volume-preserving, every information-bearing dimension of the input *must* survive somewhere in $\mathbf{z}$; the dynamic/null split together with the zero-padded round-trip squeeze incentivizes the encoder to concentrate the dynamics-relevant information into $\mathbf{z}_{\text{dyn}}$ and dump dynamics-irrelevant content (noise, off-manifold variability) into $\mathbf{z}_{\text{null}}$, where it cannot interfere with the Jacobian-MLP fit.

Induced Riemannian metric on the latent

The encoder Ψ pulls the (assumed-Euclidean) data-side metric on $\tilde{\mathbf{x}}$ forward to a Riemannian metric on $\mathbf{z}$:

$$\mathbf{R}(\mathbf{z}) = [\nabla \Psi^{-1}(\mathbf{z})]^\top [\nabla \Psi^{-1}(\mathbf{z})] = [D\Psi(\Psi^{-1}(\mathbf{z})) D\Psi(\Psi^{-1}(\mathbf{z}))^\top]^{-1}. \quad (\text{D.40})$$

Equivalently, in encoder-Jacobian form, $\mathbf{R}^{-1}(\mathbf{z}) = \mathbf{J}_E(\tilde{\mathbf{x}}) \mathbf{J}_E(\tilde{\mathbf{x}})^\top$ with $\mathbf{J}_E = D\Psi$ evaluated at $\tilde{\mathbf{x}} = \Psi^{-1}(\mathbf{z})$. Volume preservation gives $\det \mathbf{R} \equiv 1$, so the induced metric has unit determinant everywhere; isotropy of $\mathbf{R}$ (i.e., $\mathbf{R} \equiv \mathbf{I}$) corresponds exactly to the encoder being a Riemannian isometry, which is the hypothesis of Theorem 1. In practice, while we don't enforce isometry, we find that volume preservation is sufficient to reproduce relative changes in the Gramians on the synthetic systems.

Encoder geometry diagnostics: rotation angle, Haar reference, and geodesic distance

Two scalar diagnostics summarize how strongly the trained encoder rotates and warps the latent geometry (Figure 4.4B,C).

Per-layer rotation angle. Each Cayley mixer $\mathbf{Q}_\ell^k \in \text{SO}(n)$ (§D.3.1) has eigenvalues in unit-modulus conjugate pairs $e^{\pm i\theta_m}$ with $\theta_m = 2 \arctan \omega_m$, where $\pm i\omega_m$ are the purely imaginary eigenvalues of its skew generator $\mathbf{A}_\ell^k$. We summarize the rotation of a single mixer by the mean over modes of the absolute eigen-angle,

$$\bar{\theta}(\mathbf{Q}) = \frac{1}{n} \sum_{m=1}^n |\arg \lambda_m(\mathbf{Q})| \in [0, \pi], \quad (\text{D.41})$$

computed numerically as the mean of $|\text{angle}(\text{eigvals}(\mathbf{Q}))|$; $\bar{\theta} = 0$ is the identity (no mixing). Figure 4.4B reports this per Cayley mixer index ℓ through encoder depth, both for the individual mixers $\mathbf{Q}_\ell^k$ and for the cumulative composite rotation $\mathbf{Q}_\ell^k \cdots \mathbf{Q}_1^k$ (the net rotation accrued up to depth ℓ), averaged over the per-area blocks k .

Haar $\text{SO}(n)$ reference. To calibrate the rotation-angle scale we compare against the angle a *maximally mixed* rotation would produce: a matrix drawn uniformly from $\text{SO}(n)$ under the Haar measure. We sample $\mathbf{G} \sim \mathcal{N}(0, 1)^{n \times n}$, take its QR decomposition $\mathbf{G} = \mathbf{Q}\mathbf{R}$, fix the column signs by $\mathbf{Q} \leftarrow \mathbf{Q} \text{diag}(\text{sgn}(\text{diag } \mathbf{R}))$ (which makes the draw exactly Haar-uniform on $\text{O}(n)$), and negate the first column whenever $\det \mathbf{Q} < 0$ to land on the $\text{SO}(n)$ component. The reference is $\bar{\theta}(\mathbf{Q})$ of Eq. D.41 averaged over many such draws and over the per-area block dimensions n . The trained mixers sit between the identity ($\bar{\theta} = 0$) and this Haar saturation level, quantifying how far each layer has rotated relative to a fully randomized rotation of the same dimension.

Geodesic distance of the induced metric to the identity. The induced metric $\mathbf{R}(\mathbf{z})$ of §D.3.1 (equivalently its inverse $\mathbf{G} = \mathbf{J}_E \mathbf{J}_E^\top$) is symmetric positive definite (SPD). The natural distance between SPD matrices is the affine-invariant Riemannian (geodesic) distance, whose value relative to the identity is

$$d(\mathbf{I}, \mathbf{G}) = \|\log \mathbf{G}\|_F = \left(\sum_m (\log \lambda_m(\mathbf{G}))^2 \right)^{1/2}, \quad (\text{D.42})$$

where $\lambda_m(\mathbf{G})$ are the eigenvalues of $\mathbf{G}$ and $\log$ is the matrix logarithm; this is invariant to the choice of $\mathbf{R}$ versus $\mathbf{R}^{-1} = \mathbf{G}$ since $\log \mathbf{R}^{-1} = -\log \mathbf{G}$. Writing $\lambda_m(\mathbf{G}) = \sigma_m^2$ in terms of the singular values σ_m of the composite encoder Jacobian $\mathbf{J}_E$ along the d_{dyn} retained (dynamic-subspace) modes, $\log \lambda_m = 2 \log \sigma_m$. We report the per-mode root-mean-square normalization of Eq. D.42 (dividing the Frobenius norm by $\sqrt{d_{\text{dyn}}}$ so the scale is $O(1)$ in log-stretch units),

$$d_{\text{rms}} = \frac{\|\log \mathbf{G}\|_F}{\sqrt{d_{\text{dyn}}}} = 2 \sqrt{\frac{1}{d_{\text{dyn}}} \sum_{m=1}^{d_{\text{dyn}}} (\log \sigma_m)^2}, \quad (\text{D.43})$$

averaged over trajectory samples within each session and condition. $d_{\text{rms}} = 0$ exactly when all $\sigma_m = 1$, i.e. when $\mathbf{J}_E$ is orthogonal and the encoder is a local isometry (the hypothesis of Theorem 1); larger values indicate a more anisotropic, more strongly warped latent metric. Because the encoder is volume-preserving ($|\det \mathbf{J}_E| \equiv 1$, so the full set of log-singular-values sums to zero), d_{rms} reflects the anisotropy of the metric—the spread of its singular values—rather than an overall change of scale. Figure 4.4C reports the per-session d_{rms} , awake vs. anesthesia.

D.3.2 JacobianODE-parameterized dynamics

Generating predicted states with the Jacobian

The latent JacobianODE rollout is generated as follows. Given an observed conditioning window of length T_{init} (Methods: $T_{\text{init}} = 15$), the latent states $\mathbf{z}(t - (T_{\text{init}} - 1)\Delta t), \dots, \mathbf{z}(t)$ are encoded once and held fixed. The endpoint of this window, $\mathbf{z}_{\text{base}} \triangleq \mathbf{z}_{\text{dyn}}(t)$, serves as the *base point* of the rollout, and is the only point at which we invoke the double-integral functional of Proposition 2. Concretely, we compute once

$$\hat{\mathbf{f}}_{\text{base}} = \mathcal{F}[\hat{\mathbf{J}}^\theta; \mathbf{z}_{\text{dyn}}(t - (T_{\text{init}} - 1)\Delta t), \mathbf{z}_{\text{base}}, t - (T_{\text{init}} - 1)\Delta t, t], \quad (\text{D.44})$$

with the double integral G (Eq. D.7) approximated by the trapezoid rule on the straight-line path between window endpoints (`torchquad`, $N = 20$ subdivisions). All subsequent time-derivative evaluations along the rollout are then obtained by a *single* path integral from this base point, using the row-conservativity identity (Eq. 4.5):

$$\hat{\mathbf{f}}(\mathbf{z}) = \hat{\mathbf{f}}_{\text{base}} + \int_{\mathcal{C}_{\text{base} \rightarrow \mathbf{z}}} \hat{\mathbf{J}}^\theta ds, \quad (\text{D.45})$$

where $\mathcal{C}_{\text{base} \rightarrow \mathbf{z}}$ is the straight-line path from $\mathbf{z}_{\text{base}}$ to $\mathbf{z}$ (same trapezoid quadrature, $N = 20$). Concretely, at each prediction step $s = 1, \dots, T$:

1. Take the current predicted latent $\hat{\mathbf{z}}_{\text{dyn}}(t + (s - 1)\Delta t)$ (projected to the dynamic subspace if needed).
2. Evaluate the time derivative $\hat{\mathbf{f}}(\hat{\mathbf{z}}_{\text{dyn}}(t + (s - 1)\Delta t))$ via Eq. D.45: a single line integral of $\hat{\mathbf{J}}^\theta$ from $\mathbf{z}_{\text{base}}$ to the current state, added to the cached $\hat{\mathbf{f}}_{\text{base}}$.
3. Advance the latent one step by the RK4 integrator from `torchdiffeq`, producing $\hat{\mathbf{z}}_{\text{dyn}}(t + s\Delta t)$.

The number of prediction steps is T . We use $T = 30$ for most sample systems except for the partially observed task-trained RNN, for which we use $T = 20$ due to trajectory length constraints. For the neural data, we use $T = 10$ due to the challenging nature of the prediction task. The cumulative latent rollout $\{\hat{\mathbf{z}}_{\text{dyn}}(t + s\Delta t)\}_{s=1}^T$ is fed into the latent and decoded prediction losses below. The base point $\mathbf{z}_{\text{base}}$ and its cached derivative $\hat{\mathbf{f}}_{\text{base}}$ are *not* refreshed during the rollout: the double integral of Proposition 2 is paid exactly once per rollout, and every per-step time-derivative evaluation costs only one straight-line $\int \hat{\mathbf{J}}^\theta$ rather than a nested double integral. All gradients still backpropagate through θ because $\hat{\mathbf{f}}_{\text{base}}$ itself is a functional of $\hat{\mathbf{J}}^\theta$.

Generalized teacher forcing

The recursive rollout is partially anchored to the encoded ground-truth via the GTF input

$$\tilde{\mathbf{z}}_{\text{dyn}}(t + s\Delta t) = \alpha_{\text{TF}} \mathbf{z}_{\text{dyn}}(t + s\Delta t) + (1 - \alpha_{\text{TF}}) \hat{\mathbf{z}}_{\text{dyn}}(t + s\Delta t), \quad (\text{D.46})$$

which then feeds the next-step prediction $\hat{\mathbf{z}}_{\text{dyn}}(t + (s + 1)\Delta t) = \hat{F}(\tilde{\mathbf{z}}_{\text{dyn}}(t + s\Delta t))$. The mixing weight α_{TF} is annealed during training from $\alpha_{\text{init}} = 1$ (pure teacher forcing) downward toward the *theoretical bound* of Proposition 3,

$$\alpha_{\text{TF}}^* = 1 - e^{-\hat{\lambda}_{\text{max}}^{(\text{batch})}}, \quad (\text{D.47})$$

where $\hat{\lambda}_{\text{max}}^{(\text{batch})}$ is the maximum Lyapunov exponent of the model’s own linearized rollout, computed per training batch from the predicted Jacobians $\hat{\mathbf{J}}^\theta(\mathbf{z}(t_k))$ along each batch trajectory by the same discrete QR iteration used in §D.3.5. Because $\hat{\lambda}_{\text{max}}^{(\text{batch})}$ is recomputed at every batch from the current model and the current batch’s reference trajectories, the annealing target is *data-adaptive*: as the model learns more chaotic dynamics, $\hat{\lambda}_{\text{max}}^{(\text{batch})}$ grows, the bound α_{TF}^* rises toward 1, and the schedule keeps α_{TF} above the Proposition 3 threshold that guarantees bounded recursive error. Conversely, when the inferred dynamics are nearly stable, $\alpha_{\text{TF}}^* \rightarrow 0$ and the schedule permits aggressive annealing toward free-running prediction. The per-batch decay rate (`gamma_teacher_forcing`, default 0.999) controls how quickly α_{TF} approaches the current bound from above; a hard floor $\alpha_{\text{min}} = 10^{-3}$ is enforced for numerical safety so that α_{TF} never drops below this value even when the data-adaptive bound would permit it. At validation time, the trajectory `val_loss` is reported at $\alpha_{\text{TF}} = 0$ (free-running) and the one-step MASE is reported at $\alpha_{\text{TF}} = 1$ (teacher-forced).

Latent prediction loss

The latent prediction loss (Eq. 4.11) penalizes the deviation between the rolled-out latent and the encoded ground-truth latent at each prediction step:

$$\mathcal{L}_{\text{pred},\mathbf{z}}(\theta; \mathbf{x}) = \mathbb{E}_t \left[\sum_{s=1}^T \|\hat{\mathbf{z}}_{\text{dyn}}(t + s\Delta t) - \mathbf{z}_{\text{dyn}}(t + s\Delta t)\|_2^2 \right], \quad (\text{D.48})$$

where $\mathbf{z}_{\text{dyn}}(t + s\Delta t) = P_{\text{dyn}}\Psi(\tilde{\mathbf{x}}(t + s\Delta t))$ is the encoded ground-truth dynamic latent. In production we use this plain mean-squared error directly, without per-batch generalized-variance normalization or any other rescaling: because the encoder is volume-preserving (§D.3.1), the latent and the delay-embedded observation share an absolute scale, and the reconstruction, latent-prediction, and decoded-prediction MSE terms can be summed at equal weight (§D.3.6) without further normalization.

Decoded prediction loss

Eq. 4.12 introduces the decoded prediction loss, which compares the decoded prediction against the observed delay-embedded data, projected to the most-recent slice P_{recent} :

$$\mathcal{L}_{\text{pred},\mathbf{x}}(\theta; \mathbf{x}) = \mathbb{E}_t \left[\sum_{s=1}^T \|P_{\text{recent}}\Psi^{-1}([P_{\text{dyn}}\hat{\mathbf{z}}(t + s\Delta t); \mathbf{0}]) - \mathbf{x}(t + s\Delta t)\|_2^2 \right]. \quad (\text{D.49})$$

P_{recent} extracts the most-recent d_{obs} entries of the delay-embedded vector, eliminating redundant comparison against entries that overlap with the conditioning window. This loss closes the loop between the latent dynamics and the data on which they were defined: the encoder is forced to

produce a latent whose forward time evolution decodes back to a trajectory that matches the observation. Without this loss the encoder can drift to a latent on which the Jacobian-MLP fits well but whose decoder image is unconstrained.

D.3.3 Loop-closure loss

Implementation

Loop-closure regularizes $\hat{\mathbf{J}}^\theta$ to be row-conservative, exploiting the fact that the path integral of a gradient field along any closed contour vanishes. We sample loops in the dynamic-latent space as concatenations of random straight-line segments connecting randomly selected data points: for each batch we sample N_{loops} loops of $N_{\text{loop pts}} \geq 3$ data points each, closing the polygon by appending the first point to the end. The line integral along each segment is approximated by the same trapezoid rule used for the path integral G (`torchquad`, $N_{\text{interp}} = 20$ subdivisions per segment), and the regularizer is the squared 2-norm of the accumulated row vector across the loop:

$$\mathcal{L}_{\text{loop}}(\theta; \mathbf{x}) = \mathbb{E}_t \left[\left\| \oint_{\mathcal{C}_{\text{loop}}} \hat{\mathbf{J}}^\theta d\mathbf{s} \right\|_2^2 \right]. \quad (\text{D.50})$$

By default the loops are mixed across trajectories within a batch (`mix_trajectories=True`); turning this off restricts each loop to data points from a single trajectory, which we found to be slightly less effective at sampling tangent-space directions orthogonal to the local flow.

Consistency under diffeomorphism

A natural concern is that loop closure is defined in the latent rather than the data space. The following proposition verifies that the regularizer is well-posed in the sense that row-conservativity transfers across C^1 diffeomorphisms; in particular, training the latent Jacobian network against this loss is a consistent inductive bias, not a misspecified one.

Proposition 4 (Diffeomorphic transfer of row-conservativity). *Let $\mathbf{f} : \mathcal{M} \rightarrow \mathbb{R}^d$ be a C^1 vector field on an open set $\mathcal{M} \subseteq \mathbb{R}^d$ with Jacobian $\mathbf{J} = \nabla \mathbf{f}$, and let $\phi : \mathcal{M} \rightarrow \widetilde{\mathcal{M}}$ be a C^1 diffeomorphism with Jacobian $\mathbf{D} = \nabla \phi$. Define the pushforward field $\widetilde{\mathbf{f}} = \phi_* \mathbf{f} = \mathbf{D} \mathbf{f} \circ \phi^{-1}$ on $\widetilde{\mathcal{M}}$. Then for every i , the i -th row of $\widetilde{\mathbf{J}} = \nabla \widetilde{\mathbf{f}}$ is the gradient of a scalar potential on $\widetilde{\mathcal{M}}$ iff the corresponding row of $\mathbf{J}$ is on $\mathcal{M}$.*

Proof. On simply connected domains, a C^1 vector field is the gradient of a scalar potential iff its line integral along every closed contour vanishes (Poincaré lemma). Let $\mathcal{C} \subset \widetilde{\mathcal{M}}$ be a closed loop and let $\widetilde{\mathbf{r}} : [0, 1] \rightarrow \widetilde{\mathcal{M}}$ parameterize it. Pull back via ϕ^{-1} to a loop $\mathbf{r} = \phi^{-1} \circ \widetilde{\mathbf{r}}$ in $\mathcal{M}$. For row i , with $\mathbf{J}_{i,:} = \nabla u_i$ for some scalar $u_i : \mathcal{M} \rightarrow \mathbb{R}$,

$$\oint_{\mathcal{C}} \widetilde{\mathbf{J}}_{i,:} d\widetilde{\mathbf{r}} = \oint_{\phi^{-1}(\mathcal{C})} \mathbf{J}_{i,:} d\mathbf{r} = 0, \quad (\text{D.51})$$

because the line integral of a gradient field along any closed contour vanishes; conversely, every closed loop in $\mathcal{M}$ has an image loop in $\widetilde{\mathcal{M}}$ under ϕ , so the vanishing transfers in both directions. $\square$

D.3.4 Gramian computation

Continuous-time definitions

Along a reference trajectory of the linearized subsystem $\delta\dot{\mathbf{z}}^\alpha(t) = \mathbf{J}^{\alpha \rightarrow \alpha}(\mathbf{z}(t)) \delta\mathbf{z}^\alpha(t) + \mathbf{J}^{\beta \rightarrow \alpha}(\mathbf{z}(t)) \delta\mathbf{z}^\beta(t)$ of the pairwise tangent-space equation in §4.3, let $\Phi^\alpha(t_2, t_1)$ denote the within-target state-transition operator (propagator of $\delta\dot{\mathbf{z}}^\alpha = \mathbf{J}^{\alpha \rightarrow \alpha} \delta\mathbf{z}^\alpha$ on $[t_1, t_2]$), take the cross-area input matrix as $\mathbf{B}(\tau) \triangleq \mathbf{J}^{\beta \rightarrow \alpha}(\mathbf{z}(\tau))$, and take the output matrix as $\mathbf{C}(\tau) \triangleq \mathbf{J}^{\alpha \rightarrow \beta}(\mathbf{z}(\tau))$ —the *target-to-source* Jacobian block, i.e., the reverse-direction coupling through which the target’s state propagates back into the source’s dynamics. The three pairwise Gramians on $[t_0, t_1]$ are then

$$\mathbf{W}_r^{\beta \rightarrow \alpha}(t_0, t_1) \triangleq \int_{t_0}^{t_1} \Phi^\alpha(t_1, \tau) \mathbf{B}(\tau) \mathbf{B}^\top(\tau) \Phi^{\alpha \top}(t_1, \tau) d\tau, \quad (\text{D.52})$$

$$\mathbf{W}_c^{\beta \rightarrow \alpha}(t_0, t_1) \triangleq \int_{t_0}^{t_1} \Phi^\alpha(t_0, \tau) \mathbf{B}(\tau) \mathbf{B}^\top(\tau) \Phi^{\alpha \top}(t_0, \tau) d\tau, \quad (\text{D.53})$$

$$\mathbf{W}_o^{\beta \rightarrow \alpha}(t_0, t_1) \triangleq \int_{t_0}^{t_1} \Phi^{\alpha \top}(\tau, t_0) \mathbf{C}^\top(\tau) \mathbf{C}(\tau) \Phi^\alpha(\tau, t_0) d\tau. \quad (\text{D.54})$$

The *reachability* Gramian $\mathbf{W}_r^{\beta \rightarrow \alpha}$ (Eq. D.52) measures how much of target subsystem α ’s state space at the final time t_1 can be reached from a unit-energy input applied through source subsystem β over $[t_0, t_1]$; $\text{tr } \mathbf{W}_r$ summarizes the total reachable volume, and $\lambda_{\min}(\mathbf{W}_r)$ the worst-case direction. The *controllability* Gramian $\mathbf{W}_c^{\beta \rightarrow \alpha}$ (Eq. D.53) is the same input–output coupling propagated to the initial time t_0 ; equivalently, $\text{tr } \mathbf{W}_c^{-1}$ is the minimum input energy required to drive α from a unit-norm state at t_0 to the origin at t_1 . The *observability* Gramian $\mathbf{W}_o^{\beta \rightarrow \alpha}$ (Eq. D.54) measures how strongly perturbations to α ’s state at t_0 are read off by β over $[t_0, t_1]$ via the target-to-source coupling $\mathbf{C}$ —i.e., how well source area β observes target area α ’s activity. As with reach, the $\beta \rightarrow \alpha$ superscript denotes the actor-on-target direction (here, β observing α); it is the natural dual of $\mathbf{W}_r^{\alpha \rightarrow \beta}$ along the same reference trajectory. We do not report observability summaries in the chapter but include the definition for completeness because `JacobianODE/control/gramians.py` accepts $(\mathbf{A}, \mathbf{B}, \mathbf{C}) = (\exp(\mathbf{J}^{\alpha \rightarrow \alpha} \Delta t), \mathbf{J}^{\beta \rightarrow \alpha}, \mathbf{J}^{\alpha \rightarrow \beta})$ and computes all three Gramians simultaneously by the same square-root QR mechanics.

Square-root QR iteration and numerical stability

We compute the time-varying reachability, controllability, and observability Gramians along the reference trajectory by a low-rank square-root QR iteration³⁶⁵, which preserves rank across time steps and keeps the Gramian spectrum reliably computable even when the dynamic range spans many orders of magnitude (as it does for chaotic latents). Implementation lives in `JacobianODE/control/gramians.py`.

Discretize the LTV system along the trajectory at step Δt , $\mathbf{A}_k = \exp(\mathbf{J}^{\alpha \rightarrow \alpha}(\mathbf{z}(t_k))\Delta t)$, $\mathbf{B}_k = \mathbf{J}^{\beta \rightarrow \alpha}(\mathbf{z}(t_k))$. The square-root factor $\mathbf{S}_{r,k}$ of the reachability Gramian at step k is initialized to zero and updated by

$$\mathbf{Z}_k = \begin{bmatrix} \mathbf{A}_k \mathbf{S}_{r,k}^\top \\ \mathbf{B}_k \sqrt{\Delta t} \end{bmatrix}, \quad \mathbf{Z}_k^\top = \mathbf{Q}_k \mathbf{R}_k \quad (\text{reduced QR}), \quad \mathbf{S}_{r,k+1} = \mathbf{R}_k^\top. \quad (\text{D.55})$$

The reachability Gramian at horizon T is recovered as $\mathbf{W}_r = \mathbf{S}_{r,T} \mathbf{S}_{r,T}^\top$; the controllability Gramian is computed by the analogous backward sweep with $\mathbf{P}_k = \mathbf{A}_k^{-1}$ replacing $\mathbf{A}_k$.

Why log scaling is required. For chaotic latents the dynamic range of $\mathbf{S}_{r,k}$ spans many orders of magnitude: the Frobenius norm grows or shrinks at rate $\sim e^{\lambda_{\max} k \Delta t}$. Over the 10ms horizon used for the main-text Gramian summaries the per-trajectory factor is modest, but over the longer FTLE/asymptotic horizons used in §D.3.5 ($T \sim 10^4$ – 10^5 samples) the unscaled $\|\mathbf{S}_{r,k}\|_F$ leaves the `float64` representable range entirely. The eigenvalues of $\mathbf{W}_r$ inherit this dynamic range *squared*, so naively computing $\log \text{tr } \mathbf{W}_r$ or $\log \sigma_i(\mathbf{W}_r)$ on the unrescaled factor produces `inf/nan`, or quietly saturates at the smallest representable singular value, long before the underlying signal has saturated.

The log-prefactor decomposition. We avoid the overflow by representing the square-root factor as a product of an explicit log-domain scalar and a bounded matrix:

$$\mathbf{S}_{r,k} = e^{\ell_k} \tilde{\mathbf{S}}_{r,k}, \quad \ell_k \in \mathbb{R}, \quad \|\tilde{\mathbf{S}}_{r,k}\|_F \leq 1, \quad (\text{D.56})$$

where ℓ_k accumulates the exponential magnitude and $\tilde{\mathbf{S}}_{r,k}$ is the rescaled factor whose Frobenius norm is bounded by 1 by construction. The Gramian itself satisfies

$$\mathbf{W}_{r,k} = \mathbf{S}_{r,k} \mathbf{S}_{r,k}^\top = e^{2\ell_k} \tilde{\mathbf{S}}_{r,k} \tilde{\mathbf{S}}_{r,k}^\top, \quad (\text{D.57})$$

so all eigenvalue and trace summaries factor into a closed-form ℓ_k contribution plus a bounded contribution from $\tilde{\mathbf{S}}_{r,k}$.

Per-step update. At step k we form the stacked matrix

$$\mathbf{Z}_k = \begin{bmatrix} \mathbf{A}_k \tilde{\mathbf{S}}_{r,k}^\top \\ e^{-\ell_k} \mathbf{B}_k \sqrt{\Delta t} \end{bmatrix}, \quad (\text{D.58})$$

where the driving term is itself rescaled by $e^{-\ell_k}$ so that both blocks share the same log-domain reference. We then compute $\rho_k = \max(\|\mathbf{Z}_k\|_F, \epsilon)$ (floored at $\epsilon = \text{torch.finfo(float64).tiny}$ to avoid taking the log of zero on initial steps where $\tilde{\mathbf{S}}_{r,0} = \mathbf{0}$), divide $\mathbf{Z}_k$ by ρ_k , reduced-QR-factor the result, and accumulate $\log \rho_k$ into ℓ :

$$\frac{\mathbf{Z}_k}{\rho_k} = \mathbf{Q}_k \tilde{\mathbf{R}}_k, \quad \tilde{\mathbf{S}}_{r,k+1} = \tilde{\mathbf{R}}_k^\top, \quad \ell_{k+1} = \ell_k + \log \rho_k. \quad (\text{D.59})$$

Substituting Eq. D.56 into the Gramian recurrence $\mathbf{W}_{r,k+1} = \mathbf{A}_k \mathbf{W}_{r,k} \mathbf{A}_k^\top + \mathbf{B}_k \mathbf{B}_k^\top \Delta t$ and using $\mathbf{Z}_k^\top \mathbf{Z}_k = \tilde{\mathbf{R}}_k^\top \tilde{\mathbf{R}}_k$ verifies that Eq. D.59 preserves the invariant: $\mathbf{W}_{r,k+1} = e^{2\ell_k} \rho_k^2 \tilde{\mathbf{S}}_{r,k+1} \tilde{\mathbf{S}}_{r,k+1}^\top = e^{2\ell_{k+1}} \tilde{\mathbf{S}}_{r,k+1} \tilde{\mathbf{S}}_{r,k+1}^\top$. The bound $\|\tilde{\mathbf{S}}_{r,k+1}\|_F \leq 1$ follows from $\|\tilde{\mathbf{R}}_k\|_F = \|\mathbf{Q}_k \tilde{\mathbf{R}}_k\|_F = \|\mathbf{Z}_k / \rho_k\|_F \leq 1$.

Recovering log-trace and per-mode log-eigenvalues. At horizon T we recover the summaries used in the chapter (§4.4) from the pair $(\tilde{\mathbf{S}}_{r,T}, \ell_T)$. The log-trace is

$$\log \text{tr } \mathbf{W}_r(t_0, t_0 + T\Delta t) = 2\ell_T + \log \|\tilde{\mathbf{S}}_{r,T}\|_F^2, \quad (\text{D.60})$$

and the per-mode log-eigenvalues come from the SVD $\tilde{\mathbf{S}}_{r,T} = \tilde{\mathbf{U}} \text{diag}(\tilde{\sigma}_i) \tilde{\mathbf{V}}^\top$:

$$\log \lambda_i(\mathbf{W}_r) = \log \sigma_i^2(\mathbf{S}_{r,T}) = 2\ell_T + 2\log(\max(\tilde{\sigma}_i, \epsilon)). \quad (\text{D.61})$$

The inverse-log-trace $\log \text{tr } \mathbf{W}_r^{-1}$ is computed entirely in the log domain via the log-sum-exp identity on the per-mode log-eigenvalues, so the smallest eigenvalues (which dominate $\text{tr } \mathbf{W}_r^{-1}$ but are most vulnerable to underflow) are recovered without intermediate exponentiation.

Numerical safeguards. The full square-root QR iteration runs in `float64`; no intermediates are downcast. The SVD on the rescaled factor uses `torch.linalg.svd` with `driver='gesvd'` (the more numerically stable QR-based driver, preferred over `'gesdd'` for the ill-conditioned rescaled factors that arise near the smallest singular values), and singular values are clamped at ϵ before taking logarithms. The log-prefactor floor $\rho_k \geq \epsilon$ guarantees that ℓ_k stays in the `float64` range even on initial steps where $\mathbf{Z}_0 = [\mathbf{0}; \mathbf{B}_0\sqrt{\Delta t}]$ would otherwise have a vanishingly small Frobenius norm.

The horizon used for the main-text propofol analyses is $t_1 - t_0 = 10\Delta t$ at $\Delta t = 1$ ms, i.e., a 10 ms window.

Metric-adjusted Gramians

Theorem 1 guarantees that the pairwise reachability Gramian is preserved under an isometric direct-sum encoder. The NICE–Cayley encoder is volume-preserving but in general not isometric, so we additionally provide a metric-adjusted variant of the Gramian (`JacobianODE/control/gramians_metric.py`) that re-weights the input matrix $\mathbf{B}$ by an arbitrary per-step right-factor $\mathbf{B}_{\text{factor}}(\tau)$:

$$\mathbf{W}_r^{\text{metric}}(t_0, t_1) = \int_{t_0}^{t_1} \Phi^\alpha(t_1, \tau) \mathbf{B}(\tau) \mathbf{B}_{\text{factor}}(\tau) \mathbf{B}_{\text{factor}}(\tau)^\top \mathbf{B}^\top(\tau) \Phi^\alpha(t_1, \tau)^\top d\tau. \quad (\text{D.62})$$

The natural choice is $\mathbf{B}_{\text{factor}} = \mathbf{J}_E^{-\top}$ evaluated on the source-area input space at τ , which gives $\mathbf{B}_{\text{factor}} \mathbf{B}_{\text{factor}}^\top = (\mathbf{J}_E \mathbf{J}_E^\top)^{-1} = \mathbf{R}$, the induced latent-space metric (§D.3.1). Under this choice, control effort is measured in the obs-space-induced inner product on the latent control space rather than in the bare latent Euclidean inner product, which removes the residual non-isometry of the encoder from the reachability calculation. When the factors are absent ($\mathbf{B}_{\text{factor}} = \mathbf{I}$), the metric-adjusted routine reproduces the standard square-root QR iteration of §D.3.4 numerically. The metric-adjusted variant is available in the codebase as an experimental construction; in the present paper, the production propofol-cohort figures use the standard square-root QR Gramian of §D.3.4 (no metric adjustment), so the directional reach/ctrl log-trace summaries we report are invariants up to encoder-induced similarity rather than exact diffeomorphism invariants.

D.3.5 Lyapunov exponent computation

The Lyapunov spectrum of the latent dynamics is computed by discrete QR iteration¹²⁷ on the sequence of per-step generators $\mathbf{A}_k = \exp(\hat{\mathbf{J}}^\theta(\mathbf{z}(t_k))\Delta t)$. Implementation lives in `JacobianODE/jacobians/analysis.py`.

Starting from $\mathbf{Q}_0 = \mathbf{I}_{d_{\text{dyn}} \times k}$ for top- k exponents (or $k = d_{\text{dyn}}$ for the full spectrum), at each step we QR-factor

$$\mathbf{A}_k \mathbf{Q}_{k-1} = \mathbf{Q}_k \mathbf{R}_k, \quad (\text{D.63})$$

enforcing positive diagonal of $\mathbf{R}_k$ by sign correction (column-wise multiplication of $\mathbf{Q}_k$ by the sign of $\text{diag}(\mathbf{R}_k)$), and accumulate the log-diagonal:

$$\lambda_i = \frac{1}{T\Delta t} \sum_{k=1}^T \log |R_{k,ii}|, \quad i = 1, \dots, k. \quad (\text{D.64})$$

The full spectrum reported in the chapter (§4.4) is estimated per resting-state window (§D.4.3) and averaged across the windows of each condition: for each window we run the QR iteration over its valid latent trajectory, truncated to $T_{\text{max}} = 2000$ samples (2,s at the 1,kHz step, by which the leading exponents have stabilised). Each condition retains ~ 20 – 30 windows ($\sim 3 \times 10^4$ latent samples per condition in aggregate). The exponents are sorted in descending order at the end and divided by Δt to convert to continuous-time exponents. All arithmetic is in float64.

The 10-ms finite-time Lyapunov (FTLE) summaries are derived from the same per-step QR contributions, post-processed in the analysis pipeline by accumulating $\sum_i \log |R_{k,ii}|$ over non-overlapping 10-sample windows and averaging across windows to obtain the per-window generator log-determinant trace. The FTLE-max and FTLE-full-spectrum-mean reported in the chapter (§4.4) are computed at this 10-sample window length.

D.3.6 Training details

Optimizer and weight decay

Training uses AdamW⁴⁷⁶ with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$, weight decay 10^{-4} , and gradient clipping at ℓ_2 norm 1.0. Encoder and Jacobian-MLP parameters share a single optimizer with the same hyperparameters.

Learning rate schedule

The learning rate is coupled to the teacher-forcing strength via the `TeacherForcingLRScheduler` (`JacobianODE/jacobians/lightning_base.py`). Let $\alpha = \alpha_{\text{TF}}$ and let $\tilde{\alpha} = (\alpha - \alpha_{\text{min}})/(1 - \alpha_{\text{min}})$ be the renormalized TF strength on $[0, 1]$. The per-step learning rate is

$$\text{LR}(t) = \text{LR}_{\text{min}} + s(\tilde{\alpha}) \cdot (\text{LR}_{\text{start}} - \text{LR}_{\text{min}}), \quad s(\tilde{\alpha}) = \frac{\tilde{\alpha}}{\tilde{\alpha} + (1 - \tilde{\alpha}) \exp(-k \tilde{\alpha})}, \quad (\text{D.65})$$

where k is a fixed shape parameter. In production we use $k = 1$ (`k_scale: 1` in every `conf/experiment/*.yaml`), which makes $s(\tilde{\alpha})$ approximately linear in $\tilde{\alpha}$ for $\tilde{\alpha}$ near 1 and slightly compresses LR decay near $\tilde{\alpha} = 0$ relative to the bare linear schedule $k = 0$; the effect is a more gradual LR drop early in training (when α_{TF} is still close to 1) and a slightly steeper drop toward the floor late in training. The two strengths anneal together: as α_{TF} decays from $\alpha_{\text{init}} = 1$ toward α_{min} , the LR follows it down toward LR_{min} . We use $\text{LR}_{\text{start}} = 10^{-4}$, $\text{LR}_{\text{min}} = 10^{-6}$, $\alpha_{\text{min}} = 10^{-3}$.

Early stopping and two-stage protocol

Training runs until early stopping on the free-running trajectory validation loss with patience 5 epochs and a minimum of 15 epochs. For the sample-systems experiments (Lorenz fully observed, Lorenz partially observed, WMTask RNN fully observed, WMTask RNN partially observed), we use the two-stage protocol of³⁹ as implemented in `JacobianODE/jacobians/tuning/two_stage_cull.py`: Stage A trains every cell of the sweep grid for 20 epochs (`stage_a_epochs=20`), after which a culling step retains the top- $\lceil(1 - c)N\rceil$ cells ranked by best-so-far median trajectory `val_loss`, and Stage B warm-starts the retained cells from the Stage A `last.ckpt` and continues until the early-stopping criterion fires. For the propofol cohort, we ran only a single full sweep—without a Stage A cull—after pilot runs established that the best n_d always fell within $\{5, 10, 15\}$; the full 3×7 per-session grid is then small enough that the cull does not deliver wall-time savings comparable to the sample-systems regime.

D.3.7 Hyperparameter selection

Sweep grid

For the propofol cohort, each session is trained on a per-session grid over two hyperparameters: the delay-embedding depth n_d and the loop-closure regularizer weight λ_{loop} . The n_d range $\{5, 10, 15\}$ was established by pilot runs (early sweeps with a wider candidate set always selected one of these three values); the λ_{loop} range is the canonical seven-point geometric grid $\{0, 10^{-6}, \dots, 10^{-1}\}$ used in³⁹. All other architecture and training settings (Table D.2) are held fixed across the cohort. The per-session grid therefore has $3 \times 7 = 21$ cells, and the cohort comprises 21 sessions, for 441 production training runs in total.

Fixed per-session architecture and training settings are reported in Table D.2.

Selection criteria

For each session we select a single chosen run from the 21-cell sweep by applying two hard criteria (C1 and C2, Table D.3) and ranking the survivors by free-running trajectory validation loss. The criteria gate the cells before ranking rather than serving as a post-hoc tiebreaker, so a cell that achieves the

Table D.1: Per-session sweep grid for the propofol cohort. Sample-systems experiments swept a wider grid in a two-stage protocol; see Table D.4 for the chosen sample-systems runs.

Variable	Values
Delay depth n_d	$\{5, 10, 15\}$ (3-point grid)
Loop-closure weight λ_{loop}	$\{0, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ (7-point geometric)
Total cells per session	$3 \times 7 = 21$
Number of sessions	21 (10 NHP 1 + 11 NHP 2)
Total production training runs	$21 \times 21 = 441$

Table D.2: Fixed per-session architecture and training settings.

Encoder coupling layers L	16
Conditioner MLP width	1024
Conditioner depth	2 hidden layers, SiLU activation
Dynamic dim chosen by	PCA on delay-embedded train, variance threshold 0.99
Jacobian MLP widths	(2048, 2048, 4096, 4096), SiLU activation
Condition vector dim d_c	1 (binary $\{-1, +1\}$)
Latent prediction loss weight	1
Decoded prediction loss weight	1
Reconstruction loss weight	1
Trajectory sequence length	25 samples
Sequence spacing	15 samples
Prediction steps	10
Initial window T_{init}	15
Training noise (<code>obs_noise_scale</code>)	0 for the chosen production cohort
Optimizer	AdamW, weight decay 10^{-4}
LR schedule	TeacherForcingLRScheduler, $10^{-4} \rightarrow 10^{-6}$
Early stopping patience	5 epochs (free-running trajectory <code>val_loss</code>)
Minimum epochs	15

lowest trajectory `val_loss` but fails C1 or C2 is rejected. C1 (decoder-corrected one-step MASE < 1) certifies that the model’s one-step prediction beats the decoded-persistence baseline on the validation split; C2 (loop-closure $< \sqrt{d_{\text{dyn}}}$) certifies that the learned Jacobian is approximately row-conservative on the per-session dynamic subspace. We do not gate on C3 (small fast-eigenvalue fraction): although it is computed alongside C1 and C2 in `JacobianODE/jacobians/tuning/selection.py`, dropping it from the selection rule produces the identical chosen run on every session of the propofol cohort, so it does not bind in practice. The implementation reads the per-cell val metrics from the W&B logger, evaluates each criterion, and returns the chosen run name (which we recover from `chosen_extras.pkl` for downstream analysis).

Table D.3: Per-session model-selection criteria. Among the 21 cells of each session’s sweep, the chosen run is the one that minimizes the free-running trajectory `val_loss` subject to passing both criteria.

C1 (one-step accuracy)	decoder-corrected one-step MASE < 1 on the validation split. The decoder-corrected variant subtracts the reconstruction floor by comparing the predicted next-step against the <i>decoded encoded ground-truth next-step</i> rather than the raw observation; this prevents the reconstruction floor itself from disqualifying otherwise-good cells.
C2 (loop closure)	loop-closure validation loss $< \sqrt{d_{\text{dyn}}}$, with d_{dyn} the per-session dynamic-subspace dimension.
Tie-break	among cells passing C1–C2, the one minimizing free-running trajectory <code>val_loss</code> is chosen.

Sample-systems run identifiers and chosen hyperparameters

The W&B run identifiers for the four sample-systems runs reported in the chapter (§4.4) are reported in Table D.4.

Table D.4: Run IDs for the four sample-systems runs reported in the chapter (§4.4).

Run	Run ID	Native prediction horizon (steps)
Lorenz, fully observed	ns3oemeh	30
WMTask RNN, fully observed	uq0q07eo	30
Lorenz, partially observed	db6w4w0x	30
WMTask RNN, partially observed	gagriild	20

Evaluation is on the held-out test split at 10 prediction steps with $\alpha_{\text{TF}} = 0$ (free-running) for R^2 and $\alpha_{\text{TF}} = 1$ (teacher-forced) for one-step MASE. Selection criteria match the per-session criteria above (C1–C2 with best-trajectory-loss tie-break).

Per-session chosen hyperparameters for the propofol cohort

The chosen n_d landed at each of $\{5, 10, 15\}$ at least twice across the cohort, and λ_{loop} ranged from 10^{-5} to 10^{-1} ; no single sweep cell dominated, confirming that the per-session selection reflects genuine cross-session variation rather than a degenerate winner. Any analysis directories carrying the **smallenc** recipe slug (a separate encoder-size investigation that is not part of the production cohort) are excluded from this table at the extraction stage; the audit CSV emitted by the extraction script documents the exclusions.

Table D.5: Per-session chosen hyperparameters under the C1–C2 criterion across the 21 propofol-cohort sessions. The chosen $(n_d, \lambda_{\text{loop}})$ pair was selected from each session’s 3×7 sweep by minimizing the free-running trajectory `val_loss` subject to C1–C2; d_{dyn} is the per-session PCA-99% dynamic-subspace dimension on the delay-embedded train data (sum across the four areas), and depends on n_d . Session labels follow the <YYYYMMDD>-<NN> convention. Values extracted from the on-disk `chosen_extras.pkl` and per-session `report.md` artifacts; see `figures/paper/extract_chosen_hyperparams.py`.

Session	Animal	n_d	λ_{loop}	d_{dyn}
20160809-01	NHP 1	5	10^{-4}	115
20160818-02	NHP 1	5	10^{-2}	93
20160822-02	NHP 1	5	10^{-1}	97
20160826-02	NHP 1	10	10^{-3}	129
20160831-02	NHP 1	15	10^{-2}	96
20160902-02	NHP 1	10	10^{-4}	135
20160908-02	NHP 1	5	10^{-2}	128
20160912-02	NHP 1	10	10^{-3}	144
20160914-02	NHP 1	10	10^{-1}	136
20160916-02	NHP 1	5	10^{-3}	110
20160105-01	NHP 2	10	10^{-1}	176
20160107-01	NHP 2	10	10^{-3}	146
20160109-01	NHP 2	10	10^{-2}	134
20160113-01	NHP 2	10	10^{-3}	153
20160121-01	NHP 2	5	10^{-4}	107
20160123-01	NHP 2	10	10^{-5}	133
20160125-01	NHP 2	10	10^{-4}	164
20160201-01	NHP 2	10	10^{-4}	173
20160206-01	NHP 2	5	10^{-1}	120
20160210-01	NHP 2	15	10^{-5}	166
20160301-01	NHP 2	5	10^{-1}	128

D.4 Experimental details

D.4.1 Dynamical systems data

Lorenz system

We use the canonical Lorenz⁶⁰ system, $\dot{x} = \sigma(y - x)$, $\dot{y} = x(\rho - z) - y$, $\dot{z} = xy - \beta z$, with the standard chaotic parameters $\sigma = 10$, $\rho = 28$, $\beta = 8/3$. Following³⁹, trajectories are simulated using the `dysts` package³⁵⁷, which samples each dynamical system at 100 time steps per characteristic Fourier-spectrum timescale. Integration uses `scipy.integrate.solve_ivp` (default Radau, or RK45 when configured) through the `dysts` backend (`JacobianODE/dysts_sim/flows.py`). We use $n_{\text{ics}} = 32$ trajectories of $n_{\text{periods}} = 12$ Fourier periods each (1200 samples per trajectory), with initial conditions perturbed from the system’s default IC by Gaussian noise of standard deviation 0.2

(`new_ic_mode='random'`, `traj_offset_sd=0.2`). The train/test split is the standard 80/20 split across trajectories. For the fully observed variant, we expose all three dimensions (x, y, z) as the per-subsystem observation. For the partially observed variant, we observe only the x coordinate ($d_{\text{obs}} = 1$) and delay-embed before encoding. To evaluate the methods under observation noise, we additionally train with 1%, 5%, and 10% Gaussian observation noise added i.i.d. over time, where the percentage is the ratio of the Euclidean norm of the noise to the mean Euclidean norm of the data.

Task-trained working-memory selection RNN

The 128-dimensional task-trained RNN is the working-memory selection network of³⁹, inspired by³⁶⁸ (codebase: `ControlJacobians/wmtask/wmtask.py`). The RNN has continuous-time dynamics

$$\tau \dot{\mathbf{h}}(t) = -\mathbf{h}(t) + \mathbf{W}_{hh} \sigma(\mathbf{h}(t)) + \mathbf{W}_{hi} \mathbf{u}(t) + \mathbf{b}, \quad \mathbf{o}(t) = \mathbf{W}_{oh} \mathbf{h}(t), \quad (\text{D.66})$$

discretized in implementation by Euler integration as $\mathbf{h}_{k+1} = \mathbf{h}_k + (\Delta t / \tau) [-\mathbf{h}_k + (\mathbf{W}_{hi} \odot \text{mask}_{\text{in}}) \mathbf{u}_k + \mathbf{W}_{hh} \sigma(\mathbf{h}_k) + \mathbf{b}]$ with $\tau = 0.05$ s and $\Delta t = 0.02$ s. Here $\mathbf{W}_{hh} \in \mathbb{R}^{128 \times 128}$ defines the recurrent weights, $\mathbf{W}_{hi}$ maps the input into the hidden state, $\mathbf{W}_{oh} \in \mathbb{R}^{4 \times 128}$ maps the hidden state to a four-dimensional output, $\mathbf{b}$ is a bias term, and σ is the exponential linear unit (ELU) activation ($\sigma'(h) = 1$ for $h > 0$ and $\sigma'(h) = \exp(h)$ otherwise). The network has two 64-neuron areas (denoted $N_1 = N_2 = 64$ in the codebase): a “visual” area (which receives sensory input via an input mask zeroing the cognitive-area inputs) and a “cognitive” area (whose readout mask zeros the visual-area outputs). The within-area weights are initialized with greater connectivity strength than the across-area weights; because input enters only the visual area and output exits only the cognitive area, the two areas are forced to interact to solve the task. The RNN solves a working-memory selection task with trial structure (in milliseconds): fixation 500, stimuli 500, delay 1 750, cue 100, delay 2 750, response 250 (total 2850 ms); on each trial two of four colors are presented (one-hot vectors) during the stimulus epoch, and the cue then selects one of the two for the response. The RNN achieves 100% accuracy. JacobianODE training is performed on the post-cue delay + response portion of the trial, where the dynamics are state-locked to the cued color. The discrete Jacobian along the rollout is $\mathbf{J}_{\text{disc}} = \mathbf{I} + (\Delta t / \tau) (-\mathbf{I} + \mathbf{W}_{hh} \text{diag}(\sigma'(\mathbf{h})))$. For the partially observed variant, we retain 16 randomly chosen channels from each of the two areas ($d_{\text{obs}} = 32$ total). We condition the model on the discrete trial-condition variable $\mathbf{c} \in \{-1, +1\}$ corresponding to the cued color. As in the Lorenz case, we add 5% Gaussian observation noise during training;³⁹ reports that this improves Jacobian estimation across all systems considered there.

Synthetic two-area linearization (conceptual control factors)

Figure 4.1E illustrates, on a small synthetic linear system, how the pairwise reachability and controllability Gramians depend on four properties of a source→target interaction: the target area’s dynamic stability, and the magnitude, alignment, and dimensionality of the cross-area coupling block (codebase: `JacobianODE/_notebook/render_damping_vs_coupling.py`). In every case we form a source→target control system $(\mathbf{A}_{\text{tgt}}, \mathbf{B}, \mathbf{C})$ —with $\mathbf{A}_{\text{tgt}}$ the target’s within-area dynamics and $\mathbf{B} = \mathbf{C}$ the coupling block playing the role of $\mathbf{J}^{\beta \rightarrow \alpha}$ —and compute the reachability and controllability

Gramians by the same square-root QR pipeline used for the learned models (§D.3.4), over a horizon of $T = 1000$ steps at $\Delta t = 0.01$ (10s), reporting $\log \text{tr } \mathbf{W}$.

Stability and magnitude (coupled damped-oscillator areas). For the stability and magnitude factors the system is an 8-D pair of two 4-D areas. Each area’s 4×4 intrinsic block is a direct sum of two planar damped-rotation blocks $\mathbf{o}(d, \omega) = \begin{bmatrix} -d & \omega \\ -\omega & -d \end{bmatrix}$ at frequencies $\omega_1 = 1.0$ and $\omega_2 = 2.3$ and common damping d ; the coupling is isotropic, $\mathbf{B} = \mathbf{C} = g \mathbf{I}_4$. To vary *dynamic stability* we sweep the target damping $d = b$ from 0.8 down to 0.02 (40 log-spaced values; source damping fixed at 0.8, $g = 0.25$); the abscissa is the system stability $-\max_i \text{Re } \lambda_i(\mathbf{A})$, which decreases as $b \rightarrow 0$. To vary *coupling magnitude* we sweep g from 0.02 to 1.5 (50 linearly-spaced values; target damping fixed at 0.4), so that $\|\mathbf{B}\|_F = 2g$.

Alignment and dimensionality (fixed diagonal target, shaped coupling). For the alignment and dimensionality factors we fix a 4-D target with diagonal dynamics $\mathbf{A}_{\text{tgt}} = -\text{diag}(0.3, 0.6, 1.0, 1.5)$ (eigen-timescales ordered slow→fast) and shape the coupling block $\mathbf{B}$ to have a prescribed set of squared singular values $\{\sigma_m^2\}$ with left singular vectors aligned to the target eigenbasis, holding the total energy fixed at $\|\mathbf{B}\|_F^2 = \sum_m \sigma_m^2 = 1$. *Coupling alignment* is varied by sweeping the energy σ_1^2 that $\mathbf{B}$ places on the slowest target mode from 0.05 to 0.95 (30 values), spreading the remaining energy uniformly over the other modes. *Coupling dimensionality* is varied by sweeping the participation ratio $\text{PR}(\sigma^2) = (\sum_m \sigma_m^2)^2 / \sum_m \sigma_m^4$ of the coupling spectrum (from ≈ 2 to ≈ 3) along two mirrored paths that redistribute the off-leading energy toward either the slowest or the fastest minor mode; the two paths share the same PR at each step but yield different Gramians, illustrating that the dimensionality scalar alone does not determine reachability/controllability—the direction in which the coupling spreads its energy, relative to the target’s own timescales, also matters.

D.4.2 Non-human primate electrophysiology

The neural data for this study is the propofol cohort previously deposited in association with³⁷, comprising 21 sessions from two rhesus macaques (*Macaca mulatta*), denoted NHP 1 and NHP 2. NHP 1 was female, aged 8 years, ~ 6.6 kg; NHP 2 was male, aged 14 years, ~ 13.0 kg. All procedures followed the guidelines of the Massachusetts Institute of Technology Committee on Animal Care and the National Institutes of Health. The animals were pair-housed in temperature-controlled environments, were not involved in previous procedures, and were surgically implanted with a subcutaneous vascular access port (Model CP-6, Norfolk Access Technologies, Skokie, IL) at the cervicothoracic junction of the neck, with the catheter tip reaching the termination of the superior vena cava via the external jugular vein. Both NHPs were implanted with 8×8 chronic Utah arrays (“Utah arrays”, MultiPort: 1.0 mm shank length, $400 \mu\text{m}$ spacing, Blackrock Microsystems, Salt Lake City, UT), yielding 64 channels per multi-electrode array. Electrodes were placed in four areas: ventrolateral prefrontal cortex (vlPFC), frontal eye fields (FEF), posterior parietal cortex (PPC), and auditory cortex (STG). In each session, the NHP performed a non-demanding pre-anesthesia task (passive airpuff/tone classical conditioning); following this task, the NHP was infused with propofol anesthesia for 60 minutes (30 minutes at a higher loading dose of $0.58 \text{ mg kg}^{-1} \text{ min}^{-1}$ for NHP 1

and $0.285 \text{ mg kg}^{-1} \text{ min}^{-1}$ for NHP 2; 30 minutes at a lower maintenance dose of $0.32 \text{ mg kg}^{-1} \text{ min}^{-1}$ for NHP 1 and $\sim 0.075 \text{ mg kg}^{-1} \text{ min}^{-1}$ for NHP 2). The infusion was then stopped and the NHP gradually returned to a normal awake state. Propofol was intravenously infused via a computer-controlled syringe pump (PHD ULTRA 4400, Harvard Apparatus).

LFP recording and preprocessing. Local field potentials were recorded at 30 kHz and filtered online via a lowpass at 250 Hz, then downsampled to 1 kHz. Areas STG and PPC shared a common subdural ground/reference; areas FEF and vLPFC each had their own ground/reference. Across recording sessions, a synchronization test signal with locally unique temporal structure was simultaneously injected on one auxiliary analog channel of each system and used to rectify residual cross-system timing offsets (event-code shifts plus linear interpolation to a common time base). Following³⁷ we removed the across-session per-channel mean from each LFP, line-noise-removed via temporally windowed sinusoidal fits at 60 Hz and all harmonics (and at the empirically found 107.35, 214.7, 190.2, 196.8, 393.6 Hz line-noise frequencies), and low-pass filtered each LFP at 300 Hz with a third-order bidirectional Butterworth. Signals were subsequently low-pass filtered at 80 Hz. We did not remove eye-movement or muscle artifacts: intracortical recordings (in contrast to scalp EEG/MEG) are not subject to the same scalp-contact contamination, so these signals are non-cerebral by location but not by amplitude (the recording site is below the dura). All four areas were included in the analysis with no per-channel averaging within an area: the analysis models population dynamics, and channel averaging discards the information that the per-channel representation transforms over time.

Amplitude-threshold artifact rejection. To remove residual recording artifacts (e.g. electrical transients or movement-induced glitches) we applied the same amplitude-threshold rejection procedure used in the cross-anesthetic destabilization study of Eisen et al. [38] (the same procedure was introduced for the propofol cohort in³⁷). We first split each session into non-overlapping windows of the length used for analysis ($W = 15 \text{ s}$; §D.4.3). Within each window we computed, for every electrode, its maximum absolute amplitude over the full window. If two or more electrodes exceeded 1 mV within a window, that window was excluded from all subsequent analysis; any electrode that exceeded the 1 mV threshold was additionally excluded from the remainder of that session. As an added safeguard against onset transients, we also excluded the window immediately preceding any window in which an artifact was detected. Across the propofol cohort this procedure removed between 4 and 23 electrodes per session and between 6% and 44% of windows per session. The surviving per-channel, per-window validity mask is the noise-valid mask consumed by the resting-state window selection of §D.4.3; because electrodes were never averaged within an area (above), a rejected electrode is simply dropped from the population state for the remainder of the session rather than imputed.

D.4.3 Selection of resting-state windows

Resting-state windows for the propofol cohort are carved by the loader, which retains only the two protocol epochs of interest (“awake” and “maintenance dose”) and explicitly excludes the induction and loading-dose transients (these epochs are simply not assigned to any of the per-condition output

sections). Within each retained epoch, candidate windows begin at least $t_{\text{post trial}} = 3.0\text{ s}$ after each preceding behavioral-trial offset, are $W = 15\text{ s}$ long, and are advanced with stride $S = 15\text{ s}$ (non-overlapping). A candidate is kept only if at least 1.5 s of contiguous samples survive the per-channel noise-valid mask (computed from a per-session amplitude-threshold rejector); shorter contiguous blocks are coalesced where possible. Each retained window is assigned to a condition by the position of its midpoint relative to infusion start. The default per-channel decimation factor is 1 (no downsampling). Because windows are non-overlapping with stride equal to length, the resulting awake and maintenance corpora are exhaustive on the retained epochs and identical across sessions in structure; the only per-session degrees of freedom are the noise-valid channel mask and the exact number of windows that survive coalescing in each condition.

D.5 Statistical analyses

D.5.1 Sign-agreement gate

For every 4×4 source–target cell or per-area summary axis, we compute the per-NHP mean of the per-session paired difference (anesthesia – awake), separately for NHP 1 and NHP 2. A cell is admitted to significance testing only if these two per-NHP means agree in sign. The gate is a guard against pooled artifacts: a near-zero pooled difference produced by two equal-and-opposite per-NHP shifts is not a population-level signal, and the gate rejects such cells before any p -value is computed. We do not interpret cells that fail this gate even if the pooled Wilcoxon would have been significant.

D.5.2 Pooled signed-rank test

For the per-pair 4×4 grids we apply a one-sided Wilcoxon signed-rank test in the agreed-upon direction. For the per-area summary axes we apply a two-sided test on the pooled paired differences. For the session-level single-quantity metrics (R^2 , decoder-corrected MASE) we report per-NHP and pooled paired two-sided Wilcoxon. Wilcoxon was chosen over a paired t -test because the per-session paired differences are heavy-tailed in several metrics and the sample size ($n = 21$) is small enough that the assumptions of the t -test cannot be confidently verified. All Wilcoxon tests use `scipy.stats.wilcoxon` with the `zero_method='wilcox'` convention for tied pairs.

D.5.3 Benjamini–Hochberg correction across the 4×4 grid

For each 4×4 grid (reach trace, reach inverse trace, ctrl trace, ctrl inverse trace, block Frobenius norm, block participation ratio) we apply Benjamini–Hochberg FDR control across the 16 cells at $q < 0.05$. The grid-level family is the appropriate test family because the cells of each grid share a common per-session denominator and the substantive question is which cells of the grid carry the

signal. The per-area summary axes are tested separately at $p < 0.05$ because the axis-level family is single-axis.

D.6 Per-pair statistical tables

The chapter reports the 4×4 source–target heatmaps of mean differences (anesthesia – awake) for the six per-pair quantities tested in this study and flags significant cells in the figures. The following tables provide the underlying numerical detail for each cell that passes the sign-agree + Benjamini–Hochberg gate at $q < 0.05$ (§D.5.1–D.5.3). For each significant cell we list: the source→target pair, the pooled paired-difference mean $\pm$ standard error of the mean (across all 21 sessions), the per-NHP mean differences separately for NHP 1 and NHP 2 (which by the sign-agree gate must share the same sign), and the (one-sided, in the agreed direction) pooled Wilcoxon signed-rank p -value before BH correction. Cells that fail the sign-agree gate or the BH-FDR cut are omitted from each table. The six tables in this section cover the four directional Gramian summaries reported in Figure 4.5 (reach log tr $\mathbf{W}$, reach log tr $\mathbf{W}^{-1}$, ctrl log tr $\mathbf{W}$, ctrl log tr $\mathbf{W}^{-1}$) plus the per-block Jacobian summaries (block Frobenius norm, block participation ratio).

Reachability log-trace log tr $\mathbf{W}$

Per-pair (anesthesia – awake) differences in reach log tr $\mathbf{W}$ are reported in Table D.6.

Table D.6: Per-pair reach log tr $\mathbf{W}$, anesthesia – awake, $n = 21$ sessions. All cells listed pass the sign-agree + BH-FDR gate at $q < 0.05$.

source → target	pooled $\Delta \pm \text{SEM}$	per-NHP Δ (NHP 1, NHP 2)	p
PPC → PPC	-0.1006 ± 0.0130	$-0.1218, -0.0813$	9.54×10^{-7}
PPC → STG	-0.0709 ± 0.0328	$-0.1228, -0.0237$	0.01753
PPC → vIPFC	$+0.2451 \pm 0.0774$	$+0.3348, +0.1635$	0.006346
STG → PPC	-0.3917 ± 0.0555	$-0.5249, -0.2706$	4.77×10^{-7}
STG → STG	-0.0666 ± 0.0127	$-0.0824, -0.0522$	5.25×10^{-5}
STG → FEF	-0.1267 ± 0.0334	$-0.1117, -0.1404$	0.0008001
STG → vIPFC	-0.1179 ± 0.0275	$-0.1089, -0.1261$	4.20×10^{-5}
FEF → PPC	-0.3001 ± 0.0325	$-0.3361, -0.2673$	4.77×10^{-7}
FEF → STG	-0.2026 ± 0.0421	$-0.3147, -0.1008$	8.06×10^{-5}
FEF → FEF	-0.1518 ± 0.0109	$-0.1689, -0.1362$	4.77×10^{-7}
FEF → vIPFC	$+0.1330 \pm 0.0479$	$+0.1110, +0.1531$	0.009737
vIPFC → PPC	-0.2767 ± 0.0465	$-0.3887, -0.1748$	1.43×10^{-6}
vIPFC → STG	-0.3110 ± 0.0375	$-0.3667, -0.2604$	9.54×10^{-7}
vIPFC → vIPFC	-0.1292 ± 0.0155	$-0.1682, -0.0936$	1.43×10^{-6}

Reachability inverse log-trace $\log \text{tr } \mathbf{W}^{-1}$

Per-pair (anesthesia – awake) differences in reach $\log \text{tr } \mathbf{W}^{-1}$ are reported in Table D.7.

Table D.7: Per-pair reach $\log \text{tr } \mathbf{W}^{-1}$, anesthesia – awake. All cells listed pass the gate.

source $\rightarrow$ target	pooled $\Delta \pm \text{SEM}$	per-NHP Δ (NHP 1, NHP 2)	p
PPC $\rightarrow$ PPC	$+0.4176 \pm 0.1347$	$+0.7230, +0.1400$	0.0001464
PPC $\rightarrow$ STG	$+0.4167 \pm 0.0669$	$+0.3490, +0.4782$	9.06×10^{-6}
PPC $\rightarrow$ FEF	$+0.3425 \pm 0.0681$	$+0.4447, +0.2497$	9.06×10^{-6}
STG $\rightarrow$ PPC	$+0.3682 \pm 0.0339$	$+0.3707, +0.3659$	4.77×10^{-7}
STG $\rightarrow$ STG	$+0.1855 \pm 0.0557$	$+0.1444, +0.2228$	1.57×10^{-5}
STG $\rightarrow$ FEF	$+0.3242 \pm 0.0402$	$+0.4042, +0.2515$	4.77×10^{-7}
STG $\rightarrow$ vlPFC	$+0.2791 \pm 0.0278$	$+0.2322, +0.3217$	4.77×10^{-7}
FEF $\rightarrow$ PPC	$+0.3394 \pm 0.0335$	$+0.3027, +0.3727$	4.77×10^{-7}
FEF $\rightarrow$ STG	$+0.3546 \pm 0.0339$	$+0.3917, +0.3208$	4.77×10^{-7}
FEF $\rightarrow$ FEF	$+0.3347 \pm 0.1043$	$+0.2788, +0.3856$	4.77×10^{-7}
FEF $\rightarrow$ vlPFC	$+0.2154 \pm 0.0272$	$+0.2051, +0.2248$	4.77×10^{-7}
vlPFC $\rightarrow$ PPC	$+0.4035 \pm 0.0397$	$+0.4622, +0.3501$	4.77×10^{-7}
vlPFC $\rightarrow$ STG	$+0.4827 \pm 0.0617$	$+0.4959, +0.4706$	2.38×10^{-6}
vlPFC $\rightarrow$ FEF	$+0.4442 \pm 0.0747$	$+0.5966, +0.3056$	1.43×10^{-6}

Controllability log-trace $\log \text{tr } \mathbf{W}$

Per-pair (anesthesia – awake) differences in ctrl $\log \text{tr } \mathbf{W}$ are reported in Table D.8.

Controllability inverse log-trace $\log \text{tr } \mathbf{W}^{-1}$

Per-pair (anesthesia – awake) differences in ctrl $\log \text{tr } \mathbf{W}^{-1}$ are reported in Table D.9.

Jacobian block Frobenius norm $\|\hat{\mathbf{J}}_{ij}\|_F$

Per-pair (anesthesia – awake) differences in the Jacobian block Frobenius norm are reported in Table D.10.

Table D.8: Per-pair ctrl log tr $\mathbf{W}$, anesthesia – awake. All cells listed pass the gate.

source $\rightarrow$ target	pooled $\Delta \pm \text{SEM}$	per-NHP Δ (NHP 1, NHP 2)	p
PPC $\rightarrow$ PPC	-0.3193 ± 0.0475	$-0.3105, -0.3274$	3.34×10^{-6}
PPC $\rightarrow$ STG	-0.4217 ± 0.0527	$-0.4447, -0.4008$	2.38×10^{-6}
PPC $\rightarrow$ FEF	-0.2975 ± 0.0578	$-0.3062, -0.2895$	4.77×10^{-6}
STG $\rightarrow$ PPC	-0.5151 ± 0.0791	$-0.6274, -0.4131$	4.77×10^{-7}
STG $\rightarrow$ STG	-0.3830 ± 0.0427	$-0.3511, -0.4121$	4.77×10^{-7}
STG $\rightarrow$ FEF	-0.4218 ± 0.0531	$-0.4735, -0.3748$	4.77×10^{-7}
STG $\rightarrow$ vIPFC	-0.3410 ± 0.0343	$-0.2940, -0.3838$	4.77×10^{-7}
FEF $\rightarrow$ PPC	-0.4867 ± 0.0706	$-0.5920, -0.3909$	3.34×10^{-6}
FEF $\rightarrow$ STG	-0.6202 ± 0.0623	$-0.7128, -0.5361$	4.77×10^{-7}
FEF $\rightarrow$ FEF	-0.2879 ± 0.0339	$-0.3195, -0.2592$	4.77×10^{-7}
FEF $\rightarrow$ vIPFC	-0.1653 ± 0.0578	$-0.1496, -0.1796$	0.005063
vIPFC $\rightarrow$ PPC	-0.4950 ± 0.0655	$-0.5774, -0.4202$	1.43×10^{-6}
vIPFC $\rightarrow$ STG	-0.6657 ± 0.0536	$-0.6806, -0.6523$	4.77×10^{-7}
vIPFC $\rightarrow$ FEF	-0.4487 ± 0.0637	$-0.5770, -0.3320$	4.77×10^{-7}
vIPFC $\rightarrow$ vIPFC	-0.2069 ± 0.0224	$-0.2004, -0.2127$	4.77×10^{-7}

Table D.9: Per-pair ctrl log tr $\mathbf{W}^{-1}$, anesthesia – awake. All cells listed pass the gate.

source $\rightarrow$ target	pooled $\Delta \pm \text{SEM}$	per-NHP Δ (NHP 1, NHP 2)	p
PPC $\rightarrow$ PPC	$+0.5058 \pm 0.1475$	$+0.8585, +0.1853$	6.53×10^{-5}
PPC $\rightarrow$ STG	$+0.4384 \pm 0.0706$	$+0.3351, +0.5323$	1.19×10^{-5}
PPC $\rightarrow$ FEF	$+0.4495 \pm 0.0751$	$+0.5844, +0.3268$	9.54×10^{-7}
PPC $\rightarrow$ vIPFC	$+0.1757 \pm 0.0510$	$+0.1295, +0.2176$	0.001876
STG $\rightarrow$ PPC	$+0.4083 \pm 0.0362$	$+0.4392, +0.3802$	4.77×10^{-7}
STG $\rightarrow$ STG	$+0.2374 \pm 0.0583$	$+0.1913, +0.2794$	2.62×10^{-5}
STG $\rightarrow$ FEF	$+0.4461 \pm 0.0473$	$+0.5567, +0.3456$	4.77×10^{-7}
STG $\rightarrow$ vIPFC	$+0.4118 \pm 0.0283$	$+0.3845, +0.4366$	4.77×10^{-7}
FEF $\rightarrow$ PPC	$+0.3786 \pm 0.0363$	$+0.3637, +0.3922$	4.77×10^{-7}
FEF $\rightarrow$ STG	$+0.3888 \pm 0.0380$	$+0.4051, +0.3740$	4.77×10^{-7}
FEF $\rightarrow$ FEF	$+0.4758 \pm 0.1091$	$+0.4705, +0.4805$	4.77×10^{-7}
FEF $\rightarrow$ vIPFC	$+0.3386 \pm 0.0312$	$+0.3371, +0.3399$	4.77×10^{-7}
vIPFC $\rightarrow$ PPC	$+0.4511 \pm 0.0496$	$+0.5531, +0.3583$	4.77×10^{-7}
vIPFC $\rightarrow$ STG	$+0.5252 \pm 0.0600$	$+0.5085, +0.5403$	9.54×10^{-7}
vIPFC $\rightarrow$ FEF	$+0.5602 \pm 0.0801$	$+0.7317, +0.4044$	4.77×10^{-7}

Jacobian block participation ratio $\text{PR}(\sigma_{ij}^2)$

Per-pair (anesthesia – awake) differences in the Jacobian block participation ratio are reported in Table D.11.

Table D.10: Per-pair $\|\hat{\mathbf{J}}_{ij}\|_F$, anesthesia – awake. All cells listed pass the gate.

source $\rightarrow$ target	pooled $\Delta \pm \text{SEM}$	per-NHP Δ (NHP 1, NHP 2)	p
PPC $\rightarrow$ PPC	-47.4514 ± 7.7469	$-65.5622, -30.9870$	1.19×10^{-5}
PPC $\rightarrow$ STG	-25.0390 ± 2.9885	$-30.0683, -20.4668$	4.77×10^{-7}
PPC $\rightarrow$ FEF	-18.3192 ± 1.7879	$-19.1833, -17.5336$	4.77×10^{-7}
PPC $\rightarrow$ vIPFC	-13.6493 ± 1.3243	$-11.9255, -15.2164$	4.77×10^{-7}
STG $\rightarrow$ PPC	-8.1960 ± 1.6578	$-11.8104, -4.9101$	5.25×10^{-5}
STG $\rightarrow$ STG	-72.6155 ± 7.6688	$-70.2020, -74.8096$	9.54×10^{-7}
STG $\rightarrow$ FEF	-17.3419 ± 1.9623	$-20.6818, -14.3057$	4.77×10^{-7}
STG $\rightarrow$ vIPFC	-18.1553 ± 1.8810	$-14.9224, -21.0943$	4.77×10^{-7}
FEF $\rightarrow$ PPC	-2.3021 ± 1.0957	$-1.4782, -3.0511$	0.03507
FEF $\rightarrow$ STG	-13.4517 ± 1.3642	$-13.5055, -13.4027$	4.77×10^{-7}
FEF $\rightarrow$ FEF	-118.1231 ± 9.7532	$-138.5780, -99.5278$	4.77×10^{-7}
FEF $\rightarrow$ vIPFC	-10.3758 ± 1.5540	$-13.4146, -7.6131$	3.34×10^{-6}
vIPFC $\rightarrow$ STG	-14.6888 ± 1.5594	$-9.9267, -19.0181$	4.77×10^{-7}
vIPFC $\rightarrow$ FEF	-4.2621 ± 2.3600	$-5.5720, -3.0713$	0.04790
vIPFC $\rightarrow$ vIPFC	-89.6821 ± 8.0704	$-93.6591, -86.0666$	4.77×10^{-7}

Table D.11: Per-pair $\text{PR}(\sigma_{ij}^2)$, anesthesia – awake. Only six cells pass the gate (four diagonal blocks plus the FEF $\leftrightarrow$ vIPFC reciprocal pair).

source $\rightarrow$ target	pooled $\Delta \pm \text{SEM}$	per-NHP Δ (NHP 1, NHP 2)	p
PPC $\rightarrow$ PPC	-0.5508 ± 0.0991	$-0.5169, -0.5817$	6.68×10^{-6}
STG $\rightarrow$ STG	-0.3202 ± 0.0926	$-0.0498, -0.5661$	0.0001206
FEF $\rightarrow$ FEF	-0.8093 ± 0.1111	$-0.8368, -0.7842$	4.77×10^{-7}
FEF $\rightarrow$ vIPFC	-0.4312 ± 0.1403	$-0.4433, -0.4201$	0.0008001
vIPFC $\rightarrow$ FEF	-0.6870 ± 0.1330	$-0.7032, -0.6721$	1.57×10^{-5}
vIPFC $\rightarrow$ vIPFC	-0.8698 ± 0.1159	$-0.9023, -0.8402$	4.77×10^{-7}

D.7 Per-area Lyapunov: leading exponent and within-area spectra

Per-area within-area Lyapunov leading-exponent results (anesthesia – awake) are reported in Table D.12.

The single per-area cell that passes the sign-agree gate on the leading exponent is FEF; the per-area full-spectrum mean (reported in the chapter (§4.4)) passes the gate on STG, FEF, and vIPFC, indicating that the cohort-wide destabilization signal is broad-spectrum rather than concentrated in the leading mode.

Table D.12: Per-area within-area Lyapunov leading exponent $\max \lambda$ (1/s), anesthesia – awake. Per-area paired two-sided Wilcoxon; “sign-agree” indicates whether the per-NHP mean differences agree.

Area	$\max \lambda$ awake	$\max \lambda$ anesthesia	Δ	Wilcoxon p	sign-agree
PPC	-14.33 ± 2.49	-13.53 ± 2.97	$+0.81 \pm 1.10$	0.68	no
STG	-27.15 ± 3.73	-25.95 ± 3.32	$+1.20 \pm 0.73$	0.14	yes
FEF	-36.37 ± 3.44	-34.02 ± 3.47	$+2.35 \pm 0.73$	0.0025	yes
vIPFC	-28.88 ± 3.48	-30.29 ± 3.63	-1.40 ± 1.30	0.29	no

D.8 Per-NHP session bookkeeping

Per-animal session counts and per-condition window lengths are reported in Table D.13.

Table D.13: Session counts per animal and per condition. Resting-state windows: 15 min awake-baseline, 30 min maintenance-dose centered on the middle of the maintenance epoch.

Animal	sessions	awake (min)	maintenance (min)	areas
NHP 1	10	15	30	PPC, STG, FEF, vIPFC
NHP 2	11	15	30	PPC, STG, FEF, vIPFC
Total	21	—	—	—

D.9 Software and reproducibility

All numeric results can be reproduced by checking out the JacobianODE library at commit `4c6e86f8` and the MindControl analysis pipeline at the corresponding commit on the `main` branch, and running the sweep with the hyperparameter grid in Table D.1 on the sessions listed in the per-NHP bookkeeping table. PyTorch 2.7.1+cu118, torchdiffeq 0.2.5, SciPy 1.17.1 are pinned in the repository’s environment file. Training was carried out on the Engaging compute cluster at MIT.

References

- [1] C. Koch. *What is Consciousness?* June 2018.
- [2] T. Nagel. “What Is It Like to Be a Bat?” In: *Philos. Rev.* 83 (4 1974), pp. 435–450. DOI: [10.2307/2183914](https://doi.org/10.2307/2183914).
- [3] D. Chalmers. “Facing up to the problem of consciousness.” In: *Journal of Consciousness Studies* 2 (3 1995), pp. 200–219.
- [4] M. A. Cohen and D. C. Dennett. “Consciousness cannot be separated from function.” en. In: *Trends Cogn. Sci.* 15 (8 Aug. 2011), pp. 358–364. DOI: [10.1016/j.tics.2011.06.008](https://doi.org/10.1016/j.tics.2011.06.008).
- [5] W. Seager. “A brief history of the philosophical problem of consciousness.” In: *The Cambridge Handbook of Consciousness*. Cambridge University Press, May 2007. DOI: [10.1017/CBO9780511816789.003](https://doi.org/10.1017/CBO9780511816789.003).
- [6] *Aristotle: De Anima (Reeve Edition)*. en. Trans. by C. D. C. Reeve. Hackett Publishing Company, Sept. 2017.
- [7] P. Olivelle, ed. *The Early Upanishads - Annotated Text and Translation*. en. SUSA: Oxford University Press, Sept. 24, 1998.
- [8] S. K. Saksena. “Nature of consciousness in Hindu philosophy.” en. University of London, School of Oriental and African Studies (United Kingdom), 1939. 319 pp.
- [9] K. R. Rao. *Consciousness Studies: Cross-cultural Perspectives*. en. Jefferson, NC: McFarland, Feb. 28, 2005. 376 pp.
- [10] J. Shear. “Eastern methods for investigating mind and consciousness.” en. In: *The Blackwell Companion to Consciousness*. Malden, MA, USA: Blackwell Publishing, Dec. 13, 2007, pp. 697–710. DOI: [10.1002/9780470751466.ch55](https://doi.org/10.1002/9780470751466.ch55).
- [11] N. Block. “On a confusion about a function of consciousness.” In: *Behav. Brain Sci.* 18 (2 June 1995), pp. 227–247. DOI: [10.1017/S0140525X00038188](https://doi.org/10.1017/S0140525X00038188).
- [12] N. Block. “Perceptual consciousness overflows cognitive access.” en. In: *Trends Cogn. Sci.* 15 (12 Dec. 2011), pp. 567–575. DOI: [10.1016/j.tics.2011.11.001](https://doi.org/10.1016/j.tics.2011.11.001).
- [13] D. C. Dennett. *Sweet dreams: Philosophical obstacles to a science of consciousness*. en. Jean Nicod Lectures. Cambridge, MA: Bradford Books, Sept. 8, 2006. 216 pp.
- [14] A. K. Seth and T. Bayne. “Theories of consciousness.” en. In: *Nat. Rev. Neurosci.* (May 3, 2022). DOI: [10.1038/s41583-022-00587-4](https://doi.org/10.1038/s41583-022-00587-4).
- [15] B. J. Baars. *A Cognitive Theory of Consciousness*. New York: Cambridge University Press, 1988.

- [16] G. A. Mashour, P. Roelfsema, J.-P. Changeux, and S. Dehaene. “Conscious Processing and the Global Neuronal Workspace Hypothesis.” en. In: *Neuron* 105 (5 Mar. 4, 2020), pp. 776–798. DOI: [10.1016/j.neuron.2020.01.026](https://doi.org/10.1016/j.neuron.2020.01.026).
- [17] G. Tononi, M. Boly, M. Massimini, and C. Koch. “Integrated information theory: from consciousness to its physical substrate.” en. In: *Nat. Rev. Neurosci.* 17 (7 July 2016), pp. 450–461. DOI: [10.1038/nrn.2016.44](https://doi.org/10.1038/nrn.2016.44).
- [18] G. Tononi. “Consciousness as integrated information: a provisional manifesto.” en. In: *Biol. Bull.* 215 (3 Dec. 2008), pp. 216–242. DOI: [10.2307/25470707](https://doi.org/10.2307/25470707).
- [19] R. Brown, H. Lau, and J. E. LeDoux. “Understanding the Higher-Order Approach to Consciousness.” en. In: *Trends Cogn. Sci.* 23 (9 Sept. 2019), pp. 754–768. DOI: [10.1016/j.tics.2019.06.009](https://doi.org/10.1016/j.tics.2019.06.009).
- [20] M. S. A. Graziano. “The Attention Schema Theory: A Foundation for Engineering Artificial Consciousness.” In: *Frontiers in Robotics and AI* 4 (2017). DOI: [10.3389/frobt.2017.00060](https://doi.org/10.3389/frobt.2017.00060).
- [21] F. Crick and C. Koch. “Towards a neurobiological theory of consciousness.” In: *Seminars in the Neurosciences* 2 (1990), pp. 263–275.
- [22] J. Hohwy and A. K. Seth. “Predictive processing as a systematic basis for identifying the neural correlates of consciousness.” In: *Philosophy and the Mind Sciences* 1 (II 2020), p. 3.
- [23] K. Friston. “The free-energy principle: A unified brain theory?” In: *Nat. Rev. Neurosci.* 11 (2 2010), pp. 127–138. DOI: [10.1038/nrn2787](https://doi.org/10.1038/nrn2787).
- [24] S. Hameroff and R. Penrose. “Consciousness in the universe: a review of the ‘Orch OR’ theory.” en. In: *Phys. Life Rev.* 11 (1 Mar. 2014), pp. 39–78. DOI: [10.1016/j.plrev.2013.08.002](https://doi.org/10.1016/j.plrev.2013.08.002).
- [25] S. H. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering*. en. Avalon Publishing, July 29, 2014. 528 pp.
- [26] J. J. Hopfield. “Neurons with graded response have collective computational properties like those of two-state neurons.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 81 (10 May 1984), pp. 3088–3092. DOI: [10.1073/pnas.81.10.3088](https://doi.org/10.1073/pnas.81.10.3088).
- [27] H. Sompolinsky, A. Crisanti, and H. J. Sommers. “Chaos in random neural networks.” en. In: *Phys. Rev. Lett.* 61 (3 July 18, 1988), pp. 259–262. DOI: [10.1103/PhysRevLett.61.259](https://doi.org/10.1103/PhysRevLett.61.259).
- [28] G. Mongillo, O. Barak, and M. Tsodyks. “Synaptic Theory of Working Memory.” In: *Science* 319 (5869 2008), pp. 1543–1546. DOI: [10.1126/science.1150769](https://doi.org/10.1126/science.1150769).
- [29] K. Zhang. “Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: a theory.” en. In: *J. Neurosci.* 16 (6 Mar. 15, 1996), pp. 2112–2126. DOI: [10.1523/JNEUROSCI.16-06-02112.1996](https://doi.org/10.1523/JNEUROSCI.16-06-02112.1996).
- [30] Y. Burak and I. R. Fiete. “Accurate path integration in continuous attractor network models of grid cells.” en. In: *PLoS Comput. Biol.* 5 (2 Feb. 2009), e1000291. DOI: [10.1371/journal.pcbi.1000291](https://doi.org/10.1371/journal.pcbi.1000291).

- [31] A. K. Churchland, R. Kiani, and M. N. Shadlen. “Decision-making with multiple alternatives.” en. In: *Nat. Neurosci.* 11 (6 June 2008), pp. 693–702. DOI: [10.1038/nn.2123](https://doi.org/10.1038/nn.2123).
- [32] D. Sussillo and L. F. Abbott. “Generating coherent patterns of activity from chaotic neural networks.” en. In: *Neuron* 63 (4 Aug. 27, 2009), pp. 544–557. DOI: [10.1016/j.neuron.2009.07.018](https://doi.org/10.1016/j.neuron.2009.07.018).
- [33] D. Sussillo. “Neural circuits as computational dynamical systems.” en. In: *Curr. Opin. Neurobiol.* 25 (Apr. 2014), pp. 156–163. DOI: [10.1016/j.conb.2014.01.008](https://doi.org/10.1016/j.conb.2014.01.008).
- [34] M. M. Churchland, J. P. Cunningham, M. T. Kaufman, J. D. Foster, P. Nuyujukian, S. I. Ryu, and K. V. Shenoy. “Neural population dynamics during reaching.” en. In: *Nature* 487 (7405 July 5, 2012), pp. 51–56. DOI: [10.1038/nature11129](https://doi.org/10.1038/nature11129).
- [35] M. Khona and I. R. Fiete. “Attractor and integrator networks in the brain.” en. In: *Nat. Rev. Neurosci.* 23 (12 Dec. 2022), pp. 744–766. DOI: [10.1038/s41583-022-00642-0](https://doi.org/10.1038/s41583-022-00642-0).
- [36] S. Vyas, M. D. Golub, D. Sussillo, and K. V. Shenoy. “Computation Through Neural Population Dynamics.” en. In: *Annu. Rev. Neurosci.* 43 (July 8, 2020), pp. 249–275. DOI: [10.1146/annurev-neuro-092619-094115](https://doi.org/10.1146/annurev-neuro-092619-094115).
- [37] A. J. Eisen, L. Kozachkov, A. M. Bastos, J. A. Donoghue, M. K. Mahnke, S. L. Brincat, S. Chandra, J. Tauber, E. N. Brown, I. R. Fiete, and E. K. Miller. “Propofol anesthesia destabilizes neural dynamics across cortex.” en. In: *Neuron* (July 10, 2024). DOI: [10.1016/j.neuron.2024.06.011](https://doi.org/10.1016/j.neuron.2024.06.011).
- [38] A. J. Eisen, A. G. Bardon, J. J. Ballesteros, A. M. Bastos, J. A. Donoghue, M. K. Mahnke, S. L. Brincat, J. E. Roy, Y. Ishizawa, E. N. Brown, I. R. Fiete, and E. K. Miller. “Similar destabilization of neural dynamics under different general anesthetics.” en. In: *Cell Rep.* (117048 Mar. 17, 2026), p. 117048. DOI: [10.1016/j.celrep.2026.117048](https://doi.org/10.1016/j.celrep.2026.117048).
- [39] A. J. Eisen, M. Ostrow, S. Chandra, L. Kozachkov, E. K. Miller, and I. R. Fiete. “Characterizing control between interacting subsystems with deep Jacobian estimation.” In: *The Thirty-ninth Annual Conference on Neural Information Processing Systems*. Oct. 29, 2025.
- [40] A. J. Eisen, A. M. Bastos, J. A. Donoghue, S. L. Brincat, E. N. Brown, I. R. Fiete, and E. K. Miller. “Disruption of interareal control during propofol anesthesia.” en. In: *bioRxiv* (Aug. 4, 2026), p. 2026.07.28.741350. DOI: [10.64898/2026.07.28.741350](https://doi.org/10.64898/2026.07.28.741350).
- [41] L. D. Lewis, V. S. Weiner, E. A. Mukamel, J. A. Donoghue, E. N. Eskandar, J. R. Madsen, W. S. Anderson, L. R. Hochberg, S. S. Cash, E. N. Brown, and P. L. Purdon. “Rapid fragmentation of neuronal networks at the onset of propofol-induced unconsciousness.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 109 (49 Dec. 4, 2012), E3377–86. DOI: [10.1073/pnas.1210907109](https://doi.org/10.1073/pnas.1210907109).
- [42] A. M. Bastos, J. A. Donoghue, S. L. Brincat, M. Mahnke, J. Yanar, J. Correa, A. S. Waite, M. Lundqvist, J. Roy, E. N. Brown, and E. K. Miller. “Neural effects of propofol-induced unconsciousness and its reversal using thalamic stimulation.” en. In: *Elife* 10 (Apr. 27, 2021). DOI: [10.7554/eLife.60824](https://doi.org/10.7554/eLife.60824).

- [43] F. J. Flores, K. E. Hartnack, A. B. Fath, S.-E. Kim, M. A. Wilson, E. N. Brown, and P. L. Purdon. “Thalamocortical synchronization during induction and emergence from propofol-induced unconsciousness.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 114 (32 Aug. 8, 2017), E6660–E6668. DOI: [10.1073/pnas.1700148114](https://doi.org/10.1073/pnas.1700148114).
- [44] S. Palva and J. M. Palva. “New vistas for alpha-frequency band oscillations.” en. In: *Trends Neurosci.* 30 (4 Apr. 2007), pp. 150–158. DOI: [10.1016/j.tins.2007.02.001](https://doi.org/10.1016/j.tins.2007.02.001).
- [45] S. Ching, A. Cimenser, P. L. Purdon, E. N. Brown, and N. J. Kopell. “Thalamocortical model for a propofol-induced alpha-rhythm associated with loss of consciousness.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 107 (52 Dec. 28, 2010), pp. 22665–22670. DOI: [10.1073/pnas.1017069108](https://doi.org/10.1073/pnas.1017069108).
- [46] E. N. Brown, R. Lydic, and N. D. Schiff. “General anesthesia, sleep, and coma.” en. In: *N. Engl. J. Med.* 363 (27 Dec. 30, 2010), pp. 2638–2650. DOI: [10.1056/NEJMr0808281](https://doi.org/10.1056/NEJMr0808281).
- [47] Y. B. Saalman and S. Kastner. “The cognitive thalamus.” en. In: *Front. Syst. Neurosci.* 9 (Mar. 17, 2015), p. 39. DOI: [10.3389/fnsys.2015.00039](https://doi.org/10.3389/fnsys.2015.00039).
- [48] S. Dehaene and J.-P. Changeux. “Experimental and theoretical approaches to conscious processing.” en. In: *Neuron* 70 (2 Apr. 28, 2011), pp. 200–227. DOI: [10.1016/j.neuron.2011.03.018](https://doi.org/10.1016/j.neuron.2011.03.018).
- [49] S. Dehaene, C. Sergent, and J.-P. Changeux. “A neuronal network model linking subjective reports and objective physiological data during conscious perception.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 100 (14 July 8, 2003), pp. 8520–8525. DOI: [10.1073/pnas.1332574100](https://doi.org/10.1073/pnas.1332574100).
- [50] A. Babloyantz and A. Destexhe. “Low-dimensional chaos in an instance of epilepsy.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 83 (10 May 1986), pp. 3513–3517. DOI: [10.1073/pnas.83.10.3513](https://doi.org/10.1073/pnas.83.10.3513).
- [51] J. Theiler. “On the evidence for low-dimensional chaos in an epileptic electroencephalogram.” In: *Phys. Lett. A* 196 (5 Jan. 2, 1995), pp. 335–341. DOI: [10.1016/0375-9601\(94\)00856-K](https://doi.org/10.1016/0375-9601(94)00856-K).
- [52] R. L. Carhart-Harris, R. Leech, P. J. Hellyer, M. Shanahan, A. Feilding, E. Tagliazucchi, D. R. Chialvo, and D. Nutt. “The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs.” en. In: *Front. Hum. Neurosci.* 8 (Feb. 3, 2014), p. 20. DOI: [10.3389/fnhum.2014.00020](https://doi.org/10.3389/fnhum.2014.00020).
- [53] J. M. Beggs. “The criticality hypothesis: how local cortical networks might optimize information processing.” en. In: *Philos. Trans. A Math. Phys. Eng. Sci.* 366 (1864 Feb. 13, 2008), pp. 329–343. DOI: [10.1098/rsta.2007.2092](https://doi.org/10.1098/rsta.2007.2092).
- [54] L. Kozachkov, M. Lundqvist, J. J. Slotine, and E. K. Miller. “Achieving stable dynamics in neural circuits.” In: *PLoS Comput. Biol.* 16 (8 2020), pp. 1–15. DOI: [10.1371/JOURNAL.PCBI.1007659](https://doi.org/10.1371/JOURNAL.PCBI.1007659).

- [55] A. Demertzi, E. Tagliazucchi, S. Dehaene, G. Deco, P. Barttfeld, F. Raimondo, C. Martial, D. Fernández-Espejo, B. Rohaut, H. U. Voss, N. D. Schiff, A. M. Owen, S. Laureys, L. Naccache, and J. D. Sitt. “Human consciousness is supported by dynamic complex patterns of brain signal coordination.” en. In: *Sci Adv* 5 (2 Feb. 2019), eaat7603. DOI: [10.1126/sciadv.aat7603](https://doi.org/10.1126/sciadv.aat7603).
- [56] J. J. Hopfield. “Neural networks and physical systems with emergent collective computational abilities.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 79 (8 Apr. 1982), pp. 2554–2558. DOI: [10.1073/pnas.79.8.2554](https://doi.org/10.1073/pnas.79.8.2554).
- [57] M. W. Hirsch. “Convergent activation dynamics in continuous time networks.” In: *Neural Netw.* 2 (5 Jan. 1, 1989), pp. 331–349. DOI: [10.1016/0893-6080\(89\)90018-X](https://doi.org/10.1016/0893-6080(89)90018-X).
- [58] M. A. Cohen and S. Grossberg. “Absolute Stability of Global Pattern Formation and Parallel Memory Storage by Competitive Neural Networks.” In: *Advances in Psychology*. Ed. by S. Grossberg. Vol. 42. North-Holland, Jan. 1, 1987, pp. 288–308. DOI: [10.1016/S0166-4115\(08\)60913-9](https://doi.org/10.1016/S0166-4115(08)60913-9).
- [59] H. Sohn, D. Narain, N. Meirhaeghe, and M. Jazayeri. “Bayesian Computation through Cortical Latent Dynamics.” en. In: *Neuron* 103 (5 Sept. 4, 2019), 934–947.e5. DOI: [10.1016/j.neuron.2019.06.012](https://doi.org/10.1016/j.neuron.2019.06.012).
- [60] E. N. Lorenz. “Deterministic Nonperiodic Flow.” en. In: *J. Atmos. Sci.* 20 (2 Mar. 1, 1963), pp. 130–141. DOI: [10.1175/1520-0469\(1963\)020<0130:DNF>2.0.CO;2](https://doi.org/10.1175/1520-0469(1963)020<0130:DNF>2.0.CO;2).
- [61] D. Toker, I. Pappas, J. D. Lendner, J. Frohlich, D. M. Mateos, S. Muthukumaraswamy, R. Carhart-Harris, M. Paff, P. M. Vespa, M. M. Monti, F. T. Sommer, R. T. Knight, and M. D’Esposito. “Consciousness is supported by near-critical slow cortical electrodynamics.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 119 (7 Feb. 15, 2022). DOI: [10.1073/pnas.2024455119](https://doi.org/10.1073/pnas.2024455119).
- [62] A. López-González, R. Panda, A. Ponce-Alvarez, G. Zamora-López, A. Escrichs, C. Martial, A. Thibaut, O. Gosseries, M. L. Kringelbach, J. Annen, S. Laureys, and G. Deco. “Loss of consciousness reduces the stability of brain hubs and the heterogeneity of brain dynamics.” en. In: *Commun Biol* 4 (1 Sept. 6, 2021), p. 1037. DOI: [10.1038/s42003-021-02537-9](https://doi.org/10.1038/s42003-021-02537-9).
- [63] G. Solovey, L. M. Alonso, T. Yanagawa, N. Fujii, M. O. Magnasco, G. A. Cecchi, and A. Proekt. “Loss of Consciousness Is Associated with Stabilization of Cortical Activity.” en. In: *J. Neurosci.* 35 (30 July 29, 2015), pp. 10866–10877. DOI: [10.1523/JNEUROSCI.4895-14.2015](https://doi.org/10.1523/JNEUROSCI.4895-14.2015).
- [64] L. M. Alonso, A. Proekt, T. H. Schwartz, K. O. Pryor, G. A. Cecchi, and M. O. Magnasco. “Dynamical criticality during induction of anesthesia in human ECoG recordings.” en. In: *Front. Neural Circuits* 8 (Mar. 25, 2014), p. 20. DOI: [10.3389/fncir.2014.00020](https://doi.org/10.3389/fncir.2014.00020).
- [65] D. Krzemiński, M. Kamiński, A. Marchewka, and M. Bola. “Breakdown of long-range temporal correlations in brain oscillations during general anesthesia.” en. In: *Neuroimage* 159 (Oct. 1, 2017), pp. 146–158. DOI: [10.1016/j.neuroimage.2017.07.047](https://doi.org/10.1016/j.neuroimage.2017.07.047).

- [66] F. Takens. “Detecting strange attractors in turbulence.” In: *Dynamical Systems and Turbulence, Warwick 1980*. Springer Berlin Heidelberg, 1981, pp. 366–381. DOI: [10.1007/BFb0091924](https://doi.org/10.1007/BFb0091924).
- [67] S. L. Brunton, B. W. Brunton, J. L. Proctor, E. Kaiser, and J. N. Kutz. “Chaos as an intermittently forced linear system.” en. In: *Nat. Commun.* 8 (1 May 30, 2017), p. 19. DOI: [10.1038/s41467-017-00030-8](https://doi.org/10.1038/s41467-017-00030-8).
- [68] M. Kamb, E. Kaiser, S. L. Brunton, and J. N. Kutz. “Time-Delay Observables for Koopman: Theory and Applications.” In: *SIAM J. Appl. Dyn. Syst.* 19 (2 Jan. 1, 2020), pp. 886–917. DOI: [10.1137/18M1216572](https://doi.org/10.1137/18M1216572).
- [69] A. C. Costa, T. Ahamed, D. Jordan, and G. J. Stephens. “Maximally predictive states: From partial observations to long timescales.” en. In: *Chaos* 33 (2 Feb. 2023), p. 023136. DOI: [10.1063/5.0129398](https://doi.org/10.1063/5.0129398).
- [70] N. Dhir, A. R. Kosiorek, and I. Posner. “Bayesian delay embeddings for dynamical systems.” In: *NIPS Time Series Workshop 2017* (2017).
- [71] Y. Susuki and I. Mezić. “A prony approximation of Koopman Mode Decomposition.” In: *2015 54th IEEE Conference on Decision and Control (CDC)*. Dec. 2015, pp. 7022–7027. DOI: [10.1109/CDC.2015.7403326](https://doi.org/10.1109/CDC.2015.7403326).
- [72] H. Arbabi, M. Korda, and I. Mezić. “A Data-Driven Koopman Model Predictive Control Framework for Nonlinear Partial Differential Equations.” In: *2018 IEEE Conference on Decision and Control (CDC)*. Dec. 2018, pp. 6409–6414. DOI: [10.1109/CDC.2018.8619720](https://doi.org/10.1109/CDC.2018.8619720).
- [73] B. W. Brunton, L. A. Johnson, J. G. Ojemann, and J. N. Kutz. “Extracting spatial-temporal coherent patterns in large-scale neural recordings using dynamic mode decomposition.” en. In: *J. Neurosci. Methods* 258 (Jan. 30, 2016), pp. 1–15. DOI: [10.1016/j.jneumeth.2015.10.010](https://doi.org/10.1016/j.jneumeth.2015.10.010).
- [74] J. Bakarji, K. Champion, J. Nathan Kutz, and S. L. Brunton. “Discovering governing equations from partial measurements with deep delay autoencoders.” In: *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 479 (2276 Aug. 30, 2023), p. 20230422. DOI: [10.1098/rspa.2023.0422](https://doi.org/10.1098/rspa.2023.0422).
- [75] J. Axås and G. Haller. “Model reduction for nonlinearizable dynamics via delay-embedded spectral submanifolds.” In: *Nonlinear Dyn.* 111 (24 Dec. 1, 2023), pp. 22079–22099. DOI: [10.1007/s11071-023-08705-2](https://doi.org/10.1007/s11071-023-08705-2).
- [76] J. N. Juang and R. S. Pappa. “An Eigensystem Realization Algorithm (ERA) for modal parameter identification and model reduction.” In: *JPL Proc. of the Workshop on Identification and Control of Flexible Space Struct., Vol. 3* (Apr. 1, 1985).
- [77] M. A. Khodkar and P. Hassanzadeh. “A data-driven, physics-informed framework for forecasting the spatiotemporal evolution of chaotic dynamics with nonlinearities modeled as exogenous forcings.” In: *J. Comput. Phys.* 440 (Sept. 1, 2021), p. 110412. DOI: [10.1016/j.jcp.2021.110412](https://doi.org/10.1016/j.jcp.2021.110412).

- [78] D. J. Gauthier, E. Bollt, A. Griffith, and W. A. S. Barbosa. “Next generation reservoir computing.” en. In: *Nat. Commun.* 12 (1 Sept. 21, 2021), p. 5564. DOI: [10.1038/s41467-021-25801-2](https://doi.org/10.1038/s41467-021-25801-2).
- [79] J. Crutchfield and B. S. McNamara. “Equations of Motion from a Data Series.” en. In: *Complex Systems* (1987).
- [80] G. Sugihara and R. M. May. “Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series.” en. In: *Nature* 344 (6268 Apr. 19, 1990), pp. 734–741. DOI: [10.1038/344734a0](https://doi.org/10.1038/344734a0).
- [81] G. Rowlands and J. C. Sprott. “Extraction of dynamical equations from chaotic data.” In: *Physica D* 58 (1 Sept. 15, 1992), pp. 251–259. DOI: [10.1016/0167-2789\(92\)90113-2](https://doi.org/10.1016/0167-2789(92)90113-2).
- [82] H. D. I. Abarbanel, R. Brown, J. J. Sidorowich, and L. S. Tsimring. “The analysis of observed chaotic data in physical systems.” In: *Rev. Mod. Phys.* 65 (4 Oct. 1, 1993), pp. 1331–1392. DOI: [10.1103/RevModPhys.65.1331](https://doi.org/10.1103/RevModPhys.65.1331).
- [83] H. Ye, E. R. Deyle, L. J. Gilarranz, and G. Sugihara. “Distinguishing time-delayed causal interactions using convergent cross mapping.” en. In: *Sci. Rep.* 5 (Oct. 5, 2015), p. 14750. DOI: [10.1038/srep14750](https://doi.org/10.1038/srep14750).
- [84] H. Ye, R. J. Beamish, S. M. Glaser, S. C. H. Grant, C.-H. Hsieh, L. J. Richards, J. T. Schnute, and G. Sugihara. “Equation-free mechanistic ecosystem forecasting using empirical dynamic modeling.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 112 (13 Mar. 31, 2015), E1569–76. DOI: [10.1073/pnas.1417063112](https://doi.org/10.1073/pnas.1417063112).
- [85] E. R. Deyle, R. M. May, S. B. Munch, and G. Sugihara. “Tracking and forecasting ecosystem interactions in real time.” en. In: *Proc. Biol. Sci.* 283 (1822 Jan. 13, 2016). DOI: [10.1098/rspb.2015.2258](https://doi.org/10.1098/rspb.2015.2258).
- [86] J. Park, G. M. Pao, G. Sugihara, E. Stabenau, and T. Lorimer. “Empirical mode modeling.” In: *Nonlinear Dyn.* 108 (3 May 1, 2022), pp. 2147–2160. DOI: [10.1007/s11071-022-07311-y](https://doi.org/10.1007/s11071-022-07311-y).
- [87] S. Tajima, T. Yanagawa, N. Fujii, and T. Toyozumi. “Untangling Brain-Wide Dynamics in Consciousness by Cross-Embedding.” en. In: *PLoS Comput. Biol.* 11 (11 Nov. 2015), e1004537. DOI: [10.1371/journal.pcbi.1004537](https://doi.org/10.1371/journal.pcbi.1004537).
- [88] T. Ahamed, A. C. Costa, and G. J. Stephens. “Capturing the continuous complexity of behaviour in *Caenorhabditis elegans*.” en. In: *Nat. Phys.* 17 (2 Oct. 5, 2020), pp. 275–283. DOI: [10.1038/s41567-020-01036-8](https://doi.org/10.1038/s41567-020-01036-8).
- [89] W. Gilpin. “Deep reconstruction of strange attractors from time series.” In: *Proceedings of the 34th International Conference on Neural Information Processing Systems* (Vancouver, BC, Canada). NIPS’20 Article 18. Red Hook, NY, USA: Curran Associates Inc., Dec. 6, 2020, pp. 204–216.
- [90] W. Watanakesuntorn, K. Takahashi, K. Ichikawa, J. Park, G. Sugihara, R. Takano, J. Haga, and G. M. Pao. “Massively Parallel Causal Inference of Whole Brain Dynamics at Single Neuron Resolution.” In: *2020 IEEE 26th International Conference on Parallel and Distributed Systems (ICPADS)*. IEEE, Dec. 2020, pp. 196–205. DOI: [10.1109/ICPADS51040.2020.00035](https://doi.org/10.1109/ICPADS51040.2020.00035).

- [91] Z. Zhang, K. Li, and X. Hu. “Mapping nonlinear brain dynamics by phase space embedding with fMRI data.” In: *Biomed. Signal Process. Control* 82 (Apr. 1, 2023), p. 104521. DOI: [10.1016/j.bspc.2022.104521](https://doi.org/10.1016/j.bspc.2022.104521).
- [92] R. V. Raut, Z. P. Rosenthal, X. Wang, H. Miao, Z. Zhang, J.-M. Lee, M. E. Raichle, A. Q. Bauer, S. L. Brunton, B. W. Brunton, and J. Nathan Kutz. “Arousal as a universal embedding for spatiotemporal brain dynamics.” en. In: *bioRxiv* (Dec. 21, 2023), p. 2023.11.06.565918. DOI: [10.1101/2023.11.06.565918](https://doi.org/10.1101/2023.11.06.565918).
- [93] H. Mori. “Transport, Collective Motion, and Brownian Motion.” en. In: *Progr. Theoret. Phys.* 33 (3 Mar. 1965), pp. 423–455. DOI: [10.1143/ptp.33.423](https://doi.org/10.1143/ptp.33.423).
- [94] R. Zwanzig. “Nonlinear generalized Langevin equations.” In: *J. Stat. Phys.* 9 (3 Nov. 1, 1973), pp. 215–220. DOI: [10.1007/BF01008729](https://doi.org/10.1007/BF01008729).
- [95] Y. T. Lin, Y. Tian, D. Livescu, and M. Anghel. “Data-Driven Learning for the Mori–Zwanzig Formalism: A Generalization of the Koopman Learning Framework.” In: *SIAM J. Appl. Dyn. Syst.* 20 (4 Jan. 1, 2021), pp. 2558–2601. DOI: [10.1137/21M1401759](https://doi.org/10.1137/21M1401759).
- [96] A. J. Chorin, O. H. Hald, and R. Kupferman. “Optimal prediction and the Mori–Zwanzig representation of irreversible processes.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 97 (7 Mar. 28, 2000), pp. 2968–2973. DOI: [10.1073/pnas.97.7.2968](https://doi.org/10.1073/pnas.97.7.2968).
- [97] K. K. Lin and F. Lu. “Data-driven model reduction, Wiener projections, and the Koopman–Mori–Zwanzig formalism.” In: *J. Comput. Phys.* 424 (Jan. 1, 2021), p. 109864. DOI: [10.1016/j.jcp.2020.109864](https://doi.org/10.1016/j.jcp.2020.109864).
- [98] C. Hijón, P. Español, E. Vanden-Eijnden, and R. Delgado-Buscalioni. “Mori–Zwanzig formalism as a practical computational tool.” en. In: *Faraday Discuss.* 144 (2010), 301–22, discussion 323–45, 467–81. DOI: [10.1039/b902479b](https://doi.org/10.1039/b902479b).
- [99] B. O. Koopman. “Hamiltonian Systems and Transformation in Hilbert Space.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 17 (5 May 1931), pp. 315–318. DOI: [10.1073/pnas.17.5.315](https://doi.org/10.1073/pnas.17.5.315).
- [100] S. L. Brunton, M. Budišić, E. Kaiser, and J. N. Kutz. “Modern Koopman Theory for Dynamical Systems.” In: *SIAM Rev.* 64 (2 May 5, 2022), pp. 229–340. DOI: [10.1137/21M1401243](https://doi.org/10.1137/21M1401243).
- [101] I. Mezić. “Spectral Properties of Dynamical Systems, Model Reduction and Decompositions.” In: *Nonlinear Dyn.* 41 (1 Aug. 1, 2005), pp. 309–325. DOI: [10.1007/s11071-005-2824-x](https://doi.org/10.1007/s11071-005-2824-x).
- [102] H. Arbabi and I. Mezić. “Ergodic Theory, Dynamic Mode Decomposition, and Computation of Spectral Properties of the Koopman Operator.” In: *SIAM J. Appl. Dyn. Syst.* 16 (4 Jan. 1, 2017), pp. 2096–2126. DOI: [10.1137/17M1125236](https://doi.org/10.1137/17M1125236).
- [103] A. Kusaba, T. Kuboyama, K. Shin, M. Sasaki, and S. Inagaki. “A new combination of Hankel and sparsity-promoting dynamic mode decompositions and its application to the prediction of plasma turbulence.” en. In: *Jpn. J. Appl. Phys.* 61 (SA Dec. 15, 2021), SA1011. DOI: [10.35848/1347-4065/ac1c3c](https://doi.org/10.35848/1347-4065/ac1c3c).

- [104] S. M. Hirsh, S. M. Ichinaga, S. L. Brunton, J. Nathan Kutz, and B. W. Brunton. “Structured time-delay models for dynamical systems with connections to Frenet-Serret frame.” en. In: *Proc. Math. Phys. Eng. Sci.* 477 (2254 Oct. 2021), p. 20210097. DOI: [10.1098/rspa.2021.0097](https://doi.org/10.1098/rspa.2021.0097).
- [105] P. J. Schmid. “Dynamic mode decomposition of numerical and experimental data.” In: *J. Fluid Mech.* 656 (Aug. 2010), pp. 5–28. DOI: [10.1017/S0022112010001217](https://doi.org/10.1017/S0022112010001217).
- [106] C. W. Rowley, I. Mezić, S. Bagheri, P. Schlatter, and D. S. Henningson. “Spectral analysis of nonlinear flows.” In: *J. Fluid Mech.* 641 (Dec. 2009), pp. 115–127. DOI: [10.1017/S0022112009992059](https://doi.org/10.1017/S0022112009992059).
- [107] J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. Nathan Kutz. “On dynamic mode decomposition: Theory and applications.” en. In: *J. Comput. Dyn.* 1 (2 2014), pp. 391–421. DOI: [10.3934/jcd.2014.1.391](https://doi.org/10.3934/jcd.2014.1.391).
- [108] M. O. Williams, I. G. Kevrekidis, and C. W. Rowley. “A Data-Driven Approximation of the Koopman Operator: Extending Dynamic Mode Decomposition.” In: *J. Nonlinear Sci.* 25 (6 Dec. 1, 2015), pp. 1307–1346. DOI: [10.1007/s00332-015-9258-5](https://doi.org/10.1007/s00332-015-9258-5).
- [109] S. L. Brunton, J. L. Proctor, and J. N. Kutz. “Discovering governing equations from data by sparse identification of nonlinear dynamical systems.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 113 (15 Apr. 12, 2016), pp. 3932–3937. DOI: [10.1073/pnas.1517384113](https://doi.org/10.1073/pnas.1517384113).
- [110] M. O. Williams, C. W. Rowley, and I. G. Kevrekidis. “A kernel-based method for data-driven koopman spectral analysis.” In: *Journal of Computational Dynamics* 2 (2 May 1, 2016), pp. 247–265. DOI: [10.3934/jcd.2015005](https://doi.org/10.3934/jcd.2015005).
- [111] C. Folkestad, D. Pastor, I. Mezic, R. Mohr, M. Fonoberova, and J. Burdick. “Extended Dynamic Mode Decomposition with Learned Koopman Eigenfunctions for Prediction and Control.” In: *2020 American Control Conference (ACC)*. IEEE, July 2020, pp. 3906–3913. DOI: [10.23919/ACC45564.2020.9147729](https://doi.org/10.23919/ACC45564.2020.9147729).
- [112] D. J. Alford-Lago, C. W. Curtis, A. T. Ihler, and O. Issan. “Deep learning enhanced dynamic mode decomposition.” en. In: *Chaos* 32 (3 Mar. 2022), p. 033116. DOI: [10.1063/5.0073893](https://doi.org/10.1063/5.0073893).
- [113] B. Lusch, J. N. Kutz, and S. L. Brunton. “Deep learning for universal linear embeddings of nonlinear dynamics.” en. In: *Nat. Commun.* 9 (1 Nov. 23, 2018), p. 4950. DOI: [10.1038/s41467-018-07210-0](https://doi.org/10.1038/s41467-018-07210-0).
- [114] N. Takeishi, Y. Kawahara, and T. Yairi. “Learning Koopman invariant subspaces for dynamic mode decomposition.” In: *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA). NIPS’17. Red Hook, NY, USA: Curran Associates Inc., Dec. 4, 2017, pp. 1130–1140.
- [115] J. N. Kutz, X. Fu, and S. L. Brunton. “Multiresolution Dynamic Mode Decomposition.” In: *SIAM J. Appl. Dyn. Syst.* 15 (2 Jan. 1, 2016), pp. 713–735. DOI: [10.1137/15M1023543](https://doi.org/10.1137/15M1023543).

- [116] S. L. Brunton, B. W. Brunton, J. L. Proctor, and J. N. Kutz. “Koopman Invariant Subspaces and Finite Linear Representations of Nonlinear Dynamical Systems for Control.” en. In: *PLoS One* 11 (2 Feb. 26, 2016), e0150171. DOI: [10.1371/journal.pone.0150171](https://doi.org/10.1371/journal.pone.0150171).
- [117] J. Ferre, A. Rokem, E. A. Buffalo, J. Nathan Kutz, and A. Fairhall. “Non-Stationary Dynamic Mode Decomposition.” en. In: *bioRxiv* (Aug. 13, 2023), p. 2023.08.08.552333. DOI: [10.1101/2023.08.08.552333](https://doi.org/10.1101/2023.08.08.552333).
- [118] M. S. J. Solaija, S. Saleem, K. Khurshid, S. A. Hassan, and A. M. Kamboh. “Dynamic Mode Decomposition Based Epileptic Seizure Detection from Scalp EEG.” In: *IEEE Access* 6 (2018), pp. 38683–38692. DOI: [10.1109/ACCESS.2018.2853125](https://doi.org/10.1109/ACCESS.2018.2853125).
- [119] M. Bilal, M. Rizwan, S. Saleem, M. M. Khan, M. S. Alkathair, and M. Alqarni. “Automatic seizure detection using multi-resolution dynamic mode decomposition.” en. In: *IEEE Access* 7 (2019), pp. 61180–61194. DOI: [10.1109/access.2019.2915609](https://doi.org/10.1109/access.2019.2915609).
- [120] M. Ostrow, A. J. Eisen, L. Kozachkov, and I. R. Fiete. “Beyond geometry: Comparing the temporal structure of computation in neural circuits with dynamical similarity analysis.” In: *Adv. Neural Inf. Process. Syst.* proceedings.neurips.cc, 2023.
- [121] N. Marrouch, J. Slawinska, D. Giannakis, and H. L. Read. “Data-driven Koopman operator approach for computational neuroscience.” In: *Ann. Math. Artif. Intell.* 88 (11-12 Dec. 1, 2020), pp. 1155–1173. DOI: [10.1007/s10472-019-09666-2](https://doi.org/10.1007/s10472-019-09666-2).
- [122] B. van der Pol. “LXXXVIII. On “relaxation-oscillations”.” In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2 (11 Nov. 1, 1926), pp. 978–992. DOI: [10.1080/14786442608564127](https://doi.org/10.1080/14786442608564127).
- [123] K. Verheyden, T. Luzyanina, and D. Roose. “Efficient computation of characteristic roots of delay differential equations using LMS methods.” In: *J. Comput. Appl. Math.* 214 (1 Apr. 15, 2008), pp. 209–226. DOI: [10.1016/j.cam.2007.02.025](https://doi.org/10.1016/j.cam.2007.02.025).
- [124] D. Breda, S. Maset, and R. Vermiglio. “TRACE-DDE: a Tool for Robust Analysis and Characteristic Equations for Delay Differential Equations.” In: *Topics in Time Delay Systems: Analysis, Algorithms and Control*. Ed. by J. J. Loiseau, W. Michiels, S.-I. Niculescu, and R. Sipahi. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 145–155. DOI: [10.1007/978-3-642-02897-7_13](https://doi.org/10.1007/978-3-642-02897-7_13).
- [125] D. R. Chialvo. “Emergent complex neural dynamics.” en. In: *Nat. Phys.* 6 (10 Oct. 1, 2010), pp. 744–750. DOI: [10.1038/nphys1803](https://doi.org/10.1038/nphys1803).
- [126] G. Solovey, K. J. Miller, J. G. Ojemann, M. O. Magnasco, and G. A. Cecchi. “Self-Regulated Dynamical Criticality in Human ECoG.” en. In: *Front. Integr. Neurosci.* 6 (July 19, 2012), p. 44. DOI: [10.3389/fnint.2012.00044](https://doi.org/10.3389/fnint.2012.00044).
- [127] L. Dieci, R. D. Russell, and E. S. Van Vleck. “On the Computation of Lyapunov Exponents for Continuous Dynamical Systems.” In: *SIAM J. Numer. Anal.* 34 (1 1997), pp. 402–423.

- [128] A. Mohan, V. Vijesh, D. A. Thumba, and K. S. Kumar. “Recurrence Network-Based Approach to Distinguish Between Chaotic and Quasiperiodic Solution.” In: *Advances in Signal Processing and Intelligent Recognition Systems*. Springer Singapore, 2020, pp. 368–375. DOI: [10.1007/978-981-15-4828-4_30](https://doi.org/10.1007/978-981-15-4828-4_30).
- [129] D. Bai, P. S. Pennefather, J. F. MacDonald, and B. A. Orser. “The general anesthetic propofol slows deactivation and desensitization of GABA(A) receptors.” en. In: *J. Neurosci.* 19 (24 Dec. 15, 1999), pp. 10635–10646. DOI: [10.1523/JNEUROSCI.19-24-10635.1999](https://doi.org/10.1523/JNEUROSCI.19-24-10635.1999).
- [130] B. K. Murphy and K. D. Miller. “Balanced amplification: a new mechanism of selective amplification of neural activity patterns.” en. In: *Neuron* 61 (4 Feb. 26, 2009), pp. 635–648. DOI: [10.1016/j.neuron.2009.02.005](https://doi.org/10.1016/j.neuron.2009.02.005).
- [131] J. M. Tauber, S. L. Brincat, E. P. Stephen, J. A. Donoghue, L. Kozachkov, E. N. Brown, and E. K. Miller. “Propofol-mediated Unconsciousness Disrupts Progression of Sensory Signals through the Cortical Hierarchy.” en. In: *J. Cogn. Neurosci.* (Oct. 30, 2023), pp. 1–20. DOI: [10.1162/jocn_a_02081](https://doi.org/10.1162/jocn_a_02081).
- [132] R. Chaudhuri, B. Gerçek, B. Pandey, A. Peyrache, and I. Fiete. “The intrinsic attractor manifold and population dynamics of a canonical cognitive circuit across waking and sleep.” en. In: *Nat. Neurosci.* 22 (9 Sept. 2019), pp. 1512–1520. DOI: [10.1038/s41593-019-0460-x](https://doi.org/10.1038/s41593-019-0460-x).
- [133] F. Mastrogiuseppe and S. Ostojic. “Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks.” en. In: *Neuron* 99 (3 Aug. 8, 2018), 609–623.e29. DOI: [10.1016/j.neuron.2018.07.003](https://doi.org/10.1016/j.neuron.2018.07.003).
- [134] M. E. Hasselmo and M. P. Brandon. “A model combining oscillations and attractor dynamics for generation of grid cell firing.” en. In: *Front. Neural Circuits* 6 (May 28, 2012), p. 30. DOI: [10.3389/fncir.2012.00030](https://doi.org/10.3389/fncir.2012.00030).
- [135] A. Libby and T. J. Buschman. “Rotational dynamics reduce interference between sensory and memory representations.” en. In: *Nat. Neurosci.* (Apr. 5, 2021). DOI: [10.1038/s41593-021-00821-9](https://doi.org/10.1038/s41593-021-00821-9).
- [136] P. E. Holtzheimer and H. S. Mayberg. “Stuck in a rut: rethinking depression and its treatment.” en. In: *Trends Neurosci.* 34 (1 Jan. 2011), pp. 1–9. DOI: [10.1016/j.tins.2010.10.004](https://doi.org/10.1016/j.tins.2010.10.004).
- [137] D. F. Zullino and Y. Khazaal. “The “rut metaphor”: a conceptualization of attractor-shaping properties of addictive drugs.” en. In: *Subst. Use Misuse* 43 (3-4 2008), pp. 469–479. DOI: [10.1080/10826080701205042](https://doi.org/10.1080/10826080701205042).
- [138] R. L. Carhart-Harris, S. Chandaria, D. E. Erritzoe, A. Gazzaley, M. Girn, H. Kettner, P. A. M. Mediano, D. J. Nutt, F. E. Rosas, L. Roseman, C. Timmermann, B. Weiss, R. J. Zeifman, and K. J. Friston. “Canalization and plasticity in psychopathology.” en. In: *Neuropharmacology* 226 (109398 Mar. 15, 2023), p. 109398. DOI: [10.1016/j.neuropharm.2022.109398](https://doi.org/10.1016/j.neuropharm.2022.109398).

- [139] A. Juliani, A. Safron, and R. Kanai. “Deep CANALs: A deep learning approach to refining the canalization theory of psychopathology.” In: *PsyArXiv* (May 18, 2023). DOI: [10.31234/osf.io/uxmz6](https://doi.org/10.31234/osf.io/uxmz6).
- [140] U. Braun, A. Harneit, G. Pergola, T. Menara, A. Schäfer, R. F. Betzel, Z. Zang, J. I. Schweiger, X. Zhang, K. Schwarz, J. Chen, G. Blasi, A. Bertolino, D. Durstewitz, F. Pasqualetti, E. Schwarz, A. Meyer-Lindenberg, D. S. Bassett, and H. Tost. “Brain network dynamics during working memory are modulated by dopamine and diminished in schizophrenia.” en. In: *Nat. Commun.* 12 (1 June 9, 2021), p. 3478. DOI: [10.1038/s41467-021-23694-9](https://doi.org/10.1038/s41467-021-23694-9).
- [141] A. S. Mahadevan, E. J. Cornblath, D. M. Lydon-Staley, D. Zhou, L. Parkes, B. Larsen, A. Adebimpe, A. E. Kahn, R. C. Gur, R. E. Gur, T. D. Satterthwaite, D. H. Wolf, and D. S. Bassett. “Alprazolam modulates persistence energy during emotion processing in first-degree relatives of individuals with schizophrenia: a network control study.” en. In: *Mol. Psychiatry* 28 (8 Aug. 2023), pp. 3314–3323. DOI: [10.1038/s41380-023-02121-z](https://doi.org/10.1038/s41380-023-02121-z).
- [142] S. P. Singleton, A. I. Luppi, R. L. Carhart-Harris, J. Cruzat, L. Roseman, D. J. Nutt, G. Deco, M. L. Kringelbach, E. A. Stamatakis, and A. Kucyeski. “Receptor-informed network control theory links LSD and psilocybin to a flattening of the brain’s control energy landscape.” en. In: *Nat. Commun.* 13 (1 Oct. 3, 2022), pp. 1–13. DOI: [10.1038/s41467-022-33578-1](https://doi.org/10.1038/s41467-022-33578-1).
- [143] G. Ruffini, E. Lopez-Sola, J. Vohryzek, and R. Sanchez-Todo. “Neural geometrodynamics: a psychedelic perspective.” In: *bioRxiv* (Aug. 17, 2023), p. 2023.08.14.553258. DOI: [10.1101/2023.08.14.553258](https://doi.org/10.1101/2023.08.14.553258).
- [144] D. Zhou, Y. Kang, D. Cosme, M. Jovanova, X. He, A. Mahadevan, J. Ahn, O. Stanoi, J. K. Brynildsen, N. Cooper, E. J. Cornblath, L. Parkes, P. J. Mucha, K. N. Ochsner, D. M. Lydon-Staley, E. B. Falk, and D. S. Bassett. “Mindful attention promotes control of brain network dynamics for self-regulation and discontinues the past from the present.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 120 (2 Jan. 10, 2023), e2201074119. DOI: [10.1073/pnas.2201074119](https://doi.org/10.1073/pnas.2201074119).
- [145] L. Helpman, M.-F. Marin, S. Papini, X. Zhu, G. M. Sullivan, F. Schneier, M. Neria, E. Shvil, M. J. Malaga Aragon, J. C. Markowitz, M. A. Lindquist, T. Wager, M. Milad, and Y. Neria. “Neural changes in extinction recall following prolonged exposure treatment for PTSD: A longitudinal fMRI study.” en. In: *Neuroimage Clin* 12 (Oct. 10, 2016), pp. 715–723. DOI: [10.1016/j.nicl.2016.10.007](https://doi.org/10.1016/j.nicl.2016.10.007).
- [146] G. Tan, T. K. Dao, L. Farmer, R. J. Sutherland, and R. Gevirtz. “Heart rate variability (HRV) and posttraumatic stress disorder (PTSD): a pilot study.” en. In: *Appl. Psychophysiol. Biofeedback* 36 (1 Mar. 2011), pp. 27–35. DOI: [10.1007/s10484-010-9141-y](https://doi.org/10.1007/s10484-010-9141-y).
- [147] L. Compère, G. J. Siegle, S. Lazzaro, E. Riley, M. Strege, G. Canovali, S. Barb, T. Huppert, and K. Young. “Amygdala real-time fMRI neurofeedback upregulation in treatment resistant depression: Proof of concept and dose determination.” en. In: *Behav. Res. Ther.* 176 (Mar. 15, 2024), p. 104523. DOI: [10.1016/j.brat.2024.104523](https://doi.org/10.1016/j.brat.2024.104523).

- [148] V. Bonhomme, C. Staquet, J. Montupil, A. Defresne, M. Kirsch, C. Martial, A. Vanhaudenhuyse, C. Chatelle, S. K. Larroque, F. Raimondo, A. Demertzi, O. Bodart, S. Laureys, and O. Gosseries. “General Anesthesia: A Probe to Explore Consciousness.” en. In: *Front. Syst. Neurosci.* 13 (Aug. 14, 2019), p. 36. DOI: [10.3389/fnsys.2019.00036](https://doi.org/10.3389/fnsys.2019.00036).
- [149] M. T. Alkire, A. G. Hudetz, and G. Tononi. “Consciousness and anesthesia.” en. In: *Science* 322 (5903 Nov. 7, 2008), pp. 876–880. DOI: [10.1126/science.1149213](https://doi.org/10.1126/science.1149213).
- [150] E. N. Brown, P. L. Purdon, and C. J. Van Dort. “General anesthesia and altered states of arousal: a systems neuroscience analysis.” en. In: *Annu. Rev. Neurosci.* 34 (2011), pp. 601–628. DOI: [10.1146/annurev-neuro-060909-153200](https://doi.org/10.1146/annurev-neuro-060909-153200).
- [151] A. Luppi. “What anaesthesia reveals about human brains and consciousness.” In: *Nat. Hum. Behav.* 8 (5 Apr. 8, 2024), pp. 801–804. DOI: [10.1038/s41562-024-01860-5](https://doi.org/10.1038/s41562-024-01860-5).
- [152] G. A. Mashour. “Anesthesia and the neurobiology of consciousness.” en. In: *Neuron* (Mar. 27, 2024). DOI: [10.1016/j.neuron.2024.03.002](https://doi.org/10.1016/j.neuron.2024.03.002).
- [153] Y. Ding, S. Liu, K. Wang, J. Kang, W. Chen, S. Sun, Y. Hu, Y. Han, and X. Chen. “Windows to consciousness: The role of fronto-parietal connectivity in anesthesia-induced unconsciousness.” en. In: *Curr. Neuropharmacol.* (May 15, 2025). DOI: [10.2174/011570159X375644250405041050](https://doi.org/10.2174/011570159X375644250405041050).
- [154] A. G. Bardon, J. J. Ballesteros, S. L. Brincat, J. E. Roy, M. K. Mahnke, Y. Ishizawa, E. N. Brown, and E. K. Miller. “Convergent effects of different anesthetics on changes in phase alignment of cortical oscillations.” en. In: *Cell Rep.* 44 (5 May 27, 2025), p. 115685. DOI: [10.1016/j.celrep.2025.115685](https://doi.org/10.1016/j.celrep.2025.115685).
- [155] M. Boly, R. Moran, M. Murphy, P. Boveroux, M.-A. Bruno, Q. Noirhomme, D. Ledoux, V. Bonhomme, J.-F. Brichant, G. Tononi, S. Laureys, and K. Friston. “Connectivity changes underlying spectral EEG changes during propofol-induced loss of consciousness.” en. In: *J. Neurosci.* 32 (20 May 16, 2012), pp. 7082–7090. DOI: [10.1523/JNEUROSCI.3769-11.2012](https://doi.org/10.1523/JNEUROSCI.3769-11.2012).
- [156] M. M. Monti, E. S. Lutkenhoff, M. Rubinov, P. Boveroux, A. Vanhaudenhuyse, O. Gosseries, M.-A. Bruno, Q. Noirhomme, M. Boly, and S. Laureys. “Dynamic change of global and local information processing in propofol-induced loss and recovery of consciousness.” en. In: *PLoS Comput. Biol.* 9 (10 Oct. 17, 2013), e1003271. DOI: [10.1371/journal.pcbi.1003271](https://doi.org/10.1371/journal.pcbi.1003271).
- [157] J. W. Olney, J. W. Newcomer, and N. B. Farber. “NMDA receptor hypofunction model of schizophrenia.” en. In: *J. Psychiatr. Res.* 33 (6 1999), pp. 523–533. DOI: [10.1016/s0022-3956\(99\)00029-1](https://doi.org/10.1016/s0022-3956(99)00029-1).
- [158] J. Seamans. “Losing inhibition with ketamine.” en. In: *Nat. Chem. Biol.* 4 (2 Feb. 2008), pp. 91–93. DOI: [10.1038/nchembio0208-91](https://doi.org/10.1038/nchembio0208-91).
- [159] H. Homayoun and B. Moghaddam. “NMDA receptor hypofunction produces opposite effects on prefrontal cortex interneurons and pyramidal neurons.” en. In: *J. Neurosci.* 27 (43 Oct. 24, 2007), pp. 11496–11500. DOI: [10.1523/JNEUROSCI.2213-07.2007](https://doi.org/10.1523/JNEUROSCI.2213-07.2007).

- [160] F. Tian, L. D. Lewis, D. W. Zhou, G. A. Balanza, A. C. Paulk, R. Zelmann, N. Peled, D. Soper, L. A. Santa Cruz Mercado, R. A. Peterfreund, L. S. Aglio, E. N. Eskandar, G. R. Cosgrove, Z. M. Williams, R. M. Richardson, E. N. Brown, O. Akeju, S. S. Cash, and P. L. Purdon. “Characterizing brain dynamics during ketamine-induced dissociation and subsequent interactions with propofol using human intracranial neurophysiology.” en. In: *Nat. Commun.* 14 (1 Mar. 29, 2023), p. 1748. DOI: [10.1038/s41467-023-37463-3](https://doi.org/10.1038/s41467-023-37463-3).
- [161] C. Nacif-Coelho, C. Correa-Sales, L. L. Chang, and M. Maze. “Perturbation of ion channel conductance alters the hypnotic response to the alpha 2-adrenergic agonist dexmedetomidine in the locus coeruleus of the rat.” en. In: *Anesthesiology* 81 (6 Dec. 1, 1994), pp. 1527–1534. DOI: [10.1097/00000542-199412000-00029](https://doi.org/10.1097/00000542-199412000-00029).
- [162] L. E. Nelson, J. Lu, T. Guo, C. B. Saper, N. P. Franks, and M. Maze. “The alpha2-adrenoceptor agonist dexmedetomidine converges on an endogenous sleep-promoting pathway to exert its sedative effects.” en. In: *Anesthesiology* 98 (2 Feb. 2003), pp. 428–436. DOI: [10.1097/00000542-200302000-00024](https://doi.org/10.1097/00000542-200302000-00024).
- [163] P. L. Purdon, A. Sampson, K. J. Pavone, and E. N. Brown. “Clinical electroencephalography for anesthesiologists: Part I: Background and basic signatures.” en. In: *Anesthesiology* 123 (4 Oct. 2015), pp. 937–960. DOI: [10.1097/ALN.0000000000000841](https://doi.org/10.1097/ALN.0000000000000841).
- [164] L. E. Nelson, T. Z. Guo, J. Lu, C. B. Saper, N. P. Franks, and M. Maze. “The sedative component of anesthesia is mediated by GABA(A) receptors in an endogenous sleep pathway.” en. In: *Nat. Neurosci.* 5 (10 Oct. 26, 2002), pp. 979–984. DOI: [10.1038/nm913](https://doi.org/10.1038/nm913).
- [165] O. Akeju, S.-E. Kim, R. Vazquez, J. Rhee, K. J. Pavone, L. E. Hobbs, P. L. Purdon, and E. N. Brown. “Spatiotemporal dynamics of dexmedetomidine-induced electroencephalogram oscillations.” en. In: *PLoS One* 11 (10 Oct. 6, 2016), e0163431. DOI: [10.1371/journal.pone.0163431](https://doi.org/10.1371/journal.pone.0163431).
- [166] O. Akeju, K. J. Pavone, M. B. Westover, R. Vazquez, M. J. Prerau, P. G. Harrell, K. E. Hartnack, J. Rhee, A. L. Sampson, K. Habeeb, L. Gao, E. T. Pierce, J. L. Walsh, E. N. Brown, and P. L. Purdon. “A comparison of propofol- and dexmedetomidine-induced electroencephalogram dynamics using spectral and coherence analysis.” en. In: *Anesthesiology* 121 (5 Nov. 2014), pp. 978–989. DOI: [10.1097/ALN.0000000000000419](https://doi.org/10.1097/ALN.0000000000000419).
- [167] O. Akeju, A. H. Song, A. E. Hamilos, K. J. Pavone, F. J. Flores, E. N. Brown, and P. L. Purdon. “Electroencephalogram signatures of ketamine anesthesia-induced unconsciousness.” en. In: *Clin. Neurophysiol.* 127 (6 June 1, 2016), pp. 2414–2422. DOI: [10.1016/j.clinph.2016.03.005](https://doi.org/10.1016/j.clinph.2016.03.005).
- [168] M. Murphy, M.-A. Bruno, B. A. Riedner, P. Boveroux, Q. Noirhomme, E. C. Landsness, J.-F. Brichant, C. Phillips, M. Massimini, S. Laureys, G. Tononi, and M. Boly. “Propofol anesthesia and sleep: a high-density EEG study.” en. In: *Sleep* 34 (3 Mar. 1, 2011), 283–91A. DOI: [10.1093/sleep/34.3.283](https://doi.org/10.1093/sleep/34.3.283).
- [169] P. L. Purdon, E. T. Pierce, E. A. Mukamel, M. J. Prerau, J. L. Walsh, K. F. K. Wong, A. F. Salazar-Gomez, P. G. Harrell, A. L. Sampson, A. Cimenser, S. Ching, N. J. Kopell, C. Tavares-Stoeckel, K. Habeeb, R. Merhar, and E. N. Brown. “Electroencephalogram signatures of loss and recovery of consciousness from propofol.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 110 (12 Mar. 19, 2013), E1142–51. DOI: [10.1073/pnas.1221180110](https://doi.org/10.1073/pnas.1221180110).

- [170] P. E. Vlisides, T. Bel-Bahar, U. Lee, D. Li, H. Kim, E. Janke, V. Tarnal, A. B. Pichurko, A. M. McKinney, B. S. Kunkler, P. Picton, and G. A. Mashour. “Neurophysiologic correlates of ketamine sedation and anesthesia: A high-density electroencephalography study in healthy volunteers.” en. In: *Anesthesiology* 127 (1 July 2017), pp. 58–69. DOI: [10.1097/ALN.0000000000001671](https://doi.org/10.1097/ALN.0000000000001671).
- [171] J. J. Ballesteros, J. B. Briscoe, and Y. Ishizawa. “Neural signatures of α 2-Adrenergic agonist-induced unconsciousness and awakening by antagonist.” en. In: *Elife* 9 (Aug. 28, 2020). DOI: [10.7554/eLife.57670](https://doi.org/10.7554/eLife.57670).
- [172] J. J. Ballesteros, P. Huang, S. R. Patel, E. N. Eskandar, and Y. Ishizawa. “Dynamics of ketamine-induced loss and return of consciousness across primate neocortex.” en. In: *Anesthesiology* 132 (4 Apr. 1, 2020), pp. 750–762. DOI: [10.1097/aln.0000000000003159](https://doi.org/10.1097/aln.0000000000003159).
- [173] I. C. Garwood, S. Chakravarty, J. Donoghue, M. Mahnke, P. Kahali, S. Chamadia, O. Akeju, E. K. Miller, and E. N. Brown. “A hidden Markov model reliably characterizes ketamine-induced spectral dynamics in macaque local field potentials and human electroencephalograms.” en. In: *PLoS Comput. Biol.* 17 (8 Aug. 18, 2021), e1009280. DOI: [10.1371/journal.pcbi.1009280](https://doi.org/10.1371/journal.pcbi.1009280).
- [174] P. Barttfeld, L. Uhrig, J. D. Sitt, M. Sigman, B. Jarraya, and S. Dehaene. “Signature of consciousness in the dynamics of resting-state brain activity.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 112 (3 Jan. 20, 2015), pp. 887–892. DOI: [10.1073/pnas.1418031112](https://doi.org/10.1073/pnas.1418031112).
- [175] L. Uhrig, J. D. Sitt, A. Jacob, J. Tasserie, P. Barttfeld, M. Dupont, S. Dehaene, and B. Jarraya. “Resting-state dynamics as a cortical signature of anesthesia in monkeys.” en. In: *Anesthesiology* 129 (5 Nov. 2018), pp. 942–958. DOI: [10.1097/ALN.0000000000002336](https://doi.org/10.1097/ALN.0000000000002336).
- [176] A. I. Luppi, L. Uhrig, J. Tasserie, C. M. Signorelli, E. A. Stamatakis, A. Destexhe, B. Jarraya, and R. Cofre. “Local orchestration of distributed functional patterns supporting loss and restoration of consciousness in the primate brain.” en. In: *Nat. Commun.* 15 (1 Mar. 11, 2024), p. 2171. DOI: [10.1038/s41467-024-46382-w](https://doi.org/10.1038/s41467-024-46382-w).
- [177] P. Castro, A. Luppi, E. Tagliazucchi, Y. S-Perl, L. Naci, A. M. Owen, J. D. Sitt, A. Destexhe, and R. Cofré. “Dynamical structure-function correlations provide robust and generalizable signatures of consciousness in humans.” In: *Communications Biology* 7 (1 Sept. 30, 2024), p. 1224. DOI: [10.1101/2023.12.19.572402](https://doi.org/10.1101/2023.12.19.572402).
- [178] G. Hoffner, P. Castro, L. Uhrig, C. M. Signorelli, M. Dupont, J. Tasserie, A. Destexhe, R. Cofre, J. Sitt, and B. Jarraya. “Transcranial direct current stimulation modulates primate brain dynamics across states of consciousness.” en. In: *Elife* 13 (RP101688 Oct. 13, 2025). DOI: [10.7554/eLife.101688](https://doi.org/10.7554/eLife.101688).
- [179] J.-J. E. Slotine and W. Li. *Applied Nonlinear Control*. en. Prentice-Hall, 1991. 459 pp.
- [180] R. L. Carhart-Harris. “The entropic brain - revisited.” en. In: *Neuropharmacology* 142 (Nov. 2018), pp. 167–178. DOI: [10.1016/j.neuropharm.2018.03.010](https://doi.org/10.1016/j.neuropharm.2018.03.010).
- [181] G. Tononi and G. M. Edelman. “Consciousness and complexity.” en. In: *Science* 282 (5395 Dec. 4, 1998), pp. 1846–1851. DOI: [10.1126/science.282.5395.1846](https://doi.org/10.1126/science.282.5395.1846).

- [182] L. Kozachkov, M. Ennis, and J. Slotine. “RNNs of RNNs: Recursive construction of stable assemblies of recurrent neural networks.” In: *Adv. Neural Inf. Process. Syst.* (2022).
- [183] V. Zimmern. “Why Brain Criticality Is Clinically Relevant: A Scoping Review.” en. In: *Front. Neural Circuits* 14 (Aug. 26, 2020), p. 54. DOI: [10.3389/fncir.2020.00054](https://doi.org/10.3389/fncir.2020.00054).
- [184] J. O’Byrne and K. Jerbi. “How critical is brain criticality?” en. In: *Trends Neurosci.* (Sept. 9, 2022). DOI: [10.1016/j.tins.2022.08.007](https://doi.org/10.1016/j.tins.2022.08.007).
- [185] V. N. O’Reilly-Shah. “Delay coordinate embedding as neuronally implemented information processing: The state space theory of consciousness.” en. In: *J. Conscious. Stud.* 32 (1 Feb. 1, 2025), pp. 127–159. DOI: [10.53765/20512201.32.1.132](https://doi.org/10.53765/20512201.32.1.132).
- [186] C. Maschke, J. O’Byrne, M. A. Colombo, M. Boly, O. Gosseries, S. Laureys, M. Rosanova, K. Jerbi, and S. Blain-Moraes. “Critical dynamics in spontaneous EEG predict anesthetic-induced loss of consciousness and perturbational complexity.” en. In: *Commun. Biol.* 7 (1 Aug. 5, 2024), p. 946. DOI: [10.1038/s42003-024-06613-8](https://doi.org/10.1038/s42003-024-06613-8).
- [187] Y. Sanz Perl, H. Bocaccio, C. Pallavicini, I. Pérez-Ipiña, S. Laureys, H. Laufs, M. Kringelbach, G. Deco, and E. Tagliazucchi. “Nonequilibrium brain dynamics as a signature of consciousness.” en. In: *Phys. Rev. E* 104 (1-1 July 2021), p. 014411. DOI: [10.1103/PhysRevE.104.014411](https://doi.org/10.1103/PhysRevE.104.014411).
- [188] I. P. Ipiña, P. D. Kehoe, M. Kringelbach, H. Laufs, A. Ibañez, G. Deco, Y. S. Perl, and E. Tagliazucchi. “Modeling regional changes in dynamic stability during sleep and wakefulness.” en. In: *Neuroimage* 215 (July 15, 2020), p. 116833. DOI: [10.1016/j.neuroimage.2020.116833](https://doi.org/10.1016/j.neuroimage.2020.116833).
- [189] A. Schurger, I. Sarigiannidis, L. Naccache, J. D. Sitt, and S. Dehaene. “Cortical activity is more stable when sensory stimuli are consciously perceived.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 112 (16 Apr. 21, 2015), E2083–92. DOI: [10.1073/pnas.1418730112](https://doi.org/10.1073/pnas.1418730112).
- [190] D. G. Dávila, A. McKinstry-Wu, M. B. Kelz, and A. Proekt. “The administration of ketamine is associated with dose-dependent stabilization of cortical dynamics in humans.” en. In: *J. Neurosci.* 45 (20 May 14, 2025), e1545242025. DOI: [10.1523/JNEUROSCI.1545-24.2025](https://doi.org/10.1523/JNEUROSCI.1545-24.2025).
- [191] G. Sugihara, R. May, H. Ye, C.-H. Hsieh, E. Deyle, M. Fogarty, and S. Munch. “Detecting causality in complex ecosystems.” en. In: *Science* 338 (6106 Oct. 26, 2012), pp. 496–500. DOI: [10.1126/science.1227079](https://doi.org/10.1126/science.1227079).
- [192] E. R. Deyle and G. Sugihara. “Generalized theorems for nonlinear state space reconstruction.” en. In: *PLoS One* 6 (3 Mar. 31, 2011), e18295. DOI: [10.1371/journal.pone.0018295](https://doi.org/10.1371/journal.pone.0018295).
- [193] C. D. Young and M. D. Graham. “Deep learning delay coordinate dynamics for chaotic attractors from partial observable data.” en. In: *Phys. Rev. E* 107 (3-1 Mar. 30, 2023), p. 034215. DOI: [10.1103/PhysRevE.107.034215](https://doi.org/10.1103/PhysRevE.107.034215).
- [194] M. Casdagli, S. Eubank, J. D. Farmer, and J. Gibson. “State space reconstruction in the presence of noise.” In: *Physica D* 51 (1 Aug. 1, 1991), pp. 52–98. DOI: [10.1016/0167-2789\(91\)90222-U](https://doi.org/10.1016/0167-2789(91)90222-U).

- [195] R. V. Raut, Z. P. Rosenthal, X. Wang, H. Miao, Z. Zhang, J.-M. Lee, M. E. Raichle, A. Q. Bauer, S. L. Brunton, B. W. Brunton, and J. N. Kutz. “Arousal as a universal embedding for spatiotemporal brain dynamics.” en. In: *Nature* 647 (8089 Sept. 24, 2025), pp. 454–461. DOI: [10.1038/s41586-025-09544-4](https://doi.org/10.1038/s41586-025-09544-4).
- [196] K. M. Tye, E. K. Miller, F. H. Taschbach, M. K. Benna, M. Rigotti, and S. Fusi. “Mixed selectivity: Cellular computations for complexity.” en. In: *Neuron* 112 (14 July 17, 2024), pp. 2289–2303. DOI: [10.1016/j.neuron.2024.04.017](https://doi.org/10.1016/j.neuron.2024.04.017).
- [197] L. F. Abbott and P. Dayan. *Theoretical Neuroscience*. en. Dec. 1, 2021. (Visited on 06/24/2025).
- [198] X.-J. Wang. *Theoretical neuroscience: Understanding cognition*. en. 1st ed. London, England: CRC Press, Feb. 28, 2025. 562 pp.
- [199] J. A. S. Kelso. “Multistability and metastability: understanding dynamic coordination in the brain.” en. In: *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 367 (1591 Apr. 5, 2012), pp. 906–918. DOI: [10.1098/rstb.2011.0351](https://doi.org/10.1098/rstb.2011.0351).
- [200] K. Ryogo. “Statistical-Mechanical Theory of Irreversible Processes. I. General Theory and Simple Applications to Magnetic and Conduction Problems.” en. In: *Journal of the Physical Society of Japan* (Dec. 12, 2013). DOI: [10.1143/JPSJ.12.570](https://doi.org/10.1143/JPSJ.12.570).
- [201] C. Chatfield. *The analysis of time series: Theory and practice*. en. 5th ed. Monographs on Statistics and Applied Probability. London, England: Chapman and Hall, Jan. 1, 1975. 263 pp. DOI: [10.1007/978-1-4899-2925-9](https://doi.org/10.1007/978-1-4899-2925-9).
- [202] M. S. Fabus, J. W. Sleight, and C. E. Warnaby. “Effect of propofol on heart rate and its coupling to cortical slow waves in humans.” en. In: *Anesthesiology* 140 (1 Jan. 1, 2024), pp. 62–72. DOI: [10.1097/ALN.0000000000004795](https://doi.org/10.1097/ALN.0000000000004795).
- [203] M. Golesorkhi, J. Gomez-Pilar, F. Zilio, N. Berberian, A. Wolff, M. C. E. Yagoub, and G. Northoff. “The brain and its time: intrinsic neural timescales are key for input processing.” en. In: *Commun. Biol.* 4 (1 Aug. 16, 2021), p. 970. DOI: [10.1038/s42003-021-02483-6](https://doi.org/10.1038/s42003-021-02483-6).
- [204] A. Buccellato, D. Zang, F. Zilio, J. Gomez-Pilar, Z. Wang, Z. Qi, R. Zheng, Z. Xu, X. Wu, P. Bisiacchi, A. Del Felice, Y. Mao, and G. Northoff. “Disrupted relationship between intrinsic neural timescales and alpha peak frequency during unconscious states - A high-density EEG study.” en. In: *Neuroimage* 265 (119802 Jan. 1, 2023), p. 119802. DOI: [10.1016/j.neuroimage.2022.119802](https://doi.org/10.1016/j.neuroimage.2022.119802).
- [205] F. Zilio, J. Gomez-Pilar, S. Cao, J. Zhang, D. Zang, Z. Qi, J. Tan, T. Hiromi, X. Wu, S. Fogel, Z. Huang, M. R. Hohmann, T. Fomina, M. Synofzik, M. Grosse-Wentrup, A. M. Owen, and G. Northoff. “Are intrinsic neural timescales related to sensory processing? Evidence from abnormal behavioral states.” en. In: *Neuroimage* 226 (117579 Feb. 1, 2021), p. 117579. DOI: [10.1016/j.neuroimage.2020.117579](https://doi.org/10.1016/j.neuroimage.2020.117579).
- [206] Z. Huang, X. Liu, G. A. Mashour, and A. G. Hudetz. “Timescales of Intrinsic BOLD Signal Dynamics and Functional Connectivity in Pharmacologic and Neuropathologic States of Unconsciousness.” en. In: *J. Neurosci.* 38 (9 Feb. 28, 2018), pp. 2304–2317. DOI: [10.1523/JNEUROSCI.2545-17.2018](https://doi.org/10.1523/JNEUROSCI.2545-17.2018).

- [207] S. Ensel, L. Uhrig, A. Ozkirli, G. Hoffner, J. Tasserie, S. Dehaene, D. Van De Ville, B. Jarraya, and E. Pirondini. “Transient brain activity dynamics discriminate levels of consciousness during anesthesia.” en. In: *Commun. Biol.* 7 (1 June 10, 2024), p. 716. DOI: [10.1038/s42003-024-06335-x](https://doi.org/10.1038/s42003-024-06335-x).
- [208] M. Scheffer, J. Bascompte, W. A. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. van Nes, M. Rietkerk, and G. Sugihara. “Early-warning signals for critical transitions.” en. In: *Nature* 461 (7260 Sept. 3, 2009), pp. 53–59. DOI: [10.1038/nature08227](https://doi.org/10.1038/nature08227).
- [209] C. Wissel. “A universal law of the characteristic return time near thresholds.” en. In: *Oecologia* 65 (1 Dec. 1984), pp. 101–107. DOI: [10.1007/BF00384470](https://doi.org/10.1007/BF00384470).
- [210] C. Kuehn. *Multiple time scale dynamics*. en. 2015th ed. Applied mathematical sciences. Cham, Switzerland: Springer International Publishing, Feb. 25, 2015. 814 pp. DOI: [10.1007/978-3-319-12316-5](https://doi.org/10.1007/978-3-319-12316-5).
- [211] J. D. Breshears, J. L. Roland, M. Sharma, C. M. Gaona, Z. V. Freudenburg, R. Tempelhoff, M. S. Avidan, and E. C. Leuthardt. “Stable and dynamic cortical electrophysiology of induction and emergence with propofol anesthesia.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 107 (49 Dec. 7, 2010), pp. 21170–21175. DOI: [10.1073/pnas.1011949107](https://doi.org/10.1073/pnas.1011949107).
- [212] K. J. Pavone, O. Akeju, A. L. Sampson, K. Ling, P. L. Purdon, and E. N. Brown. “Nitrous oxide-induced slow and delta oscillations.” en. In: *Clin. Neurophysiol.* 127 (1 Jan. 1, 2016), pp. 556–564. DOI: [10.1016/j.clinph.2015.06.001](https://doi.org/10.1016/j.clinph.2015.06.001).
- [213] J. Kadmon and H. Sompolinsky. “Transition to Chaos in Random Neuronal Networks.” In: *Phys. Rev. X* 5 (4 Nov. 19, 2015), p. 041030. DOI: [10.1103/PhysRevX.5.041030](https://doi.org/10.1103/PhysRevX.5.041030).
- [214] K. Rajan, L. F. Abbott, and H. Sompolinsky. “Stimulus-dependent suppression of chaos in recurrent neural networks.” en. In: *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 82 (1 Pt 1 July 2010), p. 011903. DOI: [10.1103/PhysRevE.82.011903](https://doi.org/10.1103/PhysRevE.82.011903).
- [215] L. Molgedey, J. Schuchhardt, and H. G. Schuster. “Suppressing chaos in neural networks by noise.” en. In: *Phys. Rev. Lett.* 69 (26 Dec. 28, 1992), pp. 3717–3719. DOI: [10.1103/PhysRevLett.69.3717](https://doi.org/10.1103/PhysRevLett.69.3717).
- [216] J. Schuecker, S. Goedeke, and M. Helias. “Optimal Sequence Memory in Driven Random Networks.” In: *Phys. Rev. X* 8 (4 Nov. 14, 2018), p. 041029. DOI: [10.1103/PhysRevX.8.041029](https://doi.org/10.1103/PhysRevX.8.041029).
- [217] H. Ozeki, I. M. Finn, E. S. Schaffer, K. D. Miller, and D. Ferster. “Inhibitory stabilization of the cortical network underlies visual surround suppression.” en. In: *Neuron* 62 (4 May 28, 2009), pp. 578–592. DOI: [10.1016/j.neuron.2009.03.028](https://doi.org/10.1016/j.neuron.2009.03.028).
- [218] M. T. Alkire, R. J. Haier, and J. H. Fallon. “Toward a unified theory of narcosis: brain imaging evidence for a thalamocortical switch as the neurophysiologic basis of anesthetic-induced unconsciousness.” en. In: *Conscious. Cogn.* 9 (3 Sept. 2000), pp. 370–386. DOI: [10.1006/ccog.1999.0423](https://doi.org/10.1006/ccog.1999.0423).

- [219] P. Fiset, T. Paus, T. Daloze, G. Plourde, P. Meuret, V. Bonhomme, N. Hajj-Ali, S. B. Backman, and A. C. Evans. “Brain mechanisms of propofol-induced loss of consciousness in humans: a positron emission tomographic study.” en. In: *J. Neurosci.* 19 (13 July 1, 1999), pp. 5506–5513. DOI: [10.1523/jneurosci.19-13-05506.1999](https://doi.org/10.1523/jneurosci.19-13-05506.1999).
- [220] F. Gómez, C. Phillips, A. Soddu, M. Boly, P. Boveroux, A. Vanhaudenhuyse, M.-A. Bruno, O. Gosseries, V. Bonhomme, S. Laureys, and Q. Noirhomme. “Changes in effective connectivity by propofol sedation.” en. In: *PLoS One* 8 (8 Aug. 19, 2013), e71370. DOI: [10.1371/journal.pone.0071370](https://doi.org/10.1371/journal.pone.0071370).
- [221] N. S. White and M. T. Alkire. “Impaired thalamocortical connectivity in humans during general-anesthetic-induced unconsciousness.” en. In: *Neuroimage* 19 (2 Pt 1 June 2003), pp. 402–411. DOI: [10.1016/s1053-8119\(03\)00103-4](https://doi.org/10.1016/s1053-8119(03)00103-4).
- [222] J. A. Hashmi, M. L. Loggia, S. Khan, L. Gao, J. Kim, V. Napadow, E. N. Brown, and O. Akeju. “Dexmedetomidine disrupts the local and global efficiencies of large-scale brain networks.” en. In: *Anesthesiology* 126 (3 Mar. 2017), pp. 419–430. DOI: [10.1097/ALN.0000000000001509](https://doi.org/10.1097/ALN.0000000000001509).
- [223] S.-W. Ying and P. A. Goldstein. “Propofol suppresses synaptic responsiveness of somatosensory relay neurons to excitatory input by potentiating GABA(A) receptor chloride channels.” en. In: *Mol. Pain* 1 (Jan. 14, 2005), p. 2. DOI: [10.1186/1744-8069-1-2](https://doi.org/10.1186/1744-8069-1-2).
- [224] A. J. Krom, A. Marmelshtein, H. Gelbard-Sagiv, A. Tankus, H. Hayat, D. Hayat, I. Matot, I. Strauss, F. Fahoum, M. Soehle, J. Boström, F. Mormann, I. Fried, and Y. Nir. “Anesthesia-induced loss of consciousness disrupts auditory responses beyond primary cortex.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 117 (21 May 26, 2020), pp. 11770–11780. DOI: [10.1073/pnas.1917251117](https://doi.org/10.1073/pnas.1917251117).
- [225] Y. S. Xiong, J. A. Donoghue, M. Lundqvist, M. Mahnke, A. J. Major, E. N. Brown, E. K. Miller, and A. M. Bastos. “Propofol-mediated loss of consciousness disrupts predictive routing and local field phase modulation of neural activity.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 121 (42 Oct. 15, 2024), e2315160121. DOI: [10.1073/pnas.2315160121](https://doi.org/10.1073/pnas.2315160121).
- [226] Y. Ishizawa, O. J. Ahmed, S. R. Patel, J. T. Gale, D. Sierra-Mercado, E. N. Brown, and E. N. Eskandar. “Dynamics of propofol-induced loss of consciousness across primate neocortex.” en. In: *J. Neurosci.* 36 (29 July 20, 2016), pp. 7718–7726. DOI: [10.1523/JNEUROSCI.4577-15.2016](https://doi.org/10.1523/JNEUROSCI.4577-15.2016).
- [227] K. E. Schroeder, Z. T. Irwin, M. Gaidica, J. Nicole Bentley, P. G. Patil, G. A. Mashour, and C. A. Chestek. “Disruption of corticocortical information transfer during ketamine anesthesia in the primate brain.” en. In: *Neuroimage* 134 (July 1, 2016), pp. 459–465. DOI: [10.1016/j.neuroimage.2016.04.039](https://doi.org/10.1016/j.neuroimage.2016.04.039).
- [228] S. Dehaene, M. Kerszberg, and J. P. Changeux. “A neuronal model of a global workspace in effortful cognitive tasks.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 95 (24 Nov. 24, 1998), pp. 14529–14534. DOI: [10.1073/pnas.95.24.14529](https://doi.org/10.1073/pnas.95.24.14529).
- [229] D. M. Rosenthal. *Consciousness and Mind*. New York: Oxford University Press UK, 2005.

- [230] J. Tasserie, L. Uhrig, J. D. Sitt, D. Manasova, M. Dupont, S. Dehaene, and B. Jarraya. “Deep brain stimulation of the thalamus restores signatures of consciousness in a nonhuman primate model.” en. In: *Sci. Adv.* 8 (11 Mar. 18, 2022), eabl5547. DOI: [10.1126/sciadv.abl5547](https://doi.org/10.1126/sciadv.abl5547).
- [231] M. J. Redinbaugh, J. M. Phillips, N. A. Kambi, S. Mohanta, S. Andryk, G. L. Dooley, M. Afrasiabi, A. Raz, and Y. B. Saalman. “Thalamus modulates consciousness via layer-specific control of cortex.” en. In: *Neuron* 106 (1 Apr. 8, 2020), 66–75.e12. DOI: [10.1016/j.neuron.2020.01.005](https://doi.org/10.1016/j.neuron.2020.01.005).
- [232] F. Muindi, J. D. Kenny, N. E. Taylor, K. Solt, M. A. Wilson, E. N. Brown, and C. J. Van Dort. “Electrical stimulation of the parabrachial nucleus induces reanimation from isoflurane general anesthesia.” en. In: *Behav. Brain Res.* 306 (June 1, 2016), pp. 20–25. DOI: [10.1016/j.bbr.2016.03.021](https://doi.org/10.1016/j.bbr.2016.03.021).
- [233] S. Pillay, X. Liu, P. Baracskey, and A. G. Hudetz. “Brainstem stimulation increases functional connectivity of basal forebrain-paralimbic network in isoflurane-anesthetized rats.” en. In: *Brain Connect.* 4 (7 Sept. 2014), pp. 523–534. DOI: [10.1089/brain.2014.0254](https://doi.org/10.1089/brain.2014.0254).
- [234] A. I. Luppi, L. Uhrig, J. Tasserie, G. Shafiei, K. Muta, J. Hata, H. Okano, D. Golkowski, A. Ranft, R. Ilg, D. Jordan, S. Gini, Z.-Q. Liu, Y. Yee, C. M. Signorelli, R. Cofre, A. Destexhe, D. Menon, E. A. Stamatakis, C. W. Connor, A. Gozzi, B. D. Fulcher, B. Jarraya, and B. Misić. “Comprehensive profiling of anaesthetised brain dynamics across phylogeny.” In: *bioRxiv* biorxiv;2025.03.22.644729v1 (Mar. 24, 2025).
- [235] D. Toker, E. Müller, H. Miyamoto, M. S. Riga, L. Lladó-Pelfort, K. Yamakawa, F. Artigas, J. M. Shine, A. E. Hudson, N. Pouratian, and M. M. Monti. “Criticality supports cross-frequency cortical-thalamic information transfer during conscious states.” en. In: *Elife* 13 (Jan. 5, 2024). DOI: [10.7554/eLife.86547](https://doi.org/10.7554/eLife.86547).
- [236] D. Li, M. S. Fabus, and J. W. Sleight. “Brain complexities and anesthesia: Their meaning and measurement.” en. In: *Anesthesiology* 137 (3 Sept. 1, 2022), pp. 290–302. DOI: [10.1097/ALN.0000000000004293](https://doi.org/10.1097/ALN.0000000000004293).
- [237] X. S. Zhang, R. J. Roy, and E. W. Jensen. “EEG complexity as a measure of depth of anesthesia for patients.” en. In: *IEEE Trans. Biomed. Eng.* 48 (12 Dec. 2001), pp. 1424–1433. DOI: [10.1109/10.966601](https://doi.org/10.1109/10.966601).
- [238] A. G. Hudetz, X. Liu, S. Pillay, M. Boly, and G. Tononi. “Propofol anesthesia reduces Lempel-Ziv complexity of spontaneous brain activity in rats.” en. In: *Neurosci. Lett.* 628 (Aug. 15, 2016), pp. 132–135. DOI: [10.1016/j.neulet.2016.06.017](https://doi.org/10.1016/j.neulet.2016.06.017).
- [239] T. F. Varley, A. I. Luppi, I. Pappas, L. Naci, R. Adapa, A. M. Owen, D. K. Menon, and E. A. Stamatakis. “Consciousness & brain functional complexity in propofol anaesthesia.” en. In: *Sci. Rep.* 10 (1 Jan. 23, 2020), p. 1018. DOI: [10.1038/s41598-020-57695-3](https://doi.org/10.1038/s41598-020-57695-3).
- [240] M. Darracq, J. Sleight, M. I. Banks, and R. D. Sanders. “Characterising the effect of propofol on complexity and stability in the EEG power spectrum.” en. In: *Br. J. Anaesth.* 121 (6 Dec. 1, 2018), pp. 1368–1369. DOI: [10.1016/j.bja.2018.09.006](https://doi.org/10.1016/j.bja.2018.09.006).

- [241] J. B. Wilsenach, C. M. Deane, G. Reinert, and K. Warnaby. “Graph models of brain state in deep anesthesia reveal sink state dynamics of reduced spatiotemporal complexity.” en. In: *Netw. Neurosci.* 9 (4 Oct. 30, 2025), pp. 1176–1198. DOI: [10.1162/NETN.a.27](https://doi.org/10.1162/NETN.a.27).
- [242] J. Frohlich, J. N. Chiang, P. A. M. Mediano, M. Nespeca, V. Saravanapandian, D. Toker, J. Dell’Italia, J. F. Hipp, S. S. Jeste, C. J. Chu, L. M. Bird, and M. M. Monti. “Neural complexity is a common denominator of human consciousness across diverse regimes of cortical dynamics.” en. In: *Commun. Biol.* 5 (1 Dec. 15, 2022), p. 1374. DOI: [10.1038/s42003-022-04331-7](https://doi.org/10.1038/s42003-022-04331-7).
- [243] K. Kawazoe, R. McGlynn, W. Felix, R. Sevilla, S. Liao, P. Kulkarni, and C. F. Ferris. “Dose-dependent effects of esketamine on brain activity in awake mice: A BOLD phMRI study.” en. In: *Pharmacol. Res. Perspect.* 10 (6 Dec. 1, 2022), e01035. DOI: [10.1002/prp2.1035](https://doi.org/10.1002/prp2.1035).
- [244] M. R. Hudson, E. Sokolenko, T. J. O’Brien, and N. C. Jones. “NMDA receptors on parvalbumin-positive interneurons and pyramidal neurons both contribute to MK-801 induced gamma oscillatory disturbances: Complex relationships with behaviour.” en. In: *Neurobiol. Dis.* 134 (104625 Feb. 1, 2020), p. 104625. DOI: [10.1016/j.nbd.2019.104625](https://doi.org/10.1016/j.nbd.2019.104625).
- [245] M. J. Gandal, J. Sisti, K. Klook, P. I. Ortinski, V. Leitman, Y. Liang, T. Thieu, R. Anderson, R. C. Pierce, G. Jonak, R. E. Gur, G. Carlson, and S. J. Siegel. “GABAB-mediated rescue of altered excitatory-inhibitory balance, gamma synchrony and behavioral deficits following constitutive NMDAR-hypofunction.” en. In: *Transl. Psychiatry* 2 (7 July 17, 2012), e142. DOI: [10.1038/tp.2012.69](https://doi.org/10.1038/tp.2012.69).
- [246] C. Pascovich, S. Castro-Zaballa, P. A. M. Mediano, D. Bor, A. Canales-Johnson, P. Torterolo, and T. A. Bekinschtein. “Ketamine and sleep modulate neural complexity dynamics in cats.” en. In: *Eur. J. Neurosci.* 55 (6 Mar. 19, 2022), pp. 1584–1600. DOI: [10.1111/ejn.15646](https://doi.org/10.1111/ejn.15646).
- [247] E. Adam, M. Kowalski, O. Akeju, E. K. Miller, E. N. Brown, M. M. McCarthy, and N. Kopell. “Ketamine can produce oscillatory dynamics by engaging mechanisms dependent on the kinetics of NMDA receptors.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 121 (22 May 28, 2024), e2402732121. DOI: [10.1073/pnas.2402732121](https://doi.org/10.1073/pnas.2402732121).
- [248] J.-W. Kim and L. M. Monteggia. “Increasing doses of ketamine curtail antidepressant responses and suppress associated synaptic signaling pathways.” en. In: *Behav. Brain Res.* 380 (112378 Feb. 17, 2020), p. 112378. DOI: [10.1016/j.bbr.2019.112378](https://doi.org/10.1016/j.bbr.2019.112378).
- [249] T. Passie, M. Karst, B. Wiese, H. M. Emrich, and U. Schneider. “Effects of different subanesthetic doses of (S)-ketamine on neuropsychology, psychopathology, and state of consciousness in man.” en. In: *Neuropsychobiology* 51 (4 May 12, 2005), pp. 226–233. DOI: [10.1159/000085724](https://doi.org/10.1159/000085724).
- [250] J. Wang, G.-J. Noh, B.-M. Choi, S.-W. Ku, P. Joo, W.-S. Jung, S. Kim, and H. Lee. “Suppressed neural complexity during ketamine- and propofol-induced unconsciousness.” en. In: *Neurosci. Lett.* 653 (July 13, 2017), pp. 320–325. DOI: [10.1016/j.neulet.2017.05.045](https://doi.org/10.1016/j.neulet.2017.05.045).

- [251] S. Sarasso, M. Boly, M. Napolitani, O. Gosseries, V. Charland-Verville, S. Casarotto, M. Rosanova, A. G. Casali, J.-F. Brichant, P. Boveroux, S. Rex, G. Tononi, S. Laureys, and M. Massimini. “Consciousness and complexity during unresponsiveness induced by propofol, xenon, and ketamine.” en. In: *Curr. Biol.* 25 (23 Dec. 7, 2015), pp. 3099–3105. DOI: [10.1016/j.cub.2015.10.014](https://doi.org/10.1016/j.cub.2015.10.014).
- [252] M. M. Schartner, R. L. Carhart-Harris, A. B. Barrett, A. K. Seth, and S. D. Muthukumaraswamy. “Increased spontaneous MEG signal diversity for psychoactive doses of ketamine, LSD and psilocybin.” en. In: *Sci. Rep.* 7 (Apr. 19, 2017), p. 46421. DOI: [10.1038/srep46421](https://doi.org/10.1038/srep46421).
- [253] N. Farnes, B. E. Juel, A. S. Nilsen, L. G. Romundstad, and J. F. Storm. “Increased signal diversity/complexity of spontaneous EEG, but not evoked EEG responses, in ketamine-induced psychedelic state in humans.” en. In: *PLoS One* 15 (11 Nov. 23, 2020), e0242056. DOI: [10.1371/journal.pone.0242056](https://doi.org/10.1371/journal.pone.0242056).
- [254] M. Dasilva, A. Camassa, A. Navarro-Guzman, A. Pazienti, L. Perez-Mendez, G. Zamora-López, M. Mattia, and M. V. Sanchez-Vives. “Modulation of cortical slow oscillations and complexity across anesthesia levels.” en. In: *Neuroimage* 224 (117415 Jan. 1, 2021), p. 117415. DOI: [10.1016/j.neuroimage.2020.117415](https://doi.org/10.1016/j.neuroimage.2020.117415).
- [255] H. Lee, D. Golkowski, D. Jordan, S. Berger, R. Ilg, J. Lee, G. A. Mashour, U. Lee, and ReCCognition Study Group. “Relationship of critical dynamics, functional connectivity, and states of consciousness in large-scale human brain networks.” en. In: *Neuroimage* 188 (Mar. 2019), pp. 228–238. DOI: [10.1016/j.neuroimage.2018.12.011](https://doi.org/10.1016/j.neuroimage.2018.12.011).
- [256] T. F. Varley, V. Denny, O. Sporns, and A. Patania. “Topological analysis of differential effects of ketamine and propofol anaesthesia on brain dynamics.” en. In: *R Soc Open Sci* 8 (6 June 23, 2021), p. 201971. DOI: [10.1098/rsos.201971](https://doi.org/10.1098/rsos.201971).
- [257] T. F. Varley, O. Sporns, A. Puce, and J. Beggs. “Differential effects of propofol and ketamine on critical brain dynamics.” en. In: *PLoS Comput. Biol.* 16 (12 Dec. 2020), e1008418. DOI: [10.1371/journal.pcbi.1008418](https://doi.org/10.1371/journal.pcbi.1008418).
- [258] T. Fekete, D. B. Omer, K. O’Hashi, A. Grinvald, C. van Leeuwen, and O. Shriki. “Critical dynamics, anesthesia and information integration: Lessons from multi-scale criticality analysis of voltage imaging data.” en. In: *Neuroimage* 183 (Dec. 2018), pp. 919–933. DOI: [10.1016/j.neuroimage.2018.08.026](https://doi.org/10.1016/j.neuroimage.2018.08.026).
- [259] M. aan het Rot, K. A. Collins, J. W. Murrough, A. M. Perez, D. L. Reich, D. S. Charney, and S. J. Mathew. “Safety and efficacy of repeated-dose intravenous ketamine for treatment-resistant depression.” en. In: *Biol. Psychiatry* 67 (2 Jan. 15, 2010), pp. 139–145. DOI: [10.1016/j.biopsych.2009.08.038](https://doi.org/10.1016/j.biopsych.2009.08.038).
- [260] I. Ezquerra-Romano, W. Lawn, E. Krupitsky, and C. J. A. Morgan. “Ketamine for the treatment of addiction: Evidence and potential mechanisms.” en. In: *Neuropharmacology* 142 (Nov. 2018), pp. 72–82. DOI: [10.1016/j.neuropharm.2018.01.017](https://doi.org/10.1016/j.neuropharm.2018.01.017).
- [261] J. G. Keilp, S. P. Madden, J. E. Marver, A. Frawley, A. K. Burke, M. M. Herzallah, M. Gluck, J. J. Mann, and M. F. Grunebaum. “Effects of ketamine versus midazolam on neurocognition at 24 hours in depressed patients with suicidal ideation.” en. In: *J. Clin. Psychiatry* 82 (6 Nov. 2, 2021), p. 37764. DOI: [10.4088/JCP.21m13921](https://doi.org/10.4088/JCP.21m13921).

- [262] D. R. Lara, L. W. Bisol, and L. R. Munari. “Antidepressant, mood stabilizing and procognitive effects of very low dose sublingual ketamine in refractory unipolar and bipolar depression.” en. In: *Int. J. Neuropsychopharmacol.* 16 (9 Oct. 2013), pp. 2111–2117. DOI: [10.1017/S1461145713000485](https://doi.org/10.1017/S1461145713000485).
- [263] M. T. Davis, N. DellaGiogia, P. Maruff, R. H. Pietrzak, and I. Esterlis. “Acute cognitive effects of single-dose intravenous ketamine in major depressive and posttraumatic stress disorder.” en. In: *Transl. Psychiatry* 11 (1 Apr. 8, 2021), p. 205. DOI: [10.1038/s41398-021-01327-5](https://doi.org/10.1038/s41398-021-01327-5).
- [264] C. G. Abdallah, L. A. Averill, K. A. Collins, P. Geha, J. Schwartz, C. Averill, K. E. DeWilde, E. Wong, A. Anticevic, C. Y. Tang, D. V. Iosifescu, D. S. Charney, and J. W. Murrough. “Ketamine treatment and global brain connectivity in major depression.” en. In: *Neuropsychopharmacology* 42 (6 May 2017), pp. 1210–1219. DOI: [10.1038/npp.2016.186](https://doi.org/10.1038/npp.2016.186).
- [265] J. H. Krystal, C. G. Abdallah, G. Sanacora, D. S. Charney, and R. S. Duman. “Ketamine: A paradigm shift for depression research and treatment.” en. In: *Neuron* 101 (5 Mar. 6, 2019), pp. 774–778. DOI: [10.1016/j.neuron.2019.02.005](https://doi.org/10.1016/j.neuron.2019.02.005).
- [266] R. M. Berman, A. Cappiello, A. Anand, D. A. Oren, G. R. Heninger, D. S. Charney, and J. H. Krystal. “Antidepressant effects of ketamine in depressed patients.” en. In: *Biol. Psychiatry* 47 (4 Feb. 15, 2000), pp. 351–354. DOI: [10.1016/s0006-3223\(99\)00230-9](https://doi.org/10.1016/s0006-3223(99)00230-9).
- [267] C. A. Zarate Jr, J. B. Singh, P. J. Carlson, N. E. Brutsche, R. Ameli, D. A. Luckenbaugh, D. S. Charney, and H. K. Manji. “A randomized trial of an N-methyl-D-aspartate antagonist in treatment-resistant major depression.” en. In: *Arch. Gen. Psychiatry* 63 (8 Aug. 2006), pp. 856–864. DOI: [10.1001/archpsyc.63.8.856](https://doi.org/10.1001/archpsyc.63.8.856).
- [268] Y. Lee, K. Syeda, N. A. Maruschak, D. S. Cha, R. B. Mansur, I. K. Wium-Andersen, H. O. Woldeyohannes, J. D. Rosenblat, and R. S. McIntyre. “A new perspective on the anti-suicide effects with ketamine treatment: A procognitive effect: A procognitive effect.” en. In: *J. Clin. Psychopharmacol.* 36 (1 Feb. 2016), pp. 50–56. DOI: [10.1097/JCP.0000000000000441](https://doi.org/10.1097/JCP.0000000000000441).
- [269] R. B. Price, M. K. Nock, D. S. Charney, and S. J. Mathew. “Effects of intravenous ketamine on explicit and implicit measures of suicidality in treatment-resistant depression.” en. In: *Biol. Psychiatry* 66 (5 Sept. 1, 2009), pp. 522–526. DOI: [10.1016/j.biopsych.2009.04.029](https://doi.org/10.1016/j.biopsych.2009.04.029).
- [270] S. Chakravarty, J. Donoghue, A. S. Waite, M. Mahnke, I. C. Garwood, S. Gallo, E. K. Miller, and E. N. Brown. “Closed-loop control of anesthetic state in nonhuman primates.” en. In: *PNAS Nexus* 2 (10 Oct. 31, 2023), gad293. DOI: [10.1093/pnasnexus/pgad293](https://doi.org/10.1093/pnasnexus/pgad293).
- [271] S. Coeckelenbergh, S. Boelefahr, B. Alexander, L. Perrin, J. Rinehart, A. Joosten, and L. Barvais. “Closed-loop anesthesia: foundations and applications in contemporary perioperative medicine.” en. In: *J. Clin. Monit. Comput.* 38 (2 Apr. 2024), pp. 487–504. DOI: [10.1007/s10877-023-01111-4](https://doi.org/10.1007/s10877-023-01111-4).
- [272] P. Fries and E. Maris. “What to do if N is two?” en. In: *J. Cogn. Neurosci.* 34 (7 June 2, 2022), pp. 1114–1118. DOI: [10.1162/jocn_a_01857](https://doi.org/10.1162/jocn_a_01857).

- [273] L. L. Colgin, T. Denninger, M. Fyhn, T. Hafting, T. Bonnevie, O. Jensen, M.-B. Moser, and E. I. Moser. “Frequency of gamma oscillations routes flow of information in the hippocampus.” en. In: *Nature* 462 (7271 Nov. 19, 2009), pp. 353–357. DOI: [10.1038/nature08573](https://doi.org/10.1038/nature08573).
- [274] S. Ni, B. Harris, and P. Gong. “Distributed and dynamical communication: a mechanism for flexible cortico-cortical interactions and its functional roles in visual attention.” en. In: *Commun. Biol.* 7 (1 May 8, 2024), p. 550. DOI: [10.1038/s42003-024-06228-z](https://doi.org/10.1038/s42003-024-06228-z).
- [275] A. K. Engel and P. Fries. “Chapter 3 - Neuronal Oscillations, Coherence, and Consciousness.” In: *The Neurology of Consciousness (Second Edition)*. Ed. by S. Laureys, O. Gosseries, and G. Tononi. San Diego: Academic Press, Jan. 1, 2016, pp. 49–60. DOI: [10.1016/B978-0-12-800948-2.00003-0](https://doi.org/10.1016/B978-0-12-800948-2.00003-0).
- [276] E. Thompson and F. J. Varela. “Radical embodiment: neural dynamics and consciousness.” en. In: *Trends Cogn. Sci.* 5 (10 Oct. 1, 2001), pp. 418–425. DOI: [10.1016/s1364-6613\(00\)01750-2](https://doi.org/10.1016/s1364-6613(00)01750-2).
- [277] L. M. Ward. “Synchronous neural oscillations and cognitive processes.” en. In: *Trends Cogn. Sci.* 7 (12 Dec. 2003), pp. 553–559. DOI: [10.1016/j.tics.2003.10.012](https://doi.org/10.1016/j.tics.2003.10.012).
- [278] M. Siegel, T. H. Donner, and A. K. Engel. “Spectral fingerprints of large-scale neuronal interactions.” en. In: *Nat. Rev. Neurosci.* 13 (2 Jan. 11, 2012), pp. 121–134. DOI: [10.1038/nrn3137](https://doi.org/10.1038/nrn3137).
- [279] T. J. Buschman and E. K. Miller. “Top-down versus bottom-up control of attention in the prefrontal and posterior parietal cortices.” en. In: *Science* 315 (5820 Mar. 30, 2007), pp. 1860–1862. DOI: [10.1126/science.1138071](https://doi.org/10.1126/science.1138071).
- [280] G. G. Gregoriou, S. J. Gotts, H. Zhou, and R. Desimone. “High-frequency, long-range coupling between prefrontal and visual cortex during attention.” en. In: *Science* 324 (5931 May 29, 2009), pp. 1207–1210. DOI: [10.1126/science.1171402](https://doi.org/10.1126/science.1171402).
- [281] G. G. Gregoriou, S. J. Gotts, and R. Desimone. “Cell-type-specific synchronization of neural activity in FEF with V4 during attention.” en. In: *Neuron* 73 (3 Feb. 9, 2012), pp. 581–594. DOI: [10.1016/j.neuron.2011.12.019](https://doi.org/10.1016/j.neuron.2011.12.019).
- [282] E. Salinas and T. J. Sejnowski. “Correlated neuronal activity and the flow of neural information.” en. In: *Nat. Rev. Neurosci.* 2 (8 Aug. 2001), pp. 539–550. DOI: [10.1038/35086012](https://doi.org/10.1038/35086012).
- [283] P. Fries, J. H. Reynolds, A. E. Rorie, and R. Desimone. “Modulation of oscillatory neuronal synchronization by selective visual attention.” en. In: *Science* 291 (5508 Feb. 23, 2001), pp. 1560–1563. DOI: [10.1126/science.1055465](https://doi.org/10.1126/science.1055465).
- [284] M. Siegel, T. H. Donner, R. Oostenveld, P. Fries, and A. K. Engel. “Neuronal synchronization along the dorsal visual pathway reflects the focus of spatial attention.” en. In: *Neuron* 60 (4 Nov. 26, 2008), pp. 709–719. DOI: [10.1016/j.neuron.2008.09.010](https://doi.org/10.1016/j.neuron.2008.09.010).
- [285] J. van Kempen, M. A. Gieselmann, M. Boyd, N. A. Steinmetz, T. Moore, T. A. Engel, and A. Thiele. “Top-down coordination of local cortical state during selective attention.” en. In: *Neuron* 109 (5 Mar. 3, 2021), 894–904.e8. DOI: [10.1016/j.neuron.2020.12.013](https://doi.org/10.1016/j.neuron.2020.12.013).

- [286] C. M. Gray, P. König, A. K. Engel, and W. Singer. “Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties.” en. In: *Nature* 338 (6213 Mar. 23, 1989), pp. 334–337. DOI: [10.1038/338334a0](https://doi.org/10.1038/338334a0).
- [287] M. Siegel, A. K. Engel, and T. H. Donner. “Cortical network dynamics of perceptual decision-making in the human brain.” en. In: *Front. Hum. Neurosci.* 5 (Feb. 28, 2011), p. 21. DOI: [10.3389/fnhum.2011.00021](https://doi.org/10.3389/fnhum.2011.00021).
- [288] M. Siegel, T. J. Buschman, and E. K. Miller. “Cortical information flow during flexible sensorimotor decisions.” en. In: *Science* 348 (6241 June 19, 2015), pp. 1352–1355. DOI: [10.1126/science.aab0551](https://doi.org/10.1126/science.aab0551).
- [289] S. L. Brincat, J. A. Donoghue, M. K. Mahnke, S. Kornblith, M. Lundqvist, and E. K. Miller. “Interhemispheric transfer of working memories.” en. In: *Neuron* 109 (6 Mar. 17, 2021), 1055–1066.e4. DOI: [10.1016/j.neuron.2021.01.016](https://doi.org/10.1016/j.neuron.2021.01.016).
- [290] R. F. Salazar, N. M. Dotson, S. L. Bressler, and C. M. Gray. “Content-specific frontoparietal synchronization during visual working memory.” en. In: *Science* 338 (6110 Nov. 23, 2012), pp. 1097–1100. DOI: [10.1126/science.1224000](https://doi.org/10.1126/science.1224000).
- [291] A. K. Engel and W. Singer. “Temporal binding and the neural correlates of sensory awareness.” en. In: *Trends Cogn. Sci.* 5 (1 Jan. 1, 2001), pp. 16–25. DOI: [10.1016/s1364-6613\(00\)01568-0](https://doi.org/10.1016/s1364-6613(00)01568-0).
- [292] W. Singer and C. M. Gray. “Visual feature integration and the temporal correlation hypothesis.” en. In: *Annu. Rev. Neurosci.* 18 (1995), pp. 555–586. DOI: [10.1146/annurev.ne.18.030195.003011](https://doi.org/10.1146/annurev.ne.18.030195.003011).
- [293] V. N. Murthy and E. E. Fetz. “Oscillatory activity in sensorimotor cortex of awake monkeys: synchronization of local field potentials and relation to behavior.” en. In: *J. Neurophysiol.* 76 (6 Dec. 1996), pp. 3949–3967. DOI: [10.1152/jn.1996.76.6.3949](https://doi.org/10.1152/jn.1996.76.6.3949).
- [294] L. Logiaco, L. F. Abbott, and S. Escola. “Thalamic control of cortical dynamics in a model of flexible motor sequencing.” en. In: *Cell Rep.* 35 (9 June 1, 2021), p. 109090. DOI: [10.1016/j.celrep.2021.109090](https://doi.org/10.1016/j.celrep.2021.109090).
- [295] F. I. Arce-McShane, C. F. Ross, K. Takahashi, B. J. Sessle, and N. G. Hatsopoulos. “Primary motor and sensory cortical areas communicate via spatiotemporally coordinated networks at multiple frequencies.” en. In: *Proc. Natl. Acad. Sci. U. S. A.* 113 (18 May 3, 2016), pp. 5083–5088. DOI: [10.1073/pnas.1600788113](https://doi.org/10.1073/pnas.1600788113).
- [296] M. G. Perich, J. A. Gallego, and L. E. Miller. “A Neural Population Mechanism for Rapid Learning.” en. In: *Neuron* 100 (4 Nov. 21, 2018), 964–976.e7. DOI: [10.1016/j.neuron.2018.09.030](https://doi.org/10.1016/j.neuron.2018.09.030).
- [297] M. T. Kaufman, M. M. Churchland, S. I. Ryu, and K. V. Shenoy. “Cortical activity in the null space: permitting preparation without movement.” en. In: *Nat. Neurosci.* 17 (3 Mar. 2014), pp. 440–448. DOI: [10.1038/nn.3643](https://doi.org/10.1038/nn.3643).
- [298] S. L. Brincat and E. K. Miller. “Frequency-specific hippocampal-prefrontal interactions during associative learning.” en. In: *Nat. Neurosci.* 18 (4 Apr. 2015), pp. 576–581. DOI: [10.1038/nn.3954](https://doi.org/10.1038/nn.3954).

- [299] M. W. Jones and M. A. Wilson. “Theta rhythms coordinate hippocampal-prefrontal interactions in a spatial memory task.” en. In: *PLoS Biol.* 3 (12 Dec. 2005), e402. DOI: [10.1371/journal.pbio.0030402](https://doi.org/10.1371/journal.pbio.0030402).
- [300] A. G. Siapas and M. A. Wilson. “Coordinated interactions between hippocampal ripples and cortical spindles during slow-wave sleep.” en. In: *Neuron* 21 (5 Nov. 1998), pp. 1123–1128. DOI: [10.1016/s0896-6273\(00\)80629-7](https://doi.org/10.1016/s0896-6273(00)80629-7).
- [301] A. Fernández-Ruiz, A. Oliva, M. Soula, F. Rocha-Almeida, G. A. Nagy, G. Martin-Vazquez, and G. Buzsáki. “Gamma rhythm communication between entorhinal cortex and dentate gyrus neuronal assemblies.” en. In: *Science* 372 (6537 Apr. 2, 2021). DOI: [10.1126/science.abf3119](https://doi.org/10.1126/science.abf3119).
- [302] E. G. Antzoulatos and E. K. Miller. “Increases in functional connectivity between prefrontal cortex and striatum during category learning.” en. In: *Neuron* 83 (1 July 2, 2014), pp. 216–225. DOI: [10.1016/j.neuron.2014.05.005](https://doi.org/10.1016/j.neuron.2014.05.005).
- [303] J. D. Semedo, A. Zandvakili, C. K. Machens, B. M. Yu, and A. Kohn. “Cortical Areas Interact through a Communication Subspace.” en. In: *Neuron* 102 (1 Apr. 3, 2019), 249–259.e4. DOI: [10.1016/j.neuron.2019.01.026](https://doi.org/10.1016/j.neuron.2019.01.026).
- [304] J. D. Semedo, E. Gokcen, C. K. Machens, A. Kohn, and B. M. Yu. “Statistical methods for dissecting interactions between brain areas.” en. In: *Curr. Opin. Neurobiol.* 65 (Dec. 2020), pp. 59–69. DOI: [10.1016/j.conb.2020.09.009](https://doi.org/10.1016/j.conb.2020.09.009).
- [305] C. J. MacDowell, A. Libby, C. I. Jahn, S. Tafazoli, and T. J. Buschman. “Multiplexed Subspaces Route Neural Activity Across Brain-wide Networks.” en. In: *bioRxiv* (Feb. 12, 2023), p. 2023.02.08.527772. DOI: [10.1101/2023.02.08.527772](https://doi.org/10.1101/2023.02.08.527772).
- [306] S. Gu, F. Pasqualetti, M. Cieslak, Q. K. Telesford, A. B. Yu, A. E. Kahn, J. D. Medaglia, J. M. Vettel, M. B. Miller, S. T. Grafton, and D. S. Bassett. “Controllability of structural brain networks.” en. In: *Nat. Commun.* 6 (Oct. 1, 2015), p. 8414. DOI: [10.1038/ncomms9414](https://doi.org/10.1038/ncomms9414).
- [307] T.-C. Kao and G. Hennequin. “Neuroscience out of control: control-theoretic perspectives on neural circuit dynamics.” en. In: *Curr. Opin. Neurobiol.* 58 (Oct. 2019), pp. 122–129. DOI: [10.1016/j.conb.2019.09.001](https://doi.org/10.1016/j.conb.2019.09.001).
- [308] D. S. Bassett and O. Sporns. “Network neuroscience.” en. In: *Nat. Neurosci.* 20 (3 Feb. 23, 2017), pp. 353–364. DOI: [10.1038/nn.4502](https://doi.org/10.1038/nn.4502).
- [309] T.-C. Kao, M. S. Sadabadi, and G. Hennequin. “Optimal anticipatory control as a theory of motor preparation: A thalamo-cortical circuit model.” en. In: *Neuron* 109 (9 May 5, 2021), 1567–1581.e12. DOI: [10.1016/j.neuron.2021.03.009](https://doi.org/10.1016/j.neuron.2021.03.009).
- [310] M. Schimel, T.-C. Kao, K. T. Jensen, and G. Hennequin. “iLQR-VAE : control-based learning of input-driven dynamics with applications to neural data.” In: *International Conference on Learning Representations*. Oct. 6, 2021.
- [311] S. F. Muldoon, F. Pasqualetti, S. Gu, M. Cieslak, S. T. Grafton, J. M. Vettel, and D. S. Bassett. “Stimulation-based control of dynamic brain networks.” en. In: *PLoS Comput. Biol.* 12 (9 Sept. 9, 2016), e1005076. DOI: [10.1371/journal.pcbi.1005076](https://doi.org/10.1371/journal.pcbi.1005076).

- [312] K. Bouchard and A. Kumar. “Feedback Controllability is a Normative Theory of Neural Population Dynamics.” In: *Research Square* (Mar. 29, 2024). DOI: [10.21203/rs.3.rs-4102129/v1](https://doi.org/10.21203/rs.3.rs-4102129/v1).
- [313] D. Zöller, C. Sandini, M. Schaer, S. Eliez, D. S. Bassett, and D. Van De Ville. “Structural control energy of resting-state functional brain states reveals less cost-effective brain dynamics in psychosis vulnerability.” en. In: *Hum. Brain Mapp.* 42 (7 May 2021), pp. 2181–2200. DOI: [10.1002/hbm.25358](https://doi.org/10.1002/hbm.25358).
- [314] X. He, L. Caciagli, L. Parkes, J. Stiso, T. M. Karrer, J. Z. Kim, Z. Lu, T. Menara, F. Pasqualetti, M. R. Sperling, J. I. Tracy, and D. S. Bassett. “Uncovering the biological basis of control energy: Structural and metabolic correlates of energy inefficiency in temporal lobe epilepsy.” en. In: *Sci. Adv.* 8 (45 Nov. 11, 2022), eabn2293. DOI: [10.1126/sciadv.abn2293](https://doi.org/10.1126/sciadv.abn2293).
- [315] J. D. Medaglia, D. Y. Harvey, N. White, A. Kelkar, J. Zimmerman, D. S. Bassett, and R. H. Hamilton. “Network controllability in the inferior frontal gyrus relates to controlled language variability and susceptibility to TMS.” en. In: *J. Neurosci.* 38 (28 July 11, 2018), pp. 6399–6410. DOI: [10.1523/JNEUROSCI.0092-17.2018](https://doi.org/10.1523/JNEUROSCI.0092-17.2018).
- [316] Z. Wu, L. Huang, M. Wang, and X. He. “Development of the brain network control theory and its implications.” en. In: *Psychoradiology* 4 (Dec. 14, 2024), kkae028. DOI: [10.1093/psyrad/kkae028](https://doi.org/10.1093/psyrad/kkae028).
- [317] W. Cai, S. Ryali, R. Pasumarthy, V. Talasila, and V. Menon. “Dynamic causal brain circuits during working memory and their functional controllability.” en. In: *Nat. Commun.* 12 (1 June 29, 2021), p. 3314. DOI: [10.1038/s41467-021-23509-x](https://doi.org/10.1038/s41467-021-23509-x).
- [318] A. Moradi Amani, A. Tahmassebi, A. Stadlbauer, U. Meyer-Baese, V. Noblet, F. Blanc, H. Malberg, and A. Meyer-Baese. “Controllability of functional and structural brain networks.” en. In: *Complexity* 2024 (1 Jan. 1, 2024), p. 7402894. DOI: [10.1155/2024/7402894](https://doi.org/10.1155/2024/7402894).
- [319] Q. Li, L. Yao, W. You, J. Liu, S. Deng, B. Li, L. Luo, Y. Zhao, Y. Wang, Y. Wang, Q. Zhang, F. Long, J. A. Sweeney, S. Gu, F. Li, and Q. Gong. “Controllability of functional brain networks and its clinical significance in first-episode schizophrenia.” en. In: *Schizophr. Bull.* 49 (3 May 3, 2023), pp. 659–668. DOI: [10.1093/schbul/sbac177](https://doi.org/10.1093/schbul/sbac177).
- [320] Y.-Y. Liu, J.-J. Slotine, and A.-L. Barabási. “Controllability of complex networks.” en. In: *Nature* 473 (7346 May 12, 2011), pp. 167–173. DOI: [10.1038/nature10011](https://doi.org/10.1038/nature10011).
- [321] L. Parkes, J. Z. Kim, J. Stiso, J. K. Brynildsen, M. Cieslak, S. Covitz, R. E. Gur, R. C. Gur, F. Pasqualetti, R. T. Shinohara, D. Zhou, T. D. Satterthwaite, and D. S. Bassett. “A network control theory pipeline for studying the dynamics of the structural connectome.” en. In: *Nat. Protoc.* 19 (12 Dec. 29, 2024), pp. 3721–3749. DOI: [10.1038/s41596-024-01023-w](https://doi.org/10.1038/s41596-024-01023-w).
- [322] D. R. Tyner and A. D. Lewis. “Geometric Jacobian linearization and LQR theory.” en. In: *J. Geom. Mech.* 2 (4 2010), pp. 397–440. DOI: [10.3934/jgm.2010.2.397](https://doi.org/10.3934/jgm.2010.2.397).

- [323] R. E. Kass, H. Bong, M. Olarinre, Q. Xin, and K. N. Urban. “Identification of interacting neural populations: methods and statistical considerations.” en. In: *J. Neurophysiol.* 130 (3 Sept. 1, 2023), pp. 475–496. DOI: [10.1152/jn.00131.2023](https://doi.org/10.1152/jn.00131.2023).
- [324] M. G. Perich, C. Arlt, S. Soares, M. E. Young, C. P. Mosher, J. Minxha, E. Carter, U. Rutishauser, P. H. Rudebeck, C. D. Harvey, and K. Rajan. “Inferring brain-wide interactions using data-constrained recurrent neural network models.” en. In: *bioRxiv* (Mar. 11, 2021), p. 2020.12.18.423348. DOI: [10.1101/2020.12.18.423348](https://doi.org/10.1101/2020.12.18.423348).
- [325] M. G. Perich and K. Rajan. “Rethinking brain-wide interactions through multi-region ‘network of networks’ models.” en. In: *Curr. Opin. Neurobiol.* 65 (Dec. 2020), pp. 146–151. DOI: [10.1016/j.conb.2020.11.003](https://doi.org/10.1016/j.conb.2020.11.003).
- [326] A. S. Andalman, V. M. Burns, M. Lovett-Barron, M. Broxton, B. Poole, S. J. Yang, L. Grosenick, T. N. Lerner, R. Chen, T. Benster, P. Mourrain, M. Levoy, K. Rajan, and K. Deisseroth. “Neuronal Dynamics Regulating Brain and Behavioral State Transitions.” en. In: *Cell* 177 (4 May 2, 2019), 970–985.e20. DOI: [10.1016/j.cell.2019.02.037](https://doi.org/10.1016/j.cell.2019.02.037).
- [327] L. Pinto, K. Rajan, B. DePasquale, S. Y. Thiberge, D. W. Tank, and C. D. Brody. “Task-Dependent Changes in the Large-Scale Dynamics and Necessity of Cortical Regions.” en. In: *Neuron* 104 (4 Nov. 20, 2019), 810–824.e9. DOI: [10.1016/j.neuron.2019.08.025](https://doi.org/10.1016/j.neuron.2019.08.025).
- [328] M. Kleinman, C. Chandrasekaran, and J. Kao. “A mechanistic multi-area recurrent network model of decision-making.” In: *Adv. Neural Inf. Process. Syst.* 34 (Dec. 6, 2021), pp. 23152–23165.
- [329] J. Barbosa, R. Proville, C. C. Rodgers, M. R. DeWeese, S. Ostojic, and Y. Boubenec. “Early selection of task-relevant features through population gating.” en. In: *Nat. Commun.* 14 (1 Oct. 27, 2023), p. 6837. DOI: [10.1038/s41467-023-42519-5](https://doi.org/10.1038/s41467-023-42519-5).
- [330] E. Gokcen, A. I. Jasper, J. D. Semedo, A. Zandvakili, A. Kohn, C. K. Machens, and B. M. Yu. “Disentangling the flow of signals between populations of neurons.” en. In: *Nature Computational Science* 2 (8 Aug. 18, 2022), pp. 512–525. DOI: [10.1038/s43588-022-00282-5](https://doi.org/10.1038/s43588-022-00282-5).
- [331] E. Gokcen, A. I. Jasper, A. Xu, A. Kohn, C. K. Machens, and B. M. Yu. “Uncovering motifs of concurrent signaling across multiple neuronal populations.” In: *Thirty-seventh Conference on Neural Information Processing Systems*. Nov. 2, 2023.
- [332] S. Ebrahimi, J. Lecoq, O. Romyantsev, T. Tasci, Y. Zhang, C. Irimia, J. Li, S. Ganguli, and M. J. Schnitzer. “Emergent reliability in sensory cortical coding and inter-area communication.” en. In: *Nature* 605 (7911 May 2022), pp. 713–721. DOI: [10.1038/s41586-022-04724-y](https://doi.org/10.1038/s41586-022-04724-y).
- [333] J. Rodu, N. Klein, S. L. Brincat, E. K. Miller, and R. E. Kass. “Detecting multivariate cross-correlation between brain regions.” en. In: *J. Neurophysiol.* 120 (4 Oct. 1, 2018), pp. 1962–1972. DOI: [10.1152/jn.00869.2017](https://doi.org/10.1152/jn.00869.2017).
- [334] J. D. Semedo, A. I. Jasper, A. Zandvakili, A. Krishna, A. Aschner, C. K. Machens, A. Kohn, and B. M. Yu. “Feedforward and feedback interactions between visual cortical areas use different population activity patterns.” en. In: *Nat. Commun.* 13 (1 Mar. 1, 2022), p. 1099. DOI: [10.1038/s41467-022-28552-w](https://doi.org/10.1038/s41467-022-28552-w).

- [335] J. I. Glaser, M. Whiteway, J. P. Cunningham, L. Paninski, and S. W. Linderman. “Recurrent Switching Dynamical Systems Models for Multiple Interacting Neural Populations.” en. In: *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*. Oct. 22, 2020, p. 2020.10.21.349282. DOI: [10.1101/2020.10.21.349282](https://doi.org/10.1101/2020.10.21.349282).
- [336] O. Karniol-Tambour, D. M. Zoltowski, E. M. Diamanti, L. Pinto, D. W. Tank, C. D. Brody, and J. W. Pillow. “Modeling communication and switching nonlinear dynamics in multi-region neural activity.” In: *bioRxiv* (Sept. 15, 2022), p. 2022.09.13.507841. DOI: [10.1101/2022.09.13.507841](https://doi.org/10.1101/2022.09.13.507841).
- [337] S. L. Bressler and A. K. Seth. “Wiener-Granger causality: a well established methodology.” en. In: *Neuroimage* 58 (2 Sept. 15, 2011), pp. 323–329. DOI: [10.1016/j.neuroimage.2010.02.059](https://doi.org/10.1016/j.neuroimage.2010.02.059).
- [338] A. K. Seth, A. B. Barrett, and L. Barnett. “Granger causality analysis in neuroscience and neuroimaging.” en. In: *J. Neurosci.* 35 (8 Feb. 25, 2015), pp. 3293–3297. DOI: [10.1523/JNEUROSCI.4399-14.2015](https://doi.org/10.1523/JNEUROSCI.4399-14.2015).
- [339] K. J. Friston, L. Harrison, and W. Penny. “Dynamic causal modelling.” en. In: *Neuroimage* 19 (4 Aug. 1, 2003), pp. 1273–1302. DOI: [10.1016/s1053-8119\(03\)00202-7](https://doi.org/10.1016/s1053-8119(03)00202-7).
- [340] Y. Chen, H. Douglas, B. J. Medina, M. Olarinre, J. H. Siegle, and R. E. Kass. “Population burst propagation across interacting areas of the brain.” en. In: *J. Neurophysiol.* 128 (6 Dec. 1, 2022), pp. 1578–1592. DOI: [10.1152/jn.00066.2022](https://doi.org/10.1152/jn.00066.2022).
- [341] J. Calderon and G. J. Berman. “Inferring the time-varying coupling of dynamical systems with temporal convolutional autoencoders.” en. In: *eLife* (Nov. 12, 2024). DOI: [10.7554/elife.100692.1](https://doi.org/10.7554/elife.100692.1).
- [342] Z. Lu, A. Tabassum, S. Kulkarni, L. Mi, J. Nathan Kutz, E. Shea-Brown, and S.-H. Lim. “Attention for Causal Relationship Discovery from Biological Neural Dynamics.” In: *NeurIPS 2023 Workshop on Causal Representation Learning*. Nov. 12, 2023.
- [343] R. W. Brockett. “Nonlinear systems and differential geometry.” en. In: *Proc. IEEE Inst. Electr. Electron. Eng.* 64 (1 1976), pp. 61–72. DOI: [10.1109/proc.1976.10067](https://doi.org/10.1109/proc.1976.10067).
- [344] G. W. Haynes and H. Hermes. “Nonlinear controllability via lie theory.” en. In: *SIAM J. Control* 8 (4 Nov. 1970), pp. 450–460. DOI: [10.1137/0308033](https://doi.org/10.1137/0308033).
- [345] C. O. Aguilar and B. Gharesifard. “Necessary conditions for controllability of nonlinear networked control systems.” en. In: *2014 American Control Conference*. 2014 American Control Conference - ACC 2014 (Portland, OR, USA). IEEE, June 2014, pp. 5379–5383. DOI: [10.1109/acc.2014.6859170](https://doi.org/10.1109/acc.2014.6859170).
- [346] A. J. Whalen, S. N. Brennan, T. D. Sauer, and S. J. Schiff. “Observability and controllability of nonlinear networks: The role of symmetry.” en. In: *Phys. Rev. X*. 5 (1 Jan. 23, 2015), p. 011005. DOI: [10.1103/PhysRevX.5.011005](https://doi.org/10.1103/PhysRevX.5.011005).
- [347] L.-Z. Wang, R.-Q. Su, Z.-G. Huang, X. Wang, W.-X. Wang, C. Grebogi, and Y.-C. Lai. “A geometrical approach to control and controllability of nonlinear dynamical networks.” en. In: *Nat. Commun.* 7 (Apr. 14, 2016), p. 11323. DOI: [10.1038/ncomms11323](https://doi.org/10.1038/ncomms11323).

- [348] Y. Tang, H. Gao, W. Zou, and J. Kurths. “Identifying controlling nodes in neuronal networks in different scales.” en. In: *PLoS One* 7 (7 July 27, 2012), e41375. DOI: [10.1371/journal.pone.0041375](https://doi.org/10.1371/journal.pone.0041375).
- [349] G. Kumar, D. Menolascino, and S. Ching. “Input novelty as a control metric for time varying linear systems.” In: *arXiv [math.OC]* (Nov. 21, 2014).
- [350] F. Latrémolière, S. Narayanappa, and P. Vojtěchovský. “Estimating the Jacobian matrix of an unknown multivariate function from sample values by means of a neural network.” In: *arXiv [cs.LG]* (Apr. 1, 2022).
- [351] J. Lorraine and S. Hossain. “JacNet: Learning functions with structured jacobians.” In: *arXiv [cs.LG]* (Aug. 23, 2024).
- [352] H. Beik-Mohammadi, S. Hauberg, G. Arvanitidis, N. Figueroa, G. Neumann, and L. Rozo. “Neural Contractive Dynamical Systems.” In: *arXiv [cs.RO]* (Jan. 17, 2024).
- [353] W. Lohmiller and J.-J. E. Slotine. “On Contraction Analysis for Non-linear Systems.” In: *Automatica* 34 (6 June 1, 1998), pp. 683–696. DOI: [10.1016/S0005-1098\(98\)00019-3](https://doi.org/10.1016/S0005-1098(98)00019-3).
- [354] F. Hess, Z. Monfared, M. Brenner, and D. Durstewitz. “Generalized teacher forcing for learning chaotic dynamics.” In: *arXiv [cs.LG]* (June 7, 2023).
- [355] A. Iyer, S. Chandra, S. Sharma, and I. Fiete. “Flexible mapping of abstract domains by grid cells via self-supervised extraction and projection of generalized velocity signals.” In: *Neural Inf Process Syst* 37 (2024). Ed. by A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, pp. 85441–85466.
- [356] E. N. Lorenz. “Predictability: a problem partly solved.” en. In: *ECMWF Seminar on Predictability, 4-8 September 1995*. European Centre for Medium-Range Weather Forecasts, 1996.
- [357] W. Gilpin. “Chaos as an interpretable benchmark for forecasting and data-driven modelling.” In: *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Ed. by J. Vanschoren and S. Yeung. Vol. 1. 2021.
- [358] W. Gilpin. “Model scale versus domain knowledge in statistical forecasting of chaotic systems.” In: *Phys. Rev. Res.* 5 (4 Dec. 15, 2023), p. 043252. DOI: [10.1103/PhysRevResearch.5.043252](https://doi.org/10.1103/PhysRevResearch.5.043252).
- [359] P. Gómez, H. Toftevaag, and G. Meoni. “torchquad: Numerical Integration in Arbitrary Dimensions with PyTorch.” In: *J. Open Source Softw.* 6 (64 Aug. 31, 2021), p. 3439. DOI: [10.21105/joss.03439](https://doi.org/10.21105/joss.03439).
- [360] R. T. Q. Chen, Y. Rubanova, J. Bettencourt, and D. K. Duvenaud. “Neural Ordinary Differential Equations.” In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Vol. 31. Curran Associates, Inc., 2018.
- [361] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. “PyTorch: An imperative style, high-performance deep learning library.” In: *Adv. Neural Inf. Process. Syst.* abs/1912.01703 (Dec. 3, 2019).

- [362] J. Hoffman, D. A. Roberts, and S. Yaida. “Robust learning with Jacobian regularization.” In: *arXiv [stat.ML]* (Aug. 7, 2019).
- [363] A. Wikner, J. Harvey, M. Girvan, B. R. Hunt, A. Pomerance, T. Antonsen, and E. Ott. “Stabilizing machine learning prediction of dynamics: Novel noise-inspired regularization tested with reservoir computing.” en. In: *Neural Netw.* 170 (Feb. 1, 2024), pp. 94–110. DOI: [10.1016/j.neunet.2023.10.054](https://doi.org/10.1016/j.neunet.2023.10.054).
- [364] S. Schneider, R. G. Laiz, A. Filippova, M. Frey, and M. W. Mathis. “Time-series attribution maps with regularized contrastive learning.” In: *arXiv [stat.ML]* (Feb. 17, 2025).
- [365] P. J. Antsaklis and A. N. Michel. *Linear Systems*. Birkhäuser Boston, 2006. 18 pp. DOI: [10.1007/0-8176-4435-0](https://doi.org/10.1007/0-8176-4435-0).
- [366] Y. Kawano and J. M. A. Scherpen. “Empirical differential Gramians for nonlinear model reduction.” en. In: *Automatica (Oxf.)* 127 (109534 May 2021), p. 109534. DOI: [10.1016/j.automatica.2021.109534](https://doi.org/10.1016/j.automatica.2021.109534).
- [367] G. Lindmark and C. Altafini. “Minimum energy control for complex networks.” en. In: *Sci. Rep.* 8 (1 Feb. 16, 2018), p. 3188. DOI: [10.1038/s41598-018-21398-7](https://doi.org/10.1038/s41598-018-21398-7).
- [368] M. F. Panichello and T. J. Buschman. “Shared mechanisms underlie the control of working memory and attention.” en. In: *Nature* (Mar. 31, 2021). DOI: [10.1038/s41586-021-03390-w](https://doi.org/10.1038/s41586-021-03390-w).
- [369] W. Li and E. Todorov. “Iterative Linear Quadratic Regulator Design for Nonlinear Biological Movement Systems.” In: *ICINCO 2004, Proceedings of the First International Conference on Informatics in Control, Automation and Robotics, Setúbal, Portugal, August 25-28, 2004* 1 (Jan. 1, 2004), pp. 222–229.
- [370] Y. Tassa, T. Erez, and E. Todorov. “Synthesis and stabilization of complex behaviors through online trajectory optimization.” en. In: *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2012) (Vilamoura-Algarve, Portugal)*. IEEE, Oct. 2012, pp. 4906–4913. DOI: [10.1109/iros.2012.6386025](https://doi.org/10.1109/iros.2012.6386025).
- [371] S. Ouala, D. Nguyen, L. Drumetz, B. Chapron, A. Pascual, F. Collard, L. Gaultier, and R. Fablet. “Learning latent dynamics for partially observed chaotic systems.” en. In: *Chaos* 30 (10 Oct. 20, 2020), p. 103121. DOI: [10.1063/5.0019309](https://doi.org/10.1063/5.0019309).
- [372] P. Y. Lu, J. Ariño Bernad, and M. Soljačić. “Discovering sparse interpretable dynamics from partial observations.” en. In: *Commun. Phys.* 5 (1 Aug. 12, 2022), pp. 1–7. DOI: [10.1038/s42005-022-00987-z](https://doi.org/10.1038/s42005-022-00987-z).
- [373] E. Özalp, G. Margazoglou, and L. Magri. “Reconstruction, forecasting, and stability of chaotic dynamics from partial data.” en. In: *Chaos* 33 (9 Sept. 1, 2023), p. 093107. DOI: [10.1063/5.0159479](https://doi.org/10.1063/5.0159479).
- [374] S. Jaffe, A. Davydov, D. Lapsekili, A. Singh, and F. Bullo. “Learning neural contracting dynamics: Extended linearization and global guarantees.” In: *arXiv [cs.LG]* (Feb. 12, 2024).

- [375] R. U. Seydel. *Practical bifurcation and stability analysis*. en. 3rd ed. Interdisciplinary applied mathematics. New York, NY: Springer, Dec. 10, 2009. 477 pp. DOI: [10.1007/978-1-4419-1740-9](https://doi.org/10.1007/978-1-4419-1740-9).
- [376] F. Christiansen and H. H. Rugh. “Computing Lyapunov spectra with continuous Gram - Schmidt orthonormalization.” en. In: *Nonlinearity* 10 (5 Sept. 1, 1997), p. 1063. DOI: [10.1088/0951-7715/10/5/004](https://doi.org/10.1088/0951-7715/10/5/004).
- [377] V. Duchaine and C. M. Gosselin. “Investigation of human-robot interaction stability using Lyapunov theory.” en. In: *2008 IEEE International Conference on Robotics and Automation*. 2008 IEEE International Conference on Robotics and Automation (ICRA) (Pasadena, CA, USA). IEEE, May 2008, pp. 2189–2194. DOI: [10.1109/robot.2008.4543531](https://doi.org/10.1109/robot.2008.4543531).
- [378] I. R. Manchester and J.-J. E. Slotine. “Control contraction metrics: Convex and intrinsic criteria for nonlinear feedback design.” en. In: *IEEE Trans. Automat. Contr.* 62 (6 June 2017), pp. 3046–3053. DOI: [10.1109/tac.2017.2668380](https://doi.org/10.1109/tac.2017.2668380).
- [379] P. M. Wensing and J.-J. Slotine. “Beyond convexity-Contraction and global convergence of gradient descent.” en. In: *PLoS One* 15 (8 Aug. 4, 2020), e0236661. DOI: [10.1371/journal.pone.0236661](https://doi.org/10.1371/journal.pone.0236661).
- [380] N. A. M. Hootsmans and S. Dubowsky. “Large motion control of mobile manipulators including vehicle suspension characteristics.” In: *Proceedings. 1991 IEEE International Conference on Robotics and Automation*. 1991 IEEE International Conference on Robotics and Automation (Sacramento, CA, USA). IEEE Comput. Soc. Press, 2002, 2336–2341 vol.3. DOI: [10.1109/robot.1991.131751](https://doi.org/10.1109/robot.1991.131751).
- [381] S. A. A. Moosavian and E. Papadopoulos. “Modified transpose Jacobian control of robotic systems.” en. In: *Automatica (Oxf.)* 43 (7 July 1, 2007), pp. 1226–1233. DOI: [10.1016/j.automatica.2006.12.029](https://doi.org/10.1016/j.automatica.2006.12.029).
- [382] K. Tchoń, A. Ratajczak, and I. Góral. “Lagrangian Jacobian inverse for nonholonomic robotic systems.” en. In: *Nonlinear Dyn.* 82 (4 Dec. 2015), pp. 1923–1932. DOI: [10.1007/s11071-015-2288-6](https://doi.org/10.1007/s11071-015-2288-6).
- [383] C. Brudey, J. Park, J. Wiaderkiewicz, I. Kobayashi, T. A. Mellman, and P. J. Marvar. “Autonomic and inflammatory consequences of posttraumatic stress disorder and the link to cardiovascular disease.” en. In: *Am. J. Physiol. Regul. Integr. Comp. Physiol.* 309 (4 Aug. 15, 2015), R315–21. DOI: [10.1152/ajpregu.00343.2014](https://doi.org/10.1152/ajpregu.00343.2014).
- [384] G. A. Alvares, D. S. Quintana, I. B. Hickie, and A. J. Guastella. “Autonomic nervous system dysfunction in psychiatric disorders and the impact of psychotropic medications: a systematic review and meta-analysis.” en. In: *J. Psychiatry Neurosci.* 41 (2 Mar. 2016), pp. 89–104. DOI: [10.1503/jpn.140217](https://doi.org/10.1503/jpn.140217).
- [385] J. J. Goldberger, R. Arora, U. Buckley, and K. Shivkumar. “Autonomic nervous system dysfunction: JACC focus seminar.” en. In: *J. Am. Coll. Cardiol.* 73 (10 Mar. 19, 2019), pp. 1189–1206. DOI: [10.1016/j.jacc.2018.12.064](https://doi.org/10.1016/j.jacc.2018.12.064).
- [386] L. Xiong and T. W. H. Leung. “Autonomic dysfunction in neurological disorders.” en. In: *Aging (Albany NY)* 11 (7 Apr. 9, 2019), pp. 1903–1904. DOI: [10.18632/aging.101896](https://doi.org/10.18632/aging.101896).

- [387] A.-L. Barabási, N. Gulbahce, and J. Loscalzo. “Network medicine: a network-based approach to human disease.” en. In: *Nat. Rev. Genet.* 12 (1 Jan. 2011), pp. 56–68. DOI: [10.1038/nrg2918](https://doi.org/10.1038/nrg2918).
- [388] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kuttyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos. “Systematic localization of common disease-associated variation in regulatory DNA.” en. In: *Science* 337 (6099 Sept. 7, 2012), pp. 1190–1195. DOI: [10.1126/science.1222794](https://doi.org/10.1126/science.1222794).
- [389] P. B. Madhamshettiwar, S. R. Maetschke, M. J. Davis, A. Reverter, and M. A. Ragan. “Gene regulatory network inference: evaluation and application to ovarian cancer allows the prioritization of drug targets.” en. In: *Genome Med.* 4 (5 May 1, 2012), p. 41. DOI: [10.1186/gm340](https://doi.org/10.1186/gm340).
- [390] T. I. Lee and R. A. Young. “Transcriptional regulation and its misregulation in disease.” en. In: *Cell* 152 (6 Mar. 14, 2013), pp. 1237–1251. DOI: [10.1016/j.cell.2013.02.014](https://doi.org/10.1016/j.cell.2013.02.014).
- [391] P. Unger Avila, T. Padvitski, A. C. Leote, H. Chen, J. Saez-Rodriguez, M. Kann, and A. Beyrer. “Gene regulatory networks in disease and ageing.” en. In: *Nat. Rev. Nephrol.* 20 (9 Sept. 12, 2024), pp. 616–633. DOI: [10.1038/s41581-024-00849-7](https://doi.org/10.1038/s41581-024-00849-7).
- [392] J. Aru, M. Suzuki, and M. E. Larkum. “Cellular Mechanisms of Conscious Processing.” en. In: *Trends Cogn. Sci.* 24 (10 Oct. 1, 2020), pp. 814–825. DOI: [10.1016/j.tics.2020.07.006](https://doi.org/10.1016/j.tics.2020.07.006).
- [393] M. S. A. Graziano, A. Guterstam, B. J. Bio, and A. I. Wilterson. “Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories.” en. In: *Cogn. Neuropsychol.* 37 (3-4 May 2020), pp. 155–172. DOI: [10.1080/02643294.2019.1670630](https://doi.org/10.1080/02643294.2019.1670630).
- [394] J. G. Taylor. “Attention as the control system of the brain.” In: *Int. J. Gen. Syst.* 35 (3 June 1, 2006), pp. 361–376. DOI: [10.1080/03081070600661077](https://doi.org/10.1080/03081070600661077).
- [395] H. C. Hemmings Jr, M. H. Akabas, P. A. Goldstein, J. R. Trudell, B. A. Orser, and N. L. Harrison. “Emerging molecular mechanisms of general anesthetic action.” en. In: *Trends Pharmacol. Sci.* 26 (10 Oct. 1, 2005), pp. 503–510. DOI: [10.1016/j.tips.2005.08.006](https://doi.org/10.1016/j.tips.2005.08.006).
- [396] Y. Li, X. Chen, H. Zheng, S. Li, Q. Teng, C. Wang, F. Li, Z. Xie, D. Sessler, Z. Zhang, P. Xu, T.-F. Yuan, and T. Xu. “Neurophysiological connectomic signatures of consciousness during propofol-induced general anesthesia.” en. In: *Cell Rep. Med.* 7 (2 Feb. 17, 2026), p. 102581. DOI: [10.1016/j.xcrm.2025.102581](https://doi.org/10.1016/j.xcrm.2025.102581).
- [397] Z. Huang, G. A. Mashour, and A. G. Hudetz. “Propofol disrupts the functional core-matrix architecture of the thalamus in humans.” en. In: *Nat. Commun.* 15 (1 Sept. 9, 2024), p. 7496. DOI: [10.1038/s41467-024-51837-1](https://doi.org/10.1038/s41467-024-51837-1).

- [398] Z. Liang, L. Fan, B. Zhang, W. Shu, D. Li, X. Li, and T. Yu. “The changes in neural complexity and connectivity in thalamocortical and cortico-cortical systems after propofol-induced unconsciousness in different temporal scales.” en. In: *Neuroimage* 311 (121193 May 1, 2025), p. 121193. DOI: [10.1016/j.neuroimage.2025.121193](https://doi.org/10.1016/j.neuroimage.2025.121193).
- [399] P. Boveroux, A. Vanhaudenhuyse, M.-A. Bruno, Q. Noirhomme, S. Lauwick, A. Luxen, C. Degueldre, A. Plenevaux, C. Schnakers, C. Phillips, J.-F. Brichant, V. Bonhomme, P. Maquet, M. D. Greicius, S. Laureys, and M. Boly. “Breakdown of within- and between-network resting state functional magnetic resonance imaging connectivity during propofol-induced loss of consciousness.” en. In: *Anesthesiology* 113 (5 Nov. 2010), pp. 1038–1053. DOI: [10.1097/ALN.0b013e3181f697f5](https://doi.org/10.1097/ALN.0b013e3181f697f5).
- [400] X.-X. Song and B.-W. Yu. “Anesthetic effects of propofol in the healthy human brain: functional imaging evidence.” en. In: *J. Anesth.* 29 (2 Apr. 2015), pp. 279–288. DOI: [10.1007/s00540-014-1889-4](https://doi.org/10.1007/s00540-014-1889-4).
- [401] U. Lee and G. A. Mashour. “Role of network science in the study of anesthetic state transitions.” en. In: *Anesthesiology* 129 (5 Nov. 2018), pp. 1029–1044. DOI: [10.1097/ALN.0000000000002228](https://doi.org/10.1097/ALN.0000000000002228).
- [402] M. Malekmohammadi, N. AuYong, C. M. Price, E. Tsolaki, A. E. Hudson, and N. Pouratian. “Propofol-induced changes in α - β sensorimotor cortical connectivity.” en. In: *Anesthesiology* 128 (2 Feb. 2018), pp. 305–316. DOI: [10.1097/ALN.0000000000001940](https://doi.org/10.1097/ALN.0000000000001940).
- [403] S. Ensel, L. Uhrig, A. Ozkirli, G. Hoffner, J. Tasserie, S. Dehaene, D. Van De Ville, B. Jarraya, and E. Pirondini. “Transient brain activity dynamics discriminate levels of consciousness during anesthesia.” en. In: *Commun. Biol.* 7 (1 June 10, 2024), p. 716. DOI: [10.1038/s42003-024-06335-x](https://doi.org/10.1038/s42003-024-06335-x).
- [404] N. P. Franks. “General anaesthesia: from molecular targets to neuronal pathways of sleep and arousal.” en. In: *Nat. Rev. Neurosci.* 9 (5 May 2008), pp. 370–386. DOI: [10.1038/nrn2372](https://doi.org/10.1038/nrn2372).
- [405] A. I. Luppi, L. Uhrig, J. Tasserie, P. A. M. Mediano, F. E. Rosas, S. P. Singleton, D. Gutierrez-Barragan, S. Gini, P. Castro, C. M. Signorelli, D. Golkowski, A. Ranft, R. Ilg, D. Jordan, K. Muta, J. Hata, H. Okano, Z.-Q. Liu, Y. Yee, A. Destexhe, R. Cofre, D. K. Menon, A. Gozzi, B. Jarraya, and E. A. Stamatakis. “Convergent transcriptomic and connectomic controllers of information integration and its anaesthetic breakdown across mammalian brains.” en. In: *Nat. Hum. Behav.* (Jan. 28, 2026), pp. 1–26. DOI: [10.1038/s41562-025-02381-5](https://doi.org/10.1038/s41562-025-02381-5).
- [406] E. Tang, C. Giusti, G. L. Baum, S. Gu, E. Pollock, A. E. Kahn, D. R. Roalf, T. M. Moore, K. Ruparel, R. C. Gur, R. E. Gur, T. D. Satterthwaite, and D. S. Bassett. “Developmental increases in white matter network controllability support a growing diversity of brain dynamics.” en. In: *Nat. Commun.* 8 (1 Nov. 1, 2017), p. 1252. DOI: [10.1038/s41467-017-01254-4](https://doi.org/10.1038/s41467-017-01254-4).
- [407] S. Gu, P. Fotiadis, L. Parkes, C. H. Xia, R. C. Gur, R. E. Gur, D. R. Roalf, T. D. Satterthwaite, and D. S. Bassett. “Network controllability mediates the relationship between rigid structure and flexible dynamics.” en. In: *Netw. Neurosci.* 6 (1 Feb. 1, 2022), pp. 275–297. DOI: [10.1162/netn_a_00225](https://doi.org/10.1162/netn_a_00225).

- [408] C. Tu, R. P. Rocha, M. Corbetta, S. Zampieri, M. Zorzi, and S. Suweis. “Warnings and caveats in brain controllability.” en. In: *Neuroimage* 176 (Aug. 1, 2018), pp. 83–91. DOI: [10.1016/j.neuroimage.2018.04.010](https://doi.org/10.1016/j.neuroimage.2018.04.010).
- [409] S. Suweis, C. Tu, R. P. Rocha, S. Zampieri, M. Zorzi, and M. Corbetta. “Brain controllability: Not a slam dunk yet.” en. In: *Neuroimage* 200 (Oct. 15, 2019), pp. 552–555. DOI: [10.1016/j.neuroimage.2019.07.012](https://doi.org/10.1016/j.neuroimage.2019.07.012).
- [410] C. Versteeg, A. R. Sedler, J. D. McCart, and C. Pandarinath. “Expressive dynamics models with nonlinear injective readouts enable reliable recovery of latent features from neural activity.” en. In: *ArXiv* (Sept. 12, 2023), arXiv:2309.06402v1.
- [411] L. Dinh, D. Krueger, and Y. Bengio. “NICE: Non-linear independent components estimation.” In: International Conference on Learning Representations. Oct. 30, 2014. DOI: [10.48550/arXiv.1410.8516](https://doi.org/10.48550/arXiv.1410.8516).
- [412] M. Lezcano-Casado and D. Martínez-Rubio. “Cheap orthogonal constraints in neural networks: A simple parametrization of the orthogonal and unitary group.” In: International Conference on Machine Learning. Jan. 24, 2019. DOI: [10.48550/arXiv.1901.08428](https://doi.org/10.48550/arXiv.1901.08428).
- [413] A. Trockman and J. Zico Kolter. “Orthogonalizing Convolutional Layers with the Cayley Transform.” In: International Conference on Learning Representations. Apr. 14, 2021. DOI: [10.48550/arXiv.2104.07167](https://doi.org/10.48550/arXiv.2104.07167).
- [414] S. Jeong, H. G. Lee, W. M. Kim, C. W. Jeong, S. H. Lee, M. H. Yoon, and J. I. Choi. “Increase of paradoxical excitement response during propofol-induced sedation in hazardous and harmful alcohol drinkers.” en. In: *Br. J. Anaesth.* 107 (6 Dec. 1, 2011), pp. 930–933. DOI: [10.1093/bja/aer275](https://doi.org/10.1093/bja/aer275).
- [415] M. M. McCarthy, E. N. Brown, and N. Kopell. “Potential network mechanisms mediating electroencephalographic beta rhythm changes during propofol-induced paradoxical excitation.” en. In: *J. Neurosci.* 28 (50 Dec. 10, 2008), pp. 13488–13504. DOI: [10.1523/JNEUROSCI.3536-08.2008](https://doi.org/10.1523/JNEUROSCI.3536-08.2008).
- [416] J. R. Sneyd. “Excitatory events associated with propofol anaesthesia: a review.” en. In: *J. R. Soc. Med.* 85 (5 May 1992), pp. 288–291. DOI: [10.1177/014107689208500516](https://doi.org/10.1177/014107689208500516).
- [417] C. M. Gillan, M. Kosinski, R. Whelan, E. A. Phelps, and N. D. Daw. “Characterizing a psychiatric symptom dimension related to deficits in goal-directed control.” en. In: *Elife* 5 (Mar. 1, 2016). DOI: [10.7554/eLife.11305](https://doi.org/10.7554/eLife.11305).
- [418] A. D. Redish. “Addiction as a computational process gone awry.” en. In: *Science* 306 (5703 Dec. 10, 2004), pp. 1944–1947. DOI: [10.1126/science.1102384](https://doi.org/10.1126/science.1102384).
- [419] M. J. Frank, A. Santamaria, R. C. O’Reilly, and E. Willcutt. “Testing computational models of dopamine and noradrenaline dysfunction in attention deficit/hyperactivity disorder.” en. In: *Neuropsychopharmacology* 32 (7 July 2007), pp. 1583–1599. DOI: [10.1038/sj.npp.1301278](https://doi.org/10.1038/sj.npp.1301278).
- [420] T. V. Maia and M. J. Frank. “From reinforcement learning models to psychiatric and neurological disorders.” en. In: *Nat. Neurosci.* 14 (2 Feb. 2011), pp. 154–162. DOI: [10.1038/nn.2723](https://doi.org/10.1038/nn.2723).

- [421] S. H. Siddiqi, S. F. Taylor, D. Cooke, A. Pascual-Leone, M. S. George, and M. D. Fox. “Distinct symptom-specific treatment targets for circuit-based neuromodulation.” en. In: *Am. J. Psychiatry* 177 (5 May 1, 2020), pp. 435–446. DOI: [10.1176/appi.ajp.2019.19090915](https://doi.org/10.1176/appi.ajp.2019.19090915).
- [422] S. H. Siddiqi, N. S. Philip, S. T. Palm, D. M. Carreon, A. R. Arulpragasam, J. Barredo, H. Bouchard, M. A. Ferguson, J. H. Grafman, R. A. Morey, and M. D. Fox. “A potential target for noninvasive neuromodulation of PTSD symptoms derived from focal brain lesions in veterans.” en. In: *Nat. Neurosci.* 27 (11 Nov. 24, 2024), pp. 2231–2239. DOI: [10.1038/s41593-024-01772-7](https://doi.org/10.1038/s41593-024-01772-7).
- [423] J. K. Krauss, N. Lipsman, T. Aziz, A. Boutet, P. Brown, J. W. Chang, B. Davidson, W. M. Grill, M. I. Hariz, A. Horn, M. Schulder, A. Mammis, P. A. Tass, J. Volkmann, and A. M. Lozano. “Technology of deep brain stimulation: current status and future directions.” en. In: *Nat. Rev. Neurol.* 17 (2 Feb. 2021), pp. 75–87. DOI: [10.1038/s41582-020-00426-z](https://doi.org/10.1038/s41582-020-00426-z).
- [424] V. Tiruvadi, K. S. Choi, R. E. Gross, R. Butera, V. Jirsa, and H. Mayberg. “Dynamic oscillations evoked by subcallosal cingulate deep brain stimulation.” en. In: *Front. Neurosci.* 16 (Feb. 23, 2022), p. 768355. DOI: [10.3389/fnins.2022.768355](https://doi.org/10.3389/fnins.2022.768355).
- [425] S. Alagapan, K. S. Choi, S. Heisig, P. Riva-Posse, A. Crowell, V. Tiruvadi, M. Obatusin, A. Veerakumar, A. C. Waters, R. E. Gross, S. Quinn, L. Denison, M. O’Shaughnessy, M. Connor, G. Canal, J. Cha, R. Hershenberg, T. Nauvel, F. Isbaine, M. F. Afzal, M. Figeo, B. H. Kopell, R. Butera, H. S. Mayberg, and C. J. Rozell. “Cingulate dynamics track depression recovery with deep brain stimulation.” en. In: *Nature* 622 (7981 Oct. 20, 2023), pp. 130–138. DOI: [10.1038/s41586-023-06541-3](https://doi.org/10.1038/s41586-023-06541-3).
- [426] P. Bashivan, K. Kar, and J. J. DiCarlo. “Neural population control via deep image synthesis.” en. In: *Science* 364 (6439 May 3, 2019). DOI: [10.1126/science.aav9436](https://doi.org/10.1126/science.aav9436).
- [427] O. Karniol-Tambour, D. M. Zoltowski, E. Mika Diamanti, L. Pinto, C. D. Brody, D. W. Tank, and J. W. Pillow. “Modeling state-dependent communication between brain regions with switching nonlinear dynamical systems.” In: *The Twelfth International Conference on Learning Representations*. Oct. 13, 2023.
- [428] P. F. Hitchcock and M. J. Frank. “From tripping and falling to ruminating and worrying: A meta-control account of repetitive negative thinking.” en. In: *Curr. Opin. Behav. Sci.* 56 (101356 Apr. 1, 2024), p. 101356. DOI: [10.1016/j.cobeha.2024.101356](https://doi.org/10.1016/j.cobeha.2024.101356).
- [429] T. V. Maia, R. E. Cooney, and B. S. Peterson. “The neural bases of obsessive-compulsive disorder in children and adults.” en. In: *Dev. Psychopathol.* 20 (4 2008), pp. 1251–1283. DOI: [10.1017/S0954579408000606](https://doi.org/10.1017/S0954579408000606).
- [430] B. Liebenow, R. Jones, E. DiMarco, J. D. Trattner, J. Humphries, L. P. Sands, K. P. Spry, C. K. Johnson, E. B. Farkas, A. Jiang, and K. T. Kishida. “Computational reinforcement learning, reward (and punishment), and dopamine in psychiatric disorders.” en. In: *Front. Psychiatry* 13 (Oct. 20, 2022), p. 886297. DOI: [10.3389/fpsy.2022.886297](https://doi.org/10.3389/fpsy.2022.886297).
- [431] A. K. Seth, K. Suzuki, and H. D. Critchley. “An interoceptive predictive coding model of conscious presence.” en. In: *Front. Psychol.* 2 (2011), p. 395. DOI: [10.3389/fpsyg.2011.00395](https://doi.org/10.3389/fpsyg.2011.00395).

- [432] H. D. Critchley. “Neural mechanisms of autonomic, affective, and cognitive integration.” en. In: *J. Comp. Neurol.* 493 (1 Dec. 5, 2005), pp. 154–166. DOI: [10.1002/cne.20749](#).
- [433] J. Kiverstein. “The interdependence of embodied cognition and consciousness.” In: *J. Conscious. Stud.* (2016).
- [434] L. F. Barrett. “The theory of constructed emotion: an active inference account of interoception and categorization.” en. In: *Soc. Cogn. Affect. Neurosci.* 12 (1 Jan. 1, 2017), pp. 1–23. DOI: [10.1093/scan/nsw154](#).
- [435] F. J. Varela, E. Rosch, and E. Thompson. *The Embodied Mind: Cognitive science and human experience*. en. The MIT Press, Sept. 26, 1991. DOI: [10.7551/mitpress/6730.001.0001](#).
- [436] J. D. Hunter. “Matplotlib: A 2D Graphics Environment.” In: *Comput. Sci. Eng.* 9 (3 2007), pp. 90–95. DOI: [10.1109/MCSE.2007.55](#).
- [437] W. McKinney. “Data Structures for Statistical Computing in Python.” en. In: *Proceedings of the 9th Python in Science Conference*. Python in Science Conference (Austin, Texas). SciPy, 2010. DOI: [10.25080/majora-92bf1922-00a](#).
- [438] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. “Scikit-learn: Machine Learning in Python.” In: *J. Mach. Learn. Res.* 12 (85 2011), pp. 2825–2830.
- [439] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, Í. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. “SciPy 1.0: fundamental algorithms for scientific computing in Python.” en. In: *Nat. Methods* 17 (3 Mar. 2020), pp. 261–272. DOI: [10.1038/s41592-019-0686-2](#).
- [440] A. S. Dickey, A. Suminski, Y. Amit, and N. G. Hatsopoulos. “Single-unit stability using chronically implanted multielectrode arrays.” en. In: *J. Neurophysiol.* 102 (2 Aug. 2009), pp. 1331–1339. DOI: [10.1152/jn.90920.2008](#).
- [441] J. Solé-Casals and F.-B. Vialatte. “Towards Semi-Automatic Artifact Rejection for the Improvement of Alzheimer’s Disease Screening from EEG Signals.” en. In: *Sensors* 15 (8 July 23, 2015), pp. 17963–17976. DOI: [10.3390/s150817963](#).
- [442] J. A. Urigüen and B. Garcia-Zapirain. “EEG artifact removal-state-of-the-art and guidelines.” en. In: *J. Neural Eng.* 12 (3 June 2015), p. 031001. DOI: [10.1088/1741-2560/12/3/031001](#).
- [443] S. D. Muthukumaraswamy. “High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations.” en. In: *Front. Hum. Neurosci.* 7 (Apr. 15, 2013), p. 138. DOI: [10.3389/fnhum.2013.00138](#).

- [444] D. P. Brunner, R. C. Vasko, C. S. Detka, J. P. Monahan, C. F. Reynolds 3rd, and D. J. Kupfer. “Muscle artifacts in the sleep EEG: automated detection and effect on all-night EEG power spectra.” en. In: *J. Sleep Res.* 5 (3 Sept. 1996), pp. 155–164. DOI: [10.1046/j.1365-2869.1996.00009.x](https://doi.org/10.1046/j.1365-2869.1996.00009.x).
- [445] D. Breda. “On characteristic roots and stability charts of delay differential equations.” en. In: *Int. J. Robust Nonlinear Control* 22 (8 May 25, 2012), pp. 892–917. DOI: [10.1002/rnc.1734](https://doi.org/10.1002/rnc.1734).
- [446] J. B. Conway. *Functions of One Complex Variable I*. Springer New York, 1978. 14 pp. DOI: [10.1007/978-1-4612-6313-5](https://doi.org/10.1007/978-1-4612-6313-5).
- [447] B. Ai, L. Sentis, N. Paine, S. Han, A. Mok, and C.-L. Fok. “Stability and Performance Analysis of Time-Delayed Actuator Control Systems.” en. In: *J. Dyn. Syst. Meas. Control* 138 (5 Mar. 10, 2016), p. 051005. DOI: [10.1115/1.4032461](https://doi.org/10.1115/1.4032461).
- [448] P. Wahi and A. Chatterjee. “Galerkin Projections for Delay Differential Equations.” In: *ASME 2003 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference*. Vol. 127. American Society of Mechanical Engineers, Jan. 1, 2003. DOI: [10.1115/DETC2003/VIB-48570](https://doi.org/10.1115/DETC2003/VIB-48570).
- [449] C. P. Vyasarayani, S. Subhash, and T. Kalmár-Nagy. “Spectral approximations for characteristic roots of delay differential equations.” In: *International Journal of Dynamics and Control* 2 (2 June 1, 2014), pp. 126–132. DOI: [10.1007/s40435-014-0060-2](https://doi.org/10.1007/s40435-014-0060-2).
- [450] S. A. Hill. “Pharmacokinetics of drug infusions.” In: *Contin Educ Anaesth Crit Care Pain* 4 (3 June 1, 2004), pp. 76–80. DOI: [10.1093/bjaceaccp/mkh021](https://doi.org/10.1093/bjaceaccp/mkh021).
- [451] J. Schüttler and H. Ihmsen. “Population pharmacokinetics of propofol: a multicenter study.” en. In: *Anesthesiology* 92 (3 Mar. 2000), pp. 727–738. DOI: [10.1097/00000542-200003000-00017](https://doi.org/10.1097/00000542-200003000-00017).
- [452] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. Del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. “Array programming with NumPy.” en. In: *Nature* 585 (7825 Sept. 16, 2020), pp. 357–362. DOI: [10.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2).
- [453] O. Yadan. *Hydra - A framework for elegantly configuring complex applications*. 2019.
- [454] The pandas development team. *pandas-dev/pandas: Pandas*. Comp. software. 2025. DOI: [10.5281/ZENODO.17229934](https://doi.org/10.5281/ZENODO.17229934).
- [455] S. Seabold and J. Perktold. “statsmodels: Econometric and statistical modeling with python.” In: *9th Python in Science Conference*. 2010.
- [456] M. Behrends, M. D. Larson, A. E. Neice, and M. P. Bokoch. “Suppression of pupillary unrest by general anesthesia and propofol sedation.” en. In: *J. Clin. Monit. Comput.* 33 (2 Apr. 2019), pp. 317–323. DOI: [10.1007/s10877-018-0147-y](https://doi.org/10.1007/s10877-018-0147-y).
- [457] O. A. Khan, S. R. Taylor, and J. G. Jones. “Anaesthesia and saccadic eye movements: Anaesthesia and saccadic eye movements.” en. In: *Anaesthesia* 55 (9 Sept. 26, 2000), pp. 877–882. DOI: [10.1046/j.1365-2044.2000.01529.x](https://doi.org/10.1046/j.1365-2044.2000.01529.x).

- [458] H. Shitara, T. Hiroki, R. Koda, T. Sasaki, M. Kamiyama, R. Miyamoto, K. Nakase, F. Ino, T. Kachi, R. Mieda, C. Aso, Y. Yamakoshi, S. Saito, and H. Chikuda. “Relationship between muscle tone and elasticity: Simultaneous quantitative assessment using train-of-four monitoring and continuous shear wave elastography during anesthesia induction-A prospective observational study.” en. In: *Diagnostics (Basel)* 15 (3 Jan. 26, 2025), p. 293. DOI: [10.3390/diagnostics15030293](https://doi.org/10.3390/diagnostics15030293).
- [459] X.-L. Zhou, L.-J. Xing, H.-R. Liu, Y. Qian, J. Zhu, and H. Xie. “Pupil diameter changes after anesthesia with different doses of sufentanil under ultrasound monitoring.” en. In: *Int. J. Clin. Pract.* 2022 (1 July 14, 2022), p. 6320973. DOI: [10.1155/2022/6320973](https://doi.org/10.1155/2022/6320973).
- [460] H. Eilers and M. D. Larson. “The effect of ketamine and nitrous oxide on the human pupillary light reflex during general anesthesia.” en. In: *Auton. Neurosci.* 152 (1-2 Jan. 15, 2010), pp. 108–114. DOI: [10.1016/j.autneu.2009.10.004](https://doi.org/10.1016/j.autneu.2009.10.004).
- [461] T. Schreiber and A. Schmitz. “Improved Surrogate Data for Nonlinearity Tests.” en. In: *Phys. Rev. Lett.* 77 (4 July 22, 1996), pp. 635–638. DOI: [10.1103/PhysRevLett.77.635](https://doi.org/10.1103/PhysRevLett.77.635).
- [462] G. Lancaster, D. Iatsenko, A. Pidde, V. Ticcinelli, and A. Stefanovska. “Surrogate data for hypothesis testing of physical systems.” In: *Phys. Rep.* 748 (July 18, 2018), pp. 1–60. DOI: [10.1016/j.physrep.2018.06.001](https://doi.org/10.1016/j.physrep.2018.06.001).
- [463] V. Venema, F. Ament, and C. Simmer. “A Stochastic Iterative Amplitude Adjusted Fourier Transform algorithm with improved accuracy.” en. In: *Nonlinear Process. Geophys.* 13 (3 July 26, 2006), pp. 321–328. DOI: [10.5194/npg-13-321-2006](https://doi.org/10.5194/npg-13-321-2006).
- [464] P. Lockhart. *A Mathematician’s Lament*.
- [465] K. P. Burnham and D. R. Anderson. *Model selection and multimodel inference: A practical information-theoretic approach*. en. Ed. by K. P. Burnham and D. R. Anderson. 2nd ed. New York, NY: Springer, Sept. 28, 2002. 488 pp. DOI: [10.1007/b97636](https://doi.org/10.1007/b97636).
- [466] S. R. Tasbihgou, C. R. M. Barends, and A. R. Absalom. “The role of dexmedetomidine in neurosurgery.” en. In: *Best Pract. Res. Clin. Anaesthesiol.* 35 (2 July 1, 2021), pp. 221–229. DOI: [10.1016/j.bpa.2020.10.002](https://doi.org/10.1016/j.bpa.2020.10.002).
- [467] Z. Zhang, V. Ferretti, İ. Güntan, A. Moro, E. A. Steinberg, Z. Ye, A. Y. Zecharia, X. Yu, A. L. Vyssotski, S. G. Brickley, R. Yustos, Z. E. Pillidge, E. C. Harding, W. Wisden, and N. P. Franks. “Neuronal ensembles sufficient for recovery sleep and the sedative actions of $\alpha 2$ adrenergic agonists.” en. In: *Nat. Neurosci.* 18 (4 Apr. 23, 2015), pp. 553–561. DOI: [10.1038/nn.3957](https://doi.org/10.1038/nn.3957).
- [468] A. Bekker and M. K. Sturaitis. “Dexmedetomidine for neurological surgery.” en. In: *Neurosurgery* 57 (1 Suppl July 2005), 1–10, discussion 1–10. DOI: [10.1227/01.neu.0000163476.42034.a1](https://doi.org/10.1227/01.neu.0000163476.42034.a1).
- [469] D. R. Helsel, R. M. Hirsch, K. R. Ryberg, S. A. Archfield, and E. J. Gilroy. *Statistical methods in water resources*. 2020. DOI: [10.3133/tm4a3](https://doi.org/10.3133/tm4a3).
- [470] C. J. Geyer. *Stat 5102 Notes: Nonparametric Tests and Confidence Intervals*. 2007.
- [471] J. Pathak, Z. Lu, B. R. Hunt, M. Girvan, and E. Ott. “Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data.” en. In: *Chaos* 27 (12 Dec. 2017), p. 121102. DOI: [10.1063/1.5010300](https://doi.org/10.1063/1.5010300).

- [472] J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott. “Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach.” en. In: *Phys. Rev. Lett.* 120 (2 Jan. 12, 2018), p. 024102. DOI: [10.1103/PhysRevLett.120.024102](https://doi.org/10.1103/PhysRevLett.120.024102).
- [473] Z. Lu, J. Pathak, B. Hunt, M. Girvan, R. Brockett, and E. Ott. “Reservoir observers: Model-free inference of unmeasured variables in chaotic systems.” en. In: *Chaos* 27 (4 Apr. 5, 2017), p. 041102. DOI: [10.1063/1.4979665](https://doi.org/10.1063/1.4979665).
- [474] A. Banerjee, J. Pathak, R. Roy, J. G. Restrepo, and E. Ott. “Using machine learning to assess short term causal dependence and infer network links.” en. In: *Chaos* 29 (12 Dec. 26, 2019), p. 121104. DOI: [10.1063/1.5134845](https://doi.org/10.1063/1.5134845).
- [475] M. Rigotti, O. Barak, M. R. Warden, X.-J. Wang, N. D. Daw, E. K. Miller, and S. Fusi. “The importance of mixed selectivity in complex cognitive tasks.” en. In: *Nature* 497 (7451 May 30, 2013), pp. 585–590. DOI: [10.1038/nature12160](https://doi.org/10.1038/nature12160).
- [476] I. Loshchilov and F. Hutter. “Decoupled weight decay regularization.” In: International Conference on Learning Representations. 2019. DOI: [10.48550/arXiv.1711.05101](https://doi.org/10.48550/arXiv.1711.05101).